博士学位论文

基于数据挖掘的数字图书馆评估关键技术研究

赵育民

2011 年 11 月

中图分类号：
UDC 分类号：

基于数据挖掘的数字图书馆评估关键技术研究

作 者 姓 名　　赵育民
学 院 名 称　　计算机学院
指 导 教 师　　牛振东教授
答辩委员会主席　林守勋教授
申请学位级别　　工学博士
学 科 专 业　　计算机软件与理论
研 究 方 向　　数字图书馆
学位授予单位　　北京理工大学
论文答辩日期　　2011 年 11 月

Research about Digital Library Evaluation Based on Data Mining

Candidate Name： Yumin Zhao

School or Department: Computer Science and Technology

Faculty Mentor: Prof. Zhendong Niu

Chair, Thesis Committee： Prof. Shouxun Lin

Degree Applied: Doctor of Philosophy

Major： Computer Software and Theory

Degree by: Beijing Institute of Technology

The Date of Defence： November，2011

研究成果声明

本人郑重声明：所提交的学位论文是我本人在指导教师的指导下进行的研究工作获得的研究成果。尽我所知，文中除特别标注和致谢的地方外，学位论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京理工大学或其它教育机构的学位或证书所使用过的材料。与我一同工作的合作者对此研究工作所做的任何贡献均已在学位论文中作了明确的说明并表示了谢意。

特此申明。

签　　名：　　日期：

关于学位论文使用权的说明

本人完全了解北京理工大学有关保管、使用学位论文的规定，其中包括：①学校有权保管、并向有关部门送交学位论文的原件与复印件；②学校可以采用影印、缩印或其它复制手段复制并保存学位论文；③学校可允许学位论文被查阅或借阅；④学校可以学术交流为目的，复制赠送和交换学位论文；⑤学校可以公布学位论文的全部或部分内容（保密学位论文在解密后遵守此规定）。

签　　名：　　日期：

导师签名：　　日期：

摘要

随着互联网技术的快速发展，数字图书馆逐渐成为计算机知识组织与数字化领域的研究热点，被认为是未来互联网知识形式的必然发展趋势。而数字图书馆评估是对数字图书馆中各种资源的综合性研究，是保障和改善数字图书馆建设的一种技术，它主要是通过一系列活动来收集与数字图书馆相关的数据，对所收集的数据进行分析、处理，并采用科学的方法进行评价，近年来受到广泛的关注。目前，数字图书馆评估研究有两个特点。一方面，评估研究项目较多，研究之间呈现出一定的离散现象，反映出对数字图书馆评估基础理论认知的不一致；另一方面，传统的评估方法主要以专家制定指标评估和用户调查评估为主，缺乏计算机智能数据处理、分析和辅助决策的支持。在数字图书馆高速发展的今天，评估工作面临数据海量化、关系复杂化等诸多新问题，传统的评估方法已难以取得令人满意的效果。本论文围绕以上问题展开数字图书馆评估研究。

论文的主要内容及创新工作如下：

（1）提出一个数字图书馆评估四元理论模型

针对目前评估研究基础理论认知的不一致性，研讨了数字图书馆评估客体、评估目的和评估主体的基本性质，基于这些研究提出一个数字图书馆评估四元理论模型。论述了四元理论模型的内部联系以及与现有数字图书馆评估思想之间的关系。

（2）提出一种主、客观因素相融合的数字图书馆评估模型（SOI 模型）

针对目前评估模型中评估主体的不全面性，结合四元理论模型的思想，提出一种主、客观因素相结合的数字图书馆评估模型。将数字图书馆评估划分为硬性评估和软性评估两个部分，并对主客观因素融合机制、评估体系结构等方面进行分析研究。

（3）提出了一种基于 z 值的数字图书馆馆藏属性离散算法（PDOZ 算法）

首先对处于 SOI 模型核心位置的馆藏模块进行了设计，在馆藏属性中引入了体现用户感受的主观数据，作为解决馆藏质量量化问题的桥梁，并验证了这种主观数据的有效性，在进行数据挖掘的应用当中，针对传统数据离散算法无法有效表达出数字图书馆馆藏复杂属性关系的问题，提出了一种基于 z 值的离散算法（PDOZ 算法）。相比传统离散方法，该算法实现了概率分布意义的动态距离，它反映了馆藏各个属性语义距离的不一致性，增强了属性与预测之间的相关性，更利于发现属性关系，并在实验

中发现了非线性条件属性的存在。

（4）提出了一种用于数字图书馆馆藏价值预测的层叠决策树算法（SDT-Z 算法）

在对馆藏价值的数据预测当中，针对传统数据挖掘预测算法在数字图书馆馆藏评估数据集中应用不理想的问题，提出了一种层叠决策树算法。该算法引入层属性的概念，将决策树选取属性从平面信息的选择拓展到了立体信息，有效的消除了非线性条件属性对数据挖掘的干扰影响，并可以嵌套使用已有决策树算法，具备较强的灵活性，解决了传统数据挖掘预测技术在数字图书馆评估应用研究中的瓶颈。

关键词：数字图书馆评估；数据挖掘；评估模型；数据离散；决策树

Abstract

With the rapid development of Internet technology Digital Library is considered an inevitable development trend of computer information service and Internet technology. Data processing and use of digital library has been a research hotspot in the field of computer knowledge organization and digitization. Evaluation research about digital library is a comprehensive research focusing on digital library resources and has been concerned widely in recent years. Current characteristics of digital library evaluation can be summarized into two categories. One is about evaluation ideas and evaluation model. At the present stage many evaluation studies are isolated and can not improve from each other. This reflects research scotoma of the basic theory and inconsistency of cognitive in digital library evaluation. The other is about evaluation methods. Nowadays main methods for evaluation are based on evaluation index and user survey. There lacks advanced idea of computer intelligent data processing, analysis and decision support, and faces many problems difficult to solve with the rapid development of digital library, such as mass data, complicated relationship, etc. The traditional evaluation ideas and methods have been difficult to obtain satisfactory results. This dissertation conducts intensive research following above two problems.

The main research contents and innovative contributions are as follows:

(1) Propose a four meta-theories model.

We summarize present evaluation ideologies about digital library, analyze the definition of Digital Library, then we disscuss the divarication about the evaluation purpose and evaluation subject of digital library. Based on these studies we propose a four meta-theories model, expound the internal relations of the model and the relations with digital library evaluation, and based on this model we discuss our ideology about digital library evaluation.

(2) Propose a natural-integration model of digital library evaluation (SOI model)

Based on the basic theory system research of digital library evaluation, we propose a natural-integration model for digital library evaluation. We have a detailed description of SOI architecture and especially design an evaluation method for collections based on data mining.

(3) Research on attributes of digital library collections. Propose a parallel

discretization algorithm based on z-score theory (PDOZ algorithm). Discovering a complicated condition attribute relation.

This dissertation imports the subjective data which express users' feel into collections dataset as a solution for quantizing about collection's quality. Then we research the complex relationship among collections' properties, and propose a parallel discretization algorithm based on z-score theory. It use dynamic range that reflects the significance of probability distributions instead of traditional equidistant discrete methods, It reflects the dynamic semantics of each collections' attribute, improve the correlativity between normal attribute and prediction attribute, and can discover attribute relation more efficiently. At last based on PDOZ algorithm we discover a complicated condition attribute relation.

(4) Propose a stratified decision tree algorithm based on PDOZ (SDT-Z algorithm)

Based on collections' attributes research we propose a stratified decision tree algorithm based on PDOZ for prediction mining about the value of digital library collections. This algorithm solves the problem that traditional data mining algorithm can't be well applied in the digital library collections. Stratified attribute concept is imported in this algorithm. It expands the select of splitting attribute in decision tree from flat information to stereoscopic information, eliminates the influence of complicated condition attribute relation, can use nested existing decision tree algorithms, solves the bottleneck of data mining application in digital library evaluation.

Key Words: digital library evaluation; data mining; evaluation model; data discretization; decision tree

目录

图索引

表索引

第 1 章 绪论

1.1 研究背景及课题意义

数字图书馆是 20 世纪末期出现的知识组织形式的重大革新[1]，其独特的存在形式和重要作用使它受到了广泛的关注，数字图书馆技术涉及知识组织、信息检索、计算机网络等诸多领域。随着数字资源管理技术的快速发展，数字图书馆技术也在不断发生变化[2-4]，而数字图书馆评估研究是为了评估、保障和改善数字图书馆建设而被广泛关注的科研领域，它是围绕数字图书馆各方面资源的综合性研究。

数字图书馆评估是通过一系列活动来收集与数字图书馆相关的数据，对所收集的数据进行分析、处理，并采用科学的方法进行评价，从而保障和改善数字图书馆建设的一种技术。数字图书馆评估研究起源于 20 世纪 90 年代初，自 21 世纪初以来，取得了较快的发展和进步[5]。数字图书馆评估研究是对数字图书馆全面、系统的进行定量、定性考核和评价的过程，通过数字图书馆评估工作，检测其目前达到的水平并找出需要改善的方面，最终可以提高数字图书馆整体的质量，保障用户的权益，合理配置资源，避免重复浪费，有利于数字图书馆的协调发展。

目前，数字图书馆评估研究已成为数字图书馆建设中的一个重要方面，但在早期它并没有引起足够的重视，直到全球数字图书馆建设的第一波高潮的晚期（以美国的 Digital Library Initiatives 一期科研项目为代表），在科研人员进行问题总结的时候，数字图书馆评估的重要性才被大家重新认识，并得到了迅速发展。

数字图书馆评估最先是借助传统图书馆的评估方法和理念展开，随着计算机数字化技术的发展，传统图书馆与数字图书馆的差别越来越大，难以全面表现出数字图书馆的特点，于是出现了越来越多的专门针对数字图书馆评估的研究，主要分为理论框架、评估方法两个方面。

在理论框架上，一些项目在系统的角度上对数字图书馆评估进行了研究，较有代表性的有：英国高等教育经费委员会资助的 eVALUED 项目提出了基于服务维度的理论框架[6]，伊利诺大学香槟分校的 DeLIver 项目提出了基于评价环境和用户维度的理论[7]，美国图书馆研究协会的研究项目提出了基于服务、用户及维度的理论框架[8]，EQUNIOX 项目提出了基于多个维度的理论框架[9]，还有一些基于单一方面展开的评

估研究[10]，例如 Dillon 等人提出的时间框架模型、Wesson 和 Greunen 等人提出的用于界面评价的可用性指数，以及 White 提出的基于分析和评价的描述性数字参考服务模型等[11]。

在评估方法上，一般采用定性分析和定量计算相结合。数字图书馆评估方式，一般包含三个过程：数据收集、评估分析、校验评估结论[12]。数据收集使用最广泛的手段主要有：用户问卷调查法、数字图书馆日志、电话访问、个人访问、BBS 电子公告板、专题讨论、学术会议等方式。评估分析就是对收集的数据进行归纳、整理、分析和综合，一般使用数理统计分析的方式来实现。评估结论主要包括下面几种检验方式：将评估结论与实际状况进行对比检验，以确定评估结论是否符合实际情况；采用不同评估方法对同一评估客体进行评估，检验评估结论是否一致；将评估结论与以前的评估结果相互比较，确定前后的变化趋势或是否相互矛盾。

现有的理论框架研究，往往服务于具体的项目，并不具备基础理论的普适作用。数字图书馆评估研究应该建立起一个学科意义的基础理论体系，对一系列数字图书馆评估相关的基本性质和问题建立明确的界定，在确立评估模型、制定细节评估策略及算法、解决不同观点产生分歧等问题的时候，起到全面指导的作用。

传统数字图书馆评估方法一般通过专家制定指标和用户调查来进行，随着数字图书馆的高速发展，数字图书馆评估研究面临数据海量化、关系复杂化等诸多新问题，传统的评估方法已难以取得令人满意的效果。借助数据智能化处理、分析和辅助决策的支持可以很好的解决这个问题，而数字图书馆本质上是海量数据、信息丰富、格式规范的数据仓库，为进行智能化处理提供了一个良好的环境，但是现阶段数字图书馆评估思想制约了这方面的发展思路。

本文围绕数字图书馆评估基本性质、数字图书馆评估模型以及基于数据挖掘的评估方法等方面对数字图书馆评估开展研究。针对目前数字图书馆评估主流思想兼顾用户和专家全面性不足的问题，提出了一种将主观因素和客观因素相融合的数字图书馆评估思想及模型；创新性的将数据挖掘技术应用于馆藏评估当中，并针对其中的关键技术研制出了新的数据离散算法以及决策树预测算法，更准确的表达出馆藏的价值评估信息。

1.2 国内外研究现状及发展趋势

数字图书馆评估的发展一直处于演进过程，数字图书馆评估的分类也存在多种标

准，依据评估对象的不同，划分为针对数字图书馆系统的指标类型评估、针对数字图书馆项目的成果评估以及针对提供数字资源服务的实体的评估[13]。依据评估指标体系的特点，划分为业务主导型评估体系和服务主导型评估体系[14-16]，数字图书馆评估理论奠基人 Saracevic[17]依据数字图书馆评估研究角度的不同，将数字图书馆评估划分为理论型评估和实践型评估，本文按照撒拉赛维奇的划分方法对目前国内外的评估研究进行综述。

1.2.1 国外研究现状

（1）实践型评估研究

DigiQUAL[14]来源于著名的 LibQUAL+®（曾称为 LibQUAL+™）的思想，LibQUAL+®是完全基于用户感受的图书馆评估思想，而 DigiQUAL 是针对数字图书馆对 LibQUAL+®的改进，两者的理论基础都 SERVQUAL 理论，该理论主要思想是把顾客期望服务水平和感知服务水平的差作为服务质量的测量标准。DigiQUAL 的评估理念是目前数字图书馆评估主流代表思想之一。

DELOS 是最著名的数字图书馆评估研究项目之一[18]，在欧盟和美国国家科学基金的支持下，DELOS 从 20 世纪末期就开始连续的举办数字图书馆评估论坛，建立起了跨语言评估矩阵、测试套件、元图书馆，对评估标准、评估资源以及评估方法展开了系统的研究。DELOS 的数字图书馆评估标准的一级指标包括：数据藏品、技术和用户使用三个方面，在数据藏品中划分为内容、元内容以及管理，在技术中划分为用户技术、信息访问、系统结构技术以及文档技术，在用户使用中划分为用户、领域、信息搜索以及目的，通过二级指标的测度进行细化评估。在 DELOS 的数字图书馆概念模型中，数据藏品被认为是最核心的部分。

美国研究图书馆协会 ARL 于 2000 年启动了 E-Metrics 项目[19]，形成了一个由资源、使用、成本、本地数字化资源、性能测量 5 个部分组成的指标体系。ARL 在 2002 年对项目组进行了审核工作，发布了 520 页的报告，包括项目背景、第一阶段报告、第二阶段报告、教学模块、数据收集模块以及图书馆对机构影响的研究报告等几个部分。这其中深入的研讨了教学模块的重要性、数字图书馆面临的统计和测度问题以及推荐的测度和统计指标。

中部英格兰大学的图书馆服务部门研究与评价证据基地负责的 eVALUEd 项目研发了一系列的工具集对数字图书馆的服务进行评估[6]。该工具最早是在 2004 年推出，并在 2006 年收到英格兰高等教育资助委员会的支持。eVALUEd 工具套件分为 4 个部

分：如何评价，评价主题，评价工具以及定制工具。其中“如何评价”提供了关于评价过程和一般性的技术问题的信息，“评价主题”提供了评价的指标对象以及评价方法，“评价工具”提供了针对主题指标的调查问卷或评价参数，“定制工具”提供适合具体需求的新创建的评价工具。

Perseus 项目[20]从数字图书馆的用途、可用性和技术角度对项目进行系统化的评估。PEAK 项目（知识电子获取定价，Pricing Electronic Access to Knowledge）[21]使数字图书馆评估的范围扩展到了经济效益分析的领域。加利福尼亚大学的 ADL（Alexandria Digital Library Project）项目对数字图书馆评估进行了一系列关于“可用性”和“操作功能”相关的研究[22, 23]。加利福尼亚大学参与的另一个项目[24]则将评估重点放在了对社会环境及用户行为的研究。

其他影响较大的项目有：例如 DLI 二期工程（DLI-2）[25]、英国电子图书馆工程（eLib）评估[26]、测试套件（Dlib Test Suite）[27]、跨语言评估论坛 CLEF（Cross-Language EvaluationForum）[28]、元图书馆和数字图书馆计划 MDLS（Metalibrary and Digital Library Schema）[18]、XML 检索评估计划 INEX （Initiative for the Evaluation of XML Retrieval）[29, 30]、ARL（Association of Research Libraries）评估项目[31]，以上这些项目与基于 SERVQUAL 理论[32-34]的 LibQUAL+®项目[35-41]和 DigiQUAL[42]项目相比，更加侧重于专家在评估过程中的参与。

（2）理论型评估研究

Kilker 等人 [43]在评估中提出了科技社会结构的理论，这种思想认为不同社会组织对各种科学技术的发展和应用会持有不同的观念，对于数字图书馆的评估研究来说也应该遵循这个原则，数字图书馆的不同用户对数字图书馆都会有不同的理解和感受，针对不同用户对数字图书馆评估应该有着不一样的标准。它并不等同于以用户为中心的数字图书馆评估思想，而是采用支持评估观点的个性化标准。

Barton 等人[44]提出了将数字图书馆评估和数字图书馆管理相结合的理论，其思想是认为数字图书馆评估作为一种管理工具，应该同数字图书馆管理组织部门相结合，这种思想对评估行为进行了一种全新角度的审视，从管理的角度来重新定位数字图书馆评估，将评估作为管理的基础。

Mabe 等人[45]提出了将数字图书馆评估和市场营销策略相结合的理论。Sandusky 等人[46]将一个数字图书馆评估框架划分为六个部分：用户、机构、存取、内容、服务、设计与开发。Saracevic[47]提出了一种围绕数字图书馆概念为中心的评估方式。

1.2.2 国内研究现状

国内的数字图书馆评估研究起步较晚，从 2003 年以后陆续有相关研究出现，随着数字图书馆的快速发展，数字图书馆评估的研究得到了国内学者的越来越多的重视，其受关注程度迅速提升，涌现出了很多研究成果。但从目前研究成果可以看出多局限于理论研究，实证研究较少，缺乏有组织的评估研究和实践项目。目前有组织的研究活动只有教育部和 CALIS 管理中心组织的对高等学校电子资源计量进行的研究，与美国、英国、欧洲有关组织开展的范围广、参与多、积累深的评估活动相比需要加强[48]。

通过论文研究情况可以发现，国内的研究侧重于评估细节的研究，针对某一个具体案例进行拓深研究，而在评估基础理论、评估模型、评估算法等方面相对较为欠缺。

乔欢[49-51]等人从评估基础理论的角度对数字图书馆开展了系统研究，对数字图书馆进行了评估主体、评估客体、评估标准等各方面的讨论，并进行数字图书馆的评估设计。该研究认为数字图书馆评价客体界定分为不同层面。在概念层面，需要对“什么是数字图书馆”问题予以分析。数字图书馆定义的多样性反映出其研究内容的丰富和复杂性。对数字图书馆评价客体的解析将形成评估活动的动态发展过程。从评价角度看，数字图书馆评价客体有系统与要素之分。用户信息需求测度，是数字图书馆系统设计的第一步。针对不同用户群体而设计的数字图书馆，形成不同的数字图书馆类型。

孙华[52]针对建立数字资源评估体系的研究，从高校评估对数字馆藏建设的影响、数字馆藏发展政策的制定、多样化数字资源的构建、数字馆藏的质量控制等几个方面探讨了高校评估背景下图书馆数字馆藏的建设。

康云萍等人[53]针对数字图书馆著作权的价值评估进行了研究，利用收益提成率法评估数字图书馆著作权的价值，对数字图书馆著作权评估的重要性进行系统的分析，对传统著作权评估方法进行了横向比较，选用基于收益法基础上产生的收益提成率法来评估数字图书馆著作权的价值。结合 AHP 方法确定带来超额收益的相关因素，通过构建一个指标体系系统来确定无形资产收益权重，全面、综合地衡量数子图书馆著作权的价值。

王启云[54]针对高校数字图书馆建设展开了评估调查与分析。利用网络在线调查问卷与 Email 方式传送 Word 版问卷相结合的方式，开展了关于高校数字图书馆建设的评估实践，该研究的调查对象主要是数字图书馆研究者与实践者，不少为专家、学者，

且大部分都在高校图书馆工作，因此调查问卷的结果很有参考价值。

唐李杏等人[55]提出了图书馆 2.0 时代的数字资源评估概念，该研究认为随着 Web2.0 技术和理念的推动，图书馆 2.0 也应运而生，改变了图书馆提供数字资源服务的理念和方式、数字资源的评估思路和策略。其中评估对象出现了微内容化，在图书馆 2.0 时代的应用中，微内容是任意小的一个用户访问数据单元，在这种情况下，评估将以微内容为主要对象，每一个微内容都应有相应的评估，同时用户是评估工作的中心，应该成为评估的主导力量。

陈鍊[56]针对数字图书馆信息系统安全进行了评估研究，针对数字图书馆安全问题，在介绍有关数字图书馆信息系统、信息系统安全、信息系统安全评估概念和标准基础上，提出一种以主观定性评估与客观定量评估相结合、基于灰类综合评判的信息系统安全评估模型，分析该评估模型的优缺点，并给出数字图书馆信息系统安全评估实例。

连天奎[57]对图书馆藏书数字化元数据进行评估的因素与条件进行分析和探讨。该研究主要针对数字图书馆元数据开展评估分析，认为不同的数字馆藏具有不同的属性及性质，所发展的元数据就有不同的评估重点，提出了评估图书馆藏书数字化分类与检索系统的基本原则。此外还有很多工作[58-68]，针对数字图书馆的各方面进行了不同程度的总结和探索。

1.2.3 数据挖掘与数字图书馆评估

数据挖掘是从数据库中发现知识演变而来的 (Knowledge Discovery in Database.KDD)[69]，是一个数据处理的复杂过程，它可以从大量没有规律的数据中发现、抽取、挖掘出未知的、潜在的、有价值的模式或规律等知识[70, 71]。数据挖掘技术是一门跨学科的研究课题，多个领域的知识都在数据挖掘技术中发挥着重要的作用，这其中包括统计学、数据库技术、机器学习技术、模式识别理论、人工智能技术、可视化技术等[72, 73]。

数据挖掘的应用概括起来有：数据分析和预测，该技术主要是用来建立数据集合中待处理属性的预测模型；关联分析，该技术主要是用来发现数据集合中拥有强关联特征的关系模式[74, 75]；聚类分析，该技术的目的是为了发现密切相关的数据组群，依照组群的不同划分出不同的数据集合，使得同一组中的数据在与其他不同隶属集合的数据相比，尽可能的相似[76, 77]；异常检测，该技术的任务主要是识别一个数据集合当中，在某一特征或信息方面显著不同于其他数据的数据点，目标是发现脱离正常态的

异常点[78]。

在本文的研究中，涉及到的数据挖掘技术主要集中在预测分析方面，通过建立馆藏的分数预测模型，达到对馆藏质量评估的目的。下面将对数据挖掘的预测分析相关技术进行较为全面的介绍，通过算法思想、算法优缺点及适用范围等方面对相关分类预测算法进行了分析和总结，为后续章节中数据挖掘的应用打下良好基础。

1.2.3.1 预测分析技术

主要的预测分析算法有决策树算法、KNN 算法、SVM 算法、贝叶斯算法以及基于软计算的挖掘算法，分析如下：

（1）决策树算法

决策树（Decision Tree）是数据挖掘领域应用最广泛的算法之一[79, 80]，数据领域权威的国际学术组织 ICDM（the IEEE International Conference on Data Mining）在 2006 年 12 月评选出了数据挖掘领域的十大经典算法，代表决策树算法思想的 C4.5 算法[81-83]被排在十大算法的第一位，足以说明它受重视的程度。决策树起源于 20 世纪 70 年代后期和 80 年代初期，主要代表算法有：ID3[84-86]，C4.5[81, 83, 87, 88]，CART[89-91]，SLIQ[92-94]，SPRINT[95-101]等。

决策树算法的优点：容易理解，很容易转换为 IF-THEN 这种符合人的思维的阅读模式；算法效率较高，算法的复杂度较小，算法运算量随着数据集合的变大而变化是可控的；算法准确率高；擅长处理非数值型数据，在少量或者没有数值型数据集合当中，非常适合采用决策树的算法。

决策树的缺点也是很明显的：决策树算法中的关键处理模块往往是针对非数值型数据的处理，因此当数据集合中存在数值型数据，就需要进行数据改进算法，这样会降低决策树算法的高效，而且会带来预测结果准确性的降低，尤其是在数据类型更为复杂，例如存在时间顺序时，则需要更多的数据预处理算法，另外当预测的结果类别较多时，决策树算法的准确率也会大大降低，其他诸如数据缺失或数据错误等问题也会影响决策树的有效运作。

（2）KNN 算法

KNN（k-Nearest Neighbor）算法是基于统计分析的数据预测方法，它是基于实例的学习算法[102]，它的特点是理论上较为成熟，算法简单而实用[103]。KNN 算法是在 20 世纪 50 年代早期首次提出的，到 20 世纪 60 年代计算能力取得很大增强之后开始流行起来。

KNN 算法的基本思想是：取出一个记录，找出在数据特征空间中和它最邻近的 k 个记录，若这些记录大部分属于某一个类别的概率，则该记录也属于这个类别。

KNN 算法的优点是：算法思想直观，并且无需先验统计知识，属于非参数的数据预测，而且因为 KNN 算法是依赖于一个记录周围有限数据的样本，不需要直接去判别类域，因此更加适合那些分类结果交叉或重叠较多的数据集合。

KNN 算法的缺点是：计算量较大，一个记录的距离计算往往需要计算它与所有已知记录的距离，即使经过优化的计算也需要较多运算；另外，KNN 算法对于小样本数据集合效果并不好；最后需要注意的是，KNN 算法对于样本类别的平衡性要求较高，否则很容易倾向于选择含有多数样本的类别，一些算法已经有了针对性的完善，例如对类别加权处理。

（3）SVM 算法

SVM（Support Vector Machine）算法，也即支持向量机算法，它的理论基础是统计学习理论的 VC 维理论以及结构风险最小原理[104-106]。

最早出现在 1992 年，当时 Vapnik 发表了第一篇支持向量机的文章，其基础工作在 20 世纪 60 年代已经出现。

统计学习理论提出一种新的策略：结构风险最小化原则（SRM），而 SVM 正是对该统计学习理论的一种算法实现，体现了 SRM 原则。SVM 使用核函数，将数据样本向量映射到高维的特征空间中，在这个高维特征空间中构造最优分类面，达到数据分析的结果。

SVM 算法的优点有：对复杂的非线性决策边界的建模能力比较准确，并对已有经验的依赖较小，通过计算能够获得全局最优解，并具有良好的泛化能力，较为适用于小样本、非线性及高维模式的识别。

缺点就是训练时间长，计算量大，对于大的数据样本实施起来较为困难，而且经典的 SVM 算法理论默认只支持数据的二类分类预测，在实际使用当中，往往需要构造多个分类器组合来解决这一问题。

（4）贝叶斯方法

贝叶斯算法是基于统计的数据预测方法[107-111]。从 20 世纪 90 年代以来，贝叶斯学习算法的发展很快，因为概率统计和数据挖掘有着天然的联系，所以数据挖掘兴起之后，贝叶斯算法就迅速被应用到各种数据挖掘的方案当中[112-116]。

贝叶斯方法有很多优点：算法计算可以发现数据之间的因果关系，这在遇到其他

挖掘算法较难处理的不完整数据集时特别有用；同时，因为可以通过后验结果来修正前面所得到的结果，所以贝叶斯算法的准确率也较高；贝叶斯算法处理的对象属性可以是离散型的，也可以是连续型的。

该算法一个重要缺点在于过分依赖先验信息的正确性，假设先验信息出现错误，结果的准确率就会大大下降；另外在运用贝叶斯方法时，由于要对先验概率进行计算而得到后验概率，所以效率稍差；最后在理论上虽然贝叶斯算法看起来较为完美，但在实际中一般不能直接利用，而是需要知道数据集合确切的分布概率，而实际上常常是无法确切给出数据的概率分布的，所以往往做出某种假设以逼近贝叶斯定理的要求，这自然也会影响到贝叶斯算法的效果。

（5）基于软计算的预测挖掘算法

上面介绍了目前主要的预测分析技术，随着计算机技术的高速发展，一些新的数据挖掘技术不断涌现，这就产生了一个新的技术领域：软计算。软计算是混合的智能化计算方法，它是一种快速搜索较好解的计算方法，并不以精确解为目标。常见软计算包括的计算模式有：人工神经网络、遗传算法、模糊逻辑、粗糙集和混沌理论等。

遗传算法（Genetic Algorithms）本质是是一种有效解决最优化问题的方法，它是基于自然选择和自然遗传机制的搜索算法。遗传算法最早是由美国 Michigan 大学的 John Holland 和他的同事及学生在 1975 年提出的。他们认为世界上大部分生物的基本进化过程可以分为繁殖、变异、竞争、选择四个步骤，遗传算法正是模仿这种过程，它是一种用于复杂系统优化的、具有很强鲁棒性的搜索算法。

遗传算法的算法原理是：首先对要解决的问题进行编码，形成特征字符串来作为生物个体，选择若干个体组成初始种群，并计算各个个体的适应值；然后将这些个体进行进化操作，主要有选择、交叉、变异等处理，从而得到新一代的个体；接着在新一代个体的基础上再进行进化操作，循环进行这个过程，最终得到全局最优解并退出遗传算法。

遗传算法的优点在于：通用性强，理论上可以解决任意高度非线性寻优问题；算法的全局性特点确保群体进化，并得到全局最优解；算法具有很好的可扩展性，容易与其他技术相混合。

遗传算法的缺点在于：技术并非很成熟，仅仅是简单模拟，在算法理论的有些方面还没有得到严格的证明；在算法的精度、可信度、计算复杂性等方面，缺乏有效的定量分析方法；遗传算法运算当中的参数设定很多是依靠经验，缺乏通用性和理论依

据；遗传算法在局部搜索上能力较弱，它能较快接近全局最优解，但是要得到最终的最优解则会很慢。

神经网络算法（Neural Networks）是一种模仿人类神经网络行为特征而进行的数据处理和分析的方法[117-120]，该算法可以进行分布式并行信息处理，神经网络算法模仿神经元细胞对信息的传递和处理方式，根据系统的复杂程度，调整内部大量节点之间相互连接的关系，从而达到对信息的分析处理。

神经网络算法的优点有很多：它能解决内部机制复杂的问题，神经网络算法实现了一个从输入到输出的映射过程，而且利用数学理论已经证明这个映射过程可以实现任何复杂非线性映射，这使得它在解决复杂问题时拥有很强的处理能力；神经网络具有很强的自学习能力，因为它能对含有正确答案的实例数据集学习，然后提取出符合要求的合理求解规则；神经网络具有推广、概括能力。

神经网络算法的缺点有：该算法没有能力来解释自己的推理过程和推理依据；在推理过程当中，算法把问题的特征全部转换为数字，把整个推理过程演变为数值的计算，必然会失去原数据当中的一些信息；神经网络算法在数据不充分的时候就无法进行工作。 因此神经网络技术的理论和学习算法有待于进一步完善和提高。

其他还有很多软计算的数据预测技术，比如粗糙集理论，粗糙集理论是一种研究不精确、不一致和不确定性知识的数学分析工具[121-125]，其基本原理是在保持决策能力不发生变化的情况下，利用等价类的思想对知识进行约简，达到简化决策表的目的，从而得到问题的决策规则，类似的理论还有模糊逻辑[126-134]和混沌理论[135-138]。

1.2.3.2 数据挖掘在数字图书馆评估中的引入

根据本文的综述研究，关于利用数据挖掘技术来进行数字图书馆评估研究的方法，目前在国内外尚无相关的系统报道，本文根据评估工作面临数据海量化、关系复杂化等诸多新问题，将数据挖掘相关技术引入到了数字图书馆评估研究当中。

目前，数字图书馆评估有两种主流的评估思想：一种认为，数字图书馆评估需要依靠领域专家来制定相应评估指标及标准，对数字图书馆进行全面系统的量化评估，围绕这种观点衍生出了很多细化的评估体系；另外一种认为应将所有的评估任务完全交由数字图书馆的使用者来决定，也即不需要专家制定标准进行评估计算，这种思想近些年发展很快，并成为主流评估思想之一。

