浅谈数据挖掘及其在图书馆的应用

何 少 卓

(广西民族学院图书馆,广西 南宁 530006)

[关键词]数据挖掘;图书馆;应用

[摘 要]本文介绍了数据挖掘的定义、功能及方法,阐述了数据挖掘应用于图书馆工作的意义,列举了数据挖掘技术在图书馆业务工作及数字图书馆建设中的应用。

[中图分类号]G252.7 [文献标识码]B [文章编号]1005 - 6041(2004)03 - 0052 - 03

我们已经生活在一个网络化时代,通信、计算机和网络技术正改变着整个人类和社会。随着计算机硬件和软件的飞速发展,尤其是数据库技术与应用的日益普及,被收集并存储在众多数据库中且正在快速增长的庞大数据,已远远超过人类的处理和分析理解能力(在不借助功能强大的工具情况下),而成为"数据坟墓",即这些数据极少被访问,结果许多重要的决策不是基一过些基础数据而是依赖决策者的直觉而制定的,因为这些决策的制定者没有合适的工具帮助其从数据中抽取出所需的信息知识。因此,致力于此的数据挖掘作为当前数据库研究、开发和应用最活跃的分支之一,引起学术界和产业界的广泛关注。作为信息管理与服务的主要机构之一的图书馆如何运用这一新技术挖掘丰富的馆藏数据资源,为读者,尤其是为部门决策者的决策提供可靠的决策依据。本文试图在这方面作一尝试,探讨数据挖掘在图书馆信息管理与服务方面的应用。

1 数据挖掘的定义、功能及方法

数据挖掘概念的定义有若干表述,一个被普遍采用的定义是:数据挖掘,又称为数据库中知识发现(Knowledge Discovery from Database,简称 KDD),它是一个从大量数据中抽取挖掘未知的、有价值的模式或规律等知识的复杂过程。严格讲,数据挖掘是知识发现过程的一个基本步骤,是最核心的部分。整个知识发现过程由若干挖掘步骤组成,其中包括:①数据清洗,就是清除数据噪声和与挖掘主题明显无关的数据;②数据集成,就是将来自多数据源中的相关数据组合到一起;③数据转换,就是将数据转换为易于进行数据挖掘的数据储储形式;④数据挖掘,就是利用智能方法挖掘数据模式或规律知识;⑤模式评估;⑥知识表示。在通常情况下,许多人把数据挖掘和知识发现广泛地认为是同一概念,一般在科研领域中称为知识发现,而在工程领域则称为数据挖掘。

数据挖掘不仅能对过去的数据进行查询和遍历,并且能够对将来的趋势和行为进行预测并自动探测以前未发现的模式,从而很好地支持人们的决策。被挖掘出来的信息,能够用于信息管理,查询处理,决策支持,过程控制以及许多其它应用。具体地说,挖掘功能包括概念描述、关联分析、分类与预测、聚类分析、异类分析、演分分析、探索性数据分析等。

数据挖掘方法总体上分为两大类:基于数据保持类方法和基于模式提取类方法。基于数据保持类方法,就是要保存过去的数据,以便与新输入的新数据进行匹配,从而进行预测建模和分析评价;而基于模式提取类方法则不需要保存过去的数据,一旦提取模式后,就可以把过去的数据移去。现代数据挖掘方法主要依靠模式提取技术,同时,为了改善和提

[收稿时间]2004 - 05 - 15

高数据挖掘的功能、性能和效率,发展趋势是综合采用多种方法和技术,例如统计方法、关联分析和序列模式算法、神经网络分类方法、决策树、遗传算法、贝叶斯信念网络、模糊集、粗糙集方法、可视化技术等等。

数据挖掘是近年新兴的计算技术与方法,它在科学发现、商业零售以及信用管理、医学等储领域已得到广泛应用,并显示出巨大的威力。在图书情报信息处理方面,数据挖掘同样具有非同寻常的意义和发展潜力。

