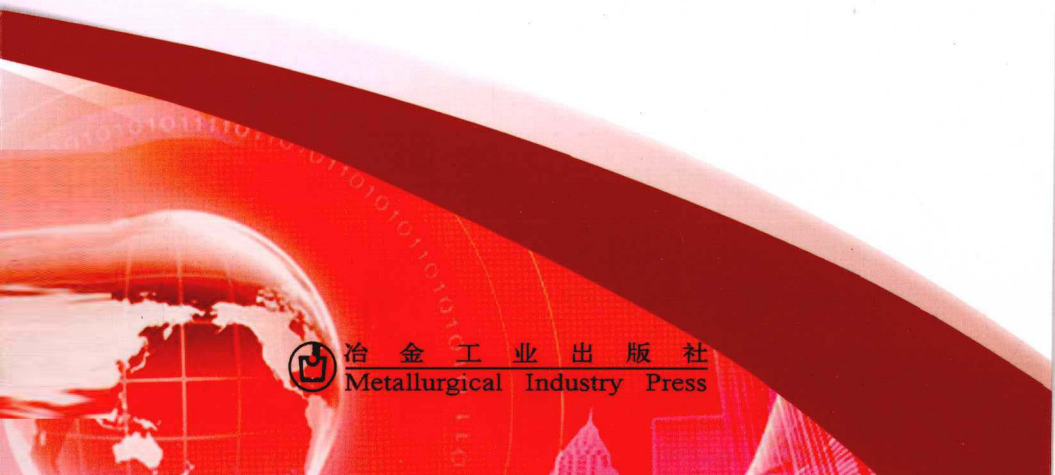
Data Mining and Management Practice

数据挖掘与管理实践

宋宇辰 孟海东 著



冶金工业出版社
Metallurgical Industry Press

数据挖掘与管理实践

宋宇辰　孟海东　著

北　京
冶 金 工 业 出 版 社
2010

内 容 提 要

本书对数据挖掘技术及其在管理决策中的应用进行了较深入的研究。书中重点介绍了聚类分析和关联分析的理论基础、算法设计、分析与对比。全书以图书馆现代化管理为主线，探索了如何对管理数据实施数据挖掘、实现管理决策的全过程，包括数据采集、数据预处理、数据挖掘与分析、挖掘结果的分析，并提出相应的决策建议；根据一系列应用实施过程，总结出图书馆现代化管理应用数据挖掘的三层决策构架，即数据层、技术层和决策层。

本书适合从事信息分析、数据挖掘的人员，企业和政府部门的管理人员，从事管理学和情报学研究的学者及相关专业的研究生阅读参考。

图书在版编目(CIP)数据

数据挖掘与管理实践/宋宇辰，孟海东著. —北京：冶金工业出版社，2010.12

ISBN 978-7-5024-5457-9

Ⅰ.①数… Ⅱ.①宋… ②孟… Ⅲ.①数据采集 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2010)第 245747 号

出 版 人 曹胜利

地　　址 北京北河沿大街嵩祝院北巷 39 号，邮编 100009

电　　话 (010) 64027926 电子信箱 yjcbs@cnmip.com.cn

责任编辑 宋 良 王雪涛 美术编辑 彭子赫 版式设计 葛新霞

责任校对 刘 倩 责任印制 张祺鑫

ISBN 978-7-5024-5457-9

北京百善印刷厂印刷；冶金工业出版社发行；各地新华书店经销

2010 年 12 月第 1 版，2010 年 12 月第 1 次印刷

148mm×210mm；6.125 印张；178 千字；181 页

20.00 元

冶金工业出版社发行部　电话：(010)64044283　传真：(010)64027893

冶金书店　地址：北京东四西大街 46 号(100010)　电话：(010)65289081(兼传真)

前　言

随着数据挖掘理论研究的不断深入和各类数据挖掘软件（工具）的成功研发，数据挖掘技术在各领域得到了越来越广泛的应用。政府、企业、高校等各部门的现代化管理产生各类数据，管理信息系统中包含大量数据。如何根据这些管理数据的特点，有针对性地进行理论研究，开发适合现代化管理应用的数据挖掘软件，如何运用这些软件挖掘出对现代化管理有意义的结果，为本部门的管理者提供决策参考，是一个值得研究的问题，也是当今社会的热门话题。

