·信息组织与服务·

数据挖掘在国内图书馆应用领域研究综述

俞锦梅

（广东中山市中等专业学校图书馆 广东中山 528458）

摘 要：文章在对国内数据挖掘图书馆应用相关文献计量分析的基础上，从数据挖掘在数字图书馆、高校图书馆以及图书馆个性化服务等方面的应用分析了其在图书馆应用的研究现状，最后从理论研究不深入、研究项目经费支持少，以及应用研究与实践结合不紧密等方面归纳总结了图书馆数据挖掘应用研究存在的问题。

关键词：数据挖掘；图书馆；应用；综述

中图分类号：G250.73 文献标识码：A DOI：10.11968/tsygb.1003-6938.2015053

Overview of the Data Mining Application Research of Domestic Library

Abstract Based on metrological analysis of the publishedpapers in Data Mining application research of domestic library , the author analyzes the research status of data mining in term of digital library, university library, the personalized information service of the library, and the methods of data mining.In the end, the problems exiting in these papers are discussed,such as the theory problem, fund problem and application problem,etc..

Key words data mining；library；application；literature research；overview

随着数据库技术应用的快速普及，图书馆信息的种类和形式越来越丰富，需要存储和传播的信息资源数量日益庞大，数据量呈现“爆发式”增长的趋势[1]。然而，面对海量数据的处理，图书馆传统的信息化管理模式和手段却显得无能为力，有些图书馆不由自主地陷入了“数据丰富，知识贫乏”的局面[2]。在这种情况下，如果将数据挖掘技术应用于图书馆服务之中，就可从大量图书馆数据中筛选出隐藏的、有用的数据，发掘表面上复杂无序信息的内在联系[3]，找出有价值的信息知识，实现“数据→信息→知识→价值”的转变。

目前，作为数据库研究、应用与开发最活跃的分支之一[4]，数据挖掘技术正在带动学术研究进步，并推动产业界的不断发展，数据挖掘也成为图书馆应用研究的一项重要课题，不断地吸引着国内外图书馆界的专家学者们的极大关注[5]。笔者尝试对检索文献进行整理归纳，综述数据挖掘在国内图书馆领域应用研究的现状及热点，分析当前研究存在的不足，以为进一步的研究应用指引方向。

1 国外数据挖掘在图书馆应用的研究现状

国外最早以数据挖掘在图书馆中的应用为主题的论文出现在 1997 年，自此之后，国外许多专家学者开始关注数据挖掘在图书馆领域的应用[6]。围绕面向图书馆的数据挖掘技术，不少学者还提出了应用理论及实现方法[7]。从发文量来看，据统计，SCI 收录数据挖掘技术方面的文章呈现出逐年递增的趋势，其目前收录的图书馆领域有关数据挖掘技术应用的文献将近 30 篇[8]。尤其是近几年来，欧洲和北美地区对数据挖掘技术在图书馆的理论与应用方面取得丰硕的成果。例如，美国加州大学 Michael cooper 教授利用数据挖掘对加州大学数字图书馆使用记录进行分析，得出了不同类型用户的逗留时间规律，他还构建了数学模型，应用时间序列以及聚类等分析方法研究图书馆用户的行为规律，并对未来的趋向进行了科学预测[9]。芝加哥大学图书馆的 Swansan 开发了 Arrowsmith 软件系统[10]。该系统可以对数据库文献信息进行深度挖

收稿日期：2015-01-20 责任编辑：魏志鹏

掘,探索文献中信息之间的内在联系,挖掘有价值的信息知识,这一成果吸引了该领域专家学者的广泛关注[11]。Papatheodorou 等人提出数据挖掘技术可用于图书馆数字化数据分析,其结果可成为图书馆管理者制定科学馆藏和管理策略的重要依据。

