**Advances in Social Sciences** 社会科学前沿, 2017, 6(11), 1394-1399
Published Online November 2017 in Hans. http://www.hanspub.org/journal/ass
https://doi.org/10.12677/ass.2017.611196

Hans汉斯

# Status Quo and Development of Technology of Library Data Mining

**Zhengya Ma**

Xi'an Technological University Library, Xi'an Shaanxi
Email: 997369313@qq.com

Received: Oct. 27th, 2017; accepted: Nov. 10th, 2017; published: Nov. 17th, 2017

## Abstract

**This paper retrieves the papers on library data mining published from 2010 to 2016, based on the academic literature in the field of academy culture database research from CNKI. Some aspects of them are analyzed such as the reader service, personalized document acquisition and digital resources, information service. This paper points out the problems existing in the research of library data mining.**

## Keywords

**Data Mining, Library, Overview**

# 图书馆数据挖掘技术研究的现状与发展

**马征亚**

西安工业大学图书馆，陕西 西安
Email: 997369313@qq.com

收稿日期：2017年10月27日；录用日期：2017年11月10日；发布日期：2017年11月17日

## 摘　要

**以中国知网的“中国学术期刊网络出版总库”为数据源，检索2010年~2016年国内图书馆数据挖掘研究论文，对其从读者个性化服务、文献采访、数字资源建设、信息服务等方面进行主题评述，指出现阶段图书馆数据挖掘研究存在的问题。**

## 关键词

**数据挖掘，图书馆，综述**

文章引用: 马征亚. 图书馆数据挖掘技术研究的现状与发展[J]. 社会科学前沿, 2017, 6(11): 1394-1399.
DOI: 10.12677/ass.2017.611196

## 1. 前言

随着计算机技术、通信技术和多媒体技术的迅速发展与综合利用，人类已经进入信息时代。如今各行各业都建立了信息化系统，每天都会产生大量的数据。这些数据通常数量很多，并且以其原始形式显示时并没有什么直接用处。数据挖掘就是通过数据库、机器学习、人工智能、统计学等领域的技术，从这些原始数据中提取出那些隐含的、未知的和有价值的潜在的信息的过程[1]。图书馆的各种信息资源(如图书借还日志、WEB 访问记录、书目检索记录等)包含着大量潜在、有价值的信息，这些信息往往能反映出读者实际的信息需求。通过数据挖掘技术所挖掘出有意义的隐藏信息，不仅有益图书馆了解读者信息需求，而且对于图书馆的各项业务如开展个性化服务、文献采访工作、数字图书馆建设、参考咨询工作等都具有重要的参考价值。近几年来由于数据挖掘在图书馆行业得到不断广泛的应用，吸引了更多的图书馆研究人员关注这个领域的发展，研究成果颇多。笔者统计了近六年我国有关图书馆数据挖掘领域的论文，对其进行整理和归纳，分析数据挖掘在国内图书馆领域研究的现状，并指出研究的不足之处，为进一步的研究指明了方向。

## 2. 数据挖掘技术在图书馆应用的研究现状分析

### 2.1. 文献计量分析

通过对中国知网“中国学术期刊网络出版总库”收录的 2010~2016 年间有关国内数据挖掘技术在图书馆应用研究方面的论文进行定量、归纳分析，以展示其研究现状。采用检索表达式“主题 = (图书馆)并且主题 = (数据挖掘)并且时间 = (2010~2016)”进行检索，得到有效记录 856 条(以上检索时间为 2016 年 11 月 24 日)。

从发表论文的年代分布看，国内数据挖掘技术在图书馆的应用研究所发表的论文数量先逐年递增，在 2014 年达到最高值后又逐年递减。发表论文的数量说明学者在这一领域的研究热情很高，数据挖掘技术在图书馆的应用具有广阔前景；2014 年以后逐年递减说明学者在这一领域的研究可能遇到了瓶颈。

