

# 基于数字图书馆检索技术的数据挖掘研究

王 预

(安徽财经大学 信息工程学院, 安徽 蚌埠 233041)

**摘 要:** 数字图书馆的研究与建设已成为国内外图书馆研究的重要课题。文中对数字图书馆概念、基本组成及作用进行了概述, 并对数据挖掘特点及其常用技术做了分析, 指出基于数字图书馆检索的数据挖掘及其实现方式。具体分析了数据挖掘在数字图书馆检索技术中的作用, 并预测了数据挖掘技术在图书馆领域的应用前景。从而为此技术的研究和推广起到了积极的作用。

**关键词:** 数字图书馆; 数据挖掘; 检索技术

中图分类号: TP391; G250.76

文献标识码: A

文章编号: 1673-629X(2006)11-0172-03

## Studies about Data Mining Based on Digital Library Retrieval Technology

WANG Yu

(College of Information Engineering, Anhui University of Finance and Economics, Bengbu 233041, China)

**Abstract:** The digital library research and the construction have become the important topic of domestic and foreign libraries research. This article presents the digital library concept, the basic composition and the function, and makes the analysis according to the data mining characteristic and its commonly used technology, pointed out based on the digital library retrieval data mining and the realization way. Specifically analyzed the data mining in the digital library retrieval technology function, and forecast the data mining technology in the library domain application prospect. So the theories can play more active role in researches and applications generally.

**Key words:** digital library; data mining; retrieval technology

### 1 数字图书馆的概念

数字图书馆是高新技术的产物, 涉及数字化、超大规模数据库、网络多媒体信息处理、信息压缩与传送、分布式处理、安全保密、可靠性、数据仓库与联机分析处理、信息抽取、数据挖掘、基于内容的检索、自然语言理解等各类技术。数字图书馆是一整套面向对象、分布式、与平台无关的数字化资源的集合<sup>[1]</sup>。广义而言, 数字图书馆包括所有数字形式的图书馆资源: 经过数字化转换的或以电子形式出版的资料, 新出版的或经过回溯性加工的资料; 包括期刊、参考工具书、专著、视频音频资料等; 各种文件格式等资源类型。它将分散于不同载体、不同地理位置的信息资源以数字化的形式贮存, 以网络化的方式互相连接, 提供及时利用, 实现资源共享, 其核心是数字化和网络化, 实质是形成有序的信息空间<sup>[2]</sup>; 它具有分布的、大规模的和有组织的数据库和知识库, 用户或用户团体可对系统内的数据仓库进行一致性访问, 获得自己所需的最终情报<sup>[1]</sup>。

### 2 数字图书馆的基本组成及作用

借助于数据挖掘技术, 数字图书馆用户能从浩瀚的数据仓库中及时准确地提取自己所需的信息。它具有可用于广域网服务的网络设备和通信条件; 一整套符合标准规范的数字图书馆赖以运作的软件系统, 主要分为信息的获取与创建、存储与管理、访问与查询、动态发布以及权限管理五大模块, 所起的作用是实现数字图书馆的维护管理和用户服务<sup>[3]</sup>。数字图书馆是传统图书馆在信息时代的发展, 它不但包含了传统图书馆的功能, 向社会公众提供相应的服务, 还融合了诸如博物馆、档案馆等信息资源的功能, 提供综合的公共信息访问服务。数字图书馆主要特征表现在“信息资源数字化, 信息传递网络化, 信息利用共享化, 信息提供知识化和信息实体虚拟化”。

### 3 数据挖掘特点及其常用技术

#### 3.1 数据挖掘的特点

数据挖掘是一种新的信息处理技术, 是指从数据集中自动抽取隐藏在数据中的有用信息的非平凡过程, 这些信息的表现形式为: 规则、概念、规律及模式等。它可帮助决策者分析历史数据及当前数据, 并从中发现隐藏的关系和模式, 预测未来可能发生的行为。数据挖掘的过程也叫

收稿日期: 2006-02-19

基金项目: 安徽省教育厅自然科学基金资助(2006KJ052B)