2 数据挖掘技术应用于图书馆信息管理与服务中的意义

2.1 为图书馆工作提供技术支持和决策管理支持

图书馆信息管理与服务的发展大致经历了三个阶段:第一是文献信息以藏为主,文献利用较为消极的馆藏模式阶段;第二是文献信息注重有效服务,馆藏文献服务与网上信息服务相结合的服务模式阶段;第三是文献信息注重经济效益,着重开拓为经济服务的新信息资源并以网上服务为主的商业模式阶段。三个阶段的发展历程表明,信息管理与服务在横向上朝着市场化发展,在纵向上朝着网络化进深。市场化的挑战和网络化的复杂都对图书馆的信息管理与服务提出了新的要求。首先,图书馆要处理和提供的信息更多、更新、更广泛、更复杂。为了避免陷入"数据丰富,但信息贫乏"的局面,图书馆有必要增强对信息的处理能力以及对信息资源的组织能力,尤其是地海量信息深层次的开发,提取表面上庞杂无序的内在联系供读者使用。其次,个性化主动信息服务将是未来信息服务的主流模式,它实现的是"信息找人,按需服务"。而其实现途径就是通过对用户的信息需要、兴趣爱好和访问历史的收集分析,建立用户模型,并将用户模型应用于网上信息的过滤和排序,从而指导用户的浏览过程和信息检索,或向用户主动推送服务。而这正是数据挖掘工具的强项。第三,图书馆日积月累产生的大量统计数据和表单,如果没有一个强有力的数据采集和处理工具介入,往往会变成"数据坟墓",失去其对图书馆精力的指导作用,而数据挖掘就是这样一种新兴的技术,可以为图书馆工作提供技术支持和决策管理支持。

2.2 优化数字图书馆的信息内容

数字图书馆是一种数据信息系统,这一系统不但拥有内容丰富,形式多样的数字化信息资源,而且依赖于现代高新技术所支持,通过采用现代高新技术,有序地组织资源,高效地满足用户的需求。目前,数字图书馆的信息内容包括大量的数字化馆藏、种类繁多的数据库、全文 WEB 资源链接以及互联网上的大量信息。这些大量的数据,只有通过组织、分析和挖掘,找出数据背后真正有价值的信息,才是用户实际需要的。这也正是数据挖掘技术所要解决的问题。采用数据挖掘技术,将其用于数字图书馆的信息发现和信息提供的全过程。从而向用户提供更优化的信息服务,并满足用户的个性化需求。因此,数据挖掘在数字图书馆信息最优化建设、信息自动化处理、信息服务质量的提升和义务拓展等方面具有广阔的应用空间;在数字图书馆向自动化、网络化、智能化方向发展过程中将一展神通。

3 数据挖掘在图书馆业务工作中的应用

3.1 指导采访工作

采访是图书馆各项业务工作的第一个环节,是图书馆藏书建设和文献资源布局的首要内容。传统图书馆信息采集多由专门采访人员独自确定,或采纳学科专家的意见,不可避免地带有极大的主观性以及个人喜好。同时,图书馆每年的文献购置费是有限的,各门学科之间如何分配,各种文献载体形式如何均衡才能使这些经费最好的发挥效益,这是一件令人头痛的事。另外,图书馆内每天产生大量可以对采访工作产生指导作用的数据,如自动化系统的流通数据,图书馆的历史采购数据,查询系统的各种查询数据等等。如何从这些大量数据中分析、统计出有用信息并非易事。传统做法只能做些模糊分析与评价,而数据挖掘技术的应用将使这些问题迎刃而解。例如,运用分类分析技术对流通记录,检索请求进行分析,按类统计文献拒借集和频繁借阅集,并以此分析出文献的利用率,及时补充短缺的文献,剔除过时的文献;运用关联分析技术,对用户每次借阅的文献进行关联分析,发现各类文献间的关联规则或比例关系,为各学科文献的采访工作提供科学、合理的分析报告和预测报告,提供必要的决策支持。另外,利用数据挖掘技术对 Internet 上无序的、非结构的数据进行采集分类,使图书馆的信息资源更加丰富。

3.2 加强书库管理

书库的频繁倒架以及图书的残破率、丢失率最高都是较常遇到的问题,如何对之防微杜渐也是值得挖掘的一个方面。通过对历年借阅数据的相关分析,相应的增长幅度较大的图书种类在上架的时候