本研究要将以上两个主流评估思想融合起来，通过第 2 章的数字图书馆评估基础理论研究解决两个思想在评估理论上的分歧，提出了主、客观因素相融合的评估模型，

将客观评估数据和主观评估数据结合起来，需要借助数据挖掘技术实现数据的处理和分析，取得潜在的价值信息。

同时，数字图书馆评估的本质就是对价值的发现和挖掘，这与数据挖掘的目的是一致的，而数字图书馆本身是一个规范数据的海量集合，这又符合数据挖掘技术应用的条件。基于这些原因，本文尝试从数据挖掘的思维去研究数字图书馆评估，建立评估体系，并研制相关的算法。

本文在研究中进行了数字图书馆评估的综合分析，并依据理论研究提出了整个数字图书馆的评估模型，在进一步的实践中，重点围绕处于评估模型中核心位置的馆藏评估开展，研制相关的算法。

数字图书馆馆藏评估的数据集中，属性的个数是相对较少的，而且其中含有较多定性类的属性，由上节的分析可知并不适合 KNN 算法最近邻计算模式，在馆藏评估中以用户评分为类别属性，其数值往往被离散成一定数量，不适合 SVM 倾向低类数目的倾向，贝叶斯算法则过分依赖先验信息的正确性，这对于数字图书馆馆藏的数据来说较难准确获取，遗传算法和神经网络算法在应用中解决的问题往往是高度非线性及复杂性，虽然馆藏属性的关系存在非线性条件属性关系（第 4 章的研究内容），但是其他关系并不复杂，使用软计算的数据挖掘预测算法，反而使问题复杂化。馆藏评估数据大部分都是定性类属性，属性数目也不多，比较适合决策树算法，在决策树算法当中，当数据集中存在定量类数据（即数值型数据），需要进行数据离散，一个好的数据离散算法能使原有的信息量尽可能的保留并发挥作用。因此，本论文在馆藏模块的评估中将围绕决策树算法来开展，并根据实际应用中的问题提出了一种新的数据离散算法以及决策树算法。另外，本文的数字图书馆评估研究是根据馆藏属性对数字图书馆进行价值预测，从这个角度上看，似乎与个性化推荐的工作目标很相近，但是个性化推荐是在复杂多变的环境中进行兴趣度的灵活匹配来进行个性化服务，而数字图书馆评估目的是得到一个稳定的评价，因此两者还是有一定区别的。

1.3 主要研究内容

本论文研究数字图书馆评估基础理论，建立体现主、客观因素相融合思想的评估模型，并对评估模型中核心的馆藏模块设计了基于数据挖掘的评估方法，围绕该方法，研究馆藏属性的复杂关系，引入数据挖掘技术并研制新的数据处理算法和决策树挖掘算法。

具体包括以下四个方面：

（1）对数字图书馆评估理论的研究

研究传统数字图书馆评估在评估思想和评估方法上的成果，对数字图书馆评估的现有情况进行总结；分析研讨数字图书馆评估的一系列基本性质和问题，包括数字图书馆评估客体、评估目的、评估主体；提出一个数字图书馆评估四元理论模型，并论述了模型内部联系以及与数字图书馆评估的关系，该模型作为数字图书馆评估工作的基础理论，体现了本文的评估思想。

（2）提出一种主、客观因素相融合的数字图书馆评估模型（SOI 模型）

对现有评估思想进行了概括，针对两大主流评估思想的不足，提出将两种评估思想融合起来形成一个有机互动的评估模型的想法。提出明确的主、客观因素相融合评估思想，依据对评估客体和评估主体的分析，将评估模型划分为硬性模块和软性模块，并对融合机制进行了论述。依据前面的分析，设计了主、客观因素相融合的数字图书馆评估模型及其体系结构。

（3）围绕数字图书馆馆藏开展属性研究，提出一种基于 z 值的并行离散算法（PDOZ 算法），发现了非线性条件属性的存在。

论述了馆藏质量难以量化的问题，提出将含有用户感受的数据与馆藏的基本信息结合起来，形成包含客观表述属性和主观感受属性的馆藏数据集合，应用数据挖掘技术进行馆藏价值的预测分析，从而将数量、质量有机的结合在一起。提出软属性的概念，在馆藏属性中引入了体现用户感受的主观数据，作为解决馆藏质量量化问题的桥梁；对数字图书馆馆藏数据的复杂属性关系展开研究，提出了一种基于 z 值的并行离散算法（PDOZ 算法），增强了属性与预测之间的相关性。基于这种离散方法研究了馆藏属性之间的关系，发现了非线性条件属性的存在。

（4）提出了一种层叠决策树算法（SDT-Z 算法）

基于馆藏属性的研究，针对数字图书馆馆藏价值的预测挖掘提出了一种层叠决策树算法，解决了传统挖掘算法在数字图书馆馆藏数据集中应用不理想的问题，该算法引入层属性的概念，将决策树选取属性从平面信息的选择拓展到了立体信息，使非线性条件属性对数据挖掘的干扰得到有效转化，并可以嵌套使用已有决策树算法，具备较强的灵活性，解决了数据挖掘在数字图书馆评估应用研究中的瓶颈。

1.4 论文的组织结构

本文的组织结构安排如下：

第 1 章为绪论。本章对选题的研究背景、意义和当前的发展状况进行了描述，介绍了数字图书馆评估和数据挖掘相关技术，讨论了国内外相关研究的现状，给出了本文的主要研究内容和论文的组织结构。

第 2 章围绕数字图书馆评估的评估思想、评估客体、评估目的、评估主体等问题开展分析研究，通过对基本性质的研讨，解决现有评估主流思想的分歧，并提出了一个数字图书馆评估四元理论模型，在四元模型中论述了内部联系以及与本文数字图书馆评估思想的关系，本章是整个论文研究的起点，也是第 3 章主、客观因素相融合数字图书馆评估模型的理论基础。

第 3 章提出了一种主、客观因素相融合的数字图书馆评估模型，详细论述了评估模块的划分、主客观因素相融合的机制以及评估模型的体系结构。

第 4 章针对数字馆藏中的属性展开了研究，引入了体现用户意愿的软属性；在属性研究中提出了一种基于 z 值优化的连续型数据离散方法，用有概率分布意义的动态距离代替传统离散方法，反映了馆藏各个属性语义的动态变化，增强了属性与预测之间的相关性，并基于这种离散方法研究了馆藏属性之间的关系，发现了非线性条件属性的存在。

第 5 章提出了一种层叠决策树算法，该算法消除了数字图书馆馆藏属性中的复杂条件属性关系的影响，解决了现有挖掘算法在数字图书馆馆藏数据应用不理想的问题。该算法思想可嵌套已有的决策树算法从而有较强的灵活性。

第 6 章对论文的内容进行了总结，结合本文的工作基础，对下一步工作进行了设想。

第 2 章 数字图书馆评估理论研究

2.1 引言

数字图书馆评估在近些年发展很快，产生了很多评估研究项目，各个研究对数字图书馆评估的理解以及处理的角度都有所不同。例如，有研究按照数字图书馆服务对象的不同来构建不同的评估体系，而有研究依照数字图书馆的功能模块的不同来构建评估体系。

可以看出，数字图书馆评估工作的具体方式和成果往往取决于该研究对于数字图书馆评估基本理论的认知理解。例如，依据什么来拆分评估体系，使用什么样的评估主体才能更好的进行数字图书馆评估，如何确定数字图书馆评估应该涉及的范围等。

本章是围绕数字图书馆评估开展的综合分析与研究，目的在于把握数字图书馆评估基本理论的特点，为提出科学合理的数字图书馆评估理论模型打下良好基础。本章首先总结了现有数字图书馆评估研究的特点，剖析了数字图书馆的定义问题，在此基础之上研讨了数字图书馆评估目的和评估主体的分歧，基于这些研究提出一个数字图书馆评估四元理论模型，并论述了模型内部联系以及与数字图书馆评估的关系，体现了本论文的评估思想，也为后续研究提出主、客观因素相融合评估模型提供了依据。

2.2 数字图书馆评估思想的分析

目前在国内外尚未建立一个得到一致认可的评估思想体系[139]，本章首先对一些具有代表性的评估思想进行评析。

（1）以评估主体不同而进行的不同指标量化评估体系

该评估体系[51]认为数字图书馆系统构成复杂，如何对主客体之间的相互关联进行准确测量，是数字图书馆评估首先要解决的问题。这种评估体系采用的数字图书馆评估手段以分级量化评估指标为主，将评估主体划分为国家、组织和个人，依照不同的主体制定相应评估指标。

该评估体系的优点在于：认识到评估主体的侧重点不同会带来评估结果的不同，具有一定的全面性。其缺点在于：这种思想在其评估模型建立当中不容易体现出来，另外无论是哪个主体，其本身都是数字图书馆的使用者，所以从根本的意义来说，所

有的评估主体都是用户，不同用户的不同评估目的和意图，这是需要重视的多角度评估的问题，但据此拆分评估主体而将评估过程分开，造成评估活动的复杂化，并不能准确的体现数字图书馆的价值。

（2）以传统图书馆的评估方法为指导的延伸评估体系

这种评估体系是针对传统图书馆的评估方法和信息系统评估标准进行适当修改，延伸到数字图书馆的评估上。

该评估体系的优点在于：因为数字图书馆是来源于传统图书馆，因此与传统图书馆有一些本质上的相同特点，于是采用已有的对传统图书馆评估的成果，可以很容易的找到切入点。该评估体系按照传统图书馆的评估标准，划分评估指标，操控比较容易，在基于传统图书馆构建的数字图书馆评估中具有较好指导意义。其缺点在于：存在一些主要用于传统图书馆服务评估的标准，并不适合数字图书馆评估，比如外借数量指标，同时数字图书馆的一些特殊属性也无法得到充分重视和评估，因此难以全面的表达出数字图书馆的特点，无法形成真正准确高效的评估体系。

（3）多方法反复以用户为中心策略的评估体系

John T. Snead 在 2005 年 11 月的“Annual Meeting of the American Society for Information Science & Technology”上提出了一种多方法反复以用户为中心的策略来进行数字图书馆评估[140]。该评估体系的核心内容是从“功能性，可用性，可操作性”三个层次对数字图书馆评估工作进行全面覆盖，涵盖了“分类能力，限制性搜索，搜索类别，再次精练检索”,“导航水平，内容展示的方式，标签等辅助功能，搜索能力”,“内容的转化形式（服务于眼睛和耳朵有障碍的用户），颜色的独立（服务于色盲用户），清晰简单的导航机制，放大设备等等”。

该评估体系的优点在于：站在三个角度上进行数字图书馆的评估，提出了一些独特的指标，针对特定用户群体具有良好的效果，有非常好的借鉴意义。其缺点在于：评估中以独特用户为重点，并没有扩展至整个用户群体，并将其融入到更加全面的评估体系中，另外该评估体系从功能性，可用性和可操作性上进行分类分析，但是三者之间没有量化比重关系。

（4）基于组件的评估体系

Ingrid Hsieh-Yee[141]提出来的一种基于组件的数字图书馆评估框架，其核心内容为以数字图书馆的基本操作作为划分依据，进行对应的剖析和评估。针对数字图书馆的三个主要对象“用户、内容和服务”进行了评估体系的分配，三个对象之间的关系是

“用户使用内容，内容在系统里面被检索，系统服务于用户”，利用这三者的关系确定了整个数字图书馆的评估体系。该评估体系通过科学分析，以用户、内容和系统服务之间的关系有效的将各部分评估工作有机组织在一起。

该评估体系的优点在于：评估体系清晰，模型简单，实现了借助三个对象的研究而涵盖评估工作的各个细节的目的，在数字图书馆建设过程中具备很好的指导作用，建设中即可进行评估和实时指导。其缺点在于：因为一开始就站在专家的分析立场来拆分整个数字图书馆的体系结构，因此在衡量自身服务系统的时候，将会出现偏差。

（5）DigiQUAL/LibQUAL+®评估体系

DigiQUAL 是以 LibQUAL+®为基础发展而来。LibQUAL+®本身既可以对传统图书馆进行评估，又可以稍加变通而应用于数字图书馆评估当中，而 DigiQUAL 一个专门研究 LibQUAL+®在数字图书馆方面应用的项目。

LibQUAL+®是近年来颇有影响的一个图书馆评估项目，这种思想的核心是完全站在用户的角度上进行分析和评估，其理论源于 SERVQUAL 理论。该评估的理论依据来自于美国市场营销学家 L.L.Berry、A.Parasuraman 和 V.A.Zeithaml 依据“全面质量管理”（TQM）理论提出的这种服务质量评估方法。其理论基础是“服务质量差距理论”——“服务质量”是“用户感受到的服务水平”与“用户所期望的服务水平”之间的差距。该评估体系通过科学分析和调查总结得出调查表项目的内容，依据“全面质量管理”制定调查表上每个问题的答案选项，然后通过 Web、Email 等方式发布，由用户自己选择对各项内容的满意程度，随后统计归纳，得出评估结果和有益信息。在强调数字图书馆服务价值的评估中，该体系具有相对省时省力却准确有效的评估价值。

该评估体系的优点在于：改变了以往以专家为中心开展工作的评估模式，专家在评估数字图书馆时，关注的是投入评估、过程评估，而 DigiQUAL/LibQUAL+®评估体系关注的是结果评估，结果评估能够表明一个数字图书馆向用户提供服务的水平，直接而准确的体现出数字图书馆建设机构的办公效率和工作效果，因此，DigiQUAL/LibQUAL+®评估体系通过服务对象的主观感受检验服务的成效的思想是值得肯定和借鉴的。其缺点在于：完全按照数字图书馆用户的主观感受来评估，并不能完全涵盖数字图书馆的各个方面，而且调查表内容的设计和层次划分，缺乏足够的理论验证，在发展的过程中经历了诸多版本的更新和完善，至今仍在不断研究和变化当中。

其他还有一些评估思想和方法，比如 AHP 法在数字图书馆综合评估中的应用[142]——这是一种根据数字图书馆的不同模块建立综合评估模型，采用层次分析法研究数字图书馆的综合评估方法，在理论上往往属于上述代表性思想之一，只是方法上略有不同。

综上所述，可以将国内外数字图书馆评估研究分为两种思想。第一种思想是以领域专家的指标评估为主，这些研究的共同点是采用评估领域的常用方法，结合数字图书馆的结构特点，对数字图书馆的各项指标进行专业打分，不同点在于指标体系的框架纷繁复杂，细节多有不同，更有一些评估研究选取的评估角度新颖，形成了独特的评估体系，但都是基于专家的指标体系为评估驱动；第二种思想是完全抛开专家指标打分，使用用户调查表等多种方式，让用户打分，进行汇总分析，以用户的主观感受为评估驱动。

第一种思想的特点可以概括为评估标准多样化。虽然各个研究者所站的角度和层面不一样，所得到的评估体系指标各不相同，但原理是一样的，就是依据某种评估角度，拆分评估对象，针对数字图书馆的各项情况进行综合分析和考评。在评估指标的制定上，后出现的评估体系往往包含了先出现的评估体系，同时也会适当的剔除一些不合理或者过时的东西，但因为数字图书馆技术更新很快的特点，几乎不可能出现一套稳定的指标评估方法。

这一种评估思想的本质是站在专家专业分析的角度，对数字图书馆的各个方面进行分析评比。因为各个专家的侧重点不同，针对同一个评估方面，采用的指标都或多或少有所争议。

第二种思想的特点是将数字图书馆的各项情况以调查表的形式发给用户，由用户根据自己的感受打分，以 DigiQUAL/LibQUAL+®思想为代表。

这一种评估思想的本质是站在用户主观感受的角度，针对数字图书馆的各项情况进行打分，这样针对数字图书馆的任何一个方面的评估结果，几乎不存在争议，因为专家没有决定权。这种完全站在用户主观感受的角度上进行数字图书馆评估的思想，因为将最终用户作为评估的主体而具有先天优势，其价值和意义得到了较为广泛的认同。

两种思想的相同点：评估涵盖的对象都很全面，比如“馆藏资源，数字图书馆相关技术，数字图书馆管理，针对用户的服务等等”，只是所采用的评估角度不同，评估方法也因此有所区别。

两种思想的不同点：一个是以专家的指标体系为评估驱动，另外一个是以用户的主观感受为评估驱动。

目前数字图书馆评估领域中，虽然以用户为中心的评估思想得到认可，但是以专家为中心的评估思想依然有其优势之处。现阶段两种评估思想都在不断的发展和演变，成为目前数字图书馆评估领域的两个主流评估思想。

2.3 数字图书馆评估客体的研讨

根据评估理论，完整的评估过程包括三个研究对象，分别为评估主体、评估客体以及评估目的，而依据数字图书馆发展很快并没有一个明确定义的特点，我们从评估客体展开论述以确定客体的基本性质，进而分析评估目的的问题，最后论述评估主体。

数字图书馆评估客体就是数字图书馆，而数字图书馆的定义，决定了数字图书馆评估研究的内容[143]，数字图书馆概念的内涵将制约评估研究的评估主体、评估模型等一系列问题的定性和决策，而数字图书馆概念的外延会影响到评估研究考察的范围。

数字图书馆的概念从正式提出至今，并没有一个公认的统一定义，那些较为严谨的定义，大都是由具体研究项目的课题小组提出来的，某种意义来说，正是数字图书馆定义的不明确导致了目前数字图书馆评估研究的很多问题和困难[47]。以下为一些比较著名的数字图书馆的定义[144]。

密执安大学的研究人员在 1990 年首次提出了“数字图书馆”的概念，并将其定义为：数字图书馆是若干联合机构的总称，它使人们能够智能地（intellectually）和实实在在地（physically）存取全球网络上以多媒体数字化格式存在的、为数巨大的且仍在不断增多的信息。

美国研究图书馆协会（ARL）给出的定义是：数字图书馆不是一个单独的实体，不能独立存在，也就是说数字图书馆是一个无限的信息空间；是以数字形式有序地存储信息，而不像现在因特网上信息组织和存储杂乱无章；存储的数字化信息有效地传播与利用，是数字化图书馆强调信息利用的关键；用户使用统一的界面在任何地方、任何时候均可获得信息。

1997 年 3 月由 NSF 赞助的关于“分散式知识工作环境”专题讨论会议报告指出，数字图书馆的概念不仅仅是一个有着信息管理工具的数字收藏的等价词，更是一个环境。它将收藏、服务和人带到一起以支持数据、信息乃至知识的全部流程，包括从创

造、传播、使用到保存的全过程。

中国国家图书馆对数字图书馆的定义是：数字图书馆是采用现代高新技术所支持的数字信息资源系统，是下一代因特网上信息资源的管理模式，它将从根本上改变目前因特网上信息分散不便使用的现状。

此外，还有很多科研机构给予了数字图书馆的定义，综合来看，几乎每个定义都是比较科学完整的描述了当时的数字图书馆状况，而随着时间的推移，随着数字图书馆的发展和内涵的丰富，越来越多的新的定义涌现出来。定义是对事物性质的论述，若一个“定义”无法对事物的本质进行准确的定性表述，那么它就不能算是一个严格意义上的定义，而应该是一种特征描绘和总结。本文从逻辑学的角度，提出了一个数字图书馆的定义。

在逻辑学中，概念的逻辑结构分为“内涵”与“外延”，内涵是指一个概念所概括的思维对象本质特有的属性的总和；外延是指一个概念所概括的思维对象的数量或范围，故本文给出定义如下：

（1）数字图书馆的核心内涵是一个知识服务系统，它唯一的目的就是更好的为用户提供知识信息服务。

（2）数字图书馆的外延是一个基于数字化知识组织和计算机网络的服务系统，它立足于数字化信息，通过任意只要合理的服务方式，与社会、经济、政治和生活有着密切的关联作用。

这个定义对数字图书馆性质的描述，排除了那些只是外在表现特点的，随着技术发展动态变化的，不隶属于本质性质范畴的诸多辅助因素，包括知识特点、服务方式、涉及内容等等因素，从根本上明确数字图书馆的性质。通过该定义，能够正确取得数字图书馆评估相关问题的分析依据。

2.4 数字图书馆评估目的的研讨

数字图书馆的目的是在其能力范围内，最大限度的发挥知识的供给服务，因此数字图书馆评估的目的可以被唯一的确定为：检验数字图书馆目的的完成程度，并给予有利于完成数字图书馆目的的指导。在实际的研究当中，由于其应用背景的干扰，数字图书馆评估目的是一个容易出现混淆的问题，关于评估目的常常出现不同的观点。“多角度推导多目的”思想是其中的一个代表。

“多角度推导多目的”思想认为在某些状况下，数字图书馆评估的目的不能一概

而论，比如分别对数字图书馆经济效益和安全服务进行考察时，两者所关注的评估目的就是不同的，也因此有学术研究提出根据不同的角度对数字图书馆评估进行研究，认为不同的角度下数字图书馆的评估目的是不同的，也就是“多角度推导多目的”理念的思路。这个问题属于对研究层次混淆的范畴，数字图书馆作为一个复杂的知识服务系统，问题之间并不一定就是并列关系，可能存在纵向层次的区分和隶属关系。“多角度推导多目的”的问题，正反映了问题层次的区分和隶属关系。

本文研究认为不同角度下评估目的存在多样性并不为错，也并不和数字图书馆评估目的的唯一性相矛盾，因为两者并不处在一个研究层次，前者隶属于后者，也服务于后者。数字图书馆的评估目的是唯一确定的，而考察角度不同的时候，其实就已经发生了前提和性质的变化，不再是“数字图书馆评估”，而演变成了“数字图书馆经济效益的评估”、“数字图书馆安全的评估”等等，使问题发生了层次上的改变，而且，这些具体角度的评估也自成系统，也会涉及到很多关联因素，同时大部分关联因素和整体“数字图书馆评估”有重合，很容易产生困惑和分歧，需要有区分的正确对待。

综上所述，本文所指的数字图书馆评估的目的是唯一的，即“检验数字图书馆目的的完成程度，并给予有利于完成数字图书馆目的的指导”。

2.5 数字图书馆评估主体的研讨

基于评估理论的指导可知整个评估行为由评估主体、评估客体、评估目的构成，评估主体是评估行为的执行者，往往决定了数字图书馆评估思想的策略。

结合 2.2 节对国内外评估思想的分析和总结可知，目前主要有两大研究流派，一方是以专家专业化指标体系进行评估，另一方是以用户主观感受数据进行评估，通过对两种评估思想实质的分析，可以发现双方就是在评估主体的确立上产生了分歧，一个是以专家为评估主体，一个是以用户为评估主体，并坚持了评估主体唯一性的原则。本文认为目前这两种唯一确定评估主体的思想都是有局限的。

当评估主体为用户的时候，虽然数字图书馆的最终目的是服务于用户，用户通过感受服务评分有着较为准确的检验效果，但用户有两个方面的问题：（1）感受局部的问题，（2）感受错误的问题。感受局部的问题，用户是以个体出现的，存在个体感受局部、狭隘的特点，容易缺少对比如数字图书馆的整体经济影响、社会影响等与用户自身相关性弱的因素，但这些影响却是数字图书馆定义的外延所决定的评估研究必须考察的范畴；感觉错误的问题，比如馆藏的问题，一个馆藏 10 万和馆藏 100 万的大

众性数字图书馆 A 和 B，可能针对某个领域的藏书是不同的，比如 A 有 2 万册计算机类资源，而 B 只有 1 万，那么在满足计算机领域的用户的时候，他们的感觉就是 B 比 A 的馆藏还要丰富，这显然是错误的。

当评估主体是专家的时候，由于评估主体和评估服务对象相脱离，缺乏真实的实践感知，造成实施评估方法、划分评估指标等方面存在较大的难度，比如对于馆藏质量的评估，单单依靠专家来制定指标是很难真实反映出来的，而此时采用基于用户感受的评估方法则会既容易又准确。

无论是以专家为评估主体，还是以用户为评估主体，两种评估体系都各自坚持了主体唯一性原则。在这种思想影响下的互相借鉴中，任一方必然会将对方的优点放在一个非主体策略来对待，尝试在己方主体的立场下变通以发挥出对应效果，而对方的方法在非主体的策略上自然就丧失了其优势所在，无法充分发挥作用。

本文认为两个评估主体并不矛盾，可以将两者结合起来，围绕具体的评估需求而合理的选择使用，该观点不同于目前评估主体唯一性原则下两种评估体系的互相借鉴，而是一个整体运作体制中两个局部机制的协作。

2.6 数字图书馆评估四元理论模型

在数字图书馆评估的分析与研究当中，涉及到了一些基础学科的理论支持，很多评估思想都不可避免的使用到了这些理论知识，这些理论是构建数字图书馆评估体系不可或缺的考虑因素，有着重要的指导作用。

本文将这些理论及之间的关系进行归纳总结，提出一个数字图书馆评估的四元理论模型，并在这个四元理论模型中介绍本文所提出的评估思想。其中四个数字图书馆元理论包括：数字图书馆理论、评估理论、认知心理学、系统论。四元理论模型如图 2.1 所示。

本文将数字图书馆评估作为一个系统的学科事物进行研究，其学科世界观的组成元素就是数字图书馆理论、认知心理学、评估理论，而学科方法论则是系统论。

（1）数字图书馆理论

数字图书馆理论是整个评估研究首先需要明确的环节，一方面，它界定了数字图书馆定义的内涵和外延，决定了数字图书馆评估研究对象的性质和范畴，另外一个方面，其内部技术和管理组织方式是后期评估工作开展的依据。

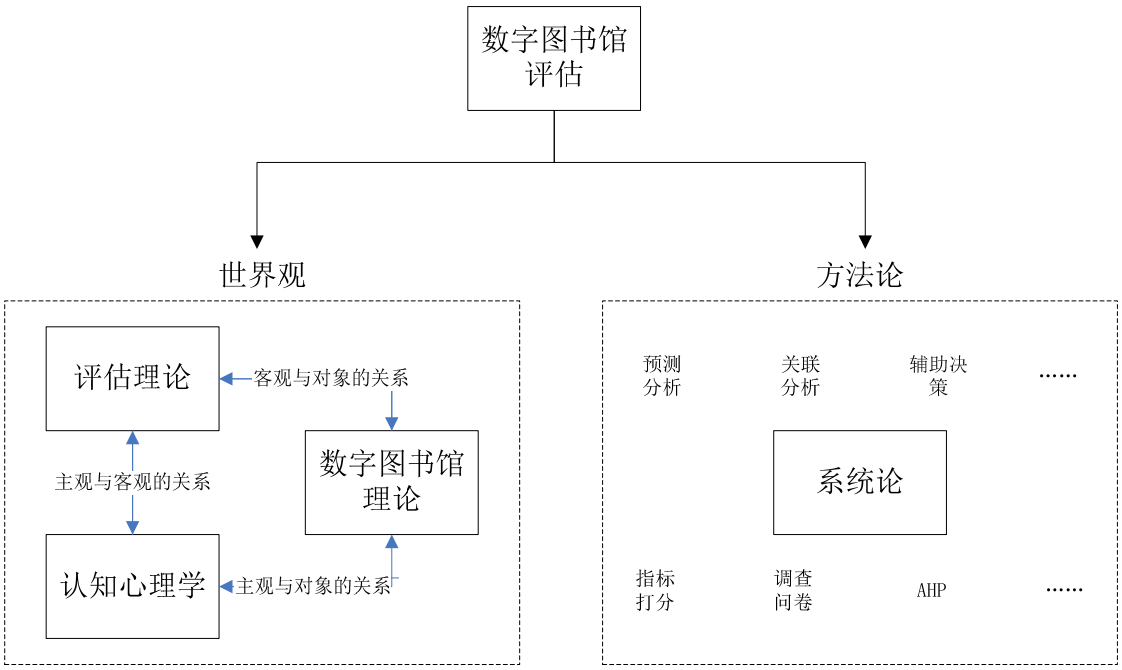


图 2.1 数字图书馆评估四元理论模型图

（2）评估理论

通过评估学科理论的研究对评估行为进行科学的分析，辨明一系列数字图书馆评估研究相关的性质问题，包括评估目的、评估常用方法、评估主体和评估客体等。

（3）认知心理学

数字图书馆的建设主体、服务客体、数字图书馆评估的评估主体、评估客体以及其他细节都体现了人的认知因素，评估研究中引入人的认知科学是学科发展必然的需求，其合理应用是解决目前评估客体分歧、评估模型构建等问题的关键之一。

认知科学使人的思想、行为、心理等认知过程具体化和科学化，认知理论是认知主体、认识对象以及把它们联系在一起的信息所构成的认知运动模型，对于数字图书馆评估研究来说，就是评估主体、评估客体以及双方围绕评估目的而联系作用所构成的认知运动模型。

（4）系统论

系统论在基础理论的研究具有很重要的指导价值，系统论是 20 世纪科学领域发生的一场革命性的变革，它作为重要的方法论基础几乎已被应用到所有的自然科学和社会科学领域，系统论的任务，不仅是认识系统的特点和规律，反映系统的层次、结构、演化，更主要的是调整系统结构，协调各要素关系，使系统达到优化的目的，这些都很符合数字图书馆评估研究的需求。

接下来对四元理论模型进行解构分析，研究它们与数字图书馆评估思想之间的关系。

其中三元理论交织组成的学科世界观与数字图书馆评估思想之间的关系如图 2.2 所示：

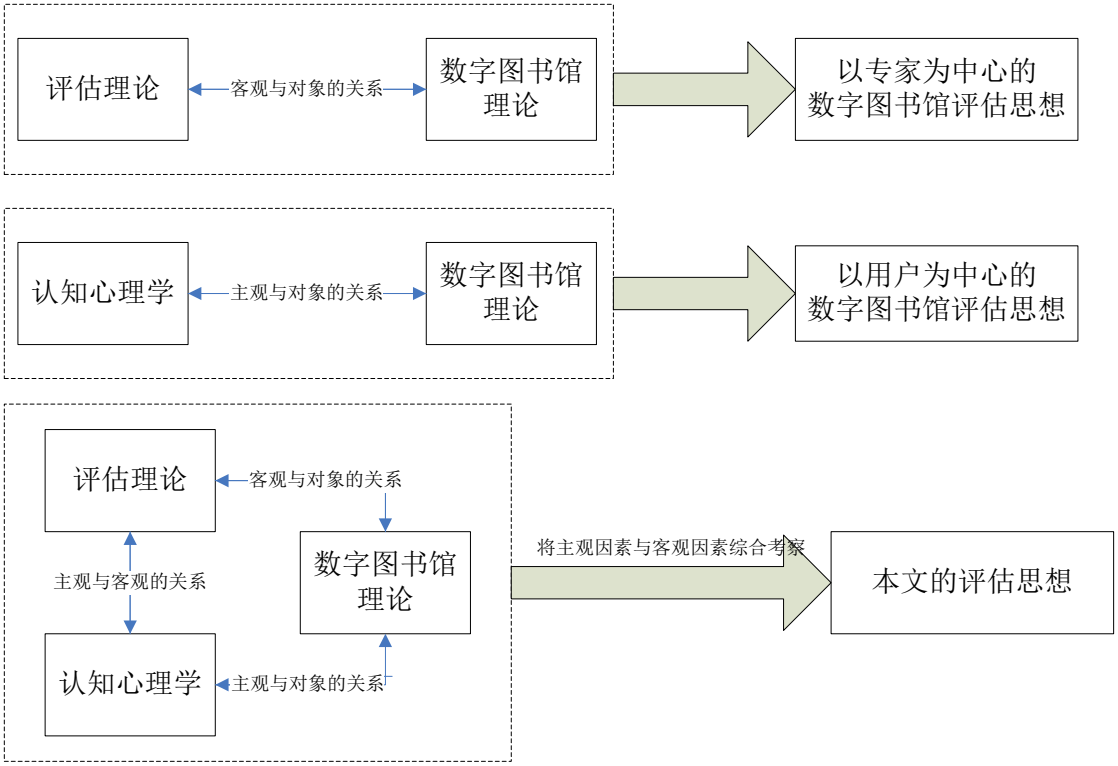


图 2.2 三元理论交织组成的学科世界观与数字图书馆评估思想的关系图

在认知心理学、评估理论、数字图书馆理论三元理论内部：

认知心理学的研究对象是人，其本质是数字图书馆评估中主观因素的研究；评估理论的研究对象是评估学科的指导理论和方法，其本质是数字图书馆评估中客观因素的研究；而数字图书馆理论是关于数字图书馆评估的研究对象的理论，其本质是对数字图书馆评估中评估对象的内涵和外延的界定。

评估理论与数字图书馆理论的结合是基于客观因素的数字图书馆评估研究基础，其代表就是以专家为核心的评估思想。

认知心理学与数字图书馆理论的结合是基于主观因素的数字图书馆评估研究基础，其代表就是以用户为中心的评估思想。LibQUAL+®的核心研究就在于调查表的设计，而调查表内容正是在 LibQUAL+®项目围绕用户认知心理深入研究中不断演变。