作者简介: 王 预(1965-), 女, 天津人, 副教授, 研究方向为情报学、信息管理。

知识发现的过程,它涉及面很广,涉及到数据库、人工智能、数理统计、可视化、并行计算等领域。其主要特点是对数据库中的大量数据进行抽取、转换、分析和其他模型化处理,并从中提取辅助决策的关键性数据<sup>[4]</sup>。它并不是用规范的数据库查询语言(如 SQL)进行查询,而是对查询的内容进行模式的总结和内在规律的搜索。传统的查询和报表处理只是得到事件发生的结果,并没有深入研究发生的原因,而数据挖掘则主要了解发生的原因,并且以一定的置信度对未来进行预测,为决策行为提供有利支持。

### 3.2 数据挖掘常用技术

目前对数据挖掘的研究主要集中在算法及其应用方面。机器学习、数理统计等方法数据挖掘进行知识学习的重要方法。统计方法应用于数据挖掘主要是进行数据评估;机器学习是人工智能的另一个分支,也称为归纳推理,它通过学习训练数据集,发现模型的参数,并找出数据中隐含的规则<sup>[5]</sup>。

1) 关联分析法。挖掘关联是通过搜索系统中的所有事物,从中找到出现条件概率较高的规律和模式。关联实际上就是数据对象之间相关性的确定,用关联找出所有能将一组数据项和另一组数据项相联系的规则,这种规则的建立并不是确定的关系,而是一个具有一定置信度的可能值,即事件发生的概率。关联分析法直观、易理解,它依据一定的可信度、支持度、期望可信度、作用度等建立相关规则。它是几种主要的数据挖掘方法之一,但对于关联度不高或相关性复杂的情况不太有效。

2) 人工神经网络(ANN)是数据挖掘中应用最广泛的技术。神经网络的数据挖掘方法是建立在自学习的数学模型基础之上,通过模仿人的神经系统来反复训练学习数据集,从分析的数据集中发现用于预测和分类的模式,可以进行趋势分析。神经网络对于复杂情况仍能得到精确的预测结果,而且可以处理类别和连续变量,但神经网络不适合处理高维变量,最大的缺点是不透明性,因为其无法解释结果是如何产生的,及其在推理过程中所用的规则。神经网络适合于结果比可理解性更重要的分类和预测的复杂情况,可用于聚类、分类和序列模式。

3) 决策树(DT)是一种树型结构的预测模型,其中树的非终端节点表示属性,叶节点表示所属的不同类别。首先,通过一批已知的训练数据建立一棵决策树。其次,利用建好的决策树,对数据进行预测。决策树的建立过程可以看成是数据规则的生成过程,决策树实现了数据规则的可视化,其输出结果也容易理解。决策树也可用于聚类、分类及序列模式。决策树一般产生直观、易理解的规则,而且分类不需太多计算时间,适于对记录分类或结果的预测,决策树方法精确度比较高,结果容易理解,效率也比较高,因而比较常用。

4) 遗传算法(GA)是一种基于生物进化理论的搜索优化技术,基本观点是“适者生存”原理,用于数据挖掘中则把任务表示为一种搜索问题,利用遗传算法强大的搜索能力

找到最优解。实际上遗传算法是模仿生物进化的过程,反复进行选择、交叉和突变等遗传操作,直至满足最优解。遗传算法可处理许多数据类型,同时可并行处理各种数据,常用于优化神经网络,解决其他技术难以解决的问题,但需要的参数太多,对许多问题编码困难,一般计算量大。

5) 联机分析处理(On-Line Analytical Processing, OLAP)主要通过多维的方式对数据进行分析、查询和报表。它不同于传统的联机事物处理(On-Line Transaction Processing, OLTP)应用。OLTP 应用主要是用来完成用户的事务处理,如民航订票系统、银行储蓄系统等,通常要进行大量的更新操作,同时对响应时间要求较高。而 OLAP 应用主要是对用户当前及历史数据进行分析,辅助领导决策,典型应用有对银行信用卡风险的分析与预测、公司市场营销策略的制定等,主要是进行大量的查询操作,对时间的要求不太严格。

6) 数据可视化(Data Visualization)系统数据量很大,很容易使分析人员面对数据不知所措,数据挖掘的可视化工具可以通过富有成效的探索起点并按恰当的隐喻来表示数据,为数据分析人员提供很好的帮助。数据可视化工具大大扩展了传统商业图形的能力,支持多维数据的可视化,从而提供了多方向同时进行数据分析的图形方法。有些工具甚至提供动画能力,使用户可以“飞越”数据,观看到数据不同层次的细节。