2 数据挖掘在图书馆应用的研究热点

2.1 数据挖掘在数字图书馆的应用

目前,数据挖掘技术主要应用于数字图书馆读者分析研究、资源建设优化,以及多媒体数字资源挖掘等几个方面。关于读者分析研究,大部分专家学者采用聚类分析方法对读者类别进行划分,而后再进一步进行关联规则分析,以对每一类读者的借阅特征进行深入挖掘,精确地掌握读者信息,更好地实现为读者提供服务;也有学者引入"读者信息域的概念",运用数据仓库技术,对读者信息进行全面挖掘,确保能对读者特征进行准确的分析。还有学者将数据挖掘应用到读者主观感受的研究之中,例如,徐原青[12]在数字图书馆总体规划的早期就引入了数据挖掘技术,通过构建数据仓库,利用 Analysis Services 2000 数据处理机制,对基于读者满意度的数据挖掘在数字图书馆中的应用进行了研究。在图书馆资源建设优化方面,潘小枫[13]从数据应用数字图书馆管理系统建设、馆藏的深层次加工,以及网络信息资源挖掘等方面提出了推进数字图书馆发展策略;有的学者提出了应用基于数据挖掘的数字图书馆馆藏建设评价方法,通过评价为优化馆藏策略提供参考;还有学者立足于对数字图书馆借阅数据进行挖掘分析,对图书馆信息资源的利用情况进行评价等角度开展研究。对于多媒体数字资源挖掘研究,李默[14]提出使用 Web 挖掘等技术构建多媒体资源用户行为分析的原型系统,采用频繁模式树算法对用户信息进行分析的方法。

2.2 数据挖掘在高校图书馆的应用

国内图书情报学的专家学者围绕数据挖掘在高校图书馆的应用开展研究[15]。比如,赵卫军[16]就数据挖掘在高校图书馆资源优化、智能化服务、信息自动化处理等方面的应用展开了讨论;王慧敏等[17]利用 SPSS 和 MATLAB 软件作为数据挖掘工具,以西安工程大学图书馆自动化管理系统的馆藏数据作为基本数据源,对西安工程大学图书馆的入库比例以及各学院借阅量排名进行对比细分,探讨数据挖掘技术在图书馆中的应用;孙健波[18]在硕士论文中,利用 k-means 算法实现了对读者和图书的聚类分析,根据聚类结果指导图书馆管理和对读者个性化服务;同时,他还对 Apriori 算法进行了改进,采用关联规则挖掘对读者数据和图书数据进行挖掘,探索那些隐藏在数据中的潜在规律。金瑶[19]对数据挖掘在高校图书馆的资源管理、信息服务,以及图书馆工作管理进行了探讨;杨光和张学潮[20]提出了利用数据挖掘技术,对图书馆信息系统中隐藏的用户相关的知识进行发掘,并以山西大学为例,对图书馆用户行为进行了分析。此外,有的学者提出了基于数据挖掘的高校图书馆图书采购计划辅助决策方法;有的学者提出利用数据挖掘构建 web 学科导航系统,对图书馆信息资源系统进行丰富;还有学者提出了基于数据挖掘技术的图书馆信息系统建设策略。

2.3 数据挖掘在图书馆个性化服务中的应用

除了将数据挖掘用于高校图书馆之外,不少专家针对数据挖掘在图书馆个性化服务方面进行了积极的探索。国内对数据挖掘在图书馆个性化服务的应用研究包括以下几个方面:个性化服务模型构建,个性化服务软件开发。吴一平[21]提出了基于智能聚合技术的图书馆个性化信息服务方法;史艳梅通过对 CMPS 系统模型的设计,实现对用户兴趣的获取;柳炳祥等探讨了粗糙集和模糊聚类算法应用到图书馆个性化服务中的方法;张英等提出了适合图书馆多媒体数据挖掘的系统框架,并且给出了对音频、图像以及视频等多媒体进行挖掘的方法;在个性参考咨询研究方面,杨亚华提出了把知识管理、知识挖掘和参考咨询服务有机结合的参考咨询服务结构;关于图书馆个性化服务软件开发,中国人民大学等高等学府率先开发了 KBDL 个性化服务系统,沈阳东软软件股份有限公司推出的东软 Internet/Intranet 应用构架平台(Neusoft Web)等软件系统,为图书馆个性化服务提供了丰富的特色应用。

2.4 数据挖掘的主要方法及软件

综合数据挖掘在图书馆应用领域文献,可以把图书馆数据挖掘方法归纳为概念描述、分类和预测、聚类分析、关联规则和偏差检测。从现有文献进行分析,用于图书馆数据挖掘的技术主要包括人工神经网络和统计分

