知识挖掘在图书馆参考咨询服务中的应用研究*

□ 张久珍 / 北京大学信息管理系 北京 100871
王冬 / 中信证券 北京 100011

摘要：文章将知识挖掘技术引入到图书馆参考咨询服务工作中，探讨了知识挖掘技术在文献推荐、学术资源导航等方面的应用，分析了知识挖掘技术在参考咨询服务中成功实施的现实瓶颈，并提出了相应的解决建议。

关键词：参考咨询服务，知识挖掘，文献推荐服务
DOI：10.3772/j.issn.1673—2286.2012.07.005

1 概述

随着数字化建设的逐步深入、数据采集技术的不断发展，人们每天获取的数据量剧增。面对如此庞大的数据库，人们的需求已经不只是简单的查询和维护，而是希望能够对这些数据进行较高层次的处理和分析，以得到关于数据总体特征和对发展趋势的预测。而这些功能是数据库技术、人工智能和统计学等无法单独完成的。世界著名未来学家约翰•奈斯比特（John Naisbitt）曾说过，“我们淹没在信息之中，但仍处于知识的饥渴中”。知识挖掘技术就是在这种背景下应运而生的，它从数据集中识别出有效的、新颖的、潜在有用的以及最终可以理解的模式，是对数据资源所包含的显性知识进行提炼，例如从各类文献中抽取理论型、方法型、事实型和数值型的知识元；对同类数值型知识元进行列表对比，发现文献的隐性关联；对知识元进行可视化处理、文献智能聚类等，使信息资源增值。

2003年，Barbara Mento和Brendan Rapple对美国研究图书馆协会（Association of Research Libraries，ARL）的124家成员馆进行了调查[1]，以了解它们使用知识挖掘技术的情况。在65家图书馆的反馈结果中，26家图书馆（占40%）采用了知识挖掘技术，38家图书馆（占58%）的图书馆认为知识挖掘是一种有价值的应用。大多数图书馆使用知识挖掘和数据仓库技术对图书馆用户流量、馆际互借、馆藏建设、文献采访、数字资源使用和网站使用模式等相关数据进行了搜集和深度分析，以优化图书馆的管理决策。例如，麻省理工学院（Massachusetts Institute of Technology）的数据仓库中收集了图书馆文献流通、馆藏管理、编目和连续出版物管理的大量数据；印第安纳大学（Indiana University）图书馆的数据仓库则收集了关于文献流通和文献采访的大量统计数据；加利福尼亚大学河滨分校（University of California, Riverside）图书馆利用知识挖掘技术建立了INFOMINE网络学术资源库；2006年，北卡州立大学（North Carolina State University）图书馆利用Endeca公司提供的ProFind服务系统，采用知识挖掘的理念，对图书馆联机公共检索目录（OPAC）的主题关联、著者关联、载体关联、用户兴趣聚类和推荐功能进行了全新的开发[2]。此图书馆的Kristin Antelman和Emily Lynema等人通过对Endeca系统和图书馆原有Web2系统的检索日志进行对比分析，发现新系统中排前列的检索结果（top results）里，

* 本文是国家社会科学基金项目《图书馆数字参考咨询服务质量评价研究》（编号：07CTQ002）研究成果。

有68%被判断为内容相关，而这一比率在Web2系统中仅为40%；通过用户使用对比试验，发现新系统用户平均节省了48%的检索时间，系统的易用性和检索的成功率等方面也有明显的提升[3]。新西兰Waikato大学的Steve Jones等在多篇工作报告中利用业务日志分析（Transaction Log Analysis，简称TLA）对数字图书馆的用户日志文件进行使用挖掘。他们集中对该图书馆的计算机科学技术报告数据库的使用数据进行了日志分析[4]。2010年，Lake Charles大学图书馆则利用知识挖掘技术进行学习指导和藏书建设[5]。

自1996年以后，ARL成员图书馆的参考咨询服务使用数量呈现出明显的下滑趋势，从1996年的155336人次下降到2006年的67697人次，其总降幅达到56.4%。在国内，北京大学图书馆网上咨询服务的使用数量是3800人次/年[6]。笔者对北京大学校内学生进行了“图书馆参考咨询服务满意度”问卷调查，结果表明，现阶段的图书馆参考咨询服务在易用性、专业性、系统性和启发性上存在欠缺，用户对其的价值预期并不高，因而图书馆参考咨询服务被大多数用户定位于知识资源获取途径的“候补”席位。可见，图书馆如果不能深化其参考咨询服务，提供个性化的推送服务，将会逐渐失去用户而陷入一个尴尬境地。而在提高服务的专业层次、进行模式创新的过程中，知识挖掘技术将充分结合图书馆自身的资源优势，提供其强大的支持功能。