## 4 基于数字图书馆检索的数据挖掘及其实现方式

由于信息的多样性,对检索方面也提出了新的要求,原有的参数描述方法以及单纯的对参数进行索引的方法,已无法满足用户的查询需求。目前许多图书馆已实现公共查询系统(OPAC)进行联网检索。但是,互联网上的绝大多数信息仍然是无序和混乱的。在网上简单的交互式对话、原始的科技资料、一般的会议记录及书刊论文等都不分等级地储存在一起。如何对网上这类信息进行删除、过滤、分级、组织,方便用户使用,将是数字图书馆建设的重要任务。例如,院校图书馆员应用 WEB 挖掘技术为本院校不同学科中的不同研究课题从 WWW 中检索相关信息。该技术可以自动地检索信息,并把信息按照课题领域分类,使之更容易访问。图书馆员可以通过为不同的课题领域建立一组特征,并以这些特征为基础进行检索和分类,从而保证得到的信息可靠且具有权威性。

对文本文件进行检索的技术正在向全文检索方向发展,新型全文检索主要有 3 种实现方式<sup>[6]</sup>。

(1) 采取用自由指定的关键字字符,直接与全文文本的一次数据高速对照检索。

(2) 对文本内容中每个词进行位置扫描后排序,然后建立以每个词(字)的离散码为表目的倒排文件。

(3) 采用 Hypertext 模型,建立全文数据库。

此外,如何为多媒体信息建立索引和以声音、图像为

基本内容进行查询的技术也得到广泛关注。

## 5 数据挖掘在数字图书馆检索技术中的作用

### 5.1 成为未来图书馆的信息中心和枢纽,为分析和挖掘信息提供良好环境

数据挖掘在数字图书馆信息检索中的应用是传统图书馆在信息时代的发展,不但包含了传统图书馆的功能,还提供综合信息访问服务,数字图书馆将成为未来图书馆的信息中心和枢纽。网络时代的数字图书馆咨询需求不再局限于简单层次的信息查询与反馈,而是转向广阔的信息源,要求咨询人员提供综合度高、附加值大的信息产品。数据挖掘作为新型的信息架构,既含有图书馆历史信息,也含有当前信息,同时还集成有外部数据,为咨询馆员提供了广阔的查询数据源。同时,数据挖掘为分析和挖掘信息提供了良好的数据环境,利用 OLAP 和信息挖掘工具,一方面咨询馆员可以从海量数据中分析出事物之间的关联,挖掘出隐藏其中的规律信息,形成满足用户需求的深层次信息产品。另一方面,还可以根据用户的历史咨询记录,分析其研究方向和兴趣所在,实现主动的个性化信息服务<sup>[6]</sup>。

### 5.2 数字图书馆是人工智能系统

研究人员目前正努力研究基于 WEB 内容的数据挖掘,开发智能化的信息检索工具。基于代理的检索方法正是这种智能化的信息检索工具,它是一个人工智能系统。首先,它可以代表某一特定用户,自动地或半自动地发现和组织基于 WEB 的信息,可根据用户基本情况,自动检索出感兴趣的信息,并组织 and 翻译好这些信息。其次,利用数据挖掘系统提供的 OLAP 工具可以对集成数据进行多维分析比较,对决策假设进行审查和验证,提高决策的可靠性和可行性,达到合理利用有限资金、优化图书馆资源配置的目的。第三,数据挖掘工具可以从数字图书馆历史数据中找出潜在模式,并在模式基础上自动进行预测,这对启发数字图书馆决策者的创新思维、应对信息化社会的挑战具有重大意义。

### 5.3 为数据挖掘的应用奠定较为完备的基础

国内图书馆系统经过多年的自动化建设,已经具备相当的物质条件和人才储备,并积累了大量数据,为数据挖掘的应用奠定了物质基础,国家也给予高度的重视并提供大量经济和政策上的支持。数据挖掘技术虽然在数字图书馆领域的应用还处于起步阶段,但基于其在数据的组织、分析和知识发现等方面的巨大优势和对信息的深层挖掘能力,将日益显示出强大的发展潜力和广阔的应用前景。