析、模糊数学、归纳学习、仿生学、公式法、可视化手段等。而在图书馆应用软件方面，数据挖掘包括通用型工具、综合数据挖掘工具，以及面向特定应用工具。

2.4.1 通用型工具

通用型工具目前应用最为广泛，其所占市场也最大，技术手段最成熟。通用的数据挖掘工具不区分具体数据的含义，所以一般采用通用的挖掘算法，处理常见的数据类型，其中包括的主要工具有 IBM 公司 Almaden 研究中心开发的 QUEST 系统，SGI 公司开发的 MineSet 系统，加拿大 Simon Fraser 大学开发的 DBMiner 系统、SAS Enterprise Miner、IBM Intelligent Miner、Oracle Darwin、SPSS Clementine、Unica PRW 等软件。

2.4.2 综合数据挖掘工具

综合数据挖掘工具反映了商业对具有多功能的决策支持工具的真实和迫切的需求。商业要求该工具能提供管理报告、在线分析处理和普通结构中的数据挖掘能力。这些综合工具包括 Cognos Scenario 和 Business Objects 等。

2.4.3 面向特定应用工具

这一部分工具正在快速发展，在这一领域的厂商设法通过提供商业方案而不是寻求方案的一种技术来区分自己和别的领域的厂商。这些工具是纵向的、贯穿这一领域的方方面面，其常用工具有重点应用在零售业的 KD1、主要应用在保险业的 Option&Choices 和针对欺诈行为探查开发的 HNC 软件。

3 存在的问题

3.1 理论研究不深入

自 20 世纪 90 年代后期以来，国外图书情报学的专家学者们就开始致力于图书馆数据挖掘相关理论研究，就图书馆的数据挖掘技术、应用理论及方法而言，不少学者具有自己独到的见解。较为典型的有：Nicholson 提出了书目挖掘(Biblio mining)的概念；May Chau 构建了图书馆数据挖掘理论模型，并研发了图书馆网上信息数据挖掘系统；Kyle Baner-jee 对数据挖掘技术应于图书馆的各种方式进行了理论探讨。可以说，关于数据挖掘理论与算法研究，国外图书馆领域已形成较为成熟的理论体系。相比之下，国内图书馆界对于数据挖掘的理论研究起步较晚，从现有的研究文献来看，大约 76%的文献只是介绍数据挖掘的方法，以及该方法在图书馆实践的应用，有的作者甚至只对其它学科文献的理论研究成果进行简单的移植，对数据挖掘在图书馆领域的理论基础及运用实践缺乏个人分析研究。总体来看，这些文章偏重于对数据挖掘技术在图书馆领域应用的定性分析，对于数据挖掘在图书馆方面的应用缺乏必要的理论研究，文章作者也并未应用计算机仿真等定量研究手段对方法使用的可行性进行分析，并且，国内图书馆界目前还没有提出具有影响力的数据挖掘模型。中国知网中仅一篇《数字图书馆数据挖掘的基础研究》对数据挖掘技术在图书馆应用的理论基础进行了简要的分析。总之，数据挖掘技术在图书馆的应用尚属于起步阶段，迄今为止，还没有形成较为系统、成熟的理论体系，国内尚未正式出版一本有关图书馆数据挖掘方面的专著，因而，对数据挖掘理论在图书馆应用方面的探讨将是长期而艰巨的任务。

3.2 应用研究不全面

数据挖掘是计算机、统计学、可视化、人工智能和机器学习等多学科相结合的产物，并已成功应用于金融、医疗、互联网、学校教育和遥感等领域。对图书馆而言，数据挖掘主要应用于图书馆个性服务、图书馆知识发现、图书馆文献资源建设、数字图书馆建设、图书馆内部工作流程优化、图书馆用户挖掘、图书馆用户行为分析等方方面面。

纵观国内数据挖掘在图书馆领域的应用研究，从发文量上看，尽管在 2007 年之后，国内相关文献的总量达到一个高潮，然而，发表在图书情报学中文核心期刊的比例不高，质量较高的论文并不多见；从发文作者的分布来看，论文研究作者大都来自高校图书馆系统，来自公共图书馆和高职院校图书馆的作者为数不多；从作者发文数量来看，发表论文数量 3 篇以上的作者只有 6 人，发表论文数量 2 篇的作者 33 人，由此可见，高产作者数量不多；从论文主题进行分析，关于数据挖掘在高职院校图书馆应用的文献不到 10 篇，大约有 98%的研究文献是以大学图书馆为背景，很难看到有科学图书馆和公共图书馆的作者的研究成果。所有这些现象都说明国内目前对公共图书馆和高职院校图书馆的数据挖掘研究并未引起足够广泛的重视。从方法应用来看，现有文献在方法应用研究方面缺乏针对性，研究者们通常局限于将常用的贝叶斯分析、聚类分析和关联分析应用到图书馆借阅、采访等业

