【引文格式】方婷.海量学术文献资源的知识挖掘技术与模式[J].中国中医药图书情报杂志,2017,41(4):7-10.
DOI: 10.3969/j.issn.2095-5707.2017.04.003

# 海量学术文献资源的知识挖掘技术与模式

方婷
安徽省芜湖市图书馆，安徽 芜湖 241000

**摘要：**知识挖掘技术是采集检索海量学术文献资源所必需具备的。文章分析了知识挖掘的四大关键技术：联机分析挖掘技术、Web 挖掘技术、数据仓库技术和信息可视化技术，重点研究了海量学术文献资源的用户知识挖掘和学科知识挖掘模式，详细探讨了海量学术文献资源知识挖掘创新模式：知识服务、Lotus 知识发现、综合知识挖掘。

**关键词：**海量学术文献资源；知识挖掘；知识发现
**中图分类号：**G250.7　　**文献标识码：**A　　**文章编号：**2095-5707(2017)04-0007-04

**Knowledge Mining Technology and Modes of Huge Amounts of Academic Literature Resources**

FANG Ting
*(Library of Wuhu City Anhui Province, Wuhu Anhui 241000, China)*

**Abstract:** In this thesis, the necessity of applying knowledge discovery techniques to collecting huge amounts of academic literature resources is analyzed at first. The four key knowledge discovery techniques are then introduced: online analytic mining (OLAM) technique, Web mining technique, data warehouse(DW) technique, and information visualization technique. User knowledge mining mode and subject knowledge mining mode of huge amounts of academic literature resources are further studied empirically. At last, innovative knowledge mining modes of huge amounts of academic literature resources are discussed in details: knowledge service mode, Lotus knowledge discovery mode, and comprehensive knowledge mining mode.

**Key words:** huge amounts of academic literature resource; knowledge mining; knowledge discovery

图书馆海量学术文献资源知识管理已经从手工管理、计算机管理向网络化管理改进，目前图书馆领域的研究热点就是智能化、个性化管理。通过对读者的查找计划，比如想要做什么、对什么感兴趣等进行推理，为读者提供所需的知识。海量学术文献资源数据库中的知识挖掘主要是用一些复杂的统计分析及模型技术，在大量的数据和信息中筛选出不清楚、不被人们理解的信息，最终转化为可利用的信息。通过挖掘技术对数据库中的数据进行处理，重新获得有价值的信息资源[1]。

知识挖掘是通过数据挖掘完成的，即新知识＝

**作者简介：**方婷，馆员，研究方向为阅读推广、少儿亲子阅读。E-mail: whtsgft@yeah.net

析——以铁岭卫生职业学院为例[J].医学信息,2011,24(9):5994.
[4] 韩云峰,娄峰阁,葛杰,等.某医学院校大学生健康信息素养调查与培养策略研究[J].预防医学论坛,2016,22(8):565-567,573.
[5] 张士靖,杜建.健康信息素养应成为中国公众健康素养促进的关键点[J].医学信息学杂志,2010,31(2):45-49.
[6] 陆建玉,周菊芝,艾雨兵,等.宁波市某高职医学院校大学生健康信息素养认知及需求调查分析[J].中国农村卫生事业管理,2015,35(3):363-366.
[7] 钱勇.医药中职校学生信息素养的现状调查与分析——以南京市医药中等专业学校为例[D].南京:南京师范大学,2013.
[8] 封竹兵.高职医学生健康素养现状调查与研究[D].南京:东南大学,2016.
[9] 陆建玉,周菊芝,周国祥,等.大数据背景下医学生健康信息素养教育模式的研究与实践[J].中国农村卫生事业管理,2015,35(2):204-206.

（收稿日期：2017-03-29）
（修回日期：2017-04-12；编辑：魏民）

数据＋已有知识＋目标。在知道读者的需求后，通过数据分析，加上原来已有的知识进行计算，帮助读者得到他们之前不知道的知识。知识挖掘使人们找寻知识不再困难。现代图书馆可以根据人们对知识的需求程度，对原理、知识信息等进行重新整理、规划，提取出最主要的、对人们有用的且容易识别的信息。本文的研究目的在于通过图书馆馆藏的海量学术文献资源进行知识挖掘，从而提高馆藏资源的开发与利用效率。

## 1　利用知识挖掘技术的必要性

### 1.1　采集海量学术文献资源信息所必需

现在人们对图书的需求量越来越多，知识面越来越广，对知识以及阅读的需求量更加多样化。图书馆购买图书的费用降低，加大了采购的难度[2]。图书馆自动化功能可以产生很多对采购有用的数据，但是只有那些数据还是不能直接进行采购，需要应用知识挖掘技术，统计各种数据，比如寻找专家、查找光盘、调查书籍借还速度等[3]。知识挖掘技术帮助采购人员得到很多有价值的信息，是采购工作不可缺少的环节。知识挖掘技术在网络上的图书馆智能化系统可以帮助采购人员通过网上合作采购书刊，起到预测信息以及决策分析的作用。