### 5.4 为数字图书馆提供了宝贵经验

数据挖掘经过多年的发展,已经形成相对成熟的技术体系,特别是在数据挖掘设计、数据抽取以及联机分析处理技术等方面都取得了令人满意的进展,为数据挖掘的应用奠定了技术基础。数据挖掘技术在发达国家的电信、零

售、制造、金融等领域已有较强的应用,并取得了巨大回报,这些成功应用的例子为数字图书馆提供了可供借鉴的宝贵经验。

### 5.5 为数字图书馆的建设提供了关键技术

鉴于数据挖掘技术在数据的组织与分析、数据挖掘、知识发现等方面存在的巨大潜力,因此数据挖掘可以为数字图书馆的建设提供关键技术。例如:元数据的界定和自动抽取,海量信息的有效存储和利用,超大规模分布式数据库的快速存取及分布式资源库互操作性的实现等,都能够借助和参考数据挖掘技术。例如,正在实施的国家 863 计划中国数字图书馆工程对数据挖掘技术在数字图书馆建设中实际应用进行了有益性尝试,工程的一个重要部分就包括建立分布式存储、集中式管理的大型数据挖掘,并对其智能化进行管理与挖掘,再通过个性化和智能化的人机交互界面实现网络信息服务<sup>[7]</sup>。

## 6 数据挖掘技术在数字图书馆领域的前景展望

### 6.1 能够为数字图书馆的决策和管理提供强有力的保障

管理水平低下是影响中国图书馆事业发展的重要因素之一,管理水平的提高很大程度上取决于决策的科学与否。传统的图书馆决策方式大多依靠经验进行决策,存在主观、片面、盲目等诸多问题,无法适应时代发展的要求。采用数据挖掘技术将能够为领导层的科学决策提供强有力的保障。数据挖掘能将涉及图书馆这一信息系统的各种内部数据和外部信息汇集起来,经过处理和转换,形成集中统一、随时可用的决策信息,防止因信息不足造成的错误决策。因此,运用数据挖掘技术实现海量数据的存储、咨询和利用,支持图书馆各层次的科学决策服务,实现高效的行业信息合作模式,是信息化带来的外部压力与图书馆内部发展机制的共同需求。

### 6.2 支持数字图书馆的业务工作,保障信息资源体系的科学性和合理性

数据挖掘技术对图书馆业务工作的支持主要体现在信息采集和信息咨询两方面,作为信息链的第一个关键环节,信息采集是整个图书馆系统高效运转的基础。随着出版物数量日益增多、载体日益丰富,图书馆信息结构、读者需求与资金利用的平衡问题很不易把握,也令采购工作的决策变得更加复杂。数据挖掘技术可以在分析内部的历史采购数据、读者数据、流通数据、反馈信息以及来自外部的各种学科信息的基础上深入了解学科的走势和读者的需求,帮助采购人员确定采购重点,保障数字图书馆信息资源体系的科学性和合理性<sup>[8]</sup>。

### 6.3 制定国家数字图书馆发展战略规划,完善数字图书馆相关的保密、版权等法律

在信息化社会中,图书馆更好的生存与发展和先进技术的运用是密不可分的。数字图书馆是一项崭新的事物,其研究和建设水平将直接关系到中国图书馆未来信息时

(下转第 178 页)

一种特殊的形式或一套表达方式,如关联规则、分类规则或分类树、回归结构和聚类集等。

### (3) 数据挖掘、结果分析表述和挖掘应用。

此阶段运用使用兴趣度度量,并与数据挖掘模块交互,以便将搜索聚焦在有趣的模式上。它可能使用兴趣度阈值过滤发现的模式。运用统计学和关联规则等方法,把挖掘分析的结果放入一个个性化数据库,当学习者下次进入系统时,系统就可根据个性化数据库提供给其符合学习需求的页面。

## 4 结束语

网络教学平台的关键是针对用户的个性特征信息,通过系统的分析和判断,给予不同的学习环境和学习内容的呈现,通过运用数据挖掘技术可以从用户数据库及用户学习行为记录中挖掘出用户对知识点的理解程度,从而实现在学习过程中对用户的学习进行记录、指导、反馈,对用户