务之中，而没有着眼于图书馆的实际业务进行针对性的分析,有的放矢,目前尚未发现粗糙集与关联规则联合数据挖掘、时空数据挖掘,以及粗糙集理论和神经网络结合的数据挖掘等方法应用于图书馆领域的研究，现有的方法在原理上缺乏创新性；另一方面,随着"云计算"和移动互联网技术的发展成熟,人类迎来了大数据时代,然而,从研究选题情况进行分析,虽然在 2011 年就有专家学者提出了数据挖掘技术在移动图书馆和云图书馆中应用是未来的发展趋势,但当前只有周艳在《现代情报》发表的《基于云平台的图书馆数据挖掘技术研究》一文对数据挖掘技术在"云图书馆"的应用进行了探讨[23]；针对手机读者的需求，重庆大学图书馆与国家图书馆等率先推出手机图书馆 WAP 网站,满足移动用户需要,但是,从中国知网现有的数据来看,只有聂飞霞在《基于数据挖掘的移动图书馆个性化图书推荐服务》一文中[24]提出了应用数据挖掘技术的移动图书馆个性化图书推荐服务模式。关于大数据和云计算相结合的数据挖掘在图书馆领域的应用研究,目前国内尚未见到相关的文献报道。

3.3 研究项目和经费支持少

在所有检索的文献中，明确标注有支持项目和支持经费的只有 14 篇。其中,国家 863 计划资助项目资助的只有一篇，国家自然科学基金和国家社会科学基金资助的论文 6 篇,总体来看,论文基金资助率仅为 3.47%,明显低于其它领域的资助水平。相对其它研究领域,此类项目支持经费不高。而科学研究与推进需要经费的支持,尤其是数据挖掘技术门槛较高，既需要具有人工智能数理统计学、计算机、数据库等专业知识和技能,同时也需要更多的经费支持,为它进一步的研究创造条件。

3.4 研究成果与图书馆管理信息系统开发联系不紧密

图外图书馆将数据挖掘的研究结合到图书馆信息系统建设之中，目前已开发出具有数据挖掘功能的图书馆管理信息系统，如新西兰克莱斯特彻奇教育学院图书馆的 MyLibrary-Christ church College of Education,华盛顿大学图书馆的 My Gateway -University of Washington Libraries，以及康奈尔大学的图书馆 My Library Cornell University Library 等等,这些系统的构建都是基于数据挖掘的思想,并且在实践中发挥了巨大的作用。相比而言,由国内图书馆开发的真正可操作性强、易于实现、能够指导实际业务的成熟产品却为数不多，仅有包括中国人民大学在内的少数几所大学图书馆自行研发了图书馆个性化服务系统——KBDL 系统。现有的文献中,大多偏重于数据挖掘理论的研究，对数据挖掘应用于图书馆信息系统及算法测试的研究较少，大部分的研究仅是局限于聚类分析、关联规则等方法,对图书馆采访数据进行相应的研究，极个别的研究者将研究的成果应用于该馆实际运作管理。从图书馆数据挖掘软件开发来看,大部分图书馆还是通过使用 Intelligent Miner、SPSS Clementine、SAS Enterprise Miner、Orange、KNIME、Weka 等数据挖掘软件对图书馆的数据进行分析、处理和挖掘,很少用于图书馆领域的专用的数据挖掘软件，现有的成果并不能有效地指导图书馆信息系统开发建设的实践，图书馆复杂数据类型挖掘(Web,Text,音频、图形图像、视频等)软件的研究在国内尚属空白。