从发表论文的来源期刊看(表 1)，856 篇公开发表的期刊论文均分布在省级以上刊物中。其中发表在核心期刊上的论文为 176 篇，占总发表量的 20.6%，被中国社会科学引文索引(CSSCI)收录的论文 87 篇，占论文总数的 10.2%。这表明其研究成果主要集中在普通刊物，学术成果价值较高的论文不多。

### 2.2. 研究方向和算法

图书馆应用数据挖掘研究主要集中在三个研究方向：关联分析、聚类分析和决策树，其中应用关联分析数据挖掘的文献量占总量的 60%，聚类分析的文献量占 30%，决策树应用比率较低，只占 10%。也有些文献进行数据挖掘时综合应用两种手段，如《基于 Weka 读者借阅行为分析》应用了聚类和关联规则两种技术[2]，《基于数据挖掘的高校图书馆个性化信息推荐方法研究》先从读者借阅图书类别入手对读者进行聚类分析，把高校图书馆中的读者细分为不同的群体，然后再进行有针对性的关联挖掘[3]。《基于数据挖掘的图书馆书目推荐服务的研究》采用决策树和聚类分析[4]。大部分关联分析采用最常用的 Apriori 算法，也有用 CARMA 算法或者基于邻接矩阵的算法；聚类分析大部分采用 K-Means 算法，也有用两步聚类模型及使用基于目标函数的模糊聚类算法。决策树采用 C4.5 算法。有的研究探讨了一些算法

**Table 1.** Shows the distribution of the domestic library data mining technology research papers
**表 1.** 国内图书馆数据挖掘技术研究论文年代分布情况

| 年份(年) | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 |
|---|---|---|---|---|---|---|---|
| 论文数(篇) | 90 | 115 | 104 | 115 | 161 | 150 | 121 |

优化的问题。

## 2.3. 使用的软件

进行数据挖掘一般要使用专业的数据挖掘软件，根据检索到的文献情况，研究人员主要使用了以下几种软件进行图书馆数据挖掘工作：

1) Weka

Weka的全名是怀卡托智能分析环境(Waikato Environment for Knowledge Analysis)，是一款基于JAVA环境下开源的机器学习(machine learning)以及数据挖掘(data mining)软件。作为数据挖掘工作平台，WEKA集合了大量能承担数据挖掘任务的机器学习算法，包括对数据进行预处理，分类，回归、聚类、关联规则以及在新的交互式界面上的可视化。

2) SPSS Clementine

它是Spss公司推出的企业级数据挖掘产品，适用于多种操作系统和多种数据源，提供包括神经网络、决策树、聚类分析、关联分析、因子分析、回归分析等在内的丰富的数据挖掘模型，通过节点的连接来完成整个数据挖掘过程。可以输出全部挖掘过程。

3) Analysis Services

Microsoft SQL Server Analysis Services (SSAS)是微软公司提供的数据挖掘平台，提供适用于多种操作系统，可视化界面，提供几乎所有的成熟数据挖掘方法，设置了全部数据挖掘过程，包括数据准备，可以连接各种不同的数据源，提供一系列数据展示技巧，统计分析功能强大，但使用昂贵，操作复杂，结果难理解。