本书对数据挖掘技术及其在管理决策中的应用进行了较深入的研究。书中重点介绍了聚类分析和关联分析的理论基础、算法设计、分析与对比。对于新聚类算法，分别做了不同尺寸和密度的簇聚类效果实验和埋藏在“噪声”中的簇聚类效果实验。结果显示，新聚类算法能够较好地处理任意形状和大小的簇，同时算法能够较好地处理不同密度的簇、埋藏在“噪声”中的簇，克服了传统密度算法在这方面的局限。新关联算法弥补了经典算法对挖掘量化关联规则的不足，改变了对参数的设置凭经验而定、盲目性较大的缺点。并在数据挖掘理论研究和算法设计上有所创新。

本书力图通过一系列实例来说明信息时代管理者如何整理工作中的数据、如何运用数据挖掘方法进行信息分析、如何运用数据挖掘结果做出符合实际的管理决策。全书以图书馆现代化管理为主线，探索了如何对图书馆数据实施数据挖掘、实现管理决策的全过程，包括数据采集、数据预处理、数据挖掘、挖掘结果分

析，并提出相应的决策建议。在信息爆炸的时代，这对现代企业的高层管理者如何面对海量复杂的数据，通过数据挖掘获取有意义的知识，做出正确决策，具有现实意义和实用价值。

根据图书馆数据挖掘的特点，本书介绍了大量的数据收集整理工作的实例，主要包括：纸质问卷的设计与数据整理、高校图书馆数据的网络收集与整理、通用图书馆集成系统中数据的抽取和整理。纸质问卷范围涵盖图书馆资源建设、利用与服务情况，数据包括：获取资料的难易程度、增加馆藏资料的急需程度和对图书馆各部门的熟悉程度等；网络收集各高校图书馆的数据包括：图书馆硬件设施、人力资源情况、图书馆藏书量和开馆时间等；通用图书馆集成系统中所抽取的数据包括：读者信息、图书借阅记录信息和读者借阅记录信息；数据内容覆盖面较广。

本书详细、深入地探索了数据挖掘技术在图书馆管理中的具体应用方法和实现过程。对调查问卷搜集的数据和网络收集的数据进行了新聚类算法分析，根据分析结果，给出读者借阅需求、科研课题需求、提高图书馆服务质量和做好图书馆人力资源规划等方面的决策建议。对通用图书馆集成系统中所抽取的数据进行了 Clementine 关联分析，对图书馆集成系统中图书大类数据和调查问卷中馆藏资料数据进行了新关联算法分析，根据实验获得的关联规则，给出图书采购、图书排架、学科建设等方面的决策建议。

与数据挖掘应用的其他同类书相比，本书中实施数据挖掘的管理部门主要是高校图书馆（目前，关于图书馆、情报类行业数据挖掘的研究专著非常少）。从信息技术的角度，本书对数据挖掘中的聚类分析和关联分析有较深入的理论研究；从管理实践的角度看，本书对管理者如何对所在部门实施数据挖掘给出了较多的参考实例。本书是跨学科研究专著。

本书的研究内容是在国家社会科学基金项目（批准号：

06XTQ011）的支持下完成的。这项研究主要依托“自治区产业信息化与产业创新研究中心”、“矿业信息系统研究室”、“数据工程实验室”完成。

感谢 Gregory O. Hare 教授在信息技术理论和算法实现方面的启迪和指导。感谢 Cathal M. Brugha 教授在信息分析和管理实践方面的许多具体建议和指导。感谢张彦春（澳大利亚）教授在数据挖掘技术方面的建议和对本书的关注。感谢陈福集教授对全书进行了校对审阅。国内许多同仁和同事对本书提出了建议并给予了许多实际帮助，在此对他们为此书付出的辛勤劳动和汗水表示诚挚的谢意。

研究生们在本书项目的研究中起了积极作用，做了大量贡献，其中“聚类算法和关联算法”的算法设计和程序实现主要由研究生宋飞燕、郝永宽、申海涛完成；研究生房宜锋对研究内容中的数据收集与整理、数据挖掘与结果分析、决策建议等做了大量工作；研究生吕文亮和吴熙做了部分数据汇总和分析工作；甄莎、周世凯、尤天龙、郭丽、张志启也参与了本书研究项目的部分工作，在此一并表示感谢。