选择的学习策略给予支持,大幅提高《大学物理》网络教学平台的教学效果,使个性化教学真正得以实现。

### 参考文献:

- [1] 林君芬,余胜泉. 关于我国网络课程现状与问题的思考[J]. 教育技术通讯,2001(1):55-59.
- [2] 梁林梅,焦建利. 我国网络课程现状的调查分析与反思[J]. 开放教育研究,2002(6):13-16.
- [3] 刘莉. 远程学习者研究现状及发展趋势——远程教育专家访谈录[J]. 中国远程教育,2003(5):7-12.
- [4] 黄萍. 高校学生网络自主学习行为的调查研究[J]. 开放教育研究,2004(6):77-80.
- [5] 舒蓓,申瑞民,王加俊. 个性化的远程学习模型[J]. 计算机工程与应用,2001(9):90-92.
- [6] 康晓东. 基于数据仓库的数据挖掘技术[M]. 北京:机械工业出版社,2004.

(上接第 155 页)

法的特有功能,今后可以继续把新的数据挖掘算法引入到入侵检测中,改善检测的准确性、可靠性。

### 参考文献:

- [1] Lee Wenke, Stolfo S J, Mok K W. A Data Mining Framework for Building Intrusion Detection Models[C]//Proceedings of the 1999 IEEE Symposium on Security and Privacy. Los Alamitos, CA: IEEE Computer Society Press, 1999:120-132.
- [2] Wong M L, Leung K S. An Efficient Data Mining Method for Learning Bayesian Networks Using an Evolutionary Algorithm - Based Hybrid Approach[J]. IEEE Transactions on Evolu-

tionary Computation, 2004, 8(4):378-404.

- [3] 张琨,徐永红,王珩,等. 用于入侵检测的贝叶斯网络[J]. 小型微型计算机系统, 2003, 24(5):913-915.
- [4] Kruegel C, Mutz D, Robertson W, et al. Bayesian Event Classification for Intrusion Detection[C]//Proceedings 19th Annual Computer Security Applications Conference. Los Alamitos, CA: IEEE Computer Society Press, 2003:14-23.
- [5] 白耀辉,陈明,王举群. 利用朴素贝叶斯方法实现异常检测[J]. 计算机工程与应用, 2005(34):131-132.
- [6] 牛建强,曹元大,阎惠. 基于数据挖掘的 CIDE 协同交换[J]. 计算机工程, 2003, 29(14):35-36.

(上接第 174 页)

代的地位与影响。信息技术极大地推动了图书馆的现代化进程,同时也带来了信息的爆炸式增长。在知识经济时代,解决好海量信息的存储、检索、开发与利用,是关系到图书馆未来的生存与发展的重大问题。因此必须制定国家数字图书馆发展战略,做好整体建设规划。同时还要制定和完善数字图书馆相关的保密、版权等方面的法律。

## 7 结束语

数据挖掘技术及其应用是目前国际上的一个研究热点,在数字图书馆领域,面对大量的信息,用数据挖掘技术找出数字图书馆用户感兴趣的信息加以组织利用,加强客户关系的管理,提高满意度。基于数字图书馆建设与应用数据挖掘的有利时期已经到来,优质的网络信息服务事业前景不可限量。综合应用数据挖掘技术和人工智能技术,获取用户知识、文献知识等各类知识,将是实现知识检索和知识管理发展的必经之路。

### 参考文献:

- [1] Poe V. Building a Data Warehouse for Decision Support[M]. [s.l.]: Prentice PTR, Prentice-Hall Inc, 1996.
- [2] 赵洗尘. 数字图书馆及其建设[J]. 现代图书情报技术, 1999(1): 28-31.
- [3] 刘霞. 关于数字图书馆建设的几个问题[J]. 图书情报知识, 1998(1):30-32.
- [4] 王珊. 数据仓库技术和联机分析处理[M]. 北京:科学出版社, 1998.
- [5] 刘海虹,刘伯莹. 数据挖掘技术[J]. 丹东纺专学报, 2001(1):15-18.
- [6] 郝先臣. 数据挖掘工具和应用中的问题[J]. 东北大学学报: 自然科学版, 2001(2):183-187.
- [7] 卢增祥. Bookmark - 智能化网络信息服务系统[J]. 高技术通讯, 1999(6):30-32.
- [8] 郑巧英,杨宗英. 图书馆自动化新论——信息管理自动化[M]. 上海:上海交通大学出版社, 1998.