### 1.2　检索海量学术文献资源信息所必需

信息检索方式从索引检索、打孔穿卡片检索已经发展到现在的计算机信息检索[4]。在大数据时代，人们需要信息检索更加完整、正确并且快速，而智能检索技术可以通过预测读者需求，快速提供检索结果[5]。知识挖掘技术通过复杂的数据分析，从海量信息中统计出对人们有用的信息，经过数据库进一步的处理，用机器学习的方法从其中得到有用信息。通过知识挖掘，特别是知识挖掘的分支——文本挖掘及 Web 挖掘技术得到文本中的有用信息，再把这些信息通过动态方式展现给读者，满足读者要求。同时，它还可以与计算机存储、组织、管理、使用的领域的本关理论、知识事实数据进行交换、联系、集合，这用户提供个性化服务。

## 2　知识挖掘的关键技术

### 2.1　联机分析挖掘技术

联机分析挖掘技术（OLAM）由联机分析处理（online analytical processing, OLAP）及数据挖掘（data mining, DM）组成，OLAP 和 DM 是通过数据仓库技术发展而来的信息分析和挖掘工具[6]，二者相辅相成。OLAP 通过多维数据进行分析，和数据仓库原有的数据取长补短。OLAP 分析主要是通过数据库对模型进行剖析、旋转等方式来得到有用的信息，给决策者提供帮助。多维结构是 OLAP 的重要组成部分，多维结构里面的维不仅仅指的是三维，还可以是超立方体和多立方体的一些结构。OLAP 有 2 个显著特征：一方面是在线性，就是能很快感应到读者的需求；另一方面是多维性，对统计分析、整理后的数据再进行多维分析。DM 数据挖掘分为 2 个步骤：建立模型和预测结果，通过对各种信息源尤其是最原始的数据进行挖掘。

### 2.2　Web 挖掘技术

现在网络在技术应用、规模复杂程度等方面都有突飞猛进的发展，包括多行业的内容信息、广泛的超链接信息还有 Web 自身页面的访问及使用信息。Web 挖掘指的是通过数据挖掘技术从 Web 文档以及 Web 活动中抽取出人们感兴趣的、对人们有帮助的信息。

Web 挖掘分为 3 种类型[7]：⑴Web 内容的挖掘。对 Web 自身页面的内容进行挖掘，主要通过对非结构化的数据，像文本、音频、视频以及图形图像等多种数据结合的信息进行挖掘。⑵Web 结构的挖掘。通过对 Web 页面之间的结构来进行挖掘，通过对 WWW 之间的组织结构以及连接间的关系来推理知识，例如 HITS 算法即属于对 Web 结构的挖掘。⑶Web 使用信息的挖掘。通过 Web 日志文件及相关数据，对用户查看 Web 时留下的记录进行挖掘，从而发现人们查找 Web 页面的规律及模式。

### 2.3　数据仓库技术

数据仓库是由面向主题的、集成的、稳定的、并且随时间变化而变化的数据组成的，对经营管理起决策性作用。一方面，数据仓库用来支持决策，对分析型数据进行处理，与现在数据库的操作模式不一样；另一方面，数据仓库对不同主题、不同来源的数据库进行重新组合。

数据仓库系统由 4 个方面组成：⑴数据源，是数据仓库系统的基本组成部分。⑵数据的存储与管理。是数据仓库系统的主要组成部分，对数据库的管理起到了关键性的作用。⑶OLAP 服务器。通过对有用的数据进行分析，按照多维模型结构进行重组，有利于朝更广、更深层次的方向发展，从而发现未来的走向。⑷前端工具，是由各种报表工具、查询工具、数据分析工具、数据挖掘工具等组成。其中数据分析工具主要是针对 OLAP 服务器的，报

表工具、数据挖掘工具是在数据仓库中使用。

### 2.4　信息可视化技术

信息可视化技术是信息科学非常重要的一部分，正向信息的知觉化、感知化方向发展。信息可视化应用领域非常广，由信息的抽取和描述、人工和机器的交换、数据的挖掘、制图学和成像学等组成[8]。信息可视化能把抽象的数据以及文字信息通过二维或者多维模式体现出来。在对信息进行可视化处理时，能够看出数据的属性，那些不容易发现的数据之间的关系也能清晰地体现出来，用户可获得更丰富的信息资源。

## 3　海量学术文献资源的知识挖掘模式

知识挖掘是从海量学术文献资源中挖掘出有用的信息，再对相同类型的数据元进行比较、处理及分类等，得出他们之间的某些关系，从而生成、获得更多的知识信息。

### 3.1　用户知识挖掘

因为读者有需求才有个性化的服务，只有足够了解读者的需求，才可能设计、构思出能够满足读者要求的更高层次的知识信息。

用户知识可以分为 3 类：可以满足用户需要的知识、本身出自用户的知识和与用户相关的知识，每种知识都有各自的特征。能满足人们要求的知识是指图书馆能够给用户提供的知识；本身出自用户的知识是指用户表达的知识信息，是在利用图书馆时向工作人员所反馈的信息；与用户相关的知识是指用户在图书馆时需要的信息。用户知识挖掘为图书馆的发展打下良好的基础，充分了解用户，把图书馆的管理水平和人们的知识需求相结合，就会形成满足人们需要的个性化服务。