著　者

2010 年 10 月

目　录

1 概　论

1.1 背景

我国图书馆自动化管理的发展已有20多年历史，自动化的进程经历了单机、多用户、局域网等几个发展时期，由小型系统发展到图书资料采、编、检综合管理系统，并逐步实现网络化、现代化。

随着数字图书馆的建设，图书馆管理业务实现了计算机管理，包括读者查询、采编、流通、借阅、期刊订购、行政人事、设备等，积累了大量的事务数据，使图书馆要处理和提供的信息更多、更新、更广泛、更复杂。传统的数据分析工具可以高效地实现数据的录入、修改、统计、查询等功能，但无法发现数据中存在的关系和规则，无法根据现有的数据预测未来的发展趋势，导致了“数据丰富，但知识贫乏”的现象，为了避免这种局面，图书馆有必要增强对信息的分析、处理能力以及对信息资源的组织能力，尤其是对海量管理信息的深层次的开发，提供表面上看来庞杂无序的信息的内在联系供管理人员和读者使用。另外，随着读者信息水平和信息要求的不断提高，向读者提供更主动的和个性化的信息服务已成为图书馆管理者的职责。

传统的经验管理已经不能适应图书馆的现代化发展，我们需要一些强有力的数据采集和处理工具介入到图书馆现代化管理中来，为图书馆工作提供技术支持和决策管理支持。而数据挖掘就是这样一种新兴的技术。图书馆的现代化管理水平在很大程度上取决于决策的科学与否。利用数据挖掘技术能够为管理层的科学决策提供强有力的保障。

1.1.1 国外研究与应用

自1995年以来，国外在数据挖掘方面做了较多的研究与应用工

作。在欧洲和北美地区数据挖掘在图书馆现代化管理中获得了较为广泛的应用。

1998 年，Schulman 研究指出，当图书馆数据量非常大或当特定用途的资料库运作时，图书馆管理者就会考虑到建置一个决策支持系统。此时利用传统人工的方式来掌握不断变动的使用者行为模式及趋势是不可能的，因此利用数据挖掘技术来了解图书馆使用者的行为，重新规划图书馆馆藏发展方向与政策的制定，并设计图书馆相关活动[1]。

Scott Nicholson 博士在《Gaining Strategic Advantage through Bibliomining: Data Mining for Management Decisions in Corporate, Special, Digital, and Traditional Libraries》[2]一文中以 Syracuse University 图书馆为例利用数据挖掘帮助图书馆做管理决策。把来源于图书馆系统中的图书信息、供货商信息、读者信息、流通信息和检索等信息，以及图书馆自动化系统外的访谈信息和电话记录等整合到数据仓库中，综合运用聚类分析、遗传算法、预测模型、决策树、神经网络和时间序列法等数据挖掘技术、数据可视化技术和统计分析技术，把最终的分析结果以报告的形式上传给管理者，供管理者参考。同时，Scott 博士谈到图书馆必须继续保护他们的读者和员工的数据记录信息，必须平衡信息保护和提供信息服务之间的关系。

美国加州大学伯克利校区信息管理与系统学院的迈克尔·库伯（Michael Cooper）教授曾对加州大学数字图书馆使用记录数据进行挖掘分析，发现不同类型的使用者所逗留的时间是不同的。库伯还设计了模型，对用户的查询时间、过程采用聚类、时间序列分析等方法进行分析，发现不同的用户在查询数量、时间结果的次数、显示结果的时间等方面具有不同的特点。通过数据分析，以便了解和掌握数字图书馆用户的特点，预测其未来趋势，从而研究数字图书馆用户的行为规律[3]。

Neumann 等人提到对于图书馆而言，读者借阅推荐服务将大有可为，用数据挖掘技术对自动化系统中的图书借阅记录文件与读者搜寻记录文件，进行分析，建立一个像亚马逊一样的顾客导向式的