4 结语

图书馆数据挖掘综合了可视化技术、智能图书馆系统和数据挖掘等方面的知识和技术，它是一个新兴的研究领域。本文通过对数据挖掘在图书馆中应用研究的回顾,从高校图书馆、数字图书馆、图书馆个性化服务[25],及数字挖掘的主要方法及软件研究等多个方面归纳总结了国内数据挖掘在图书馆领域应用的研究现状。可以看出,国内图书情报学的专家学者为数据挖掘在图书馆领域的应用做了大量的研究工作,取得了丰硕的成果。但同时也应看到：目前在该领域的研究，仍存在理论研究不够深入、应用研究不全面、研究成果与图书馆管理信息系统开发联系不紧密等问题。因此,为了能使数据挖掘更好地应用到图书馆和各项实践,将来应在理论研究上下功夫,在实践研究上求突破,同时,还要加速"一专多能"的人才培养,加大科研经费的投入力度,进一步推动研究成果向实践应用的转化。

参考文献：

[1]奉国和,奉永桃.近十年国内图书馆数据挖掘研究文献计量分析[J].图书馆论坛,2011(1):46-49.

[2]唐吉深.图书馆数据挖掘技术研究现状述评[J].图书馆界,2011(l):42-64.

[3]Michael C . Patterns of a web based library catalog[J].

Journal of the American Society for Information Science &Technology, 2001,52(2):137–148.

[4]R. Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns from the Web Data[M]. phD thesis. Dept of Computer Science.University of Minnesota. 2000.

[5]Fu Kai–Yan,LiuYan,Zhang Qin,etc. Data mining services in the university library in the application[J],Medical Information. 2011,24(1):262–264.

[6]UTHURUSAMY R. From Data Mining to Knowledge Discovery:Current Challenges and Future Directions[C]. FAYGAD U.Advances in Knowledge Discovery and Data Mining. The MIT Press,1996:561–569.

[7]HANJ, KAMBER M,TUNGAK H.Spatial Clustering Methods in Data Mining[J].A Survey Geographie Data Mining and Knowledge Diseovery,2008,8(10).

[8]习慧丹.数据挖掘研究综述[J].电脑与信息技术,2012(2):43–45.

[9]Nicholson S. Bibliomining for automated collection development in a digital library setting: Using data mining to discover Web –based scholarly research works [J]. Journal of the American Society for Information Science and Technology, 2003, 54(12): 1081–1090.

[10]高巨山.数字图书馆构建中的数据挖掘应用研究[J].图书馆工作与研究,2009,158(4):20–21.

[11]潘庆超.网格数据挖掘在信息服务质量评价中的应用[J].现代情报,2009(7):141–143.

[12]徐原青.基于读者满意度的数据挖掘在数字图书馆中的应用[J].图书馆学刊,2009(7):107–109.

[13]潘小枫.数据挖掘技术及其在数字图书馆建设中的运用[J].图书馆理论与实践,2006(4):105–106.

[14]李默.基于 web 数据挖掘技术在数字图书馆建设中的应用[J].大学图书情报学刊,2007(4):105–106.

[15]田瑞雪.国内图书馆数据挖掘技术应用研究述评[J].科技信息,2014,(1):167–232.

[16]赵卫军.数据挖掘技术在高校图书馆中的应用[J].图书馆论坛,2007(4):126–128.

[17]王慧敏,贺兴时,牛四强.数据挖掘在高校图书馆中的应用[J].西安工程大学学报,2014(2):241–245.

[18]唐杰,梅俏竹数据发掘学科发展报告[EB/OL].[2013–10–17].http://www.pinggu.org/jingji/987.html.

[19]金瑶.数据挖掘技术在高校图书馆管理系统中的应用[D].上海:华东师范大学信息学院,2010.

[20]杨光,张学潮.数据挖掘在高校图书馆用户行为分析中的应用——以山西大学图书馆为例[J].晋图学刊,2011(2):19–27.

[21]吴一平.智能聚合技术在图书馆个性化信息服务中的应用[J].图书馆工作与研究,2008(11):58–61.

[22]杨传明.基于移动代理的数据挖掘在数字图书馆中的应用研究[J].情报理论与实践,2008(3):436–439.

[23]周艳,李萍,吴雷.基于云平台的图书馆数据挖掘技术研究[J].现代情报,2012(7):46–49.

[24]聂飞霞.基于数据挖掘的移动图书馆个性化推荐服务[J].图书馆学刊,2014(5):46–49.

[25]韩丽.Agent 技术在数字图书馆个性化信息服务中的应用[J].现代情报,2008(4):104–105.

作者简介：俞锦梅,女,广东省中山市中等专业学校图书馆馆员。