2 参考咨询服务中知识挖掘技术的应用探索

2.1 图书馆文献推荐服务中的知识挖掘

信息需求属于一种复杂的心理活动，是一个层次化、时序化的过程。在一次信息获取过程中，用户的信息需求在数量上呈现出衰减的趋势，即用户最终表达出来的信息需求，只是为解决问题所需要的客观信息的很小一部分，对于信息服务机构或信息系统来说，它们很难获知用户潜在的信息需求。这部分潜在的信息需求会随着用户掌握信息资源的增多和对问题的认识深入而逐步被激发和唤醒，之后用户会向信息服务机构或信息系统进一步表达自己的信息需求。这个过程会反复、持续地进行，直至用户获得的信息资源足够用于解决其意识到的实际问题。

目前的文献推荐服务中应用的知识挖掘技术，都是针对用户直接或间接表达出来的信息需求进行文献推荐，这在很大程度上提高了参考咨询服务的个性化和主动性。但对于任务型用户和学科参考咨询馆员来说，这些挖掘结果的参考价值可能并不是很高，更准确地说，挖掘出来的文献相对于用户的信息需求针对性可能还不够强，在内容上可能也不具备启发性。这是因为图书馆任务型用户的信息需求不同于一般的商品需求，任务型用户针对某一主题的信息需求是一个时序化的、层次化的心理活动，其信息需求的满足是一个螺旋式深入的过程。因而，知识挖掘技术如果能够支持发掘出用户未意识到的信息需求，那将深刻体现出它对图书馆参考咨询服务的巨大价值。

基于用户信息需求的层次化和时序化的特点，笔者认为时序模式分析是挖掘用户隐性信息需求的出发点和基本思路。图书馆可以利用数据仓库的联机分析处理技术，对该用户的这些图书馆使用记录进行聚集和切块，获得其在这一段时间内的所有相关记录，其中包括每条记录所涉及的文献资源的关键词等。之后通过关键词信息对这些文献资源进行聚类，比如最终将这些文献资源划分为n类，再依照这些类别中核心文献的记录时间进行排序，最后生成一个文献集合的时间序列，如D_1，D_2，……，D_n类。对于这个用户来说，D_2类中的文献所蕴含的知识可能是对D_1类文献所蕴含知识的补充、细化或启发。如此就产生了一条可能成立的时间序列规则，即D_n是D_{n-1}的后继知识，D_{n-1}是D_n的前驱知识。通过对图书馆内大量用户的使用记录进行类似的抽取和深度知识挖掘，能够得到许多条可能成立的时间序列规则。然后对挖掘出来的所有这些文献类别单元（如D_n，D_{n-1}，……，D_1，P_i，P_{i-1}，……，P_1，M_k，M_{k-1}，……，M_1）进行聚类。例如，聚类的结果出现图1所示情形，则M_1和D_1属于类别C_1，M_2、D_2和P_1属于类别C_2，D_4和P_2属于类别C_3。由此可以得到如下结论：对于“C_2是C_1的后继知识”这条时间序列规则，它的支持数为2，置信度为2/2×100%=100.0%；对于“C_3是C_2的后继知识”这条时间序列规则，它的支持数为2，置信度为2/3×100%=66.7%。对于系统给定的支持阈值和置信阈值，如果这两条规则的置信度和支持度都大于相应阈值，那它们就是系统需要挖掘得到的时间序列规则。

当图书馆数据仓库内用户的使用记录数据量达到一定的规模时，就可以通过知识挖掘找出许多满足系

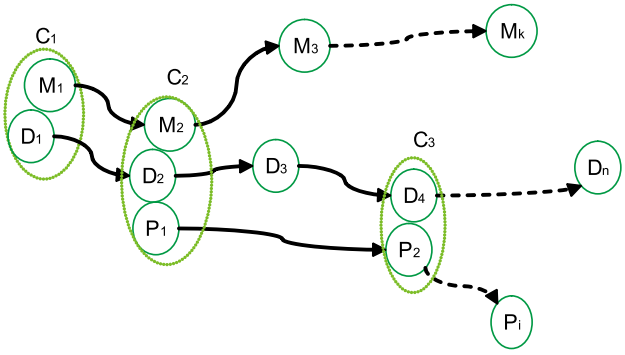

图1 文献序列聚类结果示例

统指定的置信阈值和支持阈值的时间序列规则，从而确定一条主题线索上的前驱文献集合和后继文献集合。当有用户提出信息咨询或文献推荐请求时，文献推荐系统可以根据用户咨询请求中的关键词组合或用户最近阅读文献的关键词向量，使用余弦相似度或其他聚类算法找到符合用户需求的文献集合，并利用文献时序规则确定后继文献集合。之后，结合其他用户对这些文献集合中每篇文献的加权评分，推荐系统可以生成一组评价值较高的文献，为参考咨询馆员最终做出服务决策提供参考。这样，在文献推荐系统的帮助下，面对关注某一主题领域的用户，参考咨询馆员可以以更高的概率掌握其将来可能需要的文献集合，即用户当前可能存在的隐性信息需求，从而有计划地、层次化地向用户推荐高质量的文献资源，加快用户查找文献信息以解决实际问题的进程。