4) 自行开发编程

有些文献没有写明使用的软件，使用程序语言编程完成数据挖掘。

## 2.4. 图书馆数据挖掘的研究范畴分析

### 2.4.1. 读者个性化服务中应用

数据挖掘技术在读者个性化服务中的应用，主要表现为通过对读者信息、读者借阅和Web使用记录等进行挖掘，对读者进行分类，并根据读者的历史借阅记录推送馆藏图书，开展有针对性的服务。如：杨蓉[5]提出了一种改进的Apriori算法，通过挖掘图书馆的借阅记录，挖掘出了借阅书刊间的相关性、不同类书刊被同时检索的相关性、不同类读者查阅资料的相关性等一些关联规则；郑晖[6]从Web层面研究了数字图书馆的个性化服务和数据挖掘算法，并在此研究的基础上，分析设计了基于Web的图书馆个性化推荐系统，解决了图书馆个性化信息服务不足的问题；李静[7]利用FP-growth算法来进行关联规则的数据挖掘，提出了一个在高校图书馆系统中应用数据挖掘技术的具体方案，对高校图书馆探索个性化服务有一定的意义；王哲[8]设计出了一个数字图书馆个性化服务系统整体模型，根据读者偏好、类型的聚类分析，运用关联规则方法，根据聚类的结果对读者借阅信息进行规则挖掘，找出规则模式，为读者提供个性化信息推荐服务；袁媛[9]设计出了一个馆藏推荐系统，该馆藏推荐系统成功实现了对自动化系统记录的流通数据关联规则的挖掘，为读者提供个性化图书推荐服务的目的。

#### 2.4.2. 文献采访工作中应用

数据挖掘技术在文献采访工作中的应用，主要是利用关联分析、聚类分析对读者信息、书目数据、读者借阅数据、文献检索记录等信息进行数据挖掘，发现读者与所借阅的图书之间的关系、不同读者群的借阅倾向等等，以便科学地指导采访馆员选书。数据挖掘技术在文献采访工作中的应用能使图书馆的采购工作目的性更强，订购的图书的利用率更高，还可以节约图书馆的开支。如：王婧怡[10]建立基于数据仓库的图书订购决策支持系统，设计了图书借阅数据仓库的主题星型模型，开展了联机分析处理的实证研究。结果表明，通过数据挖掘能够为图书馆订购决策提供信息支持；张宏伟[11]应用数据挖掘技术对读者借阅流通、检索查询、预约借书、博硕学位论文引文等方面的历史数据进行分析，获知读者文献需求，从数据方面为图书馆文献采访决策提供参考，提高了图书的借阅率；唐吉深[12]提出基于数据挖掘的应用型高校图书馆文献采访策略模型，并给出实现模型的技术思路；赵研科[13]利用数据挖掘中的分类、聚类和关联等方法，对影响图书采访决策的因素和数据进行挖掘，为图书采访工作提供科学依据，并且设计基于数据挖掘的高校图书采访决策系统。

#### 2.4.3. 图书馆的数字资源建设中应用

目前，全球大数据呈现爆发式增长，已经渗透到各个行业和业务职能领域。这同样也对图书馆的数字资源建设和服务产生了很深的影响。目前图书馆的数字资源建设应该包括三部分工作内容：数字资源制作、数字资源组织和管理、数字资源服务。黄筱玲等[14]以湖南大学图书馆为例，采用数据统计研究法，对其数字资源的使用情况进行调查分析。根据调查结果，对高校图书馆数字资源建设提出一些建议。何毅[15]提出可以对图书馆所拥有的数字资源进行文献计量学或网络计量学的一些统计和评价，不仅为图书馆的馆藏政策制定提供了重要依据，而且可以为用户开展创新服务。史敏鸽等[16]对长安大学图书馆用户访问信息进行分析，利用大数据挖掘技术提取用户信息需求偏好，形成用户需求与数字资源匹配趋势图。系统可以为后续采购提供辅助决策依据，为高校数字资源管理评估提供数据支撑，促进数字资源建设服务体系的建设。魏笑笑[17]以陕西省 42 所本科高校图书馆作为研究对象，对陕西高校图书馆的数据资源建设情况进行调查，分析目前陕西高校图书馆数字资源建设的特点模式以及存在的问题，并对今后的发展提出了一些建议。