在 2.5 节中对“数字图书馆评估主体”的论述中，已经阐明了本文对评估主体应

该以合适的方式进行结合的思想，反映到代表学科世界观的三元理论中，就是要将三个数字图书馆元理论结合起来，选择合适的方式将主观因素和客观因素自然的融合起来。

这种基于三元理论相结合的评估思想，正是本文所要提出的主、客观因素相融合评估思想，这种主、客观因素相融合的意义，在评估思想上体现在：突破评估主体唯一性的隐性约束将两者融合在一起；在技术实现上体现在：将数据挖掘的方法引入到数字图书馆评估中，依据具体实践的需求，将主观数据和客观数据融合起来表达出信息的价值。本文在第 3 章中将提出的评估模型就是以三元理论为思想指导，形成将客观因素和主观因素自然融合的数字图书馆评估模型。

学科方法论与数字图书馆评估思想的关系如图 2.3 所示：

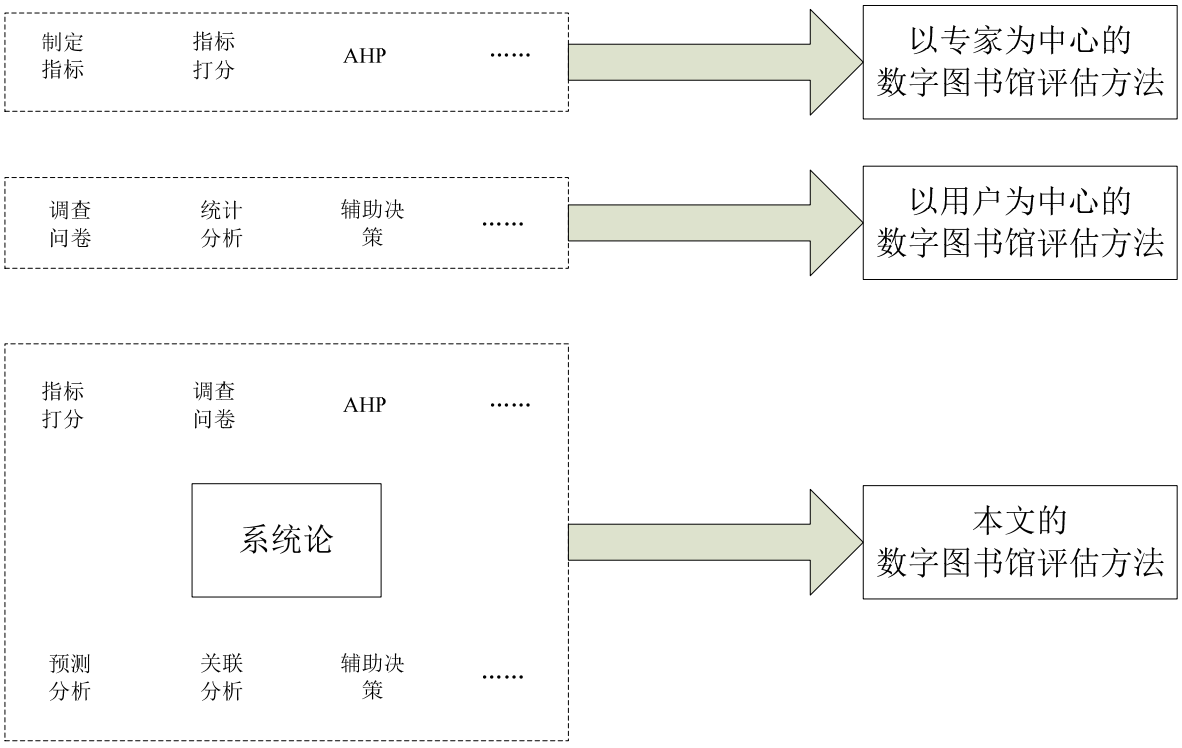


图 2.3 学科方法论与数字图书馆评估思想的关系图

系统论的研究目的在于调整系统结构和各要素关系，使系统达到最合理状态，其核心思想是寻找并建立以一定结构形式联结构成的具有某种功能的有机整体，其实质是方法的综括和指导。

在数字图书馆评估中使用系统论的思想，就是研究可用于数字图书馆评估的各种方法，并确立这些方法的运用方式和应用原则。

在以专家为中心的评估思想中，常用方法有指标制定、指标打分、层次化方法、

结构-功能分析法、信息计量法、测试法、过程评价法等等。

在以用户为中心的评估思想中，常用方法以问卷调查、网络调查为主，其核心在于调查问卷内容和统计分析策略的设计。

在主、客观因素相融合的数字图书馆思想中，拓展了方法的范围，基于前文所说学科世界观三元理论的指导思想，在将主观因素和客观因素相结合的技术实现中，会采用到基于数据挖掘的方法，对于主、客观因素相融合的数字图书馆评估体系，所有的方法都是可以使用的，具体的方法运用方式和应用原则将依照实际的需求研究而定。

总之，四元理论模型是能够将主观因素与客观因素自然融合的一种数字图书馆评估理论模型，有利于准确、全面实现数字图书馆评估目的。

2.7 本章小结

本章是围绕数字图书馆定义、数字图书馆评估目的以及评估主体等一系列问题开展的分析研究，通过对一系列基本性质的研讨，解决现有评估主流思想的分歧，提出了一个数字图书馆评估四元理论模型，在四元理论模型中论述了其内部联系以及与现有数字图书馆评估思想的关系，并初步说明了本文的数字图书馆评估思想。

本章的研究是整个论文的起点，通过理论模型的研究，对现有图书馆评估进行分析、梳理，并深入理论研究，构建一个较为系统的数字图书馆评估理论基础，也为第3章主、客观因素相融合数字图书馆评估模型提供了依据。

第 3 章 数字图书馆评估模型研究

3.1 引言

随着计算机技术的快速发展和应用，数字图书馆的特点在不断发生变化[145-155]，数字图书馆评估研究也随之不断的演化和发展。目前数字图书馆评估的基本思想主要分为两大流派：以专家为中心的评估和以用户为中心的评估。依据四元理论模型的分析可知，以专家为中心的评估是一种客观因素与评估对象相结合的评估方式，以用户为中心的评估是一种主观因素与评估对象相结合的评估方式。

本章提出一种将主、客观因素相融合的数字图书馆评估模型（Subjectivity-Objectivity-Integration Model of Digital Library Evaluation），简称 SOI 模型。由前面章节的分析为依据，SOI 模型通过对评估主体和评估客体的合理处理来兼顾目前主流评估思想的优点，弥补各自的不足，并详细论述了评估模块的划分、主客观因素相融合的机制以及评估模型的体系结构。

3.2 代表性评估模型的分析

3.2.1 基于专家指标的客观评估模型代表

基于专家指标的客观评估是根据领域专家对数字图书馆各个方面进行指标制定并打分，不同研究的指标体系框架细节多有不同，这些不同有些是对同一评估对象制定考核指标的分歧，有一些则是因为评估采用的角度不同，而产生的指标体系的多样化，但归根到底都是基于专家的指标体系为评估驱动。

本节通过对国内外相关研究中较有代表性的项目进行简要介绍，来说明这种基于专家指标的客观评估模型的特点。

国内数字图书馆的发展往往依托于传统图书馆的基础，所以高校是数字图书馆发展较为前沿的地方，有专家在高校数字图书馆建设中提出了数字图书馆评估的综合指标体系[156]。

该评估体系分为三级指标，其中的指标件的权重关系是完全依靠专家对具体指标进行调查、分析和判断，具体的指标体系模型如表 3.1 所示。

表 3.1 国内数字图书馆评估代表指标体系[156]

一级指标	二级指标	三级指标
技术	网络设施	局域网、网络综合布线、网络设备
	服务器	图书馆集成系统服务器、主页服务器、数据库服务器
	存储	存储设备、存储容量、存储技术
	软件	操作系统、数据库管理系统、集成管理系统及其他应用软件
	人员配置	自动化建设队伍、人员培训
资源	电子书	电子书数量、质量、可获取性
	电子期刊	电子期刊数量、质量、可获取性
	学位论文	博硕士学位论文数量、质量、可获取性
	自建	自建数字资源，如学位论文、学科导航等
	其他资源	文摘数据库、引文数据库、免费试用数据库等
管理	经费投入	是否列入预算、是否按预算执行、数字资源投入、软硬件设施投入
	管理制度	数字图书馆工作规范和工作细则，包括：书目数据库规范、设备管理制度、网络管理、数据库备份制度等
	办公自动化	图书馆人事、档案、业务、科研、经费、固定资产等计算机管理状况
服务	网站服务	网站服务内容、界面设计、服务水平、更新频率、访问量
	电子阅览室服务	服务内容、服务方式、服务效果
	网络用户教育	对用户进行图书馆利用宣传和培训情况、对用户进行如何使用网络数据库的宣传和培训情况
	数字参考咨询	业务量、服务时间、服务方式、服务效果、服务范围等
用户	教师感知	教师用户满意度
	学生感知	学生用户满意率

国外有代表性的评估项目有[157]：①美国存储工程对 Perseus 数字图书馆的评估，②数字化信息定价项目，③博物馆教育功能开发计划，④康奈尔大学图书馆和博物馆馆藏数字化项目，⑤科技社会结构理论，⑥加利福尼亚大学 ADL 项目，⑦加利福尼亚大学 DLI-Ⅰ项目，⑧伊利诺大学 DLI-Ⅰ项目，⑨DLI-Ⅱ的子项目 ADEPT，⑩地球科学教育数字图书馆评估，这些项目大部分在第 1 章已作简要介绍，有研究[157]对这些代表性项目中的重要指标进行了总结，形成了对比列表，如表 3.2 所示。

表 3.2 国外常用指标体系[157]

一级指标	二级指标	使用该指标的项目
系统性能	访问途径、网络通信能力	②④
	物理设施、技术先进性、系统安全性、可维护性	③
	可用性	④⑥⑨
	系统易用性	①③
	馆藏保存	④
系统功能	界面、浏览、检索、索引、用户指南	①③
	馆藏范围、数字化记录的质量和准确性	①④
系统服务	用户需求	⑩
	用户日志	⑥
	用户评价（用户的行为、习惯及社会环境、资源利用情况）	①⑥⑦⑧
	电子期刊出版商评价	②
	与其他校园服务的共享	③
系统使用效益	价格模型	②
	投入与产出、经济的可行性	③
	馆藏资源对教育所起的作用、法律问题	④
	数字图书馆的实用性、对学术成果产生的贡献、在教育环境中的价值	⑨

基于专家指标的客观评估无论指标体系如何不同，其分值的确定都体现了领域专家的客观分析，在某些方面能做出准确的数值度量，例如网络通信能力，但在另外一些方面，却难以取得好的效果，例如系统易用性。

3.2.2 基于用户感受的主观评估模型代表

基于用户感受的主观评估是完全站在用户感知的角度对数字图书馆进行评估，在整个过程中并没有专家进行干涉，工作人员仅仅负责统计用户调查的数据，进行整理和分析。

以用户主观感受为评估驱动的代表是 LibQUAL+®项目，该项目应用范围很广，它既适用于传统图书馆，也适用于数字图书馆，而且基于它的思想 ARL 提出了进行优化的 DigiQUAL 项目。本节主要围绕 LibQUAL+®项目来阐述这种评估思想。

LibQUAL+®是美国研究图书馆学会与德克萨斯 A&M 大学图书馆在 SERVQUAL 基础上建立的图书馆服务质量评估模型，LibQUAL+®项目从用户使用感受的角度出发，评价数字图书馆的各个方面，确定需要改进的内容和方向，真正体现了数字图书馆以用户为中心的人本位服务理念[158]。

据 ARL 统计，截至 2009 年 LibQUAL+®已经被应用到世界范围内共 1176 个机构图书馆，支持 17 种语言，共有 23 个国家涉及其中，是目前全球高校图书馆中应用最为广泛的服务质量评价工具[159]。从 2000 年发展至今， LibQUAL+®已经形成相对完善的理论体系。

LibQUAL+®的评估模型的流程并不复杂，如图 3.1 所示：

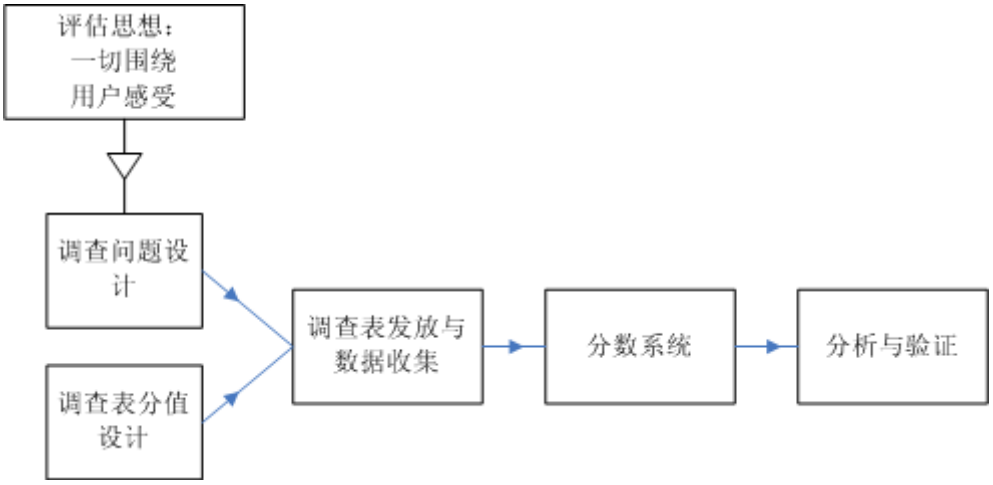


图 3.1 LibQUAL+®评估模型流程图

LibQUAL+®核心内容在于用户调查表的设计，LibQUAL+®发展的变革可以说就是调查表内容的变革，它体现了该项目组对于基于用户为核心评估的研究成果。

LibQUAL+®几乎每隔一段时间都会调整调查表的内容以及处理的策略[160]，最初该项目将 SERVQUAL 的 22 个问题扩展了 19 个问题，分别来自于资源获取、图书馆

环境两个方面，后又扩展了 15 个问题，分别来自于图书馆指导和用户自助服务两个方面，使 LibQUAL+的核心问题数量达到 56 个，随后慢慢由繁变简，发展到较为稳定的 22 个核心问题。目前 LibQUAL+®的最新成果是 LibQUAL+®精简模型，该模型要求所有参与调查的用户回答一些选定的调查问题，余下的调查问题仅被用户随机选择回答。

LibQUAL+®分数系统是很严谨的，它定义了一种解释 LibQUAL+分数的方式：容忍区[158]，容忍区被定义为针对一个调查项的最低可接受的服务水平和理想服务水平的距离，在此基础上更好的表达出用户的感受；分数系统建立了分数规范，该规范是一个单一数值或数值分布，组成一个给定群组的典型表现[161]，分数规范可以把图书馆的平均分数转换成百分等级，通过测量一段时间的分数规范，可以对单个图书馆进行纵向比较，也可以进行机构之间的比较。

3.3 主、客观因素相融合的数字图书馆评估模型

以专家为中心的评估模型和以用户为中心的评估模型之间的分歧在于评估主体的不同，而且双方带有明显的评估主体唯一性原则，这个问题在前面的章节已经得到研讨和解决，评估主体并不具有唯一性和排他性，两者可以不加排斥的结合应用，关键问题是如何协调工作，如何体现各自的优点，将两种评估思想融合起来形成一个有机互动的评估模型。

3.3.1 SOI 评估模型

对本章所提出的主、客观因素相融合的数字图书馆评估模型，从以下两个方面来论述。

一方面，关于评估主体。本文已经明确了在数字图书馆评估当中，唯一不变的是评估目的，其他评估因素都是可变的、服务于评估目的，也验证了当评估主体分别为专家和用户时各自的问题，因此，SOI 评估模型中将评估主体视为可变评估参数，融合在一起来使用。

另一方面，关于评估客体。基于专家客观分析的评估思想是以制定层级指标体系为核心，经过这么多年的发展，得到认可最多的一种一级指标划分方法是分为：馆藏、技术、管理、用户（服务）四个方面，这其中用户（服务）方面和基于用户主观感受的评估要区分开，这里的确是基于专家客观评估中对用户的一种考虑，但更多是站在

专家的角度对服务的一种评定，在进一步的二级指标分级中就出现了很多分歧。而基于用户主观感受的评估思想以用户接触到的各个方面为调查统计对象来展开评估，其项目虽然不分层级，但是依然可以归纳到馆藏、技术、管理等这些大的方面。不难发现，无论是哪种评估思想，馆藏、技术、管理这些方面是评估考察的基本部分。

因此，本课题提出将评估客体划分为两大部分：硬性评估和软性评估。其中硬性评估主要以较为客观的评估分析为主，包括三个子模块：馆藏、技术、管理，而软性评估则是完全体现用户主观感受的评估分析，也分为三个子模块：馆藏、技术、管理，即馆藏、技术、管理三个模块都要进行硬性和软性两个策略的评估求值，本文将这三个模块称为数字图书馆评估的基本模块。这样就构成了一种主、客观因素相融合数字图书馆评估思想的基本框架，基于这种评估思想，本文提出了一种主、客观因素相融合的数字图书馆评估模型——SOI 模型（Subjectivity-Objectivity-Integration Model of Digital Library Evaluation）。

主、客观因素相融合评估模型如图 3.2 所示：

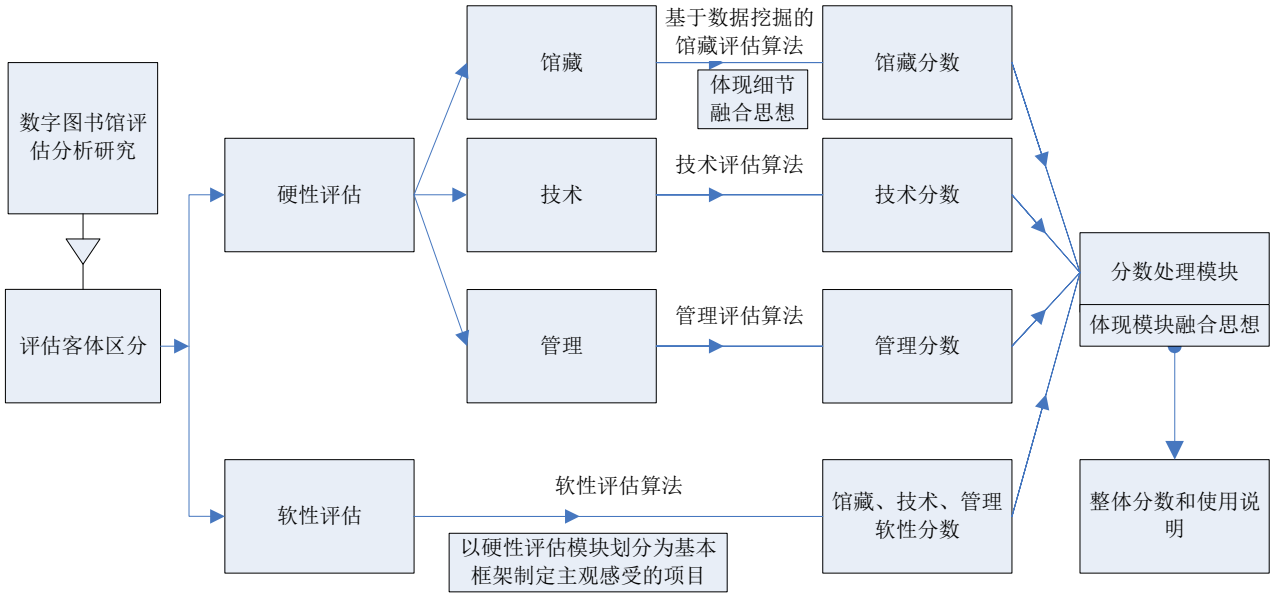


图 3.2 SOI 评估模型图

评估研究首先在数字图书馆评估分析与研究的指导下，确定评估线路，依照评估客体的特点，将评估客体划分为硬性和软性两个模块，分别对这两方面展开研究，其中硬性方面分为馆藏、技术以及管理三个模块，不同模块具有独立的评估机制，每个

模块使用基于数据挖掘的评估方法以兼顾专家客观评估数据和基于用户主观感受数据的结合，由此确定的各模块评估结果，再结合三个基本模块对应软性分数，按照模块融合策略制定的权重计算方法，取得最终的评估结果。同时它和评估中有价值的局部信息一起被最后输出，这些局部结果的信息也是被再次数据挖掘的基础，这同样是研究的重点内容之一。

依照对主、客观因素相融合评估思想的论述，将评估过程分为硬性评估和软性评估两个方面，更好的反映数字图书馆评估目的的需求，通过主、客观因素相融合的机制，将基于专家的评估思想和基于用户感受的评估思想整合在了一起。

需要注意以下两点：

（1）依照评估客体的划分来细化评估的过程，而非评估主体；

（2）在任一细分的评估客体的评估过程当中，都需要考虑到评估主体是有两个的，而至于到底是选择专家、用户、还是两者共同作用来制定评估方法，则完全依照具体的评估需求而定。

接下来，将对 SOI 评估模型的具体特点和工作方式展开论述。

3.3.2 硬性评估

硬性模块分为馆藏、技术、管理三个基本部分，类似于传统基于专家客观分析评估思想体系中的层级指标的第一级。

（1）馆藏

馆藏是数字图书馆评估的核心部分，此处被当作为硬性评估客体，本文并不完全以专家为主体的分析方法来处理它，在客观评估中会融入主观评估进行更好的价值表达。

本文在后续章节将会对馆藏评估方法进行设计和实践，研究是以中文数字图书馆的书籍类别的馆藏进行采集和应用的，其中有一个明显区别于其他馆藏的一点就是拥有一系列用户感受的数据，以验证本文的研究。

（2）技术和管理

这两个方面是一种用户不易全面感受到的却带有潜在影响能力的重要因素，所以对它的评估需要进行较为客观的分析。

这两个模块的评估特点不同于馆藏，相对比较容易确定其能力范畴，却不容易精确定性其影响方式，应该放弃传统专家评估中求全求细的方式，主要以评估基础理论为决策依据，依照关联分析方法，剔除关联性不强的辅助因素，强化关键因素的权重，

并在此基础上将指标体系转换为形式化数据，以研究其中的数据挖掘算法，得到有益结果。

3.3.3 软性评估

软性评估的以硬性评估模块划分为基本框架制定主观感受的项目，也即其研究内容同样分为馆藏、技术、管理三个基本部分，软性模块突出了用户的感受，本评估模型中软性模块主要有两个作用：

第一，体系上和基于硬性客体的评估对应整合在一起，以权重的计算形式参与到最终的评分系统当中。

第二，与相应的硬性模块对比分析，提供更加清晰、丰富的评估结果。

在进一步的深入研究中，可以尝试将这种软性数据规范化为以 XML 为基础的软性评估描述数据，增加在此基础上的数据挖掘研究，以取得有用信息。

3.3.4 主、客观因素相融合机制

主、客观因素相融合的方式，内在体现于以基本理论为依据、突破评估主体约束的评估思想上，外在就体现在了细节的处理策略上，本文在细节上采用融合策略的方式有两个层面：

（1）模块融合

三个基本模块分别进行硬性和软性两个策略的评估求值，在计分系统中，将硬性分值和软性分值进行加权处理求均值；

以专家客观评估的一级目录为基本框架制定主观感受的项目，在计分系统当中，将主观感受的软性分值权重化处理并与对应基本模块的硬性分值加权计算，最终确定各基本模块的分值。需要说明的是，这里权值的确定可以按照数字图书馆的类型进行适当的调整，可变是权值的一个原则，也是一个需要进一步深入的研究点，在模型当中，本文只是提出基本的范畴和处理的原则。

（2）细节融合

在硬性模块的评估当中，依据实际需求可以将含有用户感受的主观数据与客观的信息融合结合起来，形成包含客观表述属性和主观感受属性的数据集合，在硬性评估当中发挥用户主观感受的作用。

硬性评估当中，有一些是无法度量的参数，应该借助主观感受的结果，即在客观评估的指标当中融入主观评估的结果。比如在馆藏模块的评估当中，传统指标对于馆

藏的评估重点集中在数量，质量并不能很好的把握，而质量的重要性是毋庸置疑的，如何解决这些问题，就需要在评估当中将用户的主观感受结果融合进来，融合的方法就是评估方法创新研究的重点。

在细节融合的评估方法创新上，数据挖掘技术无疑是这种需求的最好解决方法，数据挖掘就是发现潜在的、常常不能直接推理的价值信息。

3.3.5 评估体系结构

这种主、客观因素相融合的评估模型体现了课题评估的理论思想、方法策略和模块评估算法，下面对其体系结构进行概述。

（1）设计目标

该体系基于课题评估基础理论的研究，将评估分为软性评估和硬性评估两个方面，在构建评估描述数据的基础上，实现相关模块的评估算法，取得体现各模块分数的评估值，并依据评估模型制定的原则，对评估分数进行融合、处理，使数字图书馆评估这一复杂的课题在不同角度上取得合理的分值展示，体现出数字图书馆真实的价值。

（2） 体系结构

整个体系结构分为数据支持层、逻辑业务层、实现层、表示层。具体结构图如图3.3 所示。

- 数据支持层

完成的功能是：构建良好的评估数据集，取得评估算法的形式化输入数据，并为以后的阈值改变保留修改接口。依照数字图书馆评估研究模块的不同，构造不同的数据支持集合，其中馆藏评估数据包含作者、时间、价格、出版社等客观数据，也将包含一部分基于用户感受的主观数据等，而技术和管理评估则是对现有资源的形式化描述数据，用 XML 语言来表示。

- 逻辑业务层

完成的功能：描述清晰的业务流程，确定对描述数据使用的方法，同时明确所要采取的处理策略。逻辑业务层依照数据模型中对评估客体的模块分类，逐一明确数据业务，馆藏评估算法针对馆藏评估数据集进行处理，技术和管理两个模块分别对应技术模块数据集和管理模块数据集进行处理，软性模块对应软性数据集进行处理。

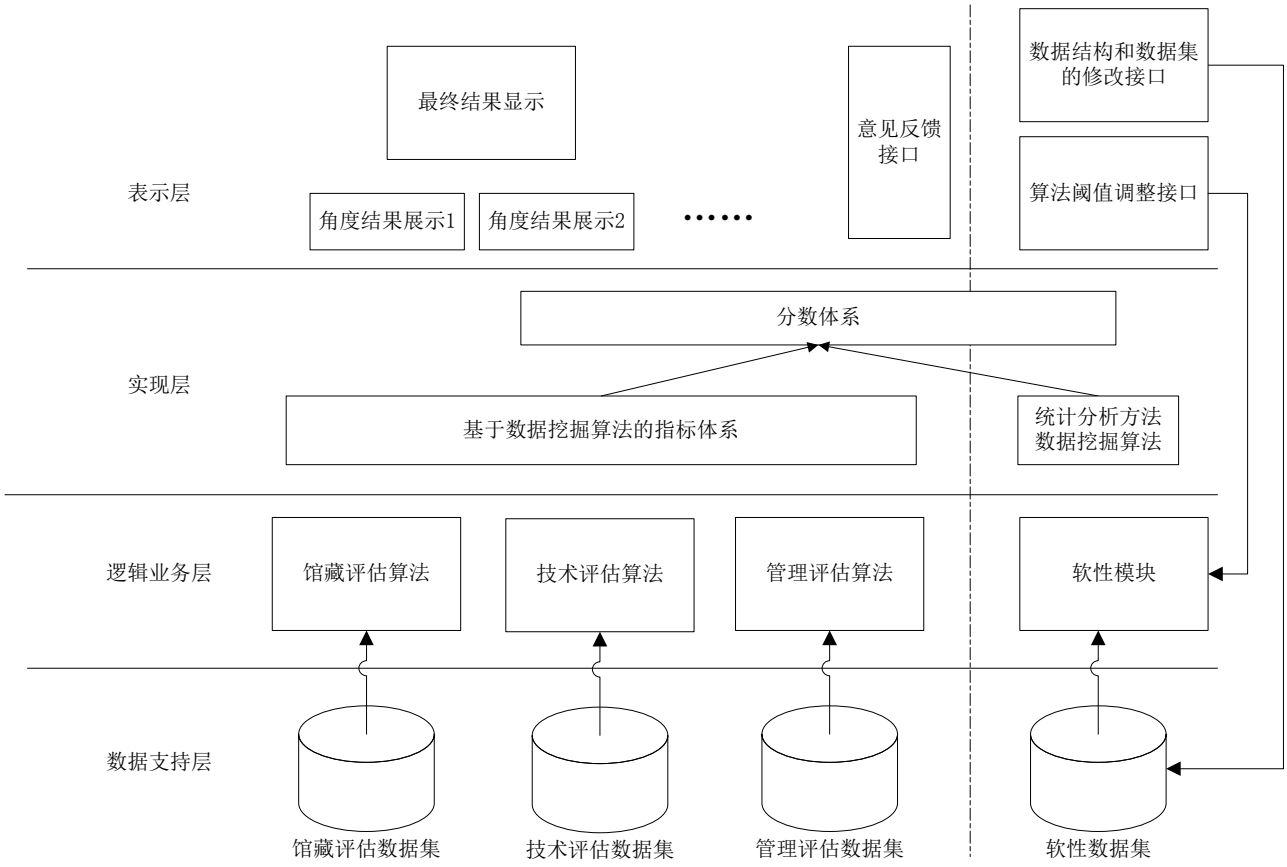


图 3.3 SOI 模型的体系结构图

- 实现层

完成的功能：通过逻辑业务层的控制策略，明确对应的处理方法，使用合适的评估方法实现逻辑业务。在计算各自分值之后，通过评估模型中确定的原则，进行分值之间的融合计算。

- 表示层

完成的功能：实现对用户的信息展示、意见反馈以及交互设置等功能。其中信息的展示分为不同角度的分数展示和综合分数展示；评估分数是唯一的，而评估角度不同带来的是系统化评估中的局部展示，也有着自己的意义和作用，因此也将获得上下文相关的展示方式；意见反馈是采集评估意见的接口，以方面修改和维护评估软件平台；交互设置是为了可以修改支持数据的形式以及调整算法阈值，以实现动态而富有弹性的评估理念。

依照这种评估模型，本文实现了相应的软件原型平台，其界面图 3.4 所示。

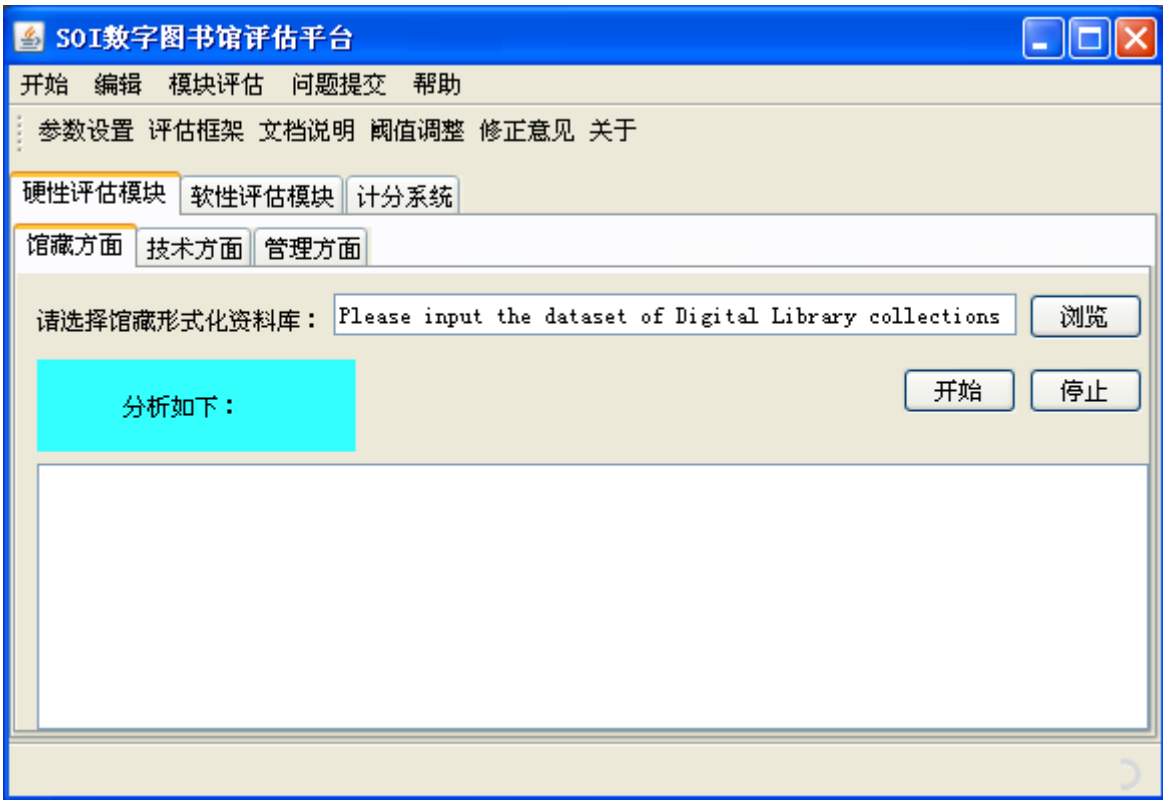


图 3.4 SOI 评估平台

在 SOI 评估原型平台中，整个评估系统经过硬性评估过程、软性评估过程、计分系统三步处理过程。

在硬性评估中，馆藏方面将馆藏的基本信息加上用户使用馆藏中一些感受信息结合起来形成评估数据集，在这个数据集上进行数据挖掘，建立馆藏价值预测模型，从而将馆藏质量属性体现出来，随后结合馆藏规模形成最终的馆藏评分。技术和管理方面，则会通过形式化指标数据，通过数据挖掘关联技术来确定之间的权重系数。