2.2 网络学术信息资源导航服务中的知识挖掘

互联网信息资源逐渐发展成为人们获取信息的重要来源，然而网络信息资源的质量参差不齐，分布散乱无序，使得人们通过网络获取信息的过程中需要花费很长的时间来评估和筛选。面对这些问题，图书馆参考咨询部门亟需开展对网络信息资源的采集、整理、组织，提供网络信息资源导航服务。这已成为当前图书馆，尤其是研究型图书馆参考咨询服务不能回避的问题，也是参考咨询服务创新研究的重要内容。

通过分析和归纳前人的研究成果，我们可以得到利用知识挖掘技术实现网络信息资源筛选的一般流程，它主要分为以下几个步骤：

（1）设定网络信息资源的筛选标准。主要从网页的作者标准、内容标准、组织标准、媒体标准和外部标准这五个大的方面来进行筛选，其具体标准与一般参考资源的采选原则基本一致，只是它更适合网络文献资源的特点。

（2）在计算机系统中设计获取网络文献资源的这些属性信息的程序。

（3）构建试验数据集，在互联网上收集一定数量的学术信息资源和非学术信息资源，并对其类别进行人工标注。

（4）通过知识挖掘和机器学习对试验数据集进行深层次分析，构建多种简约的网络资源分类模型。这一步骤中常用的知识挖掘技术包括逻辑回归、决策树和人工神经网络等。

（5）对比分析多种分类模型，从中选择出合适的模型作为智能代理应用于网络爬虫（web spider）程序中。

通过上述对自动筛选工作流程的阐述，可以看出实现网络学术资源自动采集的基本思想，即通过知识挖掘技术将参考咨询馆员头脑中的隐性知识用一种模糊规则表达出来并应用到资源采集系统中。具体说来，由于参考咨询馆员在进行网页质量评估的时候是根据网络文献资源的一些显性特征来进行分析判断的，这些分析思路往往没有形成文字性的规则，只是隐藏于参考咨询馆员的头脑中。通过德尔菲法或其他一些方法则可以获取这些分析过程中涉及的因素，再利用知识挖掘和机器学习获取针对这些因素的量化标准，之后就可以建立相应的筛选机制来自动处理网络上的大量文献资源。在这里，知识挖掘的引入大大提高了网络学术资源导航库的建设效率，使得在网络导航服务顺利开展的同时，参考咨询馆员有足够的时间来处理那些需要花费更多精力的高层次的问题。

2.3 知识挖掘在参考咨询工作中的其他应用

在参考咨询工作的其他领域，知识挖掘也可以发挥出强大的支持功能。例如在开展图书馆用户培训的过程中，图书馆参考咨询部门可以利用知识挖掘技术，对用户的背景信息和图书馆使用记录进行实时分析，通过聚类、逻辑回归或其他分类方法细分用户群体，以确定哪些用户需要或可能需要哪些方面的培训。比如，图书馆要针对某个法律数据库开展检索与使用培训时，除了在图书馆电子展板、网站主页上发布公告外，可以通

过知识挖掘找出当前或未来可能需要使用这个数据库的用户，之后通过电子邮件或个性化主页向这些用户发出培训邀请函。这种基于知识挖掘技术的培训服务营销模式，不仅能够帮助参考咨询部门确定培训课程的目标群体，还能够帮助其有针对性地设计培训内容，由此使得图书馆用户培训服务的综合效益达到最大化。

另外，基于网络使用（web usage）的知识挖掘技术还可以帮助参考咨询馆员对图书馆网站的可用性（usability）进行改善。图书馆可以从用户的网站访问路径中发现网页之间的关联，通过网页相关性分析发掘用户的偏好。例如，如果很多用户具有从网页A->网页B->网页C的访问模式，则可以认为网页A和网页C之间存在一定的关联规则，那么可以考虑设置推荐路径，在网页A上直接添加网页C的链接，由此实现图书馆网页深度的合理设置，提高用户使用图书馆网站的效率。另一方面，这种基于关联规则的挖掘能够帮助提高图书馆网站的响应速度，对于有效地调度网络代理的缓存有着显著的作用。如在上述例子中，当用户在浏览页面A时，网络代理就可以根据关联规则预先下载与该页面相关联的页面C，即用户很可能访问到的页面，从而提高网络的响应速度。由于关联规则反映的是大多数用户的习惯，因而这种机制能够有效地调度网络代理的缓存，提高图书馆网站的服务效率。