#### 2.4.4. 信息咨询工作中应用

信息咨询是图书馆服务工作的核心内容，是评价图书馆信息服务质量高低的重要标准。然而随着大数据时代的到来，如何从丰富繁多的数字资源中提取有用信息成为信息参考咨询面临的重要问题。数据挖掘技术可以去粗取精、有效组织日益增加的数字资源并将其进行整合应用。利用数据挖掘技术提升图书馆信息咨询服务质量是十分必要的。汤辉提出[18]数据挖掘技术可以应用在图书馆信息咨询的以下几个方面：用户需求分析与挖掘、资料收集甄别与结果反馈分析、用户分类管理等；姚展[19]提出了构建移动图书馆信息咨询服务体系和结合大数据与数据挖掘的图书馆信息咨询服务体系构建思想；乔幸娟[20]认为数据挖掘技术在信息参考咨询中的应用主要体现在以下两个方面：1) 读者服务方面；2) 文献检索方面。杨峰[21]提出图书馆必须利用大数据原理和技术，从大量的非结构化数据、半结构化数据中捕捉和挖掘潜在的价值，从而提高图书馆参考咨询服务的智能化水平，开拓参考咨询服务新领域；黄如花[22]认为图书馆可以利用基于密集数据的存储、采集、挖掘、分析等相关技术，提取知识资源和用户需求，将传统的被动参考咨询服务转变为主动高效的知识咨询服务。

## 3. 存在的问题

### 3.1. 理论研究不够深入

国外对数据挖掘方面的研究，始于上个世纪 90 年代。围绕面向图书馆的数据挖掘技术，研究成果颇

丰，不少学者还提出了应用理论及实现方法。较为典型的有：Nicholson [23]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。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。20 世纪特别是 2004 年以来，国内虽然有很多学者也投入到图书馆数据挖掘的研究中，但大都是利用现有的数据挖掘方法对图书馆的流通数据或者数字图书馆方面的数据进行分析，对于数据挖掘在图书馆方面的应用缺乏必要的理论研究。迄今为止，国内没有一本全面介绍数据挖掘应用到图书馆中的著作。总的来说，国内数据挖掘在图书馆中的运用研究还处于起步阶段。

### 3.2. 有价值的研究成果较少

通过前面的文献计量分析，我们知道，虽然相关文献数量以五年为一个阶段螺旋性增长，但以现代情报、情报杂志、科技情报开发与经济、农业图书情报学刊等 7 种期刊组成了图书情报领域研究数据挖掘的核心期刊，期刊整体质量不高。这主要是由于数据挖掘是一门综合性很强的学科，它汇聚了数据库、人工智能、数理统计、可视化等方面的知识。图书馆的工作人员要进行这方面的研究，不仅需要完全了解图书馆的业务知识，还要清楚的知道通过数据挖掘会给图书馆的各项工作带来什么样的变化，更重要的是还应该具备计算机、统计、数理等方面的知识，而这样的馆员少之又少。从检索结果看，2010 到 2016 年发表的 856 篇有关图书馆数据挖掘的文章中，核心期刊的篇数只有 176 篇，占总发表量的 20.6%。而且在研究人员中只有田瑞雪 1 人发表了 5 篇文章，发表 4 篇的作者只有 8 人，发表 3 篇的作者有 17 人，高产作者不多，说明这个课题就没有核心的研究群。

### 3.3. 课题研究的动力不足

相对于通讯、电子商务等商业领域，目前很多图书馆还是传统的被动服务模式，对用户的依赖度不高，没有热情去分析哪些用户是活跃度高的用户，哪些用户未来很可能流失等等，更无激情去挖掘、培养潜在用户。实际上，在大数据环境下，图书馆可以利用积累下来的一些数据，如书目检索记录、流通日志、电子资源访问日志等等，进行数据挖掘，引导文献采购和电子资源采购，提高资源的利用率；还可以通过对读者借阅行为进行挖掘，有针对性地开展图书推荐服务；利用数据挖掘技术辅助图书馆的决策管理等等。

## 4. 结语

本文通过对近六年来国内学者对数据挖掘在图书馆领域的研究现状的分析，虽然现阶段本课题的研究取得了一定的成果，但也存在理论研究不够深入、有价值的研究成果少、课题研究的动力不足等问题，因此，为了使数据挖掘技术更好地应用于图书馆，今后我们一定要更加重视理论研究，增加课题经费的投入，以便吸引更优秀的学者参与到这个领域中来。

## 参考文献 (References)

[1] Han, J.W. and Kamber, M., 著. 数据挖掘概念与技术[M]. 范明, 等, 译. 北京: 机械工业出版社, 2001.