知识挖掘非常显著的特征就是它能把一些人们没有表达清楚的信息挖掘出来，再通过这些信息点预测未来的发展走向。

知识挖掘通过对用户需要什么知识、曾经访问过的记录和某种行为，将这些知识再重新整理、分类、规划，形成一个用户知识的海洋，利用反馈出来的信息，来推测人们以后对知识的需求会往什么方向发展，根据这些推测，图书馆便能开发出符合人们要求的不同类型的知识。通过时间的不同、需要信息的种类不同，把知识种类进行归纳整理、分类划分，从而满足人们的信息需求。

### 3.2　学科知识挖掘

和用户知识挖掘相似，学科知识挖掘主要是通过挖掘技术对学科进行某种分析，得到某些不容易发现的知识，再形成面向学科的专业知识的海洋。可以观察不知道的那些知识他们之间的某种关系，从而进行整理。

学科知识挖掘由 5 个部分组成：⑴信息采集。知识挖掘的主要目的是研究和学科有关的信息。资源信息的采集大部分都是从简单到复杂，先整理出自身图书馆的信息资源，再慢慢向其他图书馆、网络机构等进行扩展。⑵信息过滤。指对这些信息进行筛选，再通过各种信息之间的关系进行归类整理。⑶信息汇总。把相同类型的学科知识信息进行整理，形成信息库。⑷信息提供。针对不同用户的不同需求提供信息。⑸反馈评价。知识挖掘非常重要的步骤。通过人们反馈的意见，对知识挖掘各方面进行改进。

## 4　海量学术文献资源知识挖掘创新模式

### 4.1　知识服务

知识服务是通过把一些知识进行重新整理、分类，组成新的知识，以满足不同用户对知识的不同需求。高校数字图书馆利用查找、整理信息的能力，来实现自身的价值。

图书馆提供知识服务，还需要有更加完善的管理机制，丰富馆藏资源，加强馆际合作，对各种信息整合、规范，提升完成知识服务的能力。

### 4.2　Lotus 知识发现

Lotus 是一种新的知识挖掘方式，主要通过知识仓库完成。前端的 Lotus 知识工作站是最基础的关于 Web 的知识信息解决，它融合了 Lotus 先进的门户技术、实时协作平台以及群组共同协作完成方案等核心的协作工具与服务，把应用、团队协作服务和个人服务融合在一起，通过把“人物、地点和事件”这三个不同的知识信息相融合，群组之间的相互协作，让工作人员通过 LotusNotes 客户端或浏览器进行访问，从而发现、使用和实现知识间的相互传递。后台的 LotusNotes 服务器可总结出“人物、地点和事件”三者的关系，从存储在企业的文档、数据库、电子邮件、Web 和 ERP（企业资源计划）等业务系统的信息库中得到自己想要的知识。

### 4.3　综合知识挖掘

双库协同知识发现（KDD&K）是数据库和知识库的综合知识挖掘系统，使知识之间相互组合以及知识增加，系统深度知识也是通过它来完成和实现深化。KDD&K 系统主要是将原来数据间的关系，

再通过分析、归纳、整理，从而发现更深层次的知识，即具有无穷尽的知识。KDD&K 系统使用很多种推理方法，发现的知识是有价值的并且人们能够理解的。因为数据和知识本来就是截然不同的 2 个概念，KDD&K 过程非常复杂并且会有很多方法和途径的，它和数据库、数据仓库、知识库的组织、以及人们对最终想要的知识类型息息相关，需要 KDD&K 必须具备很强的知识挖掘能力。

## 5 结语

知识挖掘的目的是通过最新的智能型工具找到隐藏在海量学术文献资源中的隐性知识，再形成专业的知识库和知识仓库。对图书馆馆藏的海量学术文献资源进行知识挖掘研究，有利于促进馆藏学术资源的有效开发与利用，从而进一步提升图书馆服务质量。

## 参考文献

[1] 郑立新.论图书馆员工隐性知识挖掘[J].现代情报,2010,30(3):29-31.

[2] 蔡皎洁,张玉峰.企业电子商务中客户知识挖掘模型研究[J].江苏商论,2012(8):52-55,83.

[3] 王峰,汪华方.数字图书馆信息检索技术的智能化发展趋势[J].现代情报,2008,28(11):93-95,99.

[4] 蔡皎洁.Web 环境下的语义挖掘模型研究[J].情报理论与实践,2015,38(5):121-124,111.

[5] 张春燕,覃海生.数据挖掘在多媒体资源库建设中的应用[J].无线互联科技,2015(14):138-139.

[6] 钟文一.大数据时代下的图书馆数据挖掘和情报分析研究——以中文发现系统为例[J].教育观察,2014,25(3):88-91.

[7] 庞观松,张黎莎,蒋盛益.跨语言智能学术搜索系统设计与实现[J].山东大学学报(工学版),2011,41(5):63-68.

[8] 涂寓,王志彦.浅谈资源发现系统在图书馆服务中的现状与趋势[J].图书情报论坛,2014(5):55-58.

（收稿日期：2017-03-13）

（修回日期：2017-04-11；编辑：魏民）