应根据预测的趋势预留架位;通过对注销数据的分类分析统计及与样本库比较以确认若丢失率超过一定比例的原因出在哪些方面,给出一个在制度上或人员上加强管理的建议。对于那些借阅频率较大且连续续借的书目,应以量化的方式反馈给采访部门以加重采购的力度;对罚赔款数据的挖掘则可提供对诸如特定书目的借阅期限和人员限制等的建设性建议,以提高服务质量。

4 数据挖掘在数字图书馆建设中的应用

数字图书馆建设,就是为了解决网络环境下数字化信息的组织、查询与服务问题。目前的数字图书馆是大量的数字化馆藏、数据库远程访问、全文Web资源链接和文档提交的结合体。如何更好地组织数字图书馆中的数字化信息?如何提供有效服务?这两个问题实际上也是数字图书馆的信息发现和信息提供的问题,是当前数字图书馆研究的难点之一。面对"被数据淹没,却饥饿于知识"的挑战,基于人工智能的数据开采和数据挖掘技术应运而生,并被广泛应用。

4.1 在信息发现中的应用

一是Web数据开采技术,它是针对Internet上信息的获取困难而发展起来的,其实现过程包括信息的采集、文档的识别与分类等。具体包括两种实现方式:网络智能体和智能信息捕捉器。二是多语种信息发现。它是针对全球化信息遇到的地理和语言的障碍而开发的研究项目。目前主要有几个方面的研究:多语种电子文档获取,集成机器翻译和多语种信息检索系统。三是跨学科协同检索,它的功能是可以向多个甚至几十个数据库并发请求,同时与OPAC、馆际互借、文档提交和电子资源相连,而提供给用户的是统一检索界面,并跨非学科查询返回统一结果。

4.2 在信息提供中的应用

一是在个性化服务中的应用,所谓个性化服务就是针对用户的特定需求主动地向用户提供经过集成的相对完整的信息集合或知识集合。个性化服务的形式有三种:①按照特定用户请求,为用户提供定制的Web页面,信息频道或信息栏目,实施查询代理服务;②按照特定主题,指引文献源或提供文献全文,实施个性化文献性信息服务。③按照特定主题,提供相对完整的方案知识,实施个性化决策支持服务。实现个性化服务,必须要经历以下的过程:构建个性化用户动态需求模型;搜索、挖掘,针对特定需求的相关信息;按照特定主题,将搜索、挖掘到的信息进行过滤、加工和组合,整合成要对完整的信息集合,并以在线或离线形式,主动发送到用户或服务代理,实现信息支持;按照特定主题,融合、激活相对完整的信息集合,产生新的方案知识,并以在线或离线形式,主动发送到用户,实现决策支持。在Web数据挖掘技术中,Web使用记录的挖掘就是实现这个过程的最好方法。二是信息智能推,就是根据用户输入的关键词,通过机器学习,可以识别和预测用户的兴趣和偏好,从而有针对性、及时地向用户主动推送相关知识和最新信息。推送的形式可采用频道式推送、邮件推送、网页式推送或专用式推送。三是互动式服务,其强调建立一个统一信息提供平台,让有着不同类型信息和技术的不同团体实现信息共享,并在此基础上,通过聚类和计算产生新的信息。

综上所述,数据挖掘技术是未来信息检索的主要技术。数据挖掘技术的发展为图书馆信息管理与服务水平的不断提升提供了技术支持。数据挖掘技术方兴未艾,数据挖掘技术在图书馆各项工作中的应用更是大有作为。

[参考文献]

[1]王艳 数据挖掘在数字图书馆中的应用[J]·现代图书情报技术,2002,(5).

[2]朱晓华 浅析数据挖掘技术在图书馆自动化中的应用[J] 图书馆学研究,2002,(5.)

[3]李朝葵,凌云 数据挖掘及其在图书馆中的应用[J] 情报杂志,2002,(6).

[4]张立厚,等,未来图书馆与知识发现[J] 图书馆论坛,2002,(4).

[5]朱明 数据挖掘[M]·合肥:中国科学技术大学出版社,2002,(5).

[6]韩惠琴,等 知识发现在数字图书馆中的应用[J]·大学图书馆学报,2001,(1).