[2] 储文静, 奉国和. 基于 Weka 读者借阅行为分析[J]. 情报科学, 2010, 28(3): 424-429.

[3] 刘显显. 基于数据挖掘的高校图书馆个性化信息推荐方法研究[D]: [硕士学位论文]. 沈阳: 辽宁大学, 2013.

[4] 荆月敏. 基于数据挖掘的图书馆书目推荐服务的研究[D]: [硕士学位论文]. 太原: 中北大学, 2014.

[5] 杨蓉. Apriori 算法在图书馆个性化服务中的应用研究[D]: [硕士学位论文]. 长沙: 中南大学, 2013.

[6] 郑晖. 基于 Web 挖掘的图书馆个性化推荐系统的设计与实现[D]: [硕士学位论文]. 成都: 电子科技大学, 2013.

[7] 李静. 数据挖掘技术在高校图书馆个性化服务中的应用研究[D]: [硕士学位论文]. 天津: 天津大学, 2012.

[8] 王哲. 数据挖掘技术在高校图书馆个性化服务中的应用研究[D]: [硕士学位论文]. 重庆: 重庆大学, 2012.

[9] 袁媛. 数据挖掘在高校图书馆个性化服务中的应用研究[D]: [硕士学位论文]. 郑州: 郑州大学, 2011.

[10] 王婧怡. 征订信息管理与订购决策支持系统研究[D]: [硕士学位论文]. 镇江: 江苏大学, 2010.

[11] 张宏伟. 数据挖掘在高校图书馆文献采访决策中的应用[C]//中华中医药学会. 全国中医药图书信息学术会议暨第十一届中医药院校图书馆馆长会议论文集, 2014.

[12] 唐吉深. 基于数据挖掘的应用型高校图书馆文献采访研究[J]. 农业网络信息, 2014(3): 22-24.

[13] 赵研科. 基于数据挖掘的高校图书采访决策系统设计与实现[D]: [硕士学位论文]. 长沙: 湖南大学, 2012.

[14] 黄筱玲, 李雯. 基于数据分析的高校图书馆数字资源建设优化——以湖南大学图书馆为例[J]. 高校图书馆工作, 2012, 32(1): 58-60.

[15] 何毅. 资源发现 知识导航——大数据时代图书馆的数字资源建设与服务[J]. 中国索引, 2013(4): 23-27.

[16] 史敏鸽, 孙勇, 崔玲玲. 基于用户需求的高校图书馆数字资源系统评价方法研究——以长安大学图书馆为例[J]. 情报探索, 2016(9): 80-85.

[17] 魏笑笑. 高校图书馆数字资源现状调查与模式分析——以陕西 42 所本科院校为例[J]. 农业图书情报学刊, 2016, 28(2): 32-35.

[18] 汤辉. 数据挖掘在图书馆知识咨询中的应用研究[J]. 农业图书情报学刊, 2015, 27(2): 91-93.

[19] 姚展. 高校图书馆信息咨询服务体系构建研究[J]. 通讯世界, 2016(6): 220-221.

[20] 乔幸娟. 数据挖掘技术在数字图书馆中的应用研究[J]. 农业图书情报学刊, 2014, 26(12): 118-120.

[21] 杨峰. 大数据环境下开拓参考咨询服务新领域[J]. 农业图书情报学刊, 2014, 26(12): 206-208.

[22] 黄如花, 李白杨. 数据密集型科研环境下的知识组织与导航模式研究[J]. 图书馆学研究, 2015(11): 51-55.

[23] Nicholson, S. (2003) Bibliomining for Automated Collection Development in a Digital Library Setting: Using Data Mining to Discover Web-Based Scholarly Research Works. *Journal of the American Society for Information Science and Technology*, **54**, 1081-1090. https://doi.org/10.1002/asi.10313