在软性评估中，围绕馆藏、技术、管理三个基本模块制定相关的调查选项，通过统计分析和数据挖掘的方法，不断的完善调查表内容，参考 DigiQUAL/LibQUAL+® 评估体系的计分方法形成各个基本模块的分值。

在计分系统中，一方面将三个基本模块（馆藏、技术、管理）的硬性分值和软性分值进行加权处理，并最终合计在一起形成一个有效的总分。另一个方面将每个基本模块的软硬信息进行对比分析，形成多角度的结果展示。

因为 SOI 模型中采用了用户调查表、数据挖掘等各种方法，所以平台提供了交互设置机制，用来接收用户以及专家的反馈信息，例如在馆藏硬性评估中，调整数据挖

掘算法的阈值，以获得与馆藏软性评估结果更接近的分数，并分析出现这种情况的原因；在软性评估中，根据实践的经验不断完善调查表的内容，以更好的反映用户的感受。

3.4 SOI 评估模型与主流评估模型的对比

将主流评估模型的特点与 SOI 评估模型相对比，它们之间主要不同集中在以下几个方面：

（1）在评估主体上的不同

表 3.3 评估主体的不同

评估主体	基于专家的评估模型	基于用户的评估模型	SOI 评估模型
唯一性原则	√	√	
专家	√		√
用户		√	√

SOI 评估模型在一个评估体系中，兼顾了专家和用户两种评估主体，有利于将双方的优势结合起来，在前文的主、客观因素相融合机制中也体现了这种评估主体结合的灵活性。

（2）在评估客体上的不同

表 3.4 评估客体的不同

评估客体	基于专家的评估模型	基于用户的评估模型	SOI 评估模型
馆藏	√	√	√
技术	√	√	√
管理	√	√	√
用户	√	√	

基于专家的评估模型与基于用户的评估模型都考虑到了馆藏、技术、管理、用户这四个基本方面，两者的区别在于评估选择的方法和策略完全不同，这种不同的根源在于两者评估主体选择的不同。而在 SOI 评估模型中，相比传统的评估模型少了一个“用户模块”，这是因为此时对“用户”模块的考虑，已经融于硬性模块和软性模块当中。

需要注意区分的是：SOI 评估模型中的软性模块并不是针对用户的评估，而是针

对馆藏、技术、管理三个方面，采用以用户为评估主体的策略进行的评估，在这种策略下，用户的主观感受已经融入到三个基本方面的评估当中。

（3）在评估路线上

表 3.5 评估路线的不同

评估线路	基于专家的评估模型	基于用户的评估模型	SOI 评估模型
客观	√		√
主观		√	√

SOI 评估模型将评估分为客观和主观两个评估过程，而每个部分都包含了馆藏、技术、管理三个基本模块，这相当于在评估过程中开展了两个线路。

（4）在评估结果上的不同

表 3.6 评估结果的不同

评估目的	基于专家的评估模型	基于用户的评估模型	SOI 评估模型
综合得分	√	√	√
子项分数	√	√	√
多角度展示			√
对比分析			√
意见反馈			√
阈值调整			√

SOI 评估模型，是一个开放性的评估交流平台，不同于传统基于专家的评估模型和基于用户的评估模型，SOI 评估模型的展示结果包括：综合得分、对比分析、意见反馈、阈值调整。

（5）在计分方法上

表 3.7 计分方式上的不同

计分方法	基于专家的评估模型	基于用户的评估模型	SOI 评估模型
单一计分	√	√	√
加权计分	√		√
融合计分			√

SOI 评估模型中，每个基本模块都会得到两个分数，一个是硬性评估方式的分数，

一个是软性评估方式的分数。这两个分数之间会进行加权处理成为一个分数，而三个基本模块之间又会加权处理得到最终的数字图书馆评估分数。

（6）在评估方法上

表 3.8 评估方法上的不同

评估方法	基于专家的评估模型	基于用户的评估模型	SOI 评估模型
专家打分	√		√
用户调查	√	√	√
统计分析	√	√	√
数据挖掘			√

SOI 评估模型中，评估方法引入数据挖掘技术，这种方法在馆藏模块主要体现在将主客观因素的数据结合起来进行价值挖掘，在技术和管理模块主要体现在形式化指标体系之后的关联分析，对权重的确定进行智能辅助决策。

（7）关于主客观因素融合

SOI 评估模型的硬性评估模块，意味着专家占据主要位置，但并不妨碍将用户的数据结合起来使用，例如，在馆藏的硬性评估过程中，会将主观因素的数据添加到评估数据集中，应将这种细节的软硬融合和整体评估路线的软硬融合相区分。

3.5 本章小结

本章基于第 2 章的研究指导，化解主流评估思想之间的分歧和束缚，提出了一个主、客观因素相融合的数字图书馆评估模型（SOI 模型），详细论述了评估模块的划分、主客观因素相融合的机制以及评估模型的体系结构，并比较了 SOI 模型与传统评估模型之间的区别。

第 4 章 馆藏评估及其属性离散算法研究

4.1 引言

在前文提出的 SOI 数字图书馆评估模型中，馆藏模块处于整个评估工作的核心位置，从本章开始将对数字图书馆馆藏评估展开深入研究，解决其中的问题。

首先提出了基于数据挖掘的馆藏评估思路，旨在解决传统馆藏评估中规模评估和质量评估相脱离的问题。在传统的馆藏评估中，基于专家的评估思路是以馆藏规模为主，质量评估因难以量化而被忽略；而基于用户感受的评估思路对馆藏质量有较好的把握，但因为用户感受局限和感受误差的原因，对馆藏规模往往不能准确的衡量。通过基于数据挖掘的馆藏评估方法，建立将规模评估和质量评估相结合的科学的馆藏评估体系。

为了达到这种目的，充分利用数字图书馆海量规范的馆藏信息，并提出馆藏的软属性概念，在传统馆藏属性中引入体现主观意愿的软属性，从而形成馆藏评估数据集，作为技术上解决馆藏质量量化问题的桥梁，体现了主、客观因素相融合的评估思想。

进而围绕馆藏评估数据集展开数据挖掘的应用研究，针对传统数据离散算法无法有效表达出数字图书馆馆藏复杂属性关系的问题，在研究中提出了一种基于 z 值的数据离散算法（PDOZ 算法，Parallel Discretization based on z-score），以便更好的在评估中提高数据预测的效果。该算法通过基于 z 值的有概率分布意义的动态距离代替传统数据离散方法，反映了馆藏各个属性语义的动态变化，增强了属性与预测之间的相关性，更利于发现属性关系，提高数据挖掘的性能；并基于这种离散方法研究了馆藏属性之间的关系，发现了非线性条件属性的存在，这种复杂关系是传统数据挖掘算法在馆藏评估数据集应用不理想的另一个重要原因，也是第五章研究的重点。

4.2 馆藏模块评估方法

数字馆藏是数字图书馆赖以生存的根本，在数字图书馆评估体系当中占据着最为重要的位置。根据前面章节介绍的数字图书馆评估领域两大主流思想，传统馆藏评估方法同样可以分为基于专家客观分析的馆藏评估方法和基于用户主观感受的馆藏评估方法。

基于专家客观分析的馆藏评估，是以评估指标的制定和量化为主要评估手段，指标的划分又以数字图书馆馆藏数量为核心而展开，也即在基于专家客观分析的馆藏评估当中，馆藏规模评估占据了主要的位置。在这种馆藏评估方法中缺乏对馆藏质量的衡量，因为质量的好坏程度是无法通过专家的分析直接得出分数的，因此质量评估是该方法难以克服的问题。

基于用户主观感受的馆藏评估，是一种效果评估方式。这种方式对质量评估有较好的展示，馆藏整体资源和单个资源的质量很容易就能表达出来。这种评估方法的缺点如第 2 章分析所说，存在用户感受错误的问题。这种错误来自于用户对数字图书馆整体规模的难以把握。因此，在基于用户感受为主的馆藏评估当中，馆藏效果评估占据了主要的位置，而在馆藏规模评估方面存在感知难题。

我们认为一个合理的馆藏价值评估体系需要有数量和质量两个因素的共同作用。将数量和质量结合起来，既能得到更准确的馆藏价值信息，同时也体现了主、客观因素相融合评估思想的意义。本文在 SOI 评估模型的馆藏模块中采取的策略是：基于上文的分析，结合数字图书馆馆藏资源海量规范、信息丰富的特点，将数字馆藏中含有用户感受的描述信息与基本信息结合起来，形成包含主观感受属性和客观基本属性的馆藏评估数据集合，在此基础上应用数据挖掘技术进行馆藏价值的预测分析，从而将数量、质量有机的结合在一起。

依照数字图书馆评估理论的分析研究和主、客观因素相融合评估模型的指导思想，我们提出一种基于数据挖掘的馆藏评估方法。在该方法中将基于专家客观分析的数据和基于用户主观感受的数据结合起来，同时放入数据集合当中，其方法展示如图 4.1 所示。

整个馆藏模块评估的工作包含三个阶段：

第一阶段：馆藏评估数据集构造阶段。

这个阶段需要对考察的数字图书馆馆藏资源进行数据采集，采集的数据主要是馆藏资源的描述型信息。采集数据的类型可以分为客观性数据和主观性数据。客观性数据包括资源名称、作者、价格等信息，主观性数据包括用户对馆藏的评分、是否读过、是否想读、是否推荐等信息。在数据集构造的阶段，需要采集、抽取馆藏的原始信息，并对这些数据进行初步清洗，去除残缺记录。

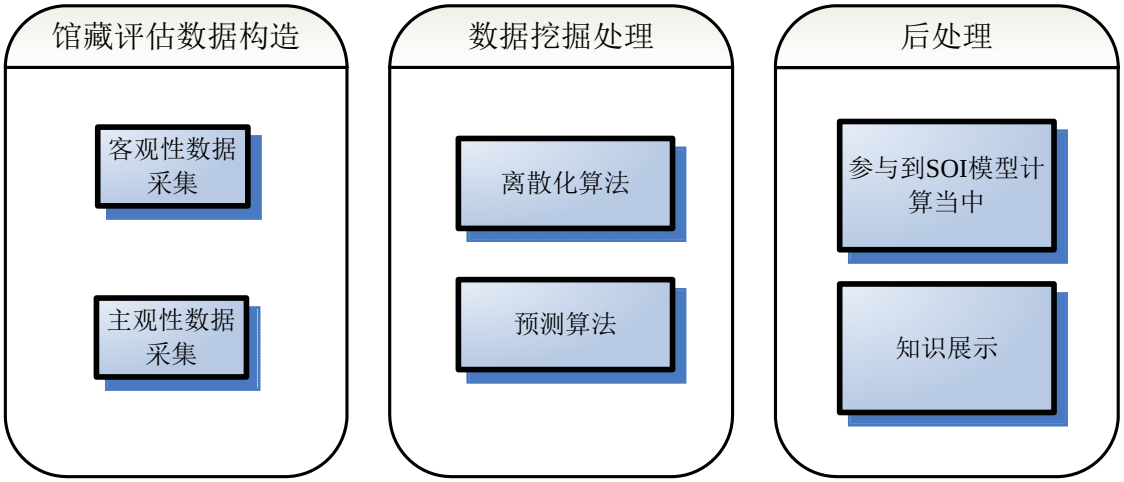


图 4.1 基于数据挖掘的馆藏评估方法图

第二阶段：数据挖掘处理阶段

这个阶段是应用数据挖掘方法对馆藏评估数据集进行应用研究，以期建立馆藏价值预测模型，形成质量评估的基础。因为数字图书馆本身的特点以及属性间的复杂关系，本阶段的研究可以归结为两个主要研究点：离散化算法和预测算法。

在应用传统数据挖掘算法中，我们发现效果并不理想，分析整个过程最容易丢失信息、影响效果的就是数值属性的离散化过程，由此产生了对数值属性离散化算法研究的需求。在离散化算法的研究过程中，我们发现了非线性的条件属性关系，针对数字图书馆馆藏属性的这种复杂特点，我们进一步研制了一种新的数据挖掘预测算法。

第三阶段：后处理阶段

这个阶段是对前面工作的综合处理，建立馆藏评估模型之后，在具体的数字图书馆中进行应用，得到馆藏评估的各方面分值，参与到 SOI 评估模型的计算当中，并为相关评估信息展示模块提供数据支持。

具体的实施步骤如下：

（1）针对具体数字图书馆的馆藏数据进行采集、整理，形成基本的馆藏数据集；

（2）通过网络调查等方式进行软性数据的收集，融入到馆藏数据集；

（3）通过数据挖掘的方法，确定出馆藏资源的预测模型；

（4）根据预测模型，对馆藏记录进行预测计算并添加为质量属性值；

（5）在馆藏规模计算过程中，将馆藏记录的累加转换为质量属性值的累加；

（6）将质量和数量结合起来作为馆藏的参数指标。

在具体的执行当中必然会涉及到对数字图书馆馆藏特点的研究，它决定了数据挖掘处理中的数据离散算法以及数据预测算法，这将是本章及下一章的主要研究内容。

4.3 馆藏属性的划分

依照上文提出的馆藏模块评估方法，我们提出将馆藏评估数据集的数据属性划分为硬属性和软属性两种。硬属性是馆藏的基本描述信息，而软属性是用户对馆藏的主观感受数据，这两个方面的属性共同构成数字图书馆馆藏评估数据集。

4.3.1 硬属性

硬属性是数字图书馆馆藏资源的各项基本属性，以书籍类别为代表，包括作者、译者、价格、页数、ISBN、出版社、出版时间、对应纸质版本的装帧等信息，这些属性是单个数字馆藏固有的、无法更改的属性，这些属性随着一本书的出版即被确定，所以称其为硬性属性。本文研究主要以数字图书馆中最具代表性的书籍类馆藏信息来组织数据集合，将其中每本书的各项属性信息抽取出来，如表 4.1 所示。

表 4.1 馆藏基本属性列表

属性名称	属性意义	测量度量
name	书名	nominal
author	作者	nominal
translator	译者	nominal
pages	页数	ordinal
price	价格	ordinal
publisher	出版社	nominal
layer	装帧	nominal
time	时间	ordinal

4.3.2 软属性

软属性是体现用户独立意愿、体验感受、主观评估等信息的数据属性，比如多少人读过了这本书，多少人想读这本书，多少人正在读这本书，用户对这本书的评价等信息。

在前文的研究当中，已经分析了主、客观因素相融合评估思想的意义，体现到馆

藏评估当中，就是要将基于专家客观评估的思想和基于用户主观感受的思想统一起来，综合作用在馆藏评估当中。硬属性就对应于客观评估的指标，而软属性就是体现用户主观感受的指标。

本文选取了以下较重要的软属性指标，如表 4.2 所示。

表 4.2 馆藏软属性列表

属性名称	属性意义	测量度量
tag	用户自定义标签	ordinal
numberHaveRead	多少人读过	ratio
numberReading	多少人正在读	ratio
numberWantread	多少人想读	ratio
numberAttention	关注人数	ratio
numberEvaluation	参与评分的人数	ratio
score	分数	ratio

这里的 score 分数是用户主观感受的结果体现，它不同于基于用户主观评估思想的馆藏得分，传统基于用户主观评估的分数是用户对于馆藏的整体感受，而这里的分数是用户对每本书籍的评价，更加具体也更加准确的反应了用户的感受。

4.3.3 硬属性的扩展

在数据挖掘的应用中，一些原始属性往往不直接拿来使用，而是转化为其他更有意义的属性参与数据分析，数字图书馆馆藏资源的属性也是如此，在深入分析之前，需要先对原始属性进行再加工，以取得着更丰富的价值信息，作为研究开始的切入点。比如中国五千年文化源远流长，博大精深，其中平仄有致，讲究韵律美感，所以本文觉得名称对于用户的主观感受也是起到一定作用的，而对大量的数字资源进行观察，也发现一些成语往往更容易取得不错的用户评估。

综上所述，针对硬属性，进行了以下的扩展：

针对书籍的名称，如表 4.3 所示。

表 4.3 名称属性的扩展

name_pure	去处副标题等信息之后的书名
name_pure_have	书名是否有附加信息
name_brackets	书名附加信息有哪些
name_length	名字长度（pure）
name_length_level	名字长度离散化
name_divisibility	名字的整除特性

针对责任人，如表 4.4 所示

表 4.4 责任人属性的扩展

numberAuthor	作者的数量
author1	第一作者
author2	第二作者
author3	第三作者
haveTranslator	是否有译者
numberTranslator	译者的数量
translator1	第一译者
translator2	第二译者
translator3	第三译者

针对其他硬属性的扩展属性，如表 4.5 所示。

表 4.5 其他硬属性的扩展

level_pages	页数水平
level_price	价格水平
level_publisher	出版社水平
level_layout	装帧水平
level_time	时间水平

4.3.4 软属性的扩展

软属性往往带有明确的意义，测量度量往往也是数值级别的，所以不需要在语义上进行延伸，只需要进行必要的离散化处理，使它们的对比效果体现出来。因此必要

的扩展如表 4.6 所示。

表 4.6 软属性的扩展

level_numberEvaluation	评估总数离散化
level_numberReading	在读数量离散化
level_numberHaveread	读过人数离散化
level_numberWantread	想读人数量的离散
level_numberThree	在读、读过、想读三者量的离散
level_score	分值的离散化

4.4 软属性的有效性实验

为了检验本文提出的软属性对于馆藏评估的有效性，本节将对相关数据进行统计分析实验，以判断软属性是否和用户的主观评分意愿相关联，即是否对用户评分的预测有一定的贡献。

4.4.1 实验数据集

本实验的数据集由数字图书馆图书类馆藏的属性所组成，数据记录共有 109801 条，包含了 48 个属性，其中主观意愿的属性主要参考于豆瓣网。（http://www.douban.com/，该网站维护着目前国内最有价值的用户对书籍的评分数据以及一系列书籍基本信息，这些数据也曾是 Google 数字图书馆对其数字馆藏评分的参考，现在 Google 数字图书馆正在致力于建立自己的馆藏用户评分数据库）

因为本实验针对软属性进行分析，因此主要考察属性有：numberEvaluation（参与评分的人数），numberReading（正在阅读该馆藏的人数），numberHaveread（已经读过该馆藏的人数），numberWantread（想读该馆藏的人数），numberThree（在读、已读、想读三者之和）以及 score（用户的评分）。

其中，这些属性都是数值型的属性，当需要进行数据的离散化时，此处采用了传统的等距方法进行处理，处理后的属性为：level_numberEvaluation，level_numberReading，level_numberHaveread，level_numberWantread，level_numberThree，level_score。

4.4.2 实验设计

实验采用数据抽样的方法，首先对数据特征进行图表分析，观察软属性与用户评分之间的关系，随后利用相关度计算模型来计算属性间相关系数，验证软属性与用户评分的相关性。

其中 Pearson 相关系数用来考察和衡量两个连续性属性之间的关系，它的计算模型如下：

$$r_{XY}=\frac{n\sum_{i=1}^{n}x_i y_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{\left[n\sum_{i=1}^{n}x_i^{\,2}-\left(\sum_{i=1}^{n}x_i\right)^2\right]\left[n\sum_{i=1}^{n}y_i^{\,2}-\left(\sum_{i=1}^{n}y_i\right)^2\right]}}$$

其中：r_{XY} 是属性 X 与属性 Y 的相关系数

n 是样本规模

x_i， y_i 分别是属性 X 与属性 Y 的具体数值

利用该相关度计算模型来检验软属性与用户评分之间是否存在显著性相关。因为目的是验证原始数据中两两属性之间的相关性，而这些数据本身就是具备序数意义的数值型数据，所以无需采用等级相关系数的方法。

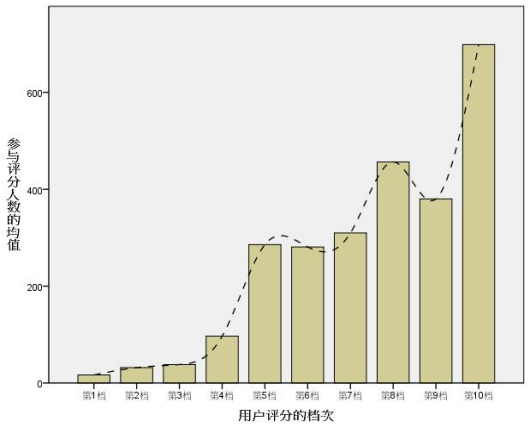
4.4.3 实验结果与分析

实验结果如图 4.2、图 4.3 所示。

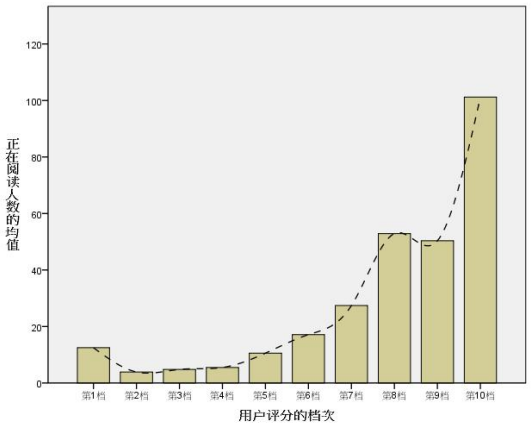
图 4.2 是将“参与评分人数”、“正在阅读人数”、“已经读过人数”、“想读人数”分别与用户评分的档次之间建立起来的关系图。

而我们可以将“正在阅读人数”、“已经读过人数”、“想读人数”三者人数之和作为用户对该数字馆藏的兴趣度的综合表达，从而建立起用户兴趣度与用户评分的档次之间的关系图。

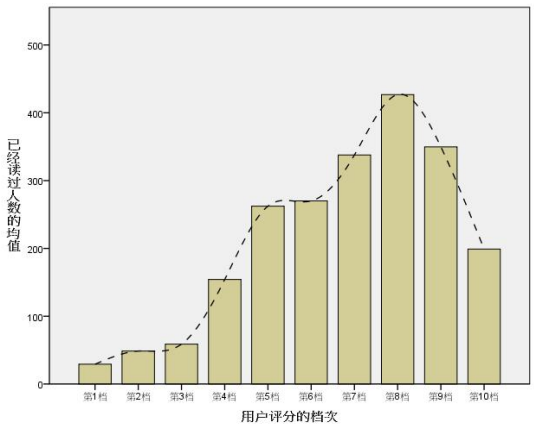
通过这两种方式可以明确看出这些体现用户主观意愿的软属性与待预测的用户评分属性之间的关系。



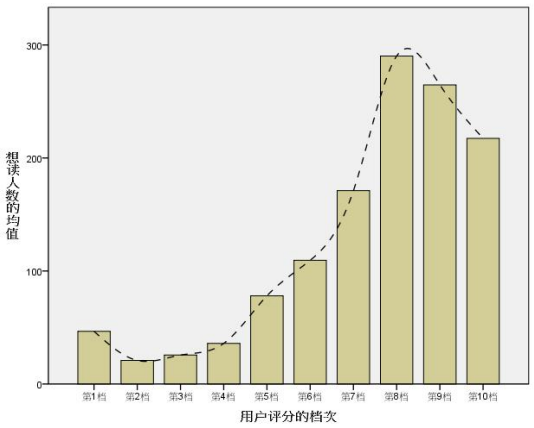
(a) 参与评分人数与用户评分的关系图



(b) 正在阅读人数与用户评分的关系图



(c) 已读人数与用户评分的关系图



(d) 想读人数与用户评分的关系图

图 4.2 主观意愿属性与用户评分属性的关系图

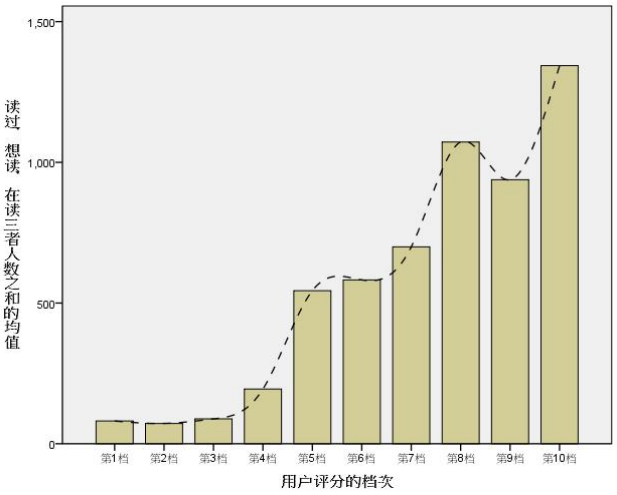


图 4.3 读过、想读、在读三者人数之和与用户评分属性的关系图

通过图 4.2 可以明显发现，软属性与待预测的用户评分属性具有明显的正关联性，

而图 4.3 也说明了，用户综合兴趣度越高，用户分值也越高，充分说明了软属性在馆藏的价值评估中具有一定贡献。

表 4.7 和表 4.8 是软属性与预测属性之间的相关性分析结果，相关分析是揭示属性之间关系的定量分析方法，在我们的属性变量当中，是进行了离散化处理，其中五个软属性为：numberEvaluation、numberReading、numberHaveread、numberWantread、numberThree，预测属性为 score。

其中表 4.7 是软属性与预测属性六者之间综合的相关分析结果，反映了两两属性之间的关系，而表 4.8 则是 5 个软属性单独与预测属性之间的相关性分析结果。

表 4.7 相关分析结果（1）

Correlations							
		numberEvaluation	numberReading	numberHaveread	numberWantread	numberThree	score
number Evaluation	Pearson Correlation	1	.627**	.560**	.581**	.976**	.043**
	Sig. (2-tailed)		.000	.000	.000	.000	.000
	N	109801	100841	109801	109801	109801	109801
number Reading	Pearson Correlation	.627**	1	.456**	.646**	.712**	.099**
	Sig. (2-tailed)	.000		.000	.000	.000	.000
	N	100841	100841	100841	100841	100841	100841
number Haveread	Pearson Correlation	.560**	.456**	1	.612**	.595**	.066**
	Sig. (2-tailed)	.000	.000		.000	.000	.000
	N	109801	100841	109801	109801	109801	109801
number Wantread	Pearson Correlation	.581**	.646**	.612**	1	.669**	.143**
	Sig. (2-tailed)	.000	.000	.000		.000	.000
	N	109801	100841	109801	109801	109801	109801
number Three	Pearson Correlation	.976**	.712**	.595**	.669**	1	.062**
	Sig. (2-tailed)	.000	.000	.000	.000		.000
	N	109801	100841	109801	109801	109801	109801
score	Pearson Correlation	.043**	.099**	.066**	.143**	.062**	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	
	N	109801	100841	109801	109801	109801	109801
**. Correlation is significant at the 0.01 level (2-tailed).							

表 4.8 相关分析结果（2）

	numberEvaluation	numberReading	numberHaveread	numberWantread	numberThree
Pearson Correlation	.043**	.099**	.066**	.143**	.062**
Sig. (2-tailed)	.000	.000	.000	.000	.000
N	109801	100841	109801	109801	109801

**. Correlation is significant at the 0.01 level (2-tailed).

从表 4.7 和表 4.8 中可以看出，每个软属性与预测属性在双侧检验中是显著性相关的，从而验证了软属性对用户评分预测贡献的有效性。

4.5 一种基于 z 值思想的并行离散算法 PDOZ

很多数据挖掘算法期望数据属性以定性的（nominal 类型、ordinal 类型）形式参与到算法设计当中，当遇到的定量属性（interval 类型、ratio 类型）的时候，往往需要进行定量数据的离散化处理，这种数据离散算法多应用于分类预测算法或关联分析算法的属性研究中，以提高数据挖掘算法的准确性。在数字图书馆评估中采集到的数据不可避免的包含有定量的连续性属性变量，比如：硬属性中的价格、时间、页码等，软属性中的关注人数、用户评分等，也需要进行必要的离散化处理。

目前，定量属性的离散算法作为数据挖掘研究中重要的一环，已经取得一定的研究成果，依据是否使用类信息可以将离散算法分为监督离散算法和非监督离散算法。但是大部分都是针对特殊案例的优化，无法普遍应用到其他案例，而且已有离散化方法大都针对单一关系设计，不能直接用于多关系环境，而现实生活中属性之间存在复杂关系的案例就有很多。在数字图书馆当中，采集到的数据属性之间就具有很强的关联性。

4.5.1 问题的出现

在有了体现主、客观因素相融合评估思想的馆藏数据之后，进行了初步的数据挖掘应用，虽然馆藏的研究需求和目的符合数据挖掘的思想，也达到了数据挖掘算法对数据的各方面要求，但应用起来效果却很不理想，于是计算了各个属性与预测属性之间的相关性，发现很多属性的相关系数很小，意味着这些属性对于预测的贡献度比较

低，出现这种现象与本文对馆藏属性的预期差别较大。那么，到底是这些馆藏属性真的对预测贡献很小，还是说馆藏属性关系太复杂？如何找到一种更好揭示数据意义的方法呢，这将是接下来研究的内容。

经过分析不难发现，在馆藏数据当中，最容易出现信息遗失的就是数值型属性数据，因为它们需要被离散后参与数据挖掘过程，而名词型的数据是不经修改的，因此对离散算法的研究和创新是解决问题首先应该考虑的。通过建立新的离散算法，使馆藏连续性属性能更好的反映出内在的语义信息。

4.5.2 基于 z 值计算的并行离散算法（PDOZ）

本课题提出了一个基于 z 值计算的并行离散算法——PDOZ 算法（Parallel Discretization based on z-score），该算法在对连续属性进行离散处理时，统筹所有属性并行处理，在整体上兼顾数据分布的特点，它从正态分布的 z 值计算中得到启发，将数据的离散进行了统一的标准化处理，用这种基于 z 值的有概率分布意义的动态距离代替传统离散方法，更加适合数字图书馆馆藏语义信息，使数据对比更有意义，更利于发现属性关系。

在一般的研究实践中，当不同数据无法直接进行比较时，解决的办法就是寻找一个中间值作为中介，进行这样的比较需要一定的标准，这就是标准值(standard scores)。在数理统计中，不同正态分布的差异性很大，其差异性的比较中最重要的标准值就是 z 值（z score），z 值就是原始数据与数据分布均值的差除以标准差所得的结果[162-165]。

z 值的计算公式：

$$z = \frac{\left(X - \bar{X}\right)}{s}$$

其中

X 是具体的数值

$\bar{X}$ 是数据分布的均值

s 是数据分布的标准差

通过公式可以看出，z 值是将数据分布中具体数值与均值的偏离用标准差进行统一处理，在比较同一数据分布的数值时，使用标准差和 z 值的意义是相当的，但在不同数据分布中，标准差之间的比较是没有意义的，而 z 值却可以进行比较，例如不同数据分布中 z 值相同的多个数据点，它们都与各自均值的距离相等。

z 值从本质上将就是以标准差为单位对原数据进行了标准化，从而可以比较不同的数据分布。

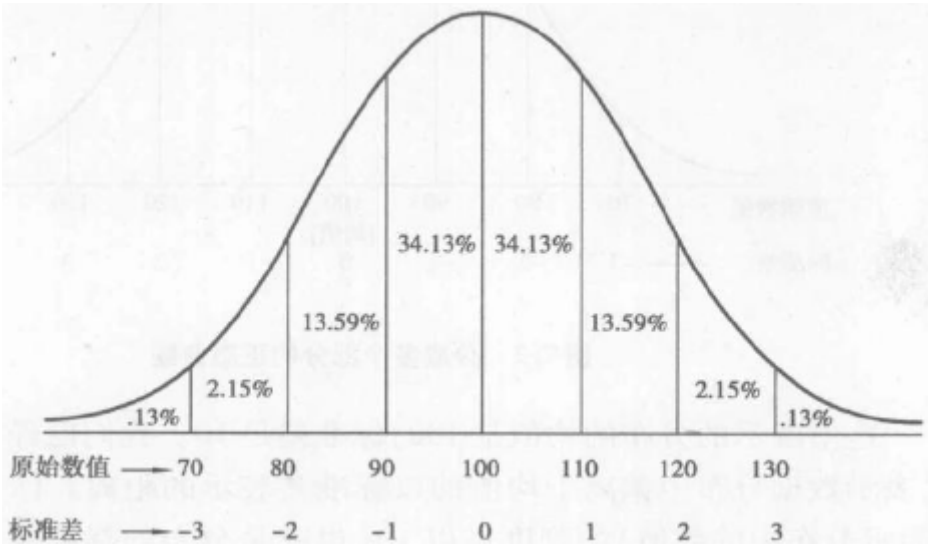


图 4.4 标准正态分布图

我们利用一个均值为 100，标准差为 10 的标准正态分布（如图 4.4），来总结一下 z 值具有的特点：

（1）每个 z 值就是对应个值偏离均值的标准差的个数。z 的绝对值越大，离均值越远。$z=0$，就是均值，$z=1$就是处于 s（偏差平均水平）处。

（2）负的 z 值在坐标轴左侧，正的 z 值在坐标轴右侧，$z=-1$和$z=1$意味偏离均值的距离是一样的。

（3）不同数据分布的 z 值具有可比性，表示数据点在所在数据分布中，偏离均值的标准差的个数。

（4）z 值代表着数据分布的概率，图中的百分数反映了这种数据分布的概率。例如 34.13%表示：数据分布中有 34.13%的数值落在了在均值和和均值以上 1 个标准差的范围内。该数值也是正态分布曲线、均值以及均值 1 个标准差之间所形成面积占曲线图总面积的比例，故可以称该数值是 z=1 或 z=-1 时的面积点。