入口网站。同时，读者也可减少搜寻和查看信息的时间，不仅增进读者服务效能，还可以对图书馆员在管理图书馆方面有很大的帮助。另外，Neumann 等人将亚马逊网站推荐系统应用在图书馆的使用者分析上，利用时间序列模式，对读者读书借阅顺序做分析，以得到使用者行为基础模式[4]。

Papatheodorou 等人提出，由数据挖掘技术分析图书馆数字化数据，可以找出读者的共同行为，以建立一套有意义的群组关系提升信息获取。而根据分析的结果也可以找出图书管理的工作项目以及建议可行的工作方式，帮助管理者重新制订符合需要的馆藏政策和提供管理决策的参考信息，找出读者兴趣，提供个性化服务等[5]。

荷兰的代夫特大学出版社与国际水路历史协会计划合办一个新刊，在创刊前，他们想掌握并分析这类刊物的撰文作家、读者、竞争对手的情况，于是在代夫特大学的要求下，代夫特大学图书馆选取了2000～2004 年6 个可能成为竞争对手的期刊，大约1600 个作者的2033 篇文章。代夫特大学图书馆通过对这些数据进行数据挖掘分析，形成报告，为代夫特大学出版社与国际水路历史协会创办的新刊吸引更多的作者前来投稿提供了依据和建议。这样的图书馆已经不只是传统的文献收藏机构，同时还成为了信息机构；图书馆的馆员也不再仅仅是传统意义上的管理员，他们已经成为了信息专家[6]。

多数国外学者针对图书馆管理提出了数据挖掘应用的构架、设计了相应的应用模型、给出了综合应用各种数据挖掘算法的建议；有一些学者利用商业挖掘工具对图书馆数据进行挖掘，获得了实际应用。但如何实施所提出的架构、模型、建议并获得实际应用还需继续探索；一些应用成果是基于某一图书馆数据的挖掘研究，数据来源较单一。到目前为止，专门针对图书馆管理数据开发数据挖掘软件并获得应用的研究较少。

1.1.2 国内研究与应用

在国内，2005 年以来出现了大量关于数据挖掘在图书馆管理中应用为主题的论文。最初研究主要集中在理论探讨和概念介绍，

近年来有学者利用数据挖掘工具对图书馆的部分数据进行处理分析，做了具有实际意义的探讨。至今国内缺少针对图书馆管理的数据挖掘软件的研发和相关算法的测试。数据挖掘在图书馆的应用研究主要包括关联分析、分类分析、序列分析和聚类分析技术。

1.1.2.1 关联分析在图书馆的应用研究

胡跟桥应用关联分析，在系统中形成借阅模式，根据这些借阅模式，向读者推荐相关的书目，从而提高图书馆的服务质量。根据关联分析的结果对图书排架进行适当的重组，并将关联分析的结果传递给图书采购部门，引导图书馆采购人员对相关图书的采购，进而提高图书的利用率[7]。

彭仪普和熊拥军对中南大学图书馆 2002～2004 年的读者数据进行挖掘，运用 Apriori 算法发现读者对文献的借阅存在着的关联、不同的学科之间的关联以及不同类型的读者对文献的借阅模式等。作者在 SQL Server 2000 数据库和 Windows 2000 系统下用 Visual Basic 6 来测试，进行关联规则的挖掘分析，得到了很多有意义的规则[8]。

上海理工大学的陈力等人针对图书馆采购的图书只是对借阅人数多的书目，往往忽视掉借阅人数不是很多，但几类书同时在学生中被借阅的行为，采用关联规则 Apriori 挖掘算法，并结合整数规划，从藏书的质量和数量结构考虑，在分析文献历史借阅信息的基础上提出了一种新的提高图书馆服务质量的方法[9]。

兰州商学院的瞿春玲利用 SQL Server 2005 关联分析实现了一个在线图书推荐服务，以提高图书馆的服务水平[10]。

魏育辉以北京工业大学焊接专业学生的图书借阅数据为对象，利用 Apriori 算法研究学生在借阅专业书籍的同时是否存在阅读其他专业书籍的趋势。通过对某一读者群在一定时期内所借阅图书的流通数据应用关联规则的挖掘分析方法，发现读者在进行专业学习时隐含的各学科知识之间的关联，对图书馆调整资源建设的学科结构、提升读者服务工作水平具有重要的指导意义[11]。