综上所述，随着图书馆参考咨询工作的具体内容不断扩展，知识挖掘技术在图书馆参考咨询服务中将有越来越大的应用空间，将帮助图书馆参考咨询部门创造出越来越多的社会价值。

3 参考咨询服务中知识挖掘的瓶颈

3.1 参考咨询服务数据仓库尚未建立

参考咨询服务中知识挖掘所需的大量数据来自于外部的各种数据库、图书馆已有的管理信息系统、图书馆网站服务器和互联网等，这些数据数量巨大，结构不同，只有建立数据仓库对这些数据进行清洗和整合，才能在参考咨询工作中有效地实施知识挖掘。当前，美国研究图书馆学会中只有一部分图书馆建立了数据仓库，且其中一些数据仓库只覆盖了馆藏资源，而不包括图书馆的使用记录，因而其中只有少数数据仓库适用于参考咨询服务中的知识挖掘。

3.2 图书馆用户的反馈信息缺乏

反馈是控制论中的基石，是信息服务实现个性化和智能化的重要条件。在当前的图书馆中，用户经常使用的信息反馈途径只有图书馆问卷调研、电子邮件和BBS留言等几种途径。这些信息反馈途径存在着进入成本高的固有缺陷，使得参与信息反馈的用户数量极其有限，久而久之，很多用户甚至不清楚图书馆有哪些信息反馈渠道。用户反馈信息的贫乏导致了图书馆工作中的一些决策缺乏有力的依据，加上这些信息反馈不够及时，图书馆的服务决策与用户的信息需求之间往往存在一定的时间偏差。

3.3 图书馆服务记录数量相对单薄

知识挖掘技术的原理决定了它的成功实施必须要有大量的数据作为基础。无论是馆藏文献资源信息，还是图书馆用户使用各种服务留下的记录，如果达不到足够的数量，系统是很难通过知识挖掘找出具有高置信度的潜在模式的。这也就是说，知识挖掘的成功实施需要图书馆长期不断地对各种业务数据进行清理并将其整合到数据仓库中。而另一方面，图书馆馆藏资源和用户的信息需求又在不断地变化着，要保证知识挖掘的结果能帮助参考咨询馆员制定及时而有效的服务策略，就需要挖掘的数据对象具有一定的时效性，也就是说，参考咨询服务数据仓库中的部分数据具有一定的有效期。这就造成了知识挖掘在技术有效性和价值有效性之间的矛盾。

在参考咨询服务中引入知识挖掘技术，需要在对参考咨询工作进行深入剖析的基础上，发掘出知识挖掘技术的切入点，而后再将其融合到信息服务过程中，以帮助参考咨询馆员做出高质高效的服务决策。相信随着对知识挖掘研究的进一步深入，知识挖掘技术将在参考咨询研究领域内受到越来越多的关注，而参考咨询服务将更加理性地发展，在不久的未来将能为图书馆用户营造一个普遍的、全面的知识环境。

参考文献

[1] MENTO B, RAPPLE B. SPEC Kit 274: Data Mining and Data Warehousing [R/OL]. Washington, D.C.: Assocation of Research Libraries. (2003). http://www.arl.org/bm~doc/spec274webbook.pdf.
[2] 数图研究笔记.书目(OPAC)的价值(二)[EB/OL]. (2006-01-27) [2012-04-20]. http://www.kevenlw.name/archives/120.
[3] ANTELMAN K, LYNEMA E, PACE A K. Toward a Twenty-First Century Library Catalog [J]. Information Technology and Libraries, 2006(9): 128-139.
[4] 吕娜.数字图书馆用户关系管理研究[D].北京:北京大学,2006:102-103.
[5] FINNELL J. Reference Question Data Mining A Systematic Approach to Library Outreach [J]. Reference & User Services Quarterly, 2010, 49(3): 278-286.
[6] 北京大学图书馆通讯(总第55期)[EB/OL]. (2006-09-28) [2012-04-20]. http://162.105.138.207/tongxun/tx2002/news17.htm.

作者简介

张久珍（1974-），女，博士，北京大学信息管理系副教授。E-mail: jiu@pku.edu.cn
王冬（1982-），男，硕士，中信证券。

Application Research on Knowledge Mining in Library Reference Services

Zhang Jiuzhen / Department of Information Management, Beijing, 100871
Wang Dong / CITIC Securities, Beijing, 100011

Abstract: This paper introduces data warehouse and knowledge mining techniques into library reference work, discusses applications in library reference services, analyses the obstacles in implementation of knowledge mining, and puts forward some solutions.

Keywords: Reference services, Knowledge mining, Literature recommendation service

（收稿日期：2012-04-21）