数字图书馆馆藏当中的数值属性大都是近似正态分布的，这符合数据的一般特点，也在后续的实验中得到验证。而在正态分布的数据差异性研究当中，标准差（Standard Deviation）是最重要的指标之一，是各数据偏离平均数的距离的平均数，标准差能反映一个数据集的离散程度。z 值正是基于标准差的标准化处理。这使本文产生了利用 z 值思想建立新的离散算法来解决信息遗失问题的想法。

解决在传统离散过程中的信息遗失，需要解决两个问题：

①单个属性要符合语义的选取离散距离；②属性之间要执行统一的标准，使属性间的关联关系不能丧失。

而基于 z 值的计算，恰恰可以用基于概率分布的动态值来表达离散的数据宽度，而所有属性都基于这种计算，就可以既满足个体的动态变化，又兼顾群体执行统一标准，接下来，将详细的说明这种算法的流程。

4.5.3 算法流程

基于 z 值计算的离散化算法流程如图 4.5 所示：

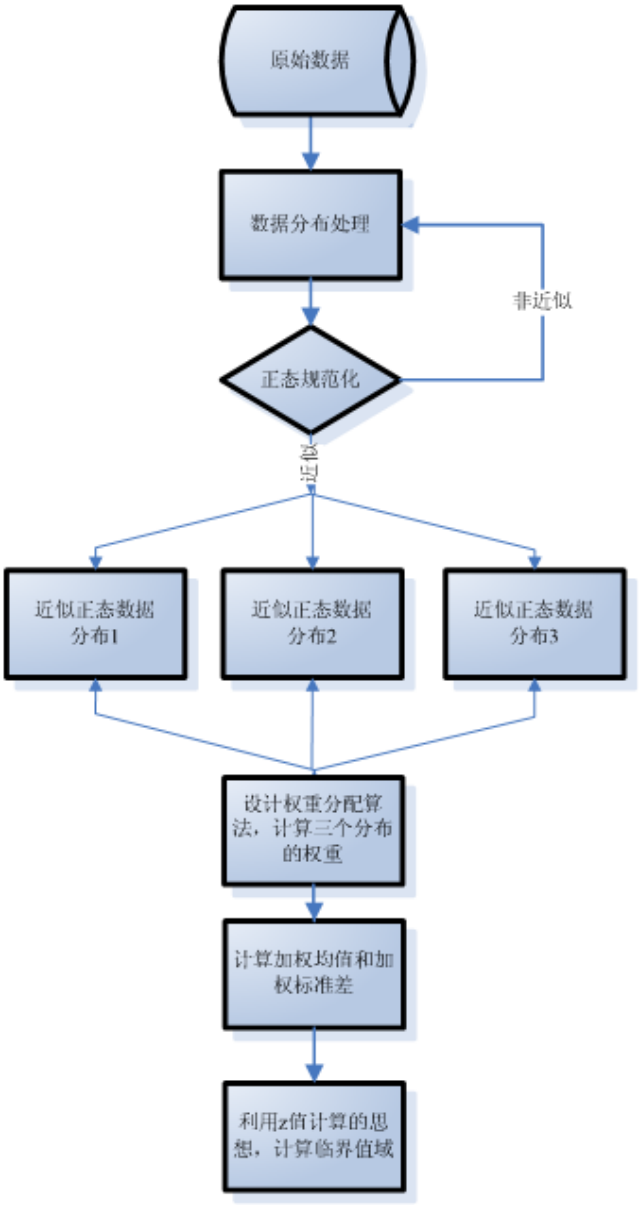


图 4.5 PDOZ 算法流程图

4.5.4 算法步骤及计算过程

可以从以下四个方面描述基于 z 值计算的并行离散算法：

（1）数据的正态分布规范化

利用均值、中位数、众数等代表性的数据参数[166-168]，通过合理调整和选取原始数据分布的两端区域，将含有异常、例外样本的原数据分布规范化为多个正态分布的近似分布；通过这一步，可以将数据进行初步的清洗，去除噪音。

（2）设计权重分配算法，计算近似正态分布群中单个分布的相对权重。

针对取得的多个近似正态分布群的属性，设计设计体现偏差对比的权重分配算法，计算近似正态分布群的各个权值。

（3）计算加权均值和加权标准差。

取出近似正态分布群的均值和标准差，利用计算得到的权重分配值，计算出唯一的加权均值和加权标准差作为有效正态分布的特征值；

（4）确定数据分布的比例区域临界值。

在加权均值和加权标准差的基础上，利用 z 值的思想，计算出数据分布的比例区域，在实际应用中确定最终的离散区域分布。

z 值的计算公式如下：

$$z_i = \frac{\left(X_i - \bar{X}\right)}{s} \tag{4.1}$$

其中：

z_i 是每个数据分布点的 z 值，

X_i 是具体的数值

$\bar{X}$ 是数据分布的均值

s 是数据分布的标准差

其中的标准差公式如下：

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \tag{4.2}$$

接下来，依照上面的四个步骤，详细阐明算法的计算过程：

（1）数据的正态分布规范化

在基于 z 值的并行离散算法当中，重要的是获得有效正态数据分布的均值和标准差，本文的策略是获取近似正态分布群，依照近似正态分布群的数值进行加权计算，求得一个合理的均值和标准差。

首先需要获取有效正态分布，这样做可以加强数据的特征化，有利于让数据集合体现出真实本质的特点。本文是选择均值、中位数以及众数分别作为数据的代表数，并以代表数结合原始数据分布，确定需要选取的数据区域上限和下限。相关的计算公式如下：

$$lower = representativeNmuber - (upper - representativeNmuber) \quad (4.3)$$

$$upper = representativeNmuber + (representativeNmuber - lower) \quad (4.4)$$

其中：

lower 是区域下限；

upper 是区域上限；

representativeNmuber 是每次被选定的代表数；

通过这两个公式，结合实际的数据分布特点可以确定出数据区域的上限或下限。

（2）设计权重分配算法，计算权重

每个近似正态分布往往都会出现不同程度的偏移，直接求均值和标准差会失去原始数据的部分准确性，所以需要一个好的计算方法获取每个近似正态分布群对最精确有效分布的贡献值，通过设计一个权重分配算法来解决这个问题。

偏度（*skewness*）是对数据分布不平衡的统计测量尺度，利用偏度作为权值分配的重要依据，可以很好的将近似群区分开[169-171]。

公式（4.5）是偏度的计算公式：

$$skewness = Mean - Median \quad (4.5)$$

这里有一个问题，可以看到偏度的计算是针对独立数据分布的，其数据值是无法在不同数据分布间直接进行比较的，本文的处理方法是：在此处将每个分布的偏度做标准差的倍数化处理，得到标准化偏差值 *sSD*，从而使不同数据分布的偏度可以对比，由此可以充分利用偏度作为比较度量的参数。

如公式（4.6）所示：

$$sSD = \frac{|Mean - Median|}{s} \quad (4.6)$$

其中：

sSD 是标准差倍数化后偏差的简写，称作标准化偏差，

Mean 是数据分布的均值

Median 是数据分布的中位数

偏度越大，其近似正态化的程度就越小，所以在对原始数据进行正态规范化的贡献度就越低，利用 *sSD* 的时候给予的权重就应该越小。

最终的权值分配方法：将最大的权重比例分配给最小偏度数据分布的数值，最小的权重比例分配给最大偏度数据分布的数值。

下面是权值分配的公式：

$$w_{\min\_skewness}=\frac{sSD_{\max}}{\sum_{i=1}^{n}sSD_i}\quad w_{med-skewness}=\frac{sSD_{med}}{\sum_{i=1}^{n}sSD_i}\quad w_{\max\_skewness}=\frac{sSD_{\min}}{\sum_{i=1}^{n}sSD_i} \tag{4.7}$$

其中：

$w_{\min\_skewness}$ 是最小偏度对应的权值

$w_{med\_skewness}$ 是中间偏度对应的权值

$w_{\max\_skewness}$ 是最大偏度对应的权值

n 是近似正态分布群的个数，它取决于代表性数据参数的个数，在本课题研究中设定为 3 个。

（3）计算加权均值和加权标准差

经过权重的分配，就可以将所求标准正态分布的标准差和均值都进行加权计算，如公式（4.8）和公式（4.9）所示。

加权均值的公式：

$$Mean=w_1Mean_1+w_2Mean_2+w_3Mean_3 \tag{4.8}$$

加权标准差的公式：

$$s=w_1s_1+w_2s_2+w_3s_3 \tag{4.9}$$

需要注意的是，均值的大小和单个近似正态分布的偏度没有绝对的关系，所以并非偏度越小，均值就越接近正确。

（4）确定数据分布的比例区域临界值。

利用 z 值计算的思想，计算临界值域。

将公式（4.8）和公式（4.9）的计算带入公式（4.10）当中，即可列出 z 值和 X_i 的关系等式。

$$z_i = \frac{(X_i - Mean)}{s} \tag{4.10}$$

最后一步依据这个关系等式，确定临界值的区域。

这里的 z 值选择依靠表 4.9 的面积点，它来自于数理统计学中正态曲线下的面积表。有前文对图 4.4 的分析知道，这些面积点代表了概率的数值。

我们选择合适的概率离散策略来选取面积点，从而获取对应的 z 值，让后通过公式 4.10 计算出对应的 X_i，最终确定属性离散的区间。

表 4.9 正态分布中 z 值对应的面积点

面积	Z 值	面积	Z 值	面积	Z 值
-40%	-1.28	-10%	-0.25	20%	0.535
-30%	-0.84	0	0	30%	0.84
-20%	-0.535	10%	0.25	40%	1.28
-17%	-0.43	17%	0.43		

面积点的选择并不固定，常用的有概率等距三分法，概率等距五分法，概率等距十分法。概率等距是依据面积点的平均分配来确定 z 值。

概率等距三分法选择的区间是：$(-50\%,-17\%]$，$(-17\%,+17\%]$，$(+17\%,+50\%]$；概率等距五分法选择的区间是：$(-50\%,-30\%]$，$(-30\%,-10\%]$，$(-10\%,+10\%]$，$(+10\%,+30\%]$，$(+30\%,+50\%]$；概率等距十分法选择的区间是：$(-50\%,-40\%]$，$(-40\%,-30\%]$，$(-30\%,-20\%]$，$(-20\%,-10\%]$，$(-10\%,0]$，$(0,-10\%]$，$(+10\%,+20\%]$，$(+20\%,+30\%]$，$(+30\%,+40\%]$，$(+50\%,+50\%]$。

面积点的选择主要依据连续型数值属性的具体特点，假如该属性的极差较大倾向于选择较多分段的离散方式，而极差较小则倾向于选择较少分段的离散方式。

4.5.5 算法总结

PDOZ 算法是借鉴数理统计中 z 值意义，z 值是用于比较不同的正态分布而引入的量化参数。世界上大部分事物的分布是符合正态分布的，也即大多事件正好在数据分布的中间，而两端却较少。虽然所有的标准正态分布都符合这一本质特点，但是各

个事物的正态分布图形却是千变万化，每个分布在集中趋势和变异性方面存在相当大的不同，因为从属不同的分布体系，分布上每一个值和其他分布的对应值比较就没有任何实际意义，在需要对它们进行比较的时候就需要一定的途径，这个途径就是通过 z 值，z 值是以标准差为单位进行了标准化的处理，从而避免了不同分布之间无法直接比较的问题，而通过自身程度的处理，使分布区间上每个值都有一个程度上的度量，从而可以和其他分布中的值进行集中趋势和变异性的比较，再结合各自分布的实际意义可以得到有价值的信息。

PDOZ 的算法思想起到了动态适应评估主体、优化评估结果的作用。在第 2 章分析过一种评估思想——“以评估主体不同而进行的不同指标量化评估体系”，该理论认为应该依据不同的评估主体拆分评估过程，本文已经分析了评估主体不足以拆分评估过程，但是也说明了一种需要重视的现象：不同评估主体侧重点不同而带来的评估结果的不同。假设存在两个数字图书馆：计算机学院的数字图书馆 A 和文学院的数字图书馆 B，两个数字图书馆的馆藏内容的特点自然会围绕服务专业的不同而产生不同，针对时间属性，在 A 中出现 1950 年代以前的数据就会是小概率事件，在 B 中反而是大概率事件，利用现有的评估离散算法，无论如何都无法兼顾两种属性在时间上的差异，但依据 PDOZ 算法基于 z 值同步离散的思想，就会自然而然的将这种动态差异以不同的离散距离区分开，依然会形成时间上相对的“古代、近代、现代”，从而将用户对时间上的感受正确的表达出来。

PDOZ 算法用这种基于 z 值的有概率分布意义的动态距离代替传统离散方法，更加适合数字图书馆馆藏语义信息，使数据对比更有意义，更利于发现属性关系。

它通过自身标准化的处理，促使属性之间的离散更好的在一起共振，有利于发现真正的关系。

4.6 应用 PDOZ 算法对主要馆藏属性的分析实验

4.6.1 数据集描述

本节的数据集同样由数字图书馆图书类馆藏的属性所组成，收集的数据记录共有 11 万条左右，包含了 48 个属性，其中主观意愿的属性主要参考于豆瓣网。我们针对具体的实验属性进行了一些初步筛选、去噪，以此保证数据集的质量，因此各个属性实验数据的数量并不相同。

在实验中选择了用户评分、书名长度、页码、价格、时间共五个数值型属性构成了一个数据集作为实验研究对象，检验基于 z 值计算的并行离散算法的有效性。其中将用户评分设定为类别属性以检验其他属性。

4.6.2 用户评分

（1）数据的正态分布规范化

首先得到原始数据的数据分布图，如图 4.6 所示。

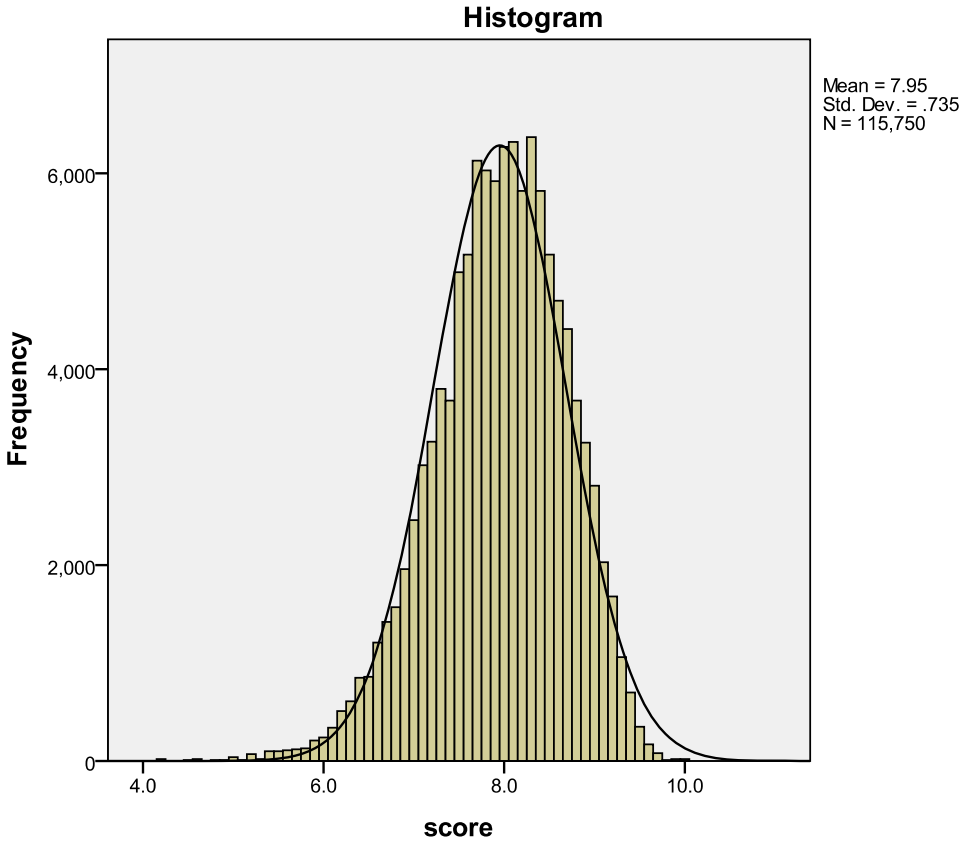


图 4.6 用户评分属性的原始数据分布图

由图 4.6 可知用户评分数据基本符合正态分布，接下来对其进行正态规范化处理来准确表达数据特点。

统计数据原始分布的代表性数据，如表 4.10 所示。

表 4.10 用户评分属性的代表性统计量

Statistics		
score		
N	Valid	115750
	Missing	0
Mean		7.950
Median		8.000
Mode		8.3
Std. Deviation		.7346
Variance		.540

利用原始数据分布的均值、中位数和众数作为代表数据，数据上限为 10，下限设定为 0～代表数的两倍，使用公式（4.3）进行规范化处理。

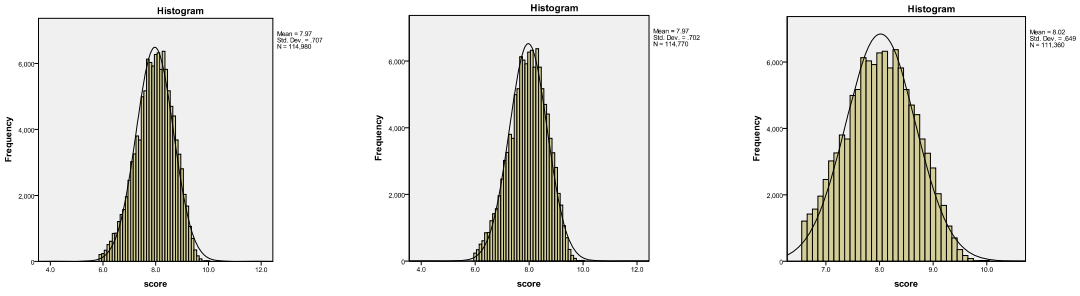
$$lower = representativeNmuber - \left(upper - representativeNmuber\right)$$

$$lower_1 = 7.95 - \left(10 - 7.95\right) = 5.90$$

$$lower_2 = 8 - \left(10 - 8\right) = 6$$

$$lower_3 = 8.3 - \left(10 - 8.3\right) = 6.60$$

依照$[5.90,10]$、$[6,10]$和$[6.60,10]$三个区间重新对数据进行分析，其 3 个近似正态数据分布图如图 4.7 所示。



(a)均值为主的近似分布　　(b)中位数为主的近似分布　　(c)众数为主的近似分布

图 4.7 用户评分属性的 3 个近似正态分布图

（2）计算权重

统计近似正态分布群的代表性数据，如表 4.11 所示。

表 4.11 用户评分属性 3 个近似分布的代表性统计量

v31time								
Statistics1			Statistics2			Statistics3		
N	Valid	114980	N	Valid	111740	N	Valid	111360
	Missing	0		Missing	0		Missing	0
Mean		7.967	Mean		7.971	Mean		8.022
Median		8.000	Median		8.000	Median		8.000
Mode		8.3	Mode		8.3	Mode		8.3
Std. Deviation		.7068	Std. Deviation		.7019	Std. Deviation		.6488
Variance		.500	Variance		.493	Variance		.421

利用公式（4.6）计算标准化偏度，利用公式（4.7）计算权值的分配：

$$sSD = \frac{|Mean - Median|}{s}$$

$$w_{\min\_skewness} = \frac{sSD_{\max}}{\sum_{i=1}^{n} sSD_i} \quad w_{med\_skewness} = \frac{sSD_{med}}{\sum_{i=1}^{n} sSD_i} \quad w_{\max\_skewness} = \frac{sSD_{\min}}{\sum_{i=1}^{n} sSD_i}$$

解得三个分布的相对权重为：

$w_{\min\_skewness} = 0.38296765$

$w_{med\_skewness} = 0.338896783$

$w_{\max\_skewness} = 0.278135568$

（3）计算加权均值和加权标准差

利用公式（4.8）和公式（4.9）计算：

加权均值的公式：

$$Mean = w_1 Mean_1 + w_2 Mean_2 + w_3 Mean_3$$

加权标准差的公式：

$$s = w_1 s_1 + w_2 s_2 + w_3 s_3$$

解得：

$$Mean = 7.989418808$$

$$s = 0.682927282$$

（4）确定数据分布的比例区域临界值。

将公式（4.8）和公式（4.9）的计算带入公式（4.10）当中，即可列出 z 值和 X_i 的关系等式。

$$z_i = \frac{(X_i - Mean)}{s}$$

利用概率等距 5 分法，从而将数值属性划分为 5 个区域，对应的面积取值、z 值取值、X_i 计算结果以及从而确定的离散范围如表 4.12 所示。

表 4.12 用户评分属性的离散区间

序号	4 个面积点	4 个 z 值	4 个 X_i	5 个范围
1	-30%	-0.84	7.415759891	(0, 7.4]
2	-10%	-0.25	7.818686987	(7.4, 7.8]
3	+10%	+0.25	8.160150628	(7.8, 8.1]
4	+30%	+0.84	8.563077725	(8.1, 8.5]
5				[8.5, 10]

4.6.3 书名长度

（1）数据的正态分布规范化。

首先得到原始数据的数据分布图，如图 4.8 所示。

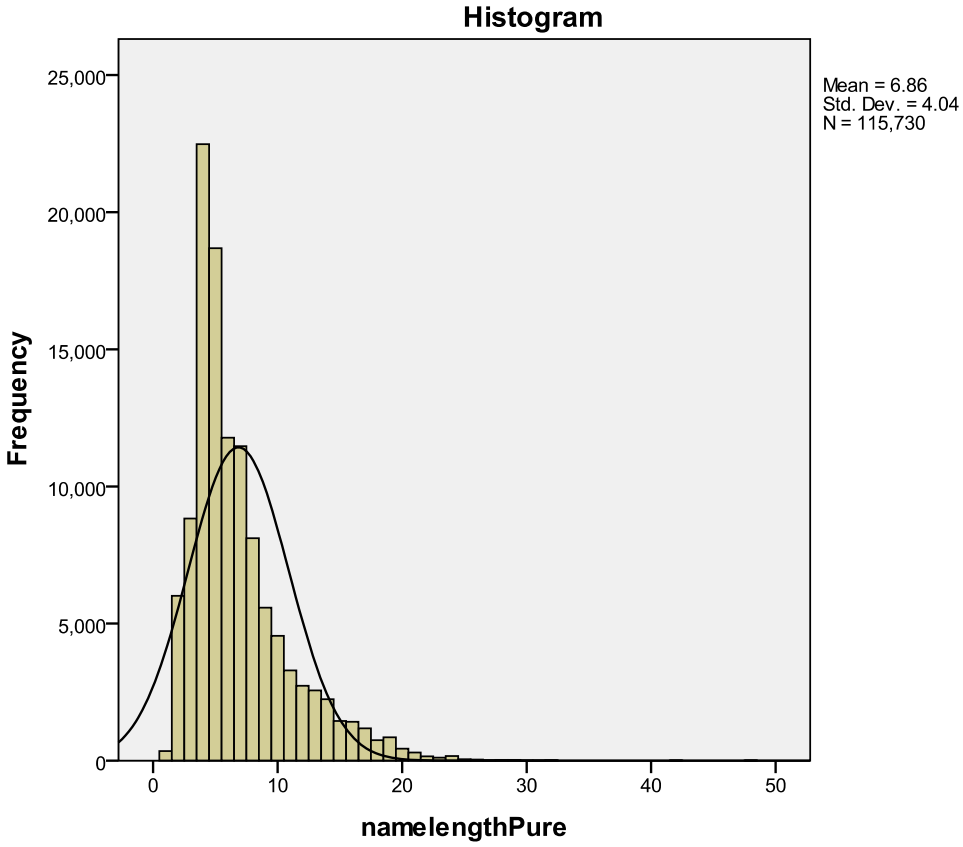


图 4.8 书名长度属性的原始数据分布图

统计该数据分布的代表性数据，如表 4.13 所示。

表 4.13 书名长度属性的代表性统计量

Statistics		
score		
N	Valid	115730
	Missing	0
Mean		6.86
Median		6.00
Mode		4
Std. Deviation		4.040
Variance		16.324

利用原数据分布的均值、中位数和众数作为代表数据，数据下限为 1，上限设定为 0～代表数的两倍，使用公式（4.4）进行正态分布的规范化处理。

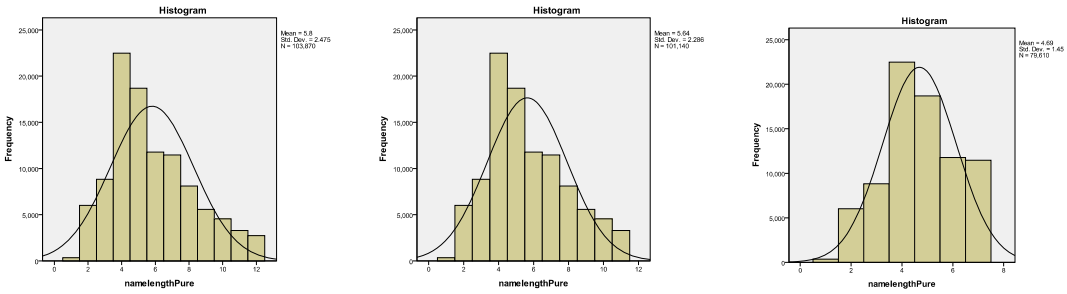
$$upper = representativeNmuber + \left(representativeNmuber - lower\right)$$

$$upper_1 = 6.9 + (6.9 - 0) = 12.72$$

$$upper_2 = 6 + (6 - 0) = 11$$

$$upper_3 = 4 + (4 - 0) = 7$$

由此为数据范围重新统计，得到 3 个数据分布图，如图 4.9 所示。



(a)均值为主的近似分布　　(b)中位数为主的近似分布　　(c)众数为主的近似分布

图 4.9 书名长度属性的 3 个近似正态分布图

从实验结果分析图中，可以得到这三个分布图都是原始分布的近似正态分布群。

（2）计算权重

统计 3 个近似分布的代表性统计量，如表 4.14 所示。

表 4.14 书名长度属性 3 个近似分布的代表性统计量

nameLengthPure								
Statistics1			Statistics2			Statistics3		
N	Valid	103870	N	Valid	101140	N	Valid	79610
	Missing	0		Missing	0		Missing	0
Mean		5.80	Mean		5.64	Mean		4.69
Median		5.00	Median		5.00	Median		5.00
Mode		4	Mode		4	Mode		4
Std. Deviation		2.475	Std. Deviation		2.286	Std. Deviation		1.450
Variance		6.127	Variance		5.228	Variance		2.103

利用公式（4.6）计算标准化偏度，利用公式（4.7）计算权值的分配：

$$sSD = \frac{\left|Mean - Median\right|}{s}$$

$$w_{\min\_skewness}=\frac{sSD_{\max}}{\sum_{i=1}^{n}sSD_i}\ w_{med\_skewness}=\frac{sSD_{med}}{\sum_{i=1}^{n}sSD_i}\ w_{\max\_skewness}=\frac{sSD_{\min}}{\sum_{i=1}^{n}sSD_i}$$

解得：

$w_{\min\_skewness}=0.395637832$

$w_{med\_skewness}=0.342678438$

$w_{\max\_skewness}=0.26168373$

（3）计算加权均值和加权标准差

利用公式（4.8）和公式（4.9）计算：

加权均值的公式：

$$Mean=w_1Mean_1+w_2Mean_2+w_3Mean_3$$

加权标准差的公式：

$$s=w_1s_1+w_2s_2+w_3s_3$$

解得：

$Mean=5.306013456$

$s=2.004704997$

（4）确定数据分布的比例区域临界值。

将公式（4.8）和公式（4.9）的计算带入公式（4.10）当中，即可列出 z 值和 X_i 的关系等式。

$$z_i=\frac{(X_i-Mean)}{s}$$

利用概率等距 5 分法，从而将数值属性划分为 5 个区域，对应的面积取值、z 值取值、 X_i 计算结果以及从而确定的离散范围如表 4.15 所示。

表 4.15 书名长度属性的离散区间

序号	4 个面积点	4 个 z 值	4 个 X_i	5 个范围
1	-30%	-0.84	3.622061258	(0, 3]
2	-10%	-0.25	4.804837207	(3, 4]
3	+10%	+0.25	5.807189705	(4, 5]
4	+30%	+0.84	6.989965654	(5, 6]
5				>6

4.6.4 页码

（1）数据的正态分布规范化。

首先得到原始数据的数据分布图，如图 4.10 所示。

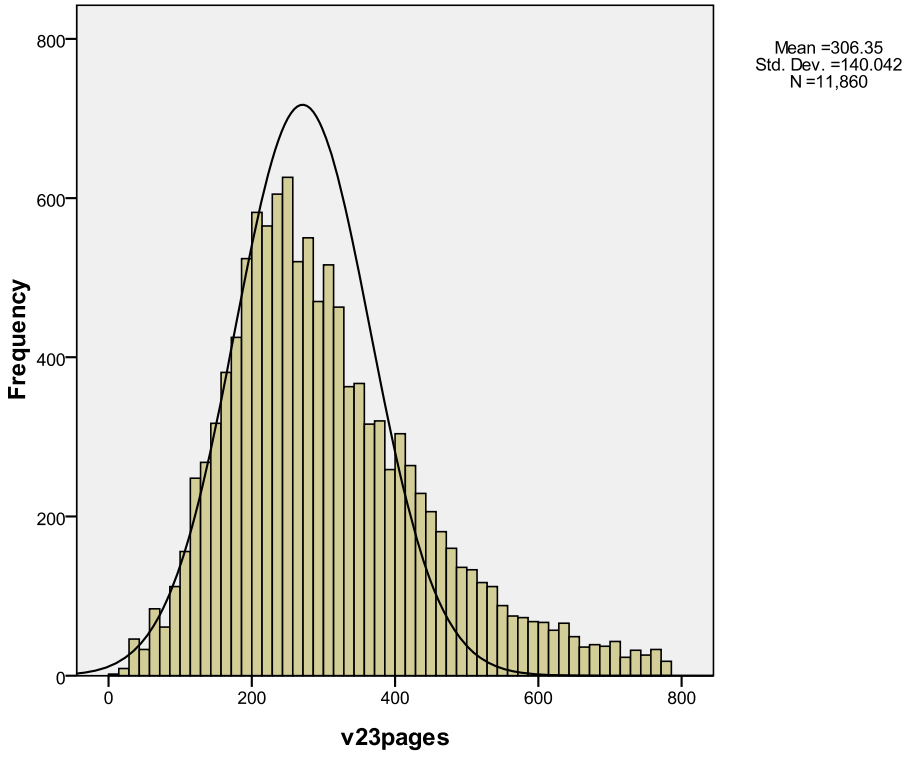


图 4.10 页码属性的原始数据分布图

得到该数据分布的代表性统计量，如表 4.16 所示。

表 4.16 页码属性的代表性统计量

Statistics		
pages		
N	Valid	111040
	Missing	0
Mean		352.33
Median		289.00
Mode		240
Std. Deviation		319.093
Variance		101820.125