关联分析算法相对其他算法简单、便于理解，应用范围广泛。关联算法又以经典 Apriori 算法或者其改进算法的应用最多。不少学者利用关联规则还建立了模型，如图书推荐模型、读者兴趣模型等。相对数据挖掘在图书馆管理中应用的其他算法来说，关联分析的应用比较成熟。

1.1.2.2 分类分析在图书馆的应用研究

分类分析中，以决策树分类算法在图书馆中的应用较多。

于光和李文峰以哈尔滨工业大学图书馆自动化系统中的用户管理为例，运用决策树的方法对整个读者流通数据库进行挖掘，了解用户访问图书馆的目的和趋势。作者运用 ID3 算法的改进算法 C4.5 算法对拥有大约 120 条记录的数据库进行挖掘，最后得出的结论是，基础类图书基本满足需求，专业类图书较少，不能满足需求；高层次学术类图书较少，即适合研究生用的书较少[12]。

浙江师范学院的吴修琴利用决策树分类算法中的 C4.5 算法，对图书馆用户分类，并对挖掘的结果及其含义进行评价，为图书馆管理提供决策支持[13]。

南开大学的王新筠，以自动化系统中书目数据中属性字段的索书号、建立日期、借出总数三个属性设计出一种决策树分类方法，分析出文献的利用率，及时补充短缺的文献，剔除过时的文献，为采购文献提供科学合理的各种分析报告及预测信息，指导采访人员对购书的种类、数量等进行科学的筛选，优化馆藏结构，为图书馆的采购决策提供支持[14]。

1.1.2.3 序列模式发现技术在图书馆的应用研究

西北大学的曹美琴以湖南工艺美术职业学院图书馆为例，利用时间序列法，找出每周、每月甚至每季、每年中读者使用图书馆的时间规律性，挖掘出的规则，可作为图书馆开馆时间延长或缩短的参考，在寒、暑假，将更加重要。以数据挖掘技术所得出的结果，将可以要求学校决策单位提供适当的人力，同时合理安排人员工作时间和图书馆的开放时间[15]。

吴红燕以季节型 ARIMA 时间序列理论和神经网络理论为基础，提出了处理具有周期性时间序列问题的季节型神经网络模型。模型根据图书借阅流量行为的非平稳时间序列的数据特点，建立了一种季节型动态时间序列的神经网络预测模式，同时为了提高模型预测精度，对实际监控数据进行剔点及光滑处理。模型充分考虑了流量行为周期性的趋势性及随机性，克服了传统时序神经网络图书借阅模型在预报中丢失序列周期性的缺点，并在一定程度上消除了数据突变等奇异点的影响，通过模型做一步预报就可以得出 T（一个周期）步预报的结果，从而避免常规预报步数增多、预报误差增大的缺点[16]。

1.1.2.4 聚类分析在图书馆的应用研究

国内图书馆领域聚类分析的研究与应用主要围绕 K-means 聚类算法而展开。

陈兴为减小 K-means 聚类结果对初值的依赖性，提高聚类的稳定性，采用聚类中心的搜索算法获得较优的初始聚类中心。在搜索过程中通过对数据随机取样，提出了基于取样思想的 K-means 改进算法。陈兴以大连妙思文献集成系统为研究对象，运用 K-means 把图书聚成三类，对应类描述为呆滞书、一般书以及热门书。根据挖掘结果，结合馆藏布局，对各种书籍进行处理。同样，可以把读者也聚成三类，对应描述为积极型、一般型和消极型读者，根据聚类结果，对不同的读者提供不同的服务[17]。

燕山大学的张付志教授将 K-means 聚类与层次聚类相结合，利用层次聚类算法的输出结果作为 K-means 聚类的初始中心变量，以解决 K-means 聚类结果的不稳定性问题，通过层次聚类和 K-means 聚类相结合的聚类方法，缩小了近邻搜索的范围，同时提高了图书推荐的准确度[18]。

陈亚东在其文章中提到使用聚类分析指导文献信息采集，优化馆藏文献结构。通过对馆藏文献的流通记录、检索请求、拒借信息及借阅频率等进行分析，为补充和优化馆藏文献提供决策支持，并可借此分析出相关文献的利用率，及时剔除过时的馆藏文献，增加