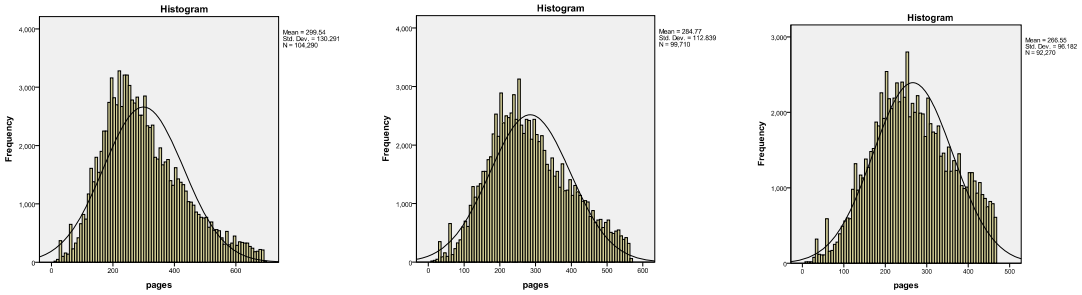
利用原数据分布的均值、中位数和众数作为代表数据，数据下限为 12，上限设定为 0～代表数的两倍，使用公式（4.4）进行正态分布的规范化处理。

$$upper = representativeNmuber + \left(representativeNmuber - lower\right)$$

$$upper_1 = 692.66$$

$$upper_2 = 566$$

$$upper_3 = 468$$



(a)均值为主的近似分布　　(b)中位数为主的近似分布　　(c)众数为主的近似分布

图 4.11 页码属性的 3 个近似正态分布图

由此为数据范围重新统计，得到 3 个数据分布图，如图 4.11 所示。从实验图中，可以得到这三个分布图都是原始分布的近似正态分布群。

（2）计算权重

首先统计出 3 个近似分布的代表性统计量，如表 4.17 所示。

表 4.17 页码属性 3 个近似分布的代表性统计量

v31time								
Statistics1			Statistics2			Statistics3		
N	Valid	104290	N	Valid	99710	N	Valid	92270
	Missing	0		Missing	0		Missing	0
Mean		299.54	Mean		284.77	Mean		266.55
Median		278.00	Median		272.00	Median		260.00
Mode		240	Mode		240	Mode		240
Std. Deviation		130.291	Std. Deviation		112.839	Std. Deviation		96.182
Variance		16975.814	Variance		12732.703	Variance		9250.932

利用公式（4.6）计算标准化偏度，利用公式（4.7）计算权值的分配：

$$sSD = \frac{|Mean - Median|}{s}$$

$$w_{\min\_skewness} = \frac{sSD_{\max}}{\sum_{i=1}^{n} sSD_i} \quad w_{med\_skewness} = \frac{sSD_{med}}{\sum_{i=1}^{n} sSD_i} \quad w_{\max\_skewness} = \frac{sSD_{\min}}{\sum_{i=1}^{n} sSD_i}$$

解得：

$w_{\min\_skewness} = 0.476993275$

$w_{med\_skewness} = 0.326522146$

$w_{\max\_skewness} = 0.196484579$

（3）计算加权均值和加权标准差

利用公式（4.8）和公式（4.9）计算：

加权均值的公式：

$$Mean = w_1 Mean_1 + w_2 Mean_2 + w_3 Mean_3$$

加权标准差的公式：

$$s = w_1 s_1 + w_2 s_2 + w_3 s_3$$

解得：

$Mean = 278.9812598$

$s = 108.3227719$

（4）确定数据分布的比例区域临界值。

将公式（4.8）和公式（4.9）的计算带入公式（4.10）当中，即可列出 z 值和 X_i 的关系等式。

$$z_i = \frac{(X_i - Mean)}{s}$$

利用概率等距 5 分法，从而将数值属性划分为 5 个区域，对应的面积取值、z 值取值、X_i 计算结果以及从而确定的离散范围如表 4.18 所示。

表 4.18 页码属性的离散区间

序号	4 个面积点	4 个 z 值	4 个 X_i	5 个范围
1	-30%	-0.84	187.9901314	(0, 187]
2	-10%	-0.25	251.9005668	(187, 251]
3	+10%	+0.25	306.0619527	(251, 306]
4	+30%	+0.84	369.9723882	(306, 369]
5				>369

4.6.5 价格

（1）数据的正态分布规范化。

首先得到原始数据的数据分布图，如图 4.12 所示。

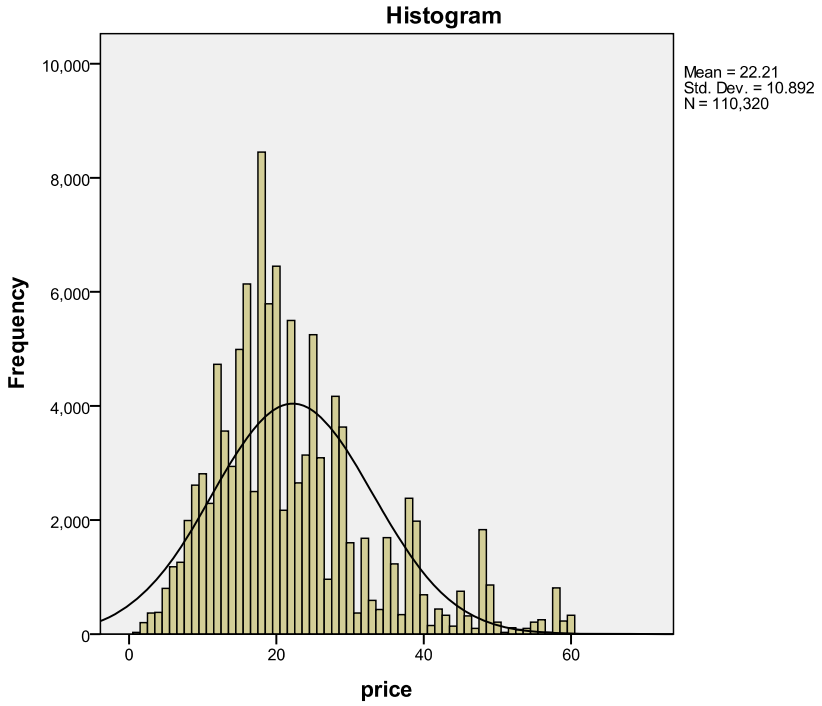


图 4.12 价格属性的原始数据分布图

统计该数据分布的代表性统计量，如表 4.19 所示。

表 4.19 价格属性的代表性统计量

Statistics		
score		
N	Valid	110320
	Missing	0
Mean		22.21
Median		20.00
Mode		18
Std. Deviation		10.892
Variance		118.637

利用原数据分布的均值、中位数和众数作为代表数据，数据下限为 1，上限设定为 1～代表数的两倍，进行正态化，即公式（4）

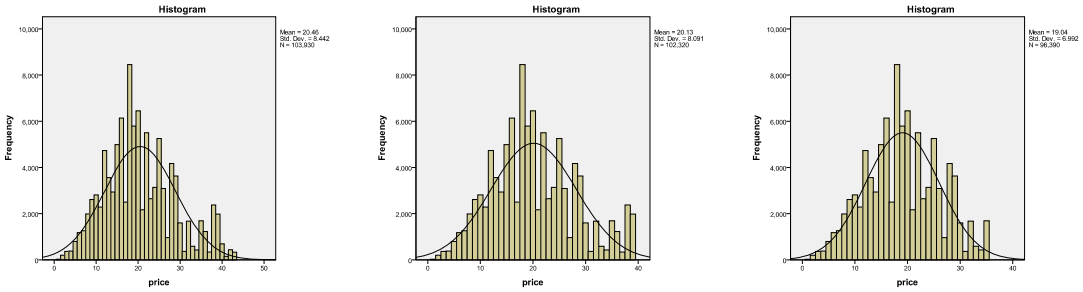
$$upper = representativeNmuber + \left(representativeNmuber - lower\right)$$

$upper_1 = 43.42$

$upper_2 = 39$

$upper_3 = 35$

由此为数据范围重新统计，得到 3 个数据分布图，如图 4.13 所示。



(a)均值为主的近似分布　　(b)中位数为主的近似分布　　(c)众数为主的近似分布

图 4.13 价格属性的 3 个近似正态分布图

从实验图中，可以得到这三个分布图都是原始分布的近似正态分布群。

（2）计算权重

首先统计出 3 个近似分布的代表性统计量，如表 4.20 所示。

表 4.20 价格属性 3 个近似分布的代表性统计量

v31time								
Statistics1			Statistics2			Statistics3		
N	Valid	103930	N	Valid	102320	N	Valid	96390
	Missing	0		Missing	0		Missing	0
Mean		20.46	Mean		20.13	Mean		19.04
Median		19.00	Median		19.00	Median		19.00
Mode		18	Mode		18	Mode		18
Std. Deviation		8.442	Std. Deviation		8.091	Std. Deviation		6.992
Variance		71.273	Variance		65.461	Variance		48.887

利用公式（4.6）计算标准化偏度，利用公式（4.7）计算权值的分配：

$$sSD = \frac{|Mean - Median|}{s}$$

$$w_{\min\_skewness} = \frac{sSD_{\max}}{\sum_{i=1}^{n} sSD_i} \quad w_{med\_skewness} = \frac{sSD_{med}}{\sum_{i=1}^{n} sSD_i} \quad w_{\max\_skewness} = \frac{sSD_{\min}}{\sum_{i=1}^{n} sSD_i}$$

解得：

$w_{\min\_skewness} = 0.543292944$

$w_{med\_skewness} = 0.438735523$

$w_{\max\_skewness} = 0.017971533$

（3）计算加权均值和加权标准差

利用公式（4.8）和公式（4.9）计算：

加权均值的公式：

$$Mean = w_1 Mean_1 + w_2 Mean_2 + w_3 Mean_3$$

加权标准差的公式：

$$s = w_1 s_1 + w_2 s_2 + w_3 s_3$$

解得：

$Mean = 19.5437413$

$s = 7.500229063$

（4）确定数据分布的比例区域临界值。

将公式（4.8）和公式（4.9）的计算带入公式（4.10）当中，即可列出 z 值和 X_i 的关系等式。

$$z_i = \frac{(X_i - Mean)}{s}$$

利用概率等距 5 分法，从而将数值属性划分为 5 个区域，对应的面积取值、z 值取值、X_i 计算结果以及从而确定的离散范围如表 4.21 所示。

表 4.21 价格属性的离散区间

序号	4 个面积点	4 个 z 值	4 个 X_i	5 个范围
1	-30%	-0.84	13.24354888	(0, 13]
2	-10%	-0.25	17.66868403	(13, 17]
3	+10%	+0.25	21.41879856	(17, 21]
4	+30%	+0.84	25.84393371	(21, 25]
5				>25

4.6.6 时间

（1）数据的正态分布规范化。

首先得到原始数据的数据分布图，如图 4.14 所示。

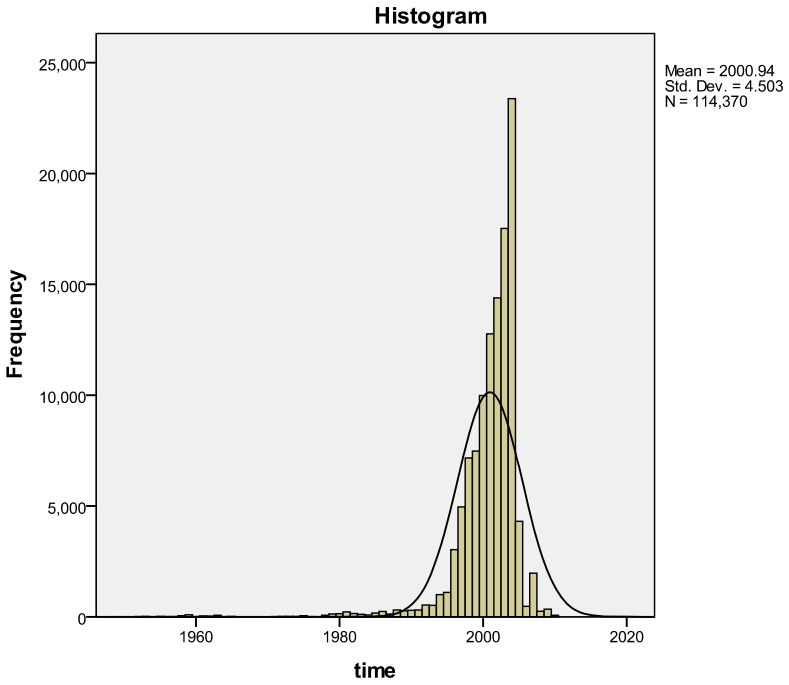


图 4.14 时间属性的原始数据分布图

统计该数据分布的几个统计量，如表 4.22 所示。

表 4.22 时间属性的代表性统计量

Statistics		
score		
N	Valid	114370
	Missing	0
Mean		2000.94
Median		2002.00
Mode		2004
Std. Deviation		4.503
Variance		20.281

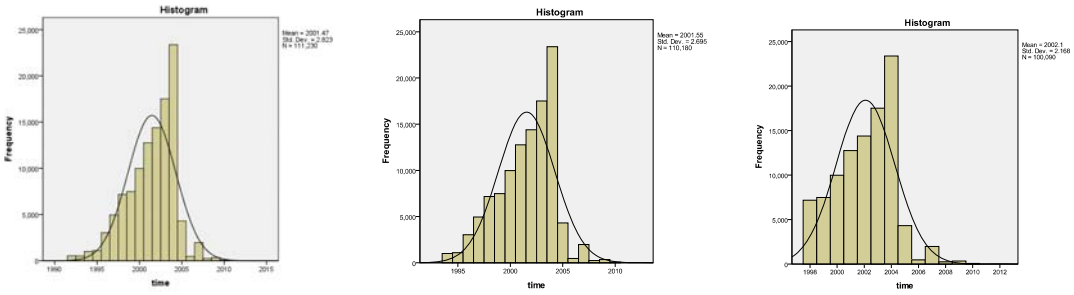
利用原数据分布的均值、中位数和众数作为代表数据，数据上限为 2010，下限设定为 0～代表数的两倍，进行正态化，即公式（4.3）

$$lower = representativeNmuber - (upper - representativeNmuber)$$

$$lower_1 = 1991.88$$

$$lower_2 = 1994$$

$$lower_3 = 1998$$



(a)均值为主的近似分布　　(b)中位数为主的近似分布　　(c)众数为主的近似分布

图 4.15 时间属性的 3 个近似正态分布图

由此为数据范围重新统计，得到 3 个数据分布图，如图 4.15 所示。从实验图中，可以得到这三个分布图都是原始分布的近似正态分布群。

（2）计算权重

首先统计出 3 个近似分布的代表性统计量，如表 4.23 所示。

表 4.23 时间属性 3 个近似分布的代表性统计量

time								
Statistics1			Statistics2			Statistics3		
N	Valid	111230	N	Valid	110180	N	Valid	100090
	Missing	0		Missing	0		Missing	0
Mean		2001.47	Mean		2001.55	Mean		2002.10
Median		2002.00	Median		2002.00	Median		2002.00
Mode		2004	Mode		2004	Mode		2004
Std. Deviation		2.823	Std. Deviation		2.695	Std. Deviation		2.168
Variance		7.967	Variance		7.266	Variance		4.701

利用公式（4.6）计算标准化偏度，利用公式（4.7）计算权值的分配：

$$sSD = \frac{|Mean - Median|}{s}$$

$$w_{\min\_skewness}=\frac{sSD_{\max}}{\sum_{i=1}^{n}sSD_i}\quad w_{med\_skewness}=\frac{sSD_{med}}{\sum_{i=1}^{n}sSD_i}\quad w_{\max\_skewness}=\frac{sSD_{\min}}{\sum_{i=1}^{n}sSD_i}$$

解得：

$w_{\min\_skewness}=0.468369551$

$w_{med\_skewness}=0.416559848$

$w_{\max\_skewness}=0.115070602$

（3）计算加权均值和加权标准差

利用公式（4.8）和公式（4.9）计算：

加权均值的公式：

$$Mean=w_1Mean_1+w_2Mean_2+w_3Mean_3$$

加权标准差的公式：

$$s=w_1s_1+w_2s_2+w_3s_3$$

解得：　$Mean=2001.798398$　　$s=2.462898284$

（4）确定数据分布的比例区域临界值。

将公式（4.8）和公式（4.9）的计算带入公式（4.10）当中，即可列出 z 值和 X_i 的关系等式。

$$z_i=\frac{(X_i-Mean)}{s}$$

利用概率等距 5 分法，从而将数值属性划分为 5 个区域，对应的面积取值、z 值取值、X_i 计算结果以及从而确定的离散范围如表 4.24 所示。

表 4.24 时间属性的离散区间

序号	4 个面积点	4 个 z 值	4 个 X_i	5 个范围
1	-30%	-0.84	1999.729563	(0, 1999]
2	-10%	-0.25	2001.182673	(1999, 2001]
3	+10%	+0.25	2002.414122	(2001, 2002]
4	+30%	+0.84	2003.867232	(2002, 2003]
5				>2003

4.6.7 相关性检验

依据 PDOZ 算法离散后的数据重新计算，得到四个属性与用户主观分数的相关系数，进行汇总分析，得到利用 PDOZ 算法后实验属性相关系数提升效果图，如图 4.16 所示，其中等距离散的宽度是指离散区间的个数。

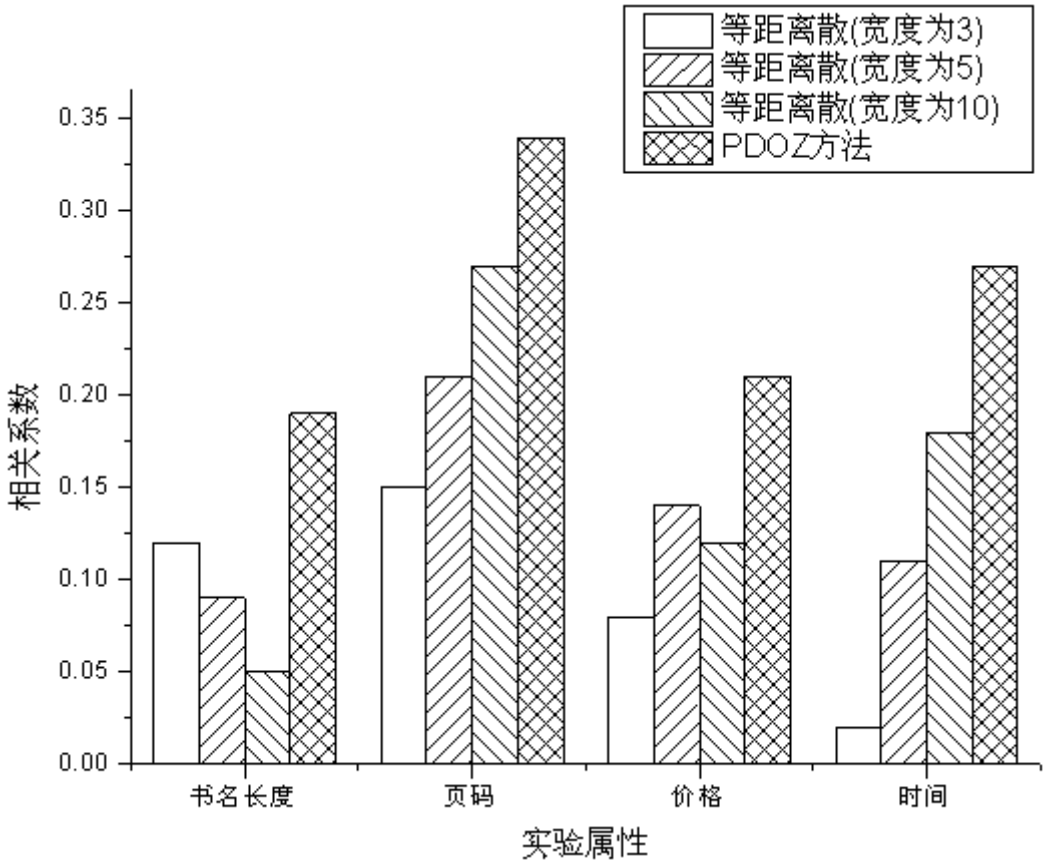


图 4.16 利用 PDOZ 算法后实验属性相关系数提升效果图

具体数据列表如表 4.25 所示。

表 4.25 利用传统离散和 PDOZ 离散计算相关系数对比列表

	书名长度	页码	价格	时间
等距离散(w=3)	0.12	0.15	0.08	0.02
等距离散(w=5)	0.09	0.21	0.14	0.11
等距离散(w=10)	0.05	0.27	0.12	0.18
PDOZ 方法	0.19	0.34	0.21	0.27

通过四组数据的相关性对比分析，可以清楚看到PDOZ算法的离散效果在与分类

属性的关联性上，好于常规的等距离散方法，更加适应于数字图书馆评估数据。

4.6.8 条件属性关系的验证

在验证了PDOZ算法的有效性之后，可以发现了价格属性虽然提升了很大的相关性，但是依然处于较低区间，这与直观感觉存在很大的不同，在接下来的实验中，针对这种现象进行了分析。

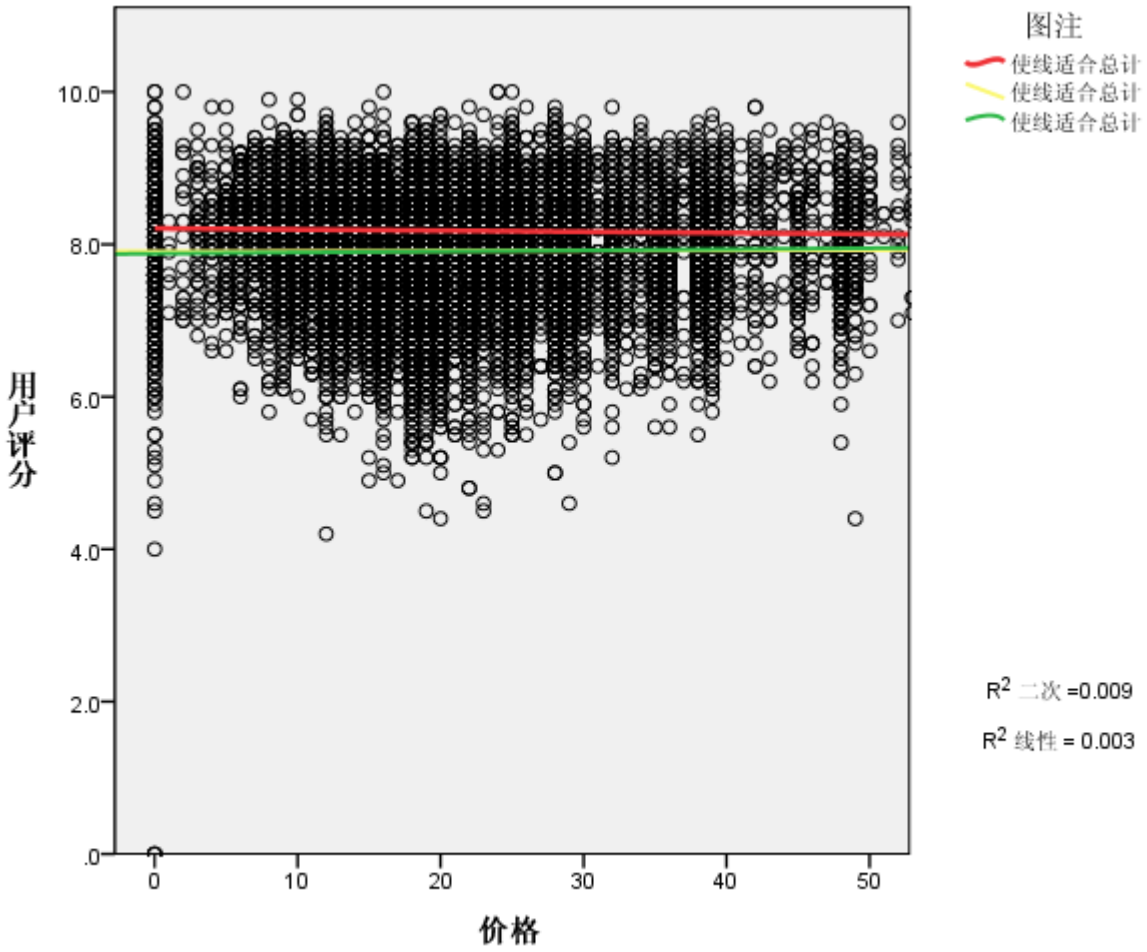
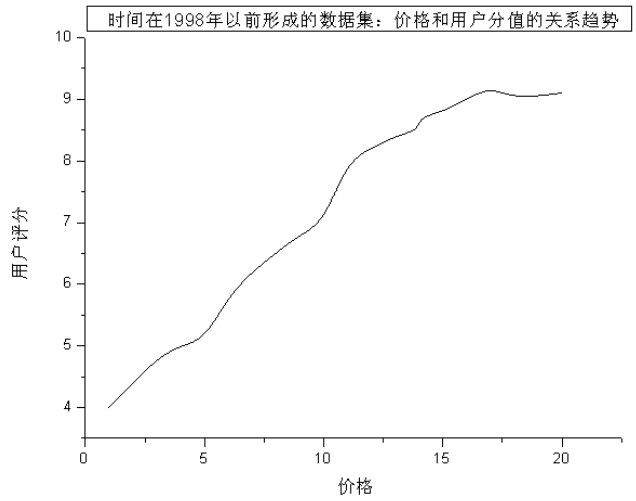
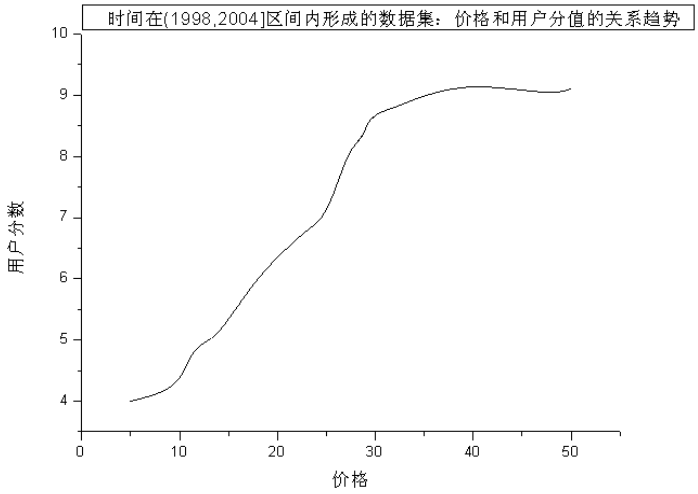


图 4.17 价格属性与用户评分属性的关系图

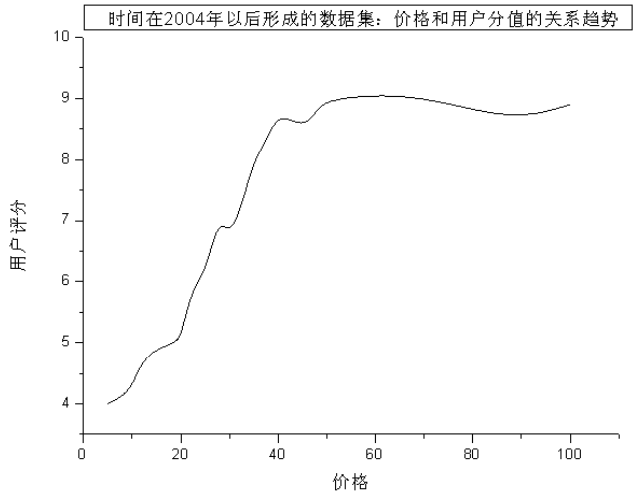
价格属性在人为判断来看应该是具有强相关特点，但是通过图4.17分析，本文发现，在最具代表性的数据当中（50元以内书籍），价格和用户评分没有关联，不同级别的数据拟合线也证明了这个结论，所以本文暂时认为价格对与馆藏价值的贡献为零。但在局部测量当中往往发现会有偶尔的高关联性，选取时间作为条件属性，价格为观察属性，重新进行分析，在PDOZ算法的作用下，可以发现：



(a) 1998 年以前



(b) 1998 年到 2004 年之间



(c) 2004 年以后

图 4. 18 价格属性在不同时间区间中与用户评分的关系趋势图

通过图 4.18 可以清晰的看到，当以时间属性来划分数据集合后，在单个数据子集中，价格明显和用户评分出现了线性关系。这种现象，本文称存在了“非线性条件属性”关系。本文将“非线性条件属性”关系定义为属性之间没有线性或其他明确的数学关系。假设一个数据集的属性集合为$\{A,B,C,\cdots,Z\}$（其中Z是类别属性，属性B是属性A的非线性条件属性），常存在这种情况：A和B之间没有相关性，而且两者与类别属性Z之间的相关性也都很弱，此时若直接应用现有数据挖掘算法，A会被当作弱区分能力的属性，而其实A可能与Z之间并非弱相关，A对于类别预测可能是有重要作用的，是具有很强的数据区分能力的，但是需要依据属性B进行数据集的拆分（因为B同样弱相关于Z，所以利用传统决策树算法是无法直接选择出B作为分裂节点的），在子集中体现出了很强的区分能力，此时这种潜在的强相关属性就会忽视而导致数据挖掘的效果大打折扣。

在数字图书馆的馆藏数据集中，时间属性是价格属性的条件属性，直接计算会发现价格和用户评分之间的相关系数是很低的，但是依据年代进行数据集的划分，可以明显发现在数据子集中价格与用户评分之间的相关性大大提升，主要原因在于书籍价格是具有时代特征的，同一时代书籍的价格之间才有可比性，因此价格属性在时间属性的条件下，可以发挥出对用户评分的较强贡献。

4.7 本章小结

本章首先针对体现课题评估思想的馆藏评估算法进行了分析，提出了软属性的概念，在馆藏属性中引入了体现用户感受的主观数据，作为解决馆藏质量量化问题的桥梁；然后对数字图书馆馆藏数据的复杂属性关系展开研究，提出了一种基于 z 值的并行离散算法作为数字馆藏分析的基础，该算法解决的是数字图书馆评估数据中连续性数据的离散问题。该算法的优点在于可对比、同步并行处理的特性，可对比体现在将蕴含 z 值标准化的处理思想借鉴过来，从而使不同属性处于一个对比单位平台下，同步并行处理体现在将一个属性之间复杂关联的数据集依据同一个策略进行了处理，同步并不等于同样的离散尺度和标准，而是将蕴含 z 值标准化的处理思想借鉴过来，用这种基于 z 值的有概率分布意义的动态距离代替传统离散方法，反映了馆藏各个属性语义的动态变化，增强了属性与预测之间的相关性，更利于发现属性关系，并基于这种离散方法研究了馆藏属性之间的关系，发现了非线性条件属性的存在。

第 5 章 用于馆藏评估预测的层叠决策树算法研究

5.1 引言

在 SOI 评估模型中，馆藏模块评估的优点在于将质量评估和规模评估结合在一起，其中质量评估是基于馆藏属性研究对馆藏评估分数的预测，而传统预测技术直接应用的效果并不理想，使我们需要开展用于馆藏评估的预测算法研究。

本章提出了一种层叠决策树算法（SDT-Z 算法，Stratified Decision Tree based on PDOZ），该算法解决现有决策树算法在数字图书馆馆藏评估数据集中应用不理想的问题。在前面章节的研究中，发现了数字图书馆馆藏评估数据集中存在非线性条件属性关系，传统的决策树算法在这种关系下无法发挥作用，SDT-Z 算法引入层属性的概念，将决策树选取属性从平面信息的选择拓展到了立体信息，使非线性条件属性对数据挖掘的干扰得到有效转化，并可以嵌套使用已有决策树算法，具备较强的灵活性，解决了数据挖掘在数字图书馆评估应用研究中的瓶颈。

5.2 决策树算法在复杂属性中的应用

在数据挖掘领域的预测算法当中，决策树的方法是最实用而且应用也是最广泛的数据预测技术之一，决策树的算法原理在于找出当前数据集里面最具区分不同数据能力的分裂属性，然后依照该分裂属性的取值把数据集合划分为许多个子集，每一个子集递归调用分裂属性的选取过程，直到所有子集都尽可能的包含同一类别的数据，从而形成的一个决策树预测模型，并依此对新的待处理数据集进行预测分类。

决策树算法过程当中，分支节点属性（分裂属性）的选取算法处于核心位置。例如 ID3 算法依据信息论中的互信息理论进行信息增益的计算，并将计算结果作为属性区分数据能力的度量，互信息的计算过程依赖于取值数量较多的属性，造成了取值较多的属性更容易被选择为“最优”，于是在随后的 C4.5 算法中利用信息增益率的方法解决了这个问题，之后又出现了多种决策树算法，每种算法的适用条件略有不同，但算法思想基本上是一致的，决策树分支节点属性的选取算法比较出名的有：J.R.Quinlan 提出了信息增益标准[162, 163]、L.Breiman 等在 CART 系统中提出的 Gini-Index 标准[164, 165]、J.Mingers 提出了 χ^2 统计标准[166]、K.Kira 等提出了 Relief 标准[167, 168]、S. J.Hong

等提出了 CM 标准[169]等。

出现这么多分支节点属性的选取算法正是因为数据属性的关系是非常复杂的，根据属性之间相互依赖的关系，将这些节点属性选取算法分为两个大类：独立属性假设的算法和非独立属性假设的算法。独立属性假设的策略是假设各个属性之间是相互独立的关系，而非独立属性假设的策略是假设各个属性之间互有影响。

独立属性假设策略的特点是工作效率相对较高，而且在实际应用当中往往简单有效；非独立属性假设策略的特点是追求理论上最真实的属性关系。要注意到一点，在非独立属性假设的研究当中，属性之间的影响往往是线性的或者有明确映射关系的，而在数字图书馆当中，属性之间存在非线性条件关系，利用现有的选取策略并不适用，而且非线性条件关系一旦被选择出来，其他属性关系之间是独立的，所以应该寻找一种合理而高效的方法来解决这个问题。

5.3 层叠决策树算法 SDT-Z

针对数字图书馆中馆藏资源的数据特点，在属性分析基础之上提出一种解决非线性条件属性关系影响的数据预测算法——SDT-Z 算法（Stratified Decision Tree based on PDOZ）。

5.3.1 算法思想

如上所述，决策树算法作为数据挖掘中的一种重要方法，主要用于数据的分类和预测。决策树算法的基本思想是在数据集中选取一个属性作为类别属性，其他属性作为普通属性，接着按照一定的运算规则在普通属性集当中选择符合要求的一个来作为决策树的一个分支节点，并依据该属性的不同属性值对数据集进行拆分，然后递归的调用运算规则进行这个过程，直到每一个数据子集都唯一地对应一个类别或者满足临界条件。

本章提出的 SDT-Z 算法也体现了决策树的上述基本思想，主要不同在于分支节点属性的选取方法。现有决策树算法选择分裂属性的算法虽然很多，但面对非线性复杂条件属性时效果却并不理想，主要原因在于数字图书馆馆藏数据之间的这种条件属性关系具有数学关系不明确、不固定的特点。

SDT-Z 决策树算法的分支节点属性的选取引入了“分层”的概念，不仅针对平面信息，在立体层信息的获取上对属性进行控制，从而消除条件属性的影响。

可以将算法中分支节点的类别区分为以下两种：

（1）层级分支节点

层级分支节点是指在方法当中起到化解条件属性影响的节点，对应于分层属性。这种分层属性可能存在多个，依照分层能力排序并拆分数据集。

（2）区间分支节点

区间分支节点是指在通过层级节点消除条件属性影响之后，在底层进行子集划分的节点，可以理解为传统决策树的分裂节点。在层级节点处理之后，在内部利用已有的分裂方法（比如信息熵）来进行决策树构建。

5.3.2 SDT-Z 算法描述

综上所述，基于 PDOZ 的层叠决策树算法可以描述为：

SDT 算法描述

输入：（1）原始数据集合;

（2）相关系数阈值;

（3）提升阈值;

BEGIN

枚举出所有非类别属性，假设总数为 n，分别计算其与类别属性的相关系数;

对计算的相关系数进行排序，取出相关系数小于阈值的 m 个属性，将该属性放入待选分层属性簇中;

WHILE(!isEMPTY(待选分层属性簇))

在待选分层属性簇中取出一个属性 $A_i\left(1\le i\le m\right)$*，按照其属性值拆分数据集;*

在每个子数据集中，重新计算其余 m-1 个待选分层属性与类别属性之间的相关系数;

IF(存在明显的相关系数提升而且相关系数突破提升阈值限定)

则设定属性 $A_i\left(1\le i\le m\right)$ *为确定分层属性;*

ELSE

则不记录;

END WHILE

形成了真正的分层属性簇;

在分层属性簇中（假设数目为 k），依据贡献度（提升的属性数目和相关值）进行从大到小排序 $\{A_1,A_2,\cdots A_k\}$ *;*

依次取出分层属性，进行数据集的拆分，其中A_1作为第一层根属性，接着按照A_2在第二层进行数据集拆分，如此循环完毕，由分层属性簇形成了层级决策树；

在层级决策树的叶子节点对应的子数据集中嵌套应用已有的高效的独立属性假设的决策树挖掘算法，形成传统决策树群；

合并整理决策树群；

End

5.3.3 SDT-Z 算法的流程图

概括来说，SDT-Z 算法的流程如图 5.1 所示：

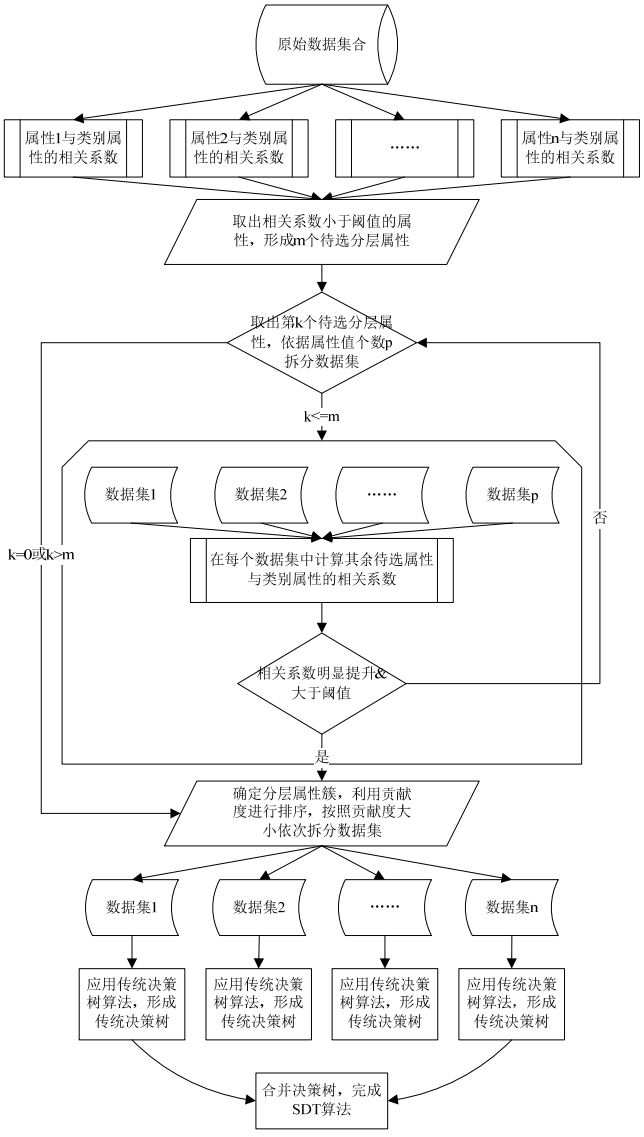


图 5.1 SDT 算法流程图

5.3.4 SDT-Z 算法的具体实施步骤

实施方式的具体步骤包括：

（1）枚举出所有非类别属性，假设总数为 n，分别计算其与类别属性的相关系数。

（2）取出相关系数小于阈值的 m 个属性，放入待选分层属性簇中。

对计算的相关系数进行排序，取出相关系数小于阈值的 m 个属性，这里的阈值是对相关系数的控制，设置太低就可能丢失具有隐蔽性的条件属性，而设置太高则会将过多的非条件属性放入待选分层属性簇，影响方法的效率，阈值的选取要根据具体数据集的特点进行调整，相关系数取值区间为$[0,1]$，在相关系数计算中，相关系数在 0.3 以下属于弱相关或无关，0.6 以上属于强相关，中间区间属于中度相关。假如数据集中无关属性过多，可以适当减小阈值以达到过滤部分无关属性的目的，反之可以提高阈值以获取属性。另外，可以根据实验最终结果的优劣来不断调正该阈值以达到最理想的效果。

（3）在待选分层属性簇中取出一个属性$A_i\left(1\le i\le m\right)$，按照其属性值拆分数据集。

（4）在数据子集中，重新计算各属性相关系数，根据各属性相关系数提升情况，判断该待选分层属性A_i是否为真正的分层属性。

在每个数据子集中，重新计算其余 m-1 个待选分层属性与类别属性之间的相关系数，如果存在明显的相关系数提升而且相关系数提升的属性数目突破提升阈值，则设定属性$A_i\left(1\le i\le m\right)$为确定分层属性，否则不记录。

这里的提升阈值是确定某属性是否作为真正的分层属性的判别值，每当数据子集中有属性的相关系数提升至 0.3 以上，就计数 1 次，第四步完成的时候，就形成了属性$A_i\left(1\le i\le m\right)$的最终的累加计数值，将该累加计数值与提升阈值进行比较。提升阈值的设定与具体的数据集有关，一般设定为待选分层属性值的个数。

（5）重复第三步和第四步，得到真正的分层属性簇。

（6）在分层属性簇中（假设数目为 k），依据贡献度进行从大到小排序$\left\{A_1,A_2,\cdots A_k\right\}$。

本实施例中的贡献度为第四步中得到的属性的最终累加计数值。

（7）依次取出分层属性，进行数据集的拆分，由分层属性簇形成了层级决策树。

其中A_1作为第一层根属性，接着依据A_2在第二层进行数据集拆分，如此循环完毕，形成了层级决策树。

（8）在层级决策树叶子节点对应的数据子集中嵌套应用已有的挖掘方法，形成传统决策树群，运行完毕，形成完整的层叠决策树，用于对待处理数据集进行预测分类及数据挖掘。

5.3.5 SDT-Z 算法的补充说明

SDT-Z 算法使用“层级节点”的策略，不但可以消除非线性条件属性的负面影响，而且使层级处理之后的数据集群转化为了常态数据集合，从而可以选取已有的高效挖掘算法嵌套应用其中，根据数据集的特点，选取的灵活性很大，从而也使 SDT 算法具备了很强的适应性。

关于 SDT-Z 算法，有以下问题需要阐明：

第一，在第 4 章里详细阐明了 PDOZ 算法，在本章提出的层叠决策树算法当中，是基于这种算法的，但并非所有的存在非线性条件属性关系的数据集都必须使用 PDOZ，这主要取决于具体数据集的属性特点，在数字图书馆馆藏当中存在同步离散的需要，例如在时间年代上恰好和价格存在关系。另外在应用 PDOZ 算法当中，也并非全部的属性都需要同步，尤其是当分类属性是数值型的时候，这个由具体的数据意义来决定。

第二，在 SDT-Z 算法中，利用 PDOZ 数据同步离散算法时，在 z 参数上一般依据 3 段、5 段、10 段来进行选择，这并不意味着 z 参数只能是这三种情况。理论上，越多区间段参与到算法当中，就越能发现复杂属性的规律，但是在时间复杂度上的代价是很大的，效果也不一定最好，这个需要根据具体的数据特点来确定。

第三，SDT-Z 算法具有普适性，SDT-Z 算法的优势在于能自适应于存在条件属性的复杂环境，又能充分利用现有决策树算法的实用和高效，SDT-Z 算法是针对数字图书馆馆藏属性的研究而研发出来的，但是这个算法并不局限于数字图书馆领域，它适合类似的存在条件属性的复杂关系的数据挖掘，尤其是在存在连续性属性的时候，可以通过 PDOZ 将这些数据同步离散，效果会更好。现实生活当中，存在很多这类数据，例如，包含年代属性的电脑销量数据，某台电脑在当时属于一流质量和顶级品牌，在 1990 年以前、1990 年～2000 年和现在都并不一定对销量的贡献是一致或者是呈现线性的关系，因为 90 年代以前电脑可能并不普及，而现在因为各种电子设备的冲击，电脑对于很多人来说，其所需功能都有时尚的替代品，比如苹果的 iPAD。但是，当针对不同职业的人来说，却呈现出了一致性或者线性的关系（可以直接应用现有决策树算法），比如对于从事计算机软件开发公司，电脑的质量和品牌是决定电脑在该行

业中销量的核心属性。

5.4 实验与分析

5.4.1 数据集描述

本章的数据集同样来自于豆瓣网，为了验证 SDT-Z 算法的效果，抽取馆藏数据集中的作者数量、有无译者、价格、时间、整除特点、用户倾向（即读过人数、在读人数、想读人数三者之和）、用户评分的等级等属性组合在一起构建实验数据集来验证 SDT-Z 算法的有效性。设计四个实验，均在数据集为 20000,40000,60000,80000,100000 五种规模级别上进行。

其中用户评分的等级作为预测的类别属性。这些属性除了价格和时间外都是与分数有强相关的关系，而价格和时间属性是作为非线性条件属性的干扰存在，以此来检验 SDT-Z 算法的使用效果，其中的连续性属性需要以 PDOZ 算法进行同步并行离散。

为了保证算法的准确性，数据集在“用户评分的等级”属性上，进行了均衡取样，每次实验，各个评分等级对应的数据记录个数保持相等。

5.4.2 实验设计

围绕 SDT-Z 算法的研究，设计以下实验来检验算法的可行性以及算法的实际效果。

实验 1：分别设定非线性条件属性在数据集中存在和不存在两种情况，运用传统的决策树算法进行数据预测，将结果进行分析，检验传统决策树算法对含有非线性条件属性的数据集的挖掘效果。

实验 2：利用实验 1 中相对效果最好的决策树算法对原始的数据进行预测分析实验，然后利用 SDT-Z 方法对原始的数据进行预测分析实验，将两者进行对比，检验效果差别。

实验 3：分别利用 SDT 和 SDT-Z 方法对数字图书馆馆藏数据进行预测分析的对比实验，检验应用第 4 章的 PDOZ 算法是否有助于 SDT 算法的效果提升。

5.4.3 实验结果分析

（1）在实验 1 中分别设定非线性条件属性在数据集中存在和不存在两种情况，运用传统的决策树算法进行数据预测，在准确率和召回率上进行结果对比。

实验 1 中，共选取了 5 种常用决策树算法，每个算法都分别在数据集为

20000,40000,60000,80000,100000 五种规模级别上进行了实验，并保证每次实验的数据集是均衡数据集，最后取 5 次实验结果的均值参与到最终的效果对比。

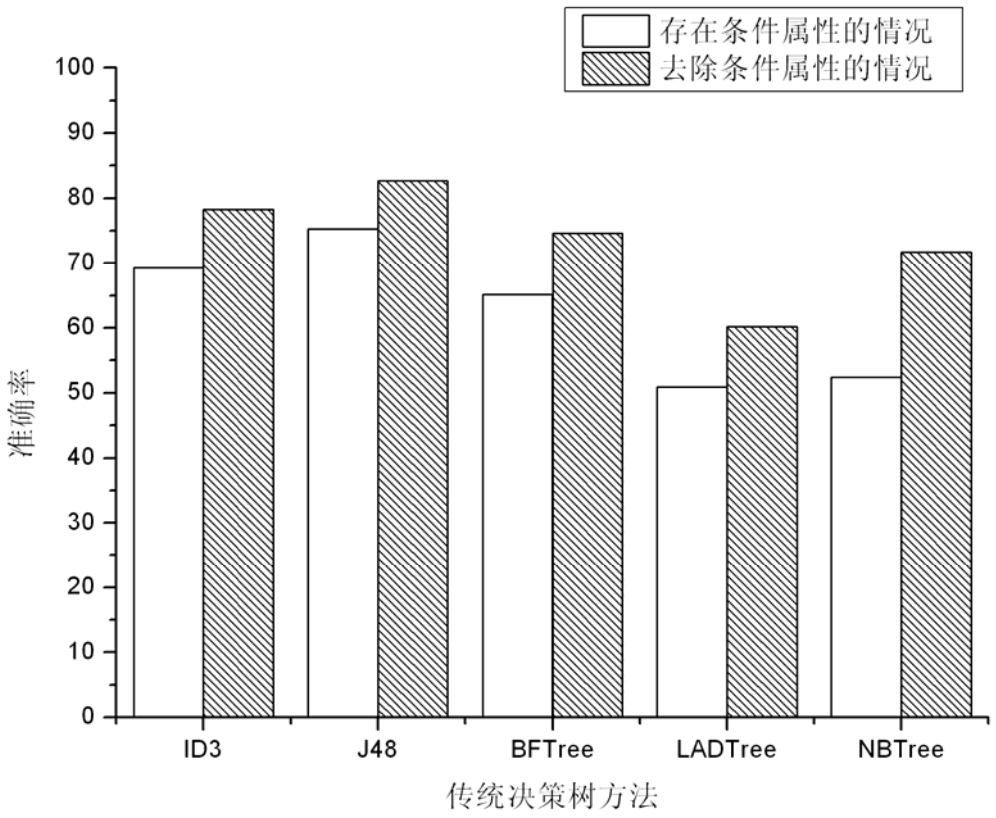


图 5.2 不同决策树算法在条件属性有无前提下应用的准确率对比图

通过图 5.2 所示，可以看到在去除非线性条件属性后，准确率有了明显的提升，但大部分都在 80%以下，并不理想，具体提升数值如表 5.1 所示。

表 5.1 五种常用决策树算法在条件属性存在和去除之后准确率的提升

算法	ID3	J48	BFTree	LADTree	NBTree
存在条件	69.3	75.3	65.2	50.9	52.4
去除条件	78.3	82.7	74.6	60.2	71.7
提升数值	9	7.4	9.4	9.3	19.3
提升比例	13.0%	9.8%	14.4%	18.3%	36.8%

从表 5.1 中可以看出，去处非线性条件属性之后，性能的提升还是比较明显的，以 ID3 算法的提升为最多，在效果的绝对值上 J48 的效果是最好的。

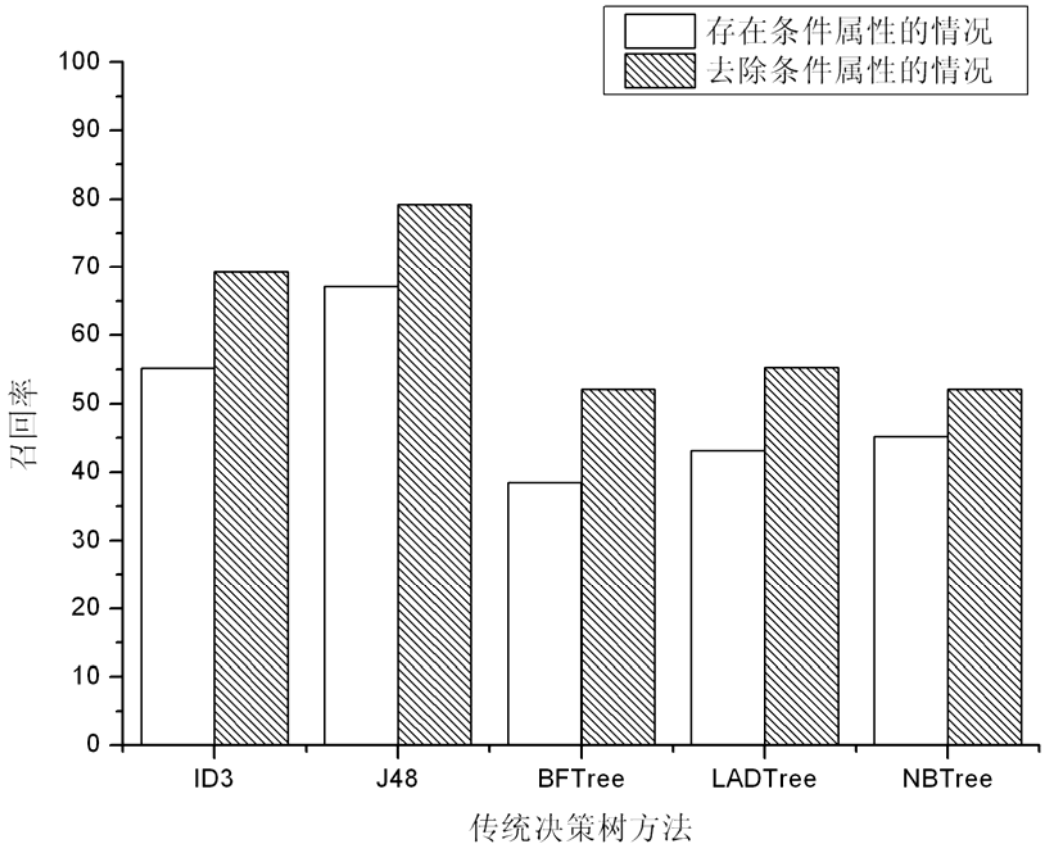


图 5.3 不同决策树算法在条件属性有无前提下应用的召回率对比图

通过图 5.3 可以看到召回率的提升也是明显的，甚至超过了准确率的提升幅度，但是在绝对数量却几乎都在准确率之下。

表 5.2 五种常用决策树算法在条件属性存在和去除之后召回率的提升

算法	ID3	J48	BFTree	LADTree	NBTree
存在条件	55.2	67.1	38.4	43.1	45.2
去除条件	69.3	79.2	52.1	55.3	52.1
提升数值	14.1	12.1	13.7	12.2	6.9
提升比例	25.5%	18%	35.7%	28.3%	15.3%

通过表 5.2，可以准确的看到，提升的幅度都是超过准确率的，但是有一个很重要的原因是提升基数很低，即在效果的绝对值上，召回率是很低的，几乎是不能应用在具体的实践当中的。

（2）实验 2 是将传统决策树算法中相对效果最好的算法和 SDT-Z 算法进行比较，

分别将两者应用在数字图书馆馆藏数据集中进行预测分析实验，进行效果的对比检验。

通过实验 1 可知在传统决策树算法中 J48 的在准确率和召回率的效果都是最好，因此在下面实验 2 中重点比较了 SDT-Z 算法和 J48 算法的效果。

实验 2 中，同样在数据集为 20000,40000,60000,80000,100000 五种规模级别上进行了实验，并保证每次实验的数据集是均衡数据集。

实验结果展示如图 5.4、图 5.5 所示：

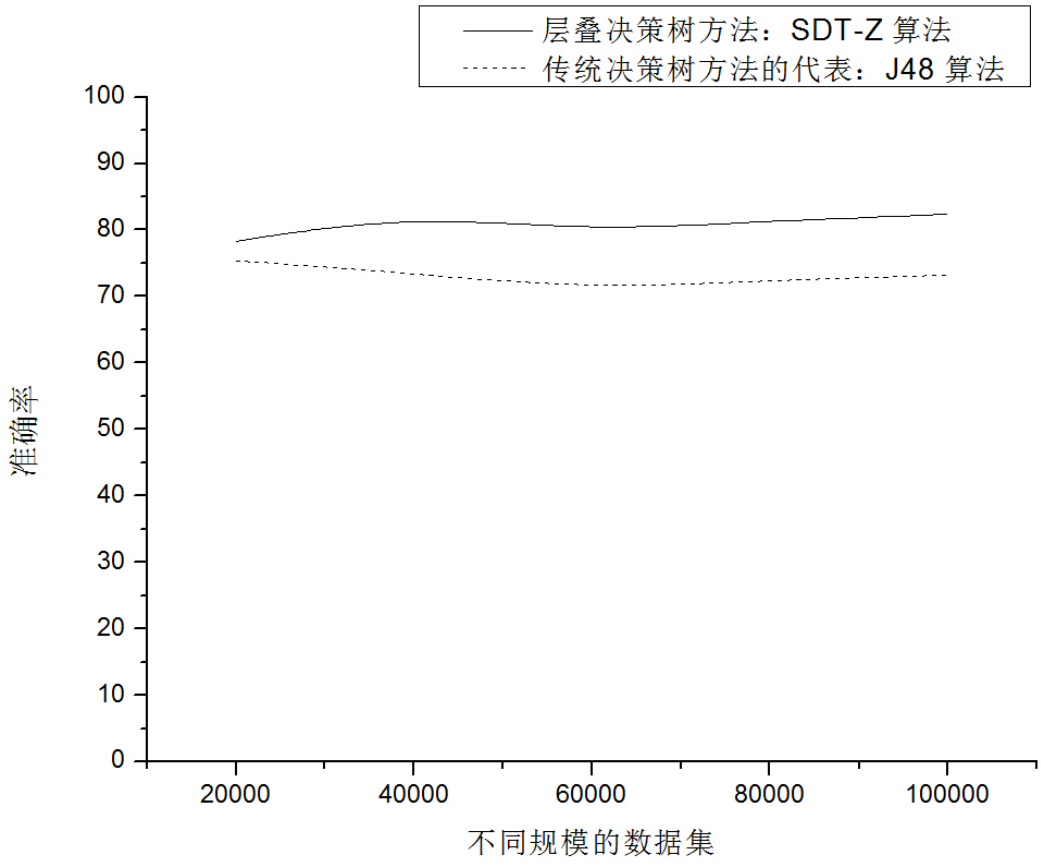


图 5.4 SDT-Z 算法和 J48 算法在准确率上的对比图

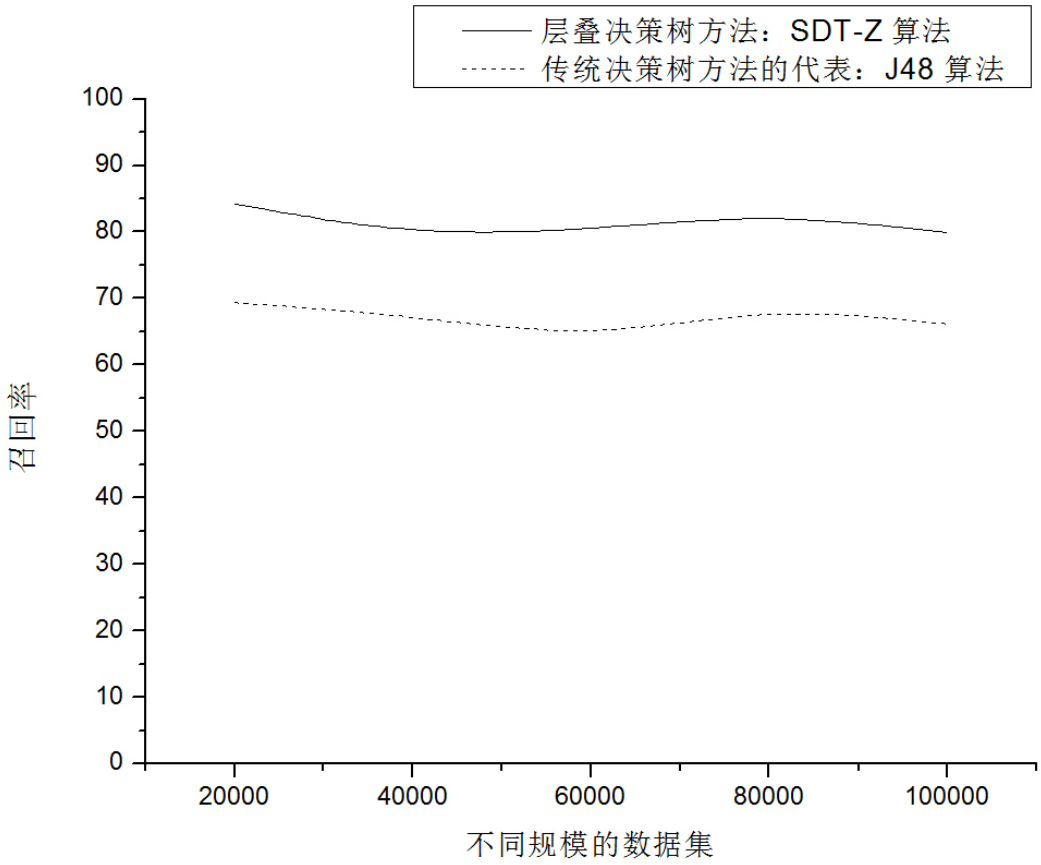


图 5.5 SDT-Z 算法和 J48 算法在召回率上的对比图

从上面两个图示中可以看到，利用 SDT-Z 算法还是有较大性能上的提升，这说明在数字图书馆馆藏数据集中，SDT-Z 算法是行之有效的。需要注意的是，这里使用的软性数据集主要参考于豆瓣网的数据，而该网站是开放性互联网站，并没有对用户进行约束和分类，因此软数据的质量并不能得到严格的保障，在实际的应用当中，可以依据用户的信誉值或贡献值等属性对用户进行可信度类别的划分，具备较高可信度的用户的意愿和评分应该被认为是高质量软性数据来源。当我们人工挑选高质量的用户数据进行泛化试验的时候，发现准确率和召回率的提升是更加明显的，可以达到 90%以上，说明该算法对于这种具有特殊数据特征的馆藏集是很有效的，这种对用户数据可信度的判定及相关工作是未来值得研究的方向之一。

5.5 本章小结

在基于数据挖掘的馆藏评估研究当中，建立馆藏质量的预测模型是很有意义的，

有利于改善传统馆藏评估中只重数量的情况，而在应用数据挖掘算法进行馆藏分值预测时，数字图书馆数据集的复杂关系使现有数据挖掘方法无法取得理想效果，在本章的研究中提出了一种层叠决策树算法，该算法有效的转化了非线性条件属性的复杂关系，其算法思想上具有较强的灵活性，理论上适应所有具有复杂关系的数据集，而且层内可以兼容现有的决策树算法，解决了数据挖掘在数字图书馆评估应用研究中的瓶颈。

第 6 章 结论及进一步的工作

6.1 论文总结

本文是围绕数字图书馆评估展开的系统研究。首先分析了数字图书馆评估研究，明确了评估的一系列基本性质，在此基础上提出了主、客观因素相融合的评估模型，并设计了基于数据挖掘的馆藏评估方法，该方法的研究实践中主要存在两个问题，一是数据离散的问题，二是数据预测的问题。针对以上两个问题，在随后的研究中提出了一种基于 z 值的并行离散算法，有效提高了馆藏属性与预测属性之间的相关系数，并发现了非线性的条件属性关系，最终提出了一种新的数据挖掘算法来解决这种复杂特点下的数据预测问题。

本文取得的主要创新点表现在以下几个方面：

（1） 提出一个数字图书馆评估四元理论模型

总结了现有数字图书馆评估研究的特点，将目前主流评估思想进行区分。分别分析了数字图书馆评估客体、评估目的以及评估主体的问题，明确了评估客体的基本性质、辨析了评估目的的混淆观点，并研讨了评估主体的特点，基于这些研究提出一个数字图书馆评估四元理论模型，并论述了四元理论模型的内部联系以及与数字图书馆评估思想之间的关系。该模型是数字图书馆评估工作的基本理论指导，体现了本文的评估思想。

（2） 提出一种主观因素与客观因素相融合的数字图书馆评估模型（SOI 模型）及其体系结构

基于数字图书馆评估研究的指导，化解主流思想分歧 ，提出一种将主观因素和客观因素相融合的数字图书馆评估模型，论述了这种评估思想、融合机制及其体系结构。

（3） 围绕数字图书馆馆藏开展属性研究，提出了一种基于 z 值的并行离散算法（PDOZ 算法），并发现了非线性条件属性的存在

论文首先在馆藏属性中引入了体现用户感受的主观数据，作为解决馆藏质量量化问题的桥梁；然后对数字图书馆馆藏数据的复杂属性关系

展开研究，提出了一种基于 z 值的并行离散算法（PDOZ 算法），这种基于 z 值的有概率分布意义的动态距离代替传统离散方法，反映了馆藏各个属性语义的动态变化，增强了属性与预测之间的相关性，更利于发现属性关系，并基于这种离散方法研究了馆藏属性之间的关系，发现了非线性条件属性的存在。

（4） 提出了一种层叠决策树算法 SDT-Z

基于馆藏属性的研究，针对数字图书馆馆藏价值的预测挖掘提出了一种层叠决策树算法，解决了传统挖掘算法在数字图书馆馆藏数据集中应用不理想的问题，该算法引入层属性的概念，将决策树选取属性从平面信息的选择拓展到了立体信息，使非线性条件属性对数据挖掘的干扰得到有效转化，并可以嵌套使用已有决策树算法，具备较强的灵活性，解决了数据挖掘在数字图书馆评估应用研究中的瓶颈。

6.2 进一步的工作

本文在数字图书馆评估研究过程中，尝试用主、客观因素相融合的评估思想和数据挖掘技术来解决数字图书馆评估问题，根据已经取得的研究成果和目前的工作进展，本文提出下一步的工作方向：

（1）在 SOI 评估模型中，论文仅针对其中馆藏模块进行的数据挖掘实践应用就引出了第 4、5 章的内容，而技术模块和管理模块只是提出基本的范畴和处理的原则，需要在进一步的研究中确定如何在这些模块当中发挥客观评估和主观评估相融合的优势。

（2）在第 5 章层叠决策树算法的研究中，利用分层属性来消除条件属性影响的策略比较清晰实用，分层属性的选取算法在条件属性不多的情况下效率较高，若条件属性较多，则算法的复杂度就会大幅增加，需要解决效率问题。

（3）在基于数据挖掘的馆藏评估研究中，本论文是围绕中文数字图书馆的书籍类馆藏进行的研究，其成果如何泛化到其他类别的数字馆藏，都是值得深入研究的工作。

参考文献

[1]He L. A Review on Overseas Evaluation of Digital Library[J]. Journal of Library Science in China, 2010, 36(190): 88-94.

[2]Bishop A. P.,Van House N. A.,Buttenfield B. P. Digital library use: Social practice in design and evaluation[M]. The MIT Press, 2003:12-34.

[3]Arms W. Y. Digital libraries[M]. The MIT Press, 2001:18-32.

[4]Lesk M. Practical digital libraries: Books, bytes, and bucks[M]. Morgan Kaufmann, 1997:28-45.

[5]Dobreva M. Evaluation of Digital Libraries: An Insight into Useful Applications and Methods[J]. Library review, 2011, 60(2):166-168.

[6]McNicol S. The eVALUEd toolkit: a framework for the qualitative evaluation of electronic information services[J]. Vine, 2004, 34(4):172-175.

[7]Larsen, Ronald L. The DLib Test Suite and Metrics Working Group: harvesting the experience from the Digital Library Initiative[EB/OL]. (2002)[2011-3-1]. http://www.dlib.org/metrics/public.

[8]Thompson R. L.,Thompson B.,Heath F. M.,et al. The Search for New Measures: The ARL LibQUAL+ Project-A Preliminary Report[J]. Portal: Libraries and the Academy, 2001, 1(1):103-112.

[9]Brophy, Peter, et al. EQUINOX: Library Performance Measurement and Quality Management System: Performance Indicators for Electronic Library Services[EB/OL]. (2000.11)[2011-3-1]. http://equinox.dcu.ie/reports/pilist.html.

[10]李贺，沈旺，国佳．国外数字图书馆评价研究现状分析[J]．中国图书馆学报，2010, 36(190):88-94.

[11]Zhang Y. Developing a holistic model for digital library evaluation[J]. Journal of the American Society for Information Science and Technology, 2010, 61(1):88-110.

[12]黄万红，陈实．数字图书馆评估研究[J]．浙江万里学院学报, 2005, 18(3):167-169.

[13]Ayers K. Enhancing digital information access in public libraries[J]. Proceedings of the American Society for Information Science and Technology, 2006, 43(1):1-25.

[14]Kyrillidou M., Giersch S. Developing the DigiQUAL protocol for digital library evaluation. Proceedings of the ACM/IEEE Joint Conference on Digital Libraries[C]. IEEE, 2005:172-173.

[15]Fuhr N., Hansen P., Mabe M., et al. Digital libraries: A generic classification and evaluation scheme.

Research and Advanced Technology for Digital Libraries, 5th European Conference, ECDL 2001 Proceedings[C]. Springer-Verlag, 2001:187-199.

[16]Nicholson S. A conceptual framework for the holistic measurement and cumulative evaluation of library services[J]. Proceedings of the American Society for Information Science and Technology, 2004, 41(1):496-506.

[17]Saracevic T., Covi L. Challenges for digital library evaluation. the Annual Meeting of the American Society for Information Science[C]. Citeseer, 2000:341-350.

[18]Borgman C., S lvberg I., Kovacs L. Fourth DELOS workshop. evaluation of digital libraries: Testbeds, measurements, and metrics[M]. Budapest: Hungarian Academy of Sciences, 2002:7-23.

[19]Blixrud J. C. Measures for electronic use: the ARL e-metrics project[J]. Statistics in Practice-Measuring & Managing, 2002:73-84.

[20]Marchionini G. Evaluating digital libraries: A longitudinal and multifaceted view[J]. Library Trends, 2000, 49(2):304-333.

[21]MacKie-Mason J. K., Jankovich A. PEAK: pricing electronic access to knowledge. First Elsevier Electronic Subscriptions Conference[C]. Elsevier, 1996:281-295.

[22]Hill L. L., Dolin R., Frew J., et al. User evaluation: summary of the methodologies and results for the Alexandria Digital Library, University of California at Santa Barbara. Proceedings of the 60th Annual Meeting of the American Society for Information Science (ASIS)[C]. Inf. Today, 1997:225-243.

[23]Hill L. L., Carver L., Larsgaard M., et al. Alexandria digital library: user evaluation studies and system design[J]. Journal of the American Society for Information Science, 2000, 51(3):246-259.

[24]Van House N. A. User needs assessment and evaluation for the UC Berkeley electronic environmental library project: a preliminary report. Proceedings of ACM Digital Libraries[C]. ACM, 1995:11-13.

[25]Leazer G. H., Gilliland-Swetland A. J., Borgman C. L. Evaluating the use of a geographic digital library in undergraduate classrooms: ADEPT. Proceedings of the Fifth ACM Conference on Digital Libraries[C]. ACM, 2000:248-249.

[26]于良芝. 英国电子图书馆项目评价[J]. 津图学刊, 2001, 34(1):7-11.

[27]Arms W. Y., Jane G., Lagoze C., et al. The D-Lib Test Suite[J]. D-Lib Magazine, 1999, 5(2): 31-50.

[28]Peters C., Braschler M., Choukri K., et al. Evaluation of Cross-Language Information Retrieval Systems. Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001[C]. Springer-Verlag,

2001:1-10.

[29]Dopichaj P. The university of Kaiserslautern at INEX 2006. 5th International Workshop of the Initiative for the Evaluation of XML Retrieval[C]. Springer-Verlag, 2007:223-232.

[30]Gery M., Largeron C. UJM at INEX 2009 Ad Hoc Track. 8th International Workshop of the Initiative for the Evaluation of XML Retrieval[C]. Springer Verlag, 2010: 88-94.

[31]Poll R. Performance measures for library networked services and resources[J]. Electronic Library, 2001, 19(5):307-315.

[32]Buttle F. SERVQUAL: review,critique,research agenda[J]. European Journal of marketing, 1996, 30(1):8-32.

[33]Saleh F., Ryan C. Analysing service quality in the hospitality industry using the SERVQUAL model[J]. Service Industries Journal, 1991, 11(3):324-345.

[34]Asubonteng P., McCleary K. J., Swan J. E. SERVQUAL revisited: a critical review of service quality[J]. Journal of Services Marketing, 1996, 10(6):62-81.

[35]刘盈盈. 基于网络的图书馆服务质量评估工具-LIBQUAL DIGIQUAL MINES[J]. Library and Information Service, 2006, 50(9):39-45.

[36]Cook C. The maturation of assessment in academic libraries: the role of LibQUAL^{+tm}[J]. Performance Measurement and Metrics, 2002, 3 (2):18-26.

[37]Thompson B., Cook C., Kyrillidou M. Stability of library service quality benchmarking norms across time and cohorts: A LibQUAL^{+TM} study[C]. Nanyang Technological University, 2006:52-60.

[38]Thompson R. L., Cook C., Thompson B. Reliability and structure of LibQUAL+ scores: measuring perceived library service quality[J]. Portal: Libraries and the Academy, 2002, 2(1):3-12.

[39]Wen-tao W. E. I., Ming G. A. O. Discussion on Library's Service Quality Evaluation Method—LibQUAL[J]. Sci-Tech Information Development & Economy, 2006:7-15.

[40]Asemi A., Kazempour Z., Rizi H. A. Using LibQUAL^{+TM} to improve services to libraries: A report on academic libraries of Iran experience[J]. Electronic Library, 2010, 28(4):568-579.

[41]Friesen M., Zaher S., Lesack P. LibQUAL+ 2009 Subset[EB/OL]. (2010)[2011-3-1]. http://hdl.handle.net/10573/42233.

[42]Kyrillidou M., Giersch S. Pilot testing the DigiQUAL^{tm} protocol: lessons learned. 2006 IEEE/ACM 6th Joint Conference on Digital Libraries[C]. IEEE, 2006:369-369.

[43]Kilker J., Gay G. The Social Construction of a Digital Library: A Case Study Examining

Implications for Evaluation[J]. Information Technology and Libraries, 1998, 17(2):60-70.

[44]Barton J. Measurement, management and the digital library[J]. Library Review, 2004, 53(3):138-141.

[45]Mabe M. Digital Library Classification and Evaluation: A Publisher’s View Of the Work of the DELOS Evaluation Forum[C]. Citeseer, 2004:21-26.

[46]Sandusky R. J. Digital Library Attributes: Framing Research and Results. Workshop on Usability of Digital Libraries at JCDL[C]. Citeseer, 2002:35-38.

[47]Saracevic T. Digital library evaluation: Toward evolution of concepts[J]. Library Trends, 2000, 49(2):350-369.

[48]吴建华, 数字图书馆评价方法[M]. 科学出版社, 2009:38-50.

[49]乔欢, 马亚平. 数字图书馆评价客体解析[J]. 大学图书馆学报, 2005, 23(5):7-12.

[50]乔欢, 王燕. 数字图书馆系统与服务质量评价[J]. 情报杂志, 2009, 28 (B12):196-199.

[51]乔欢. 数字图书馆评价研究[J]. 国家图书馆学刊, 2004, 49(3):49-54.

[52]孙华. 论高校评估背景下图书馆数字馆藏的建设[J]. 现代情报, 2008, 28(9):82-83.

[53]康云萍, 张文德. 利用收益提成率法评估数字图书馆著作权的价值[J]. 图书情报工作, 2008, 52(7):91-91.

[54]王启云. 高校数字图书馆建设评估调查与分析[J]. 图书与情报, 2008, (6):95-97.

[55]唐李杏, 张盛强. 图书馆 2.0 时代的数字资源评估[J]. 图书馆论坛, 2008, (4):60-62.

[56]陈鍊. 数字图书馆信息系统安全评估[J]. 图书情报工作, 2008, 52(2):141-144.

[57]连天奎. 对图书馆藏书数字化元数据进行评估的因素与条件[J]. 河南图书馆学刊, 2008, 28(1):111-112.

[58]牛振东, 赵四友. 数字图书馆体系结构的发展[J]. 现代图书情报技术, 2003, 100(3):20-23.

[59]吴清强, 韩涛. 数字图书馆评价研究综述[J]. 现代图书情报技术, 2006, 138(6):22-25.

[60]罗琳. 数字图书馆信息服务效率研究[J]. 图书情报知识,2005, 105(6):23-27.

[61]张宏玲. 国外数字馆藏使用及服务绩效评价指标体系述评[J]. 大学图书馆学报, 2005, 23(6):63-69.

[62]邱燕燕. 论馆藏评价范式的转变[J]. 图书馆理论与实践, 2005, (4):11-13.

[63]庞蓓. 论数字参考服务的评价方法[J]. 情报科学, 2005, 23(9):1418-1423.

[64]索传军. 论数字馆藏利用绩效分析与评价[J]. 图书馆, 2005, (3):58-61.

[65]常春,张桂英. 农业古籍数字图书馆项目评价方案[J]. 现代情报, 2005, 25(11):57-59.

[66]陈红梅. 试论图书馆知识服务评估与反馈机制的建立[J]. 情报探索, 2005, (1):91-92.

[67]吴懿咏,王军. 数字图书馆评价综述[J]. 数字图书馆论坛, 2007, (10):39-50.

[68]刘炜,楼向英,张春景. 数字图书馆评估研究[J]. LIBRARY AND INFORMATION SERVICE, 2007, 51(5):21-24.

[69]Fayyad U. M.,Piatetsky-Shapiro G.,Smyth P. From data mining to knowledge discovery: an overview. Advances in knowledge discovery and data mining[C]. American Association for Artificial Intelligence 1996:1-34.

[70]Witten I. H.,Frank E. Data mining: practical machine learning tools and techniques with Java implementations[J]. ACM SIGMOD Record, 2002,31(1):76-77.

[71]Han J.,Kamber M. Data mining: concepts and techniques[M]. Morgan Kaufmann, 2006:126-141.

[72]Keim D. A. Information visualization and visual data mining[J]. IEEE transactions on Visualization and Computer Graphics, 2002, 8(1):1-8.

[73]Fayyad U. M.,Wierse A.,Grinstein G. G. Information visualization in data mining and knowledge discovery[M]. Morgan Kaufmann Pub, 2002:31-48.

[74]蔡伟杰, 张晓辉, 朱建秋, 等. 关联规则挖掘综述[J]. 计算机工程, 2001, 27(5):31-33.

[75]张朝晖,陆玉昌. 发掘多值属性的关联规则[J]. 软件学报, 1998, 9(11):801-805.

[76]钱卫宁,周傲英. 从多角度分析现有聚类算法[J]. 软件学报, 2002, 13(8):1382-1394.

[77]张敏,于剑. 基于划分的模糊聚类算法[J]. 软件学报, 2004, 15(6):858-868.

[78]李炎,李皓. 异常检测算法分析[J]. 计算机工程, 2002, 28(6):5-6.

[79]刘小虎,李生. 决策树的优化算法[J]. 软件学报, 1998, 9(10):797-800.

[80]唐华松,姚辉文. 数据挖掘中决策树算法的探讨[J]. 计算机应用研究, 2001, 18(8):18-19.

[81]Quinlan J. R. C4. 5: programs for machine learning[M]. Morgan Kaufmann, 1993:31-37.

[82]Quinlan J. R. Learning logical definitions from relations[J]. Machine Learning, 1990, 5(3):239-266.

[83]Ruggieri S. Efficient C4.5[J]. IEEE transactions on knowledge and data engineering, 2002, 14(2):438-444.

[84]Cheng J.,Fayyad U. M.,Irani K. B.,et al. Improved decision trees: A generalized version of ID3[J]. Proceedings of ML, 1988:100-106.

[85]张桂杰,王帅. 决策树分类 ID3 算法研究[J]. 吉林师范大学学报:自然科学版, 2008, 29(003):135-137.

[86]翟俊海,张素芳,王熙照. ID3 算法的理论基础[J]. 兰州大学学报:自然科学版, 2007, 43(6):66-69.

[87]Quinlan J. R. Bagging,boosting,and C4.5. Proceedings of the National Conference on Artificial Intelligence[C]. AAAI, Menlo Park, CA, United States, 1996:725-730.

[88]Quinlan J. R. Improved use of continuous attributes in C4.5[J]. Journal of Artificial Intelligence Research, 1996, 4:77-90.

[89]冯少荣. 决策树算法的研究与改进[J]. 厦门大学学报 (自然科学版), 2007, 46 (4):10-12.

[90]刘勇洪,牛铮,王长耀. 基于 MODIS 数据的决策树分类方法研究与应用[J]. 遥感学报, 2005, 9(4):405-412.

[91]Crawford S. L. Extensions to the CART algorithm[J]. International Journal of Man-Machine Studies, 1989, 31(2):197-217.

[92]Mehta M.,Agrawal R.,Rissanen J. SLIQ: A fast scalable classifier for data mining. Advances in Database Technology—EDBT'96[C]. Springer-Verlag, 1996:18-32.

[93]Chandra B.,Varghese P. P. On improving efficiency of SLIQ decision tree algorithm. Proceedings of International Joint Conference on Neural Networks[C]. IEEE, 2007:66-71.

[94]李波. 基于 SLIQ 分类算法的数据挖掘技术及其在企业 CRM 中的应用[J]. 计算机工程与应用, 2002, 38(21):29-31.

[95]Shafer J.,Agrawal R.,Mehta M. SPRINT: A scalable parallel classifier for data mining. Proceedings of the 22nd VLDB Conference[C]. Citeseer, 1996:544-555.

[96]Khoshgoftaar T. M.,Seliya N. Software quality classification modeling using the SPRINT decision tree algorithm. Proceedings 14th IEEE International Conference on Tools with Artificial Intelligence[C]. IEEE Comput. Soc, 2002: 365-374.

[97]Ankerst M.,Elsen C.,Ester M.,et al. Visual classification: an interactive approach to decision tree construction. Proceedings: KDD[C]. ACM, 1999:392-396.

[98]魏红宁. 基于 SPRINT 方法的并行决策树分类研究[J]. 计算机应用, 2005, 25(1):39-41.

[99]魏红宁. SPRINT 决策树分类器中的数据存储方法[J]. 计算机应用, 2004, 24(6):95-96.

[100]魏红宁,颜治平. SPRINT 决策树方法中 I/O 分析及优化研究[J]. 计算机与数字工程, 2007, 35(6):49-51.

[101]王威. 基于决策树的数据挖掘算法优化研究[D]. 西南交通大学, 2005.

[102]Cover T.,Hart P. Nearest neighbor pattern classification. Information Theory[J]. IEEE Transactions on, 1967, 13 (1):21-27.

[103]桑应宾. 基于 K 近邻的分类算法研究[D]. 重庆大学, 2009.

[104]张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(1):32-42.

[105]张浩然,韩正之,李昌刚. 支持向量机[J]. 计算机科学, 2002, 29(12):135-137.

[106]邓乃扬,田英杰. 数据挖掘中的新方法:支持向量机[M]. 科学出版社, 2004:28-41.

[107]Heckerman D. Bayesian networks for data mining[J]. Data mining and knowledge discovery, 1997 ,1(1):79-119.

[108]Cheeseman P.,Stutz J. Bayesian classification (AutoClass): Theory and results. Advances in knowledge discovery and data mining[C]. American Association for Artificial Intelligence, 1996:153-180.

[109]Cios K. J.,Pedrycz W.,winiarski R.,et al. Data mining methods for knowledge discovery[M]. Kluwer Academic Publishers, 1998:38-51.

[110]Hanson R.,Stutz J.,Cheeseman P.,et al. Bayesian classification theory[M]. Citeseer, 1991:12-19.

[111]Cheeseman P.,Kelly J.,Self M.,et al. AutoClass: A Bayesian classification system[M]. Morgan Kaufmann, 1988:54-64.

[112]Flach P. A.,Lachiche N. Naive Bayesian classification of structured data[J]. Machine Learning, 2004, 57(3):233-269.

[113]Hanson R.,Stutz J.,Cheeseman P. Bayesian classification with correlation and inheritance[M]. Citeseer, 1991:692-698.

[114]周颜军,王双成,王辉. 基于贝叶斯网络的分类器研究[J]. 东北师大学报自然科学版, 2003, 35(2):21-27.

[115]宫秀军,刘少辉. 一种增量贝叶斯分类模型[J]. 计算机学报, 2002, 25(6):645-650.

[116]胡玉胜,崔晓瑜. 基于贝叶斯网络的不确定性知识的推理方法[J]. 计算机集成制造系统, 2001, 7(12):65-68.

[117]阎平凡,张长水. 人工神经网络与模拟进化计算[M]. 清华大学出版社, 2005:47-62.

[118]张立明. 人工神经网络的模型及其应用[M]. 复旦大学出版社, 1993:12-35.

[119]高隽. 人工神经网络原理及仿真实例[M]. 机械工业出版社, 2003:34-47.

[120]王伟. 人工神经网络原理[M]. 北京航空航天大学出版社, 1995:14-29.

[121]Pawlak Z.,Polkowski L.,Skowron A. Rough set theory[M]. Springer, 1998:1-10.

[122]Lin T. Y.,Liu Q. Rough approximate operators: axiomatic rough set theory[M]. Springer-Verlag. 1993:256-260.

[123]Polkowski L.,Skowron A. Rough sets in knowledge discovery: Methodology and applications[M].

Physica Verlag, 1998:61-74.

[124]Bonikowski Z.,Bryniarski E.,Wybraniec-Skardowska U. Extensions and intentions in the rough set theory[J]. Information Sciences, 1998, 107(1-4):149-167.

[125]Lin T. Y.,Cercone N. Rough sets and data mining: analysis for imprecise data[M]. Kluwer Academic Publishers, 1997:41-49.

[126]Klir G. J.,Yuan B. Fuzzy sets and fuzzy logic: theory and applications[M]. Prentice Hall PTR Upper Saddle River,NJ,USA, 1995:31-42.

[127]Zadeh L. A. Fuzzy logic,neural networks,and soft computing[J]. Communications of the ACM, 1994, 37(3):77-84.

[128]Hajek P. Metamathematics of fuzzy logic[M]. Springer, 1998:18-25.

[129]Lin C. T.,Lee C. S. G. Neural-network-based fuzzy logic control and decision system[J]. IEEE Transactions on computers, 1991, 40(12):1320-1336.

[130]Kartalopoulos S. V. Understanding neural networks and fuzzy logic: basic concepts and applications[M]. Institute of Electrical and Electronics Engineers, 1996:21-33.

[131]Klir G. J.,Yuan B. Fuzzy sets and fuzzy logic[M]. Prentice Hall New Jersey, 1995:7-12.

[132]Yager R. R.,Zadeh L. A. An introduction to fuzzy logic applications in intelligent systems[M]. Kluwer Academic Pub, 1992:17-21.

[133]Mendel J. M. Fuzzy logic systems for engineering: a tutorial[J]. Proceedings of the IEEE, 1995, 83(3):345-377.

[134]Bandemer H.,Gottwald S. Fuzzy sets,fuzzy logic,fuzzy methods with applications[M]. J. Wiley, 1995:23-27.

[135]Levy D. Chaos theory and strategy: theory,application,and managerial implications[J]. Strategic management journal, 1994, 15(S2):167-178.

[136]Li Z.,Halang W. A.,Chen G. Integration of fuzzy logic and chaos theory[M]. Springer-Verlag New York Inc, 2006:18-23.

[137]郭刚,史忠科,戴冠中. 依据混沌理论进行非线性系统建模变量个数的最优选取[J]. CONTROL AND DECISION, 2000, 15(2):233-235.

[138]孙海云,曹庆东. 混沌时间序列建模及预测[J]. SYSTEMS ENGINEERING, 2001, 21(5):106-110.

[139]Vullo G. A Global Approach to Digital Library Evaluation[J]. Liber Quarterly, 2010, 20(2):169-172.

[140]Bertot J. C.,Snead J. T.,Jaeger P. T.,et al. Functionality,usability,and accessibility: Iterative user-centered evaluation strategies for digital libraries[J]. Performance Measurement and Metrics, 2006, 7(1):17-28.

[141]Hsieh-Yee I. Digital Library Evaluation: Progress & Next Steps. Annual Meeting of the American Society for Information Science and Technology Charlotte[C]. ACM, 2005:1-10.

[142]徐婧. AHP 法在数字图书馆综合评价中的应用[J]. 图书馆论坛, 2006, 26(1):238-240.

[143]Borgman C. L. What are digital libraries? Competing visions[J]. Information processing and management, 1999, 35(3):227-243.

[144]肖汀. 数字图书馆概念探讨[J]. 情报探索, 2003, 87(3):10-12.

[145]Hahn J.,Twidale M.,Gutierrez A.,et al. Methods for Applied Mobile Digital Library Research: A Framework for Extensible Wayfinding Systems[J]. The Reference Librarian, 2011, 52(1):106-116.

[146]Cox A. M. Digital Library Economics: An Academic Perspective[J]. Program: electronic library and information systems, 2010, 45(1):121-122.

[147]Dobreva M. Evaluation of Digital Libraries: An Insight into Useful Applications and Methods[J]. Library review, 2010, 60(2):166-168.

[148] Borner K, Chaomei Chen. Visual interfaces to digital libraries: motivation, utilization, and socio-technical challenges[MC]. Springer-Verlag, 2002:1-9.

[149]Serrano-Guerrero J.,Herrera-Viedma E.,Olivas J. A.,et al. A google wave-based fuzzy recommender system to disseminate information in University Digital Libraries 2.0[J]. Information Sciences, 2011, 181(9):1503-1516.

[150]Nurnberg P.,Wiil U.,Leggett J. Structuring facilities in digital libraries. Research and Advanced Technology for Digital Libraries[C]. Springer-Verlag, 1998:295-313.

[151]Manghi P.,Pagano P.,Ioannidis Y. Second workshop on very large digital libraries: in conjunction with the european conference on digital libraries Corfu,Greece[J]. ACM SIGMOD Record, 2009, 38(4):46-48.

[152]Xie I.,Cool C. Understanding help seeking within the context of searching digital libraries[J]. Journal of the American Society for Information Science and Technology, 2009, 60(3):477-494.

[153]Payette S.,Lagoze C. Flexible and extensible digital object and repository architecture (FEDORA). Research and Advanced Technology for Digital Libraries[C]. Springer-Verlag, 2009:41-59.

[154]Porcel C.,Moreno J. M.,Herrera-Viedma E. A multi-disciplinar recommender system to advice

research resources in University Digital Libraries[J]. Expert Systems with Applications, 2009, 36(10):12520-12528.

[155]Armani B.,Catania B.,Bertino E.,et al. Repository management in an intelligent indexing approach for multimedia digital libraries. Foundations of Intelligent Systems[C]. Springer Verlag, 2009:431-440.

[156]王启云. 高校数字图书馆建设评估指标体系研究[J]. 大学图书馆学报, 2008, 5:74-81.

[157]张玲, 孙坦, 黄国彬. 国外数字图书馆评价实践综述[J]. LIBRARY AND INFORMATION SERVICE, 2006, 50(12):131-134.

[158]张艳芳. 回眸近十年国外学者之 LibQUAL+®研究[J]. 图书馆学研究, 2010, (12):2-6.

[159]唐琼, 张玫, 曾颖, 等. 基于 LibQUAL^{+TM} 的广东高校图书馆服务质量评价[J]. 大学图书馆学报, 2006, 24(2):63-69.

[160]杨广锋, 赵红. LibQUAL+的发展与实践[J]. 图书馆建设, 2009, (9):81-84.

[161]Sanders J. R. The program evaluation standards: how to assess evaluations of educational programs[M]. Sage Publications Inc, 1994:61-72.

[162]何晓群. 现代统计分析方法与应用[M]. 中国人民大学出版社, 1998:21-40.

[163]萨尔金德, 史玲玲. 爱上统计学[M]. 重庆大学出版社, 2008:10-29.

[164]Kendall M. G. The advanced theory of statistics (2nd Ed)[M]. Hafner Publishing, 1946:78-90.

[165]Montgomery D. C., Runger G. C. Applied statistics and probability for engineers[M]. Wiley, 2010:12-34.

[166]贾俊平. 统计学[M]. 清华大学出版社, 2006:103-121.

[167]Argyrous G. Statistics for research: with a guide to SPSS[M]. Sage Publications Ltd, 2011:21-48.

[168]Field A. P. Discovering statistics using SPSS[M]. SAGE publications Ltd, 2009:78-91.

[169]黄良文, 陈仁恩. 统计学原理[M]. 中央广播电视大学出版社, 2006:30-49.

[170]魏宗舒. 概率论与数理统计教程[M]. 高等教育出版社, 2008:60-78.

[171]薛薇. 统计分析与 SPSS 的应用[M]. 中国人民大学出版社, 2008:52-65.

[172]Kent J. T. Information gain and a general measure of correlation[J]. Biometrika, 1983, 70(1):163-173.

[173]Quinlan J. R. Induction of decision trees[J]. Machine Learning, 1986, 1(1):81-106.

[174]Breiman L. Classification and regression trees[M]. Chapman & Hall/CRC, 1984:16-23.

[175]Quinlan J. R. Learning decision tree classifiers[J]. ACM Computing Surveys (CSUR), 1996, 28(1):71-72.

[176]Mingers J. Expert systems-rule induction with statistical data[J]. The Journal of the Operational Research Society, 1987, 38(1):39-47.

[177]Kira K.,Rendell L. A. A practical approach to feature selection. Proceedings of the ninth international workshop on Machine learning[C]. Morgan Kaufmann Publishers Inc, 1992:249-256.

[178]Kira K.,Rendell L. A. The feature selection problem: Traditional methods and a new algorithm. Proceedings Tenth National Conference on Artificial Intelligence[C]. AAAI Press, 1992:129-134.

[179]Hong S. J. Use of Contextual Information for Feature Ranking and Discretization[J]. IEEE Transactions on Knowledge and Data Engineering, 1997, 9(5):718-730.

攻读学位期间发表论文与研究成果清单

[1] Yumin Zhao, Zhendong Niu, and Xueping Peng. Research on Data Mining Technologies for Complicated Attributes Relationship in Digital Library Collections，Applied Mathematics & Information Sciences. (SCI 刊源，已录用)

[2] Yumin Zhao, Zhendong Niu, Yueping Peng, and Lin Dai. A Discretization Algorithm of Numerical Attributes for Digital library Evaluation based on Data Mining Technology，13th International Conference on Asia-Pacific Digital Libraries, ICADL 2011. (EI Compendex: 20114514498615，数字图书馆领域三大会议之一)

[3] Yumin Zhao, Zhendong Niu, and Lin Dai. Evaluation Algorithm about Digital library Collections based on Data Mining Technology，12th International Conference on Asia-Pacific Digital Libraries, ICADL 2010. (EI Compendex: 20103313149280，数字图书馆领域三大会议之一)

[4] Yumin Zhao, Zhendong Niu, Yujuan Cao, and Lin Dai. Research on Evaluation of Digital Library，2010 International Conference on Data Storage and Data Engineering. (EI Compendex: 20102112956802)

[5] Yumin Zhao, Zhendong Niu, Kun Zhao, Lin Dai, Guixian Xu, and Weiqiang Wang. What and Why about Divarication in Research of Digital Library Evaluation，2010 IEEE International Conference on Computer Engineering and Technology. (EI Compendex: 20104313316761)

[6] Yumin Zhao, Zhendong Niu, Kun Zhao, Lin Dai, Guixian Xu, and Weiqiang Wang，A Natural-integration Model of Digital Library Evaluation，2010 IEEE International Conference on Computer Engineering and Technology. (EI Compendex: 20104313316793)

[7] Kun Zhao, Zhendong Niu, Yumin Zhao, and Jun Yang. Group-based search in unstructured peer-to-peer networks. 2009 IEEE Global Telecommunications Conference, GLOBECOM 2009.(EI compendex: 20101812901159)

[8] Weiqiang Wang, Zhendong Niu, Yumin Zhao, Yujuan Cao, and Kun Zhao. Parameter Estimation based on MCMC Methods in PM2.5 and Traffic，2nd IEEE International Conference on Information Management and Engineering.(EI compendex: 20102913087053)

[9] Kun zhao, Zhengdong Niu, Yumin Zhao and Jun Yang. Search with index replication in power-law like peer-to-peer networks，2010 IEEE International Conference on Computer Engineering and

Technology (EI Compendex: 20104313316763)

[10] Kun zhao, Zhengdong Niu, and Yumin Zhao. Degree biased random walk in unstructured peer-to-peer networks，2010 IEEE International Conference on Computer Engineering and Technology (EI Compendex: 20104313316764)

[11] Yujuan Cao, Zhendong Niu, Liuling Dai, and Yuming Zhao. Extraction of informative blocks from web pages. ALPIT 2008, 7th International Conference on Advanced Language Processing and Web Information Technology (EI Compendex: 20083911591710)

攻读学位期间申请的国家专利

[1] 专利名称：一种层叠决策树构建方法，受理号：201110111344.6

[2] 专利名称：一种基于 MCMC 的优化信息检索方法，受理号：201010520341.3

[3] 专利名称：一种基于小世界特性的中文近似网页去重方法，受理号：200910083711.9

攻读学位期间参加的国家基金与科研项目

[1] 跨语言文本自动分类关键技术研究（国家自然科学基金）

[2] 数字图书馆的个性化服务研究（霍英东教育基金项目）

[3] 数字图书馆体系结构的研究（国家社科重点基金项目）

[4] 中联部电子政务——数字图书馆系统 （中联部项目）

[5] CETV 网络化评价系统（CETV 项目）

致谢

在论文就要完工付梓之际，我谨向我的导师牛振东教授致以深深的感谢。在牛老师的指导下，我才得以取得现在的微薄成绩。在我多年的博士生涯当中，牛老师用敏锐的学术思维能力，给予了我很多学习上的提点，让我在学术创新的道路上找到了方向，摆脱了困惑，也树立了信心。最让我受益的是老师的工作态度，目睹牛老师每天如一日的忘我工作态度，感染我有勇气去迎接困难，而老师对我们大家生活中的无微不至，让我在远离家乡和父母的北京体会到了令人温暖的感情。

感谢宋瀚涛老师、樊孝忠老师，这两位长辈的为人师表和慈祥耐心给予了我很多帮助，使我心中由衷的充满了感激。在未来的工作和学习中，我会以两位老师为楷模，努力创造更大的价值。

感谢刘辉老师、张春霞老师、袁武老师、张继老师和金福生老师给予我的热情帮助与勉励。感谢徐晓梅、施重阳、高翀、赵堃、王维强、曹玉娟、胥桂仙、彭学平、江鹏、杨青、李学进、牛科、谷培培、赵向宇、黄胜等。他们是我的良师益友，在学习和生活当中给予了我很多帮助和支持，还有很多同学在学习和生活上都曾给我帮助和关心，在这里一并表示感谢。

感谢计算机学院的领导和老师在我攻读博士学位期间给予我的关心和支持。

由衷感谢各位学界前辈在百忙之中评阅本文并给出宝贵评审意见，帮助我总结工作中的不足之处。

最后，我要感谢我的父母和我的妻子，面对他们无私的爱和无私的付出，我感到深深的幸福，深深的感谢你们，我会永远与你们承担和分享生命中的一切。