文章编号:1004-1478(2012)03-0045-04

数据挖掘技术在高校图书馆信息管理系统中的应用

周建华

(郑州轻工业学院 图书馆,河南 郑州 450002)

摘要: 针对目前图书馆对读者信息的获取停留在比较浅显的层面,缺乏深层次的信息加工和数据的综合分析等问题,基于数据挖掘技术,通过数据清理、数据整合,并加上规约算法,对图书馆信息管理的数据进行挖掘和预测. 实践表明,该方法可实现对读者的借阅行为、图书需求和阅读兴趣等信息的挖掘和预测,以便调整和提高图书管理策略.

关键词: 数据挖掘; 图书馆信息管理系统; 关联规则; 借阅服务

中图分类号: TP391 **文献标志码:** A

Application of data mining technology in university library information management system

ZHOU Jian-hua

(*Library, Zhengzhou Univ. of Light Ind., Zhengzhou* 450002, *China*)

Abstract: Aiming at the problem that readers information extract stays at a realtively simple level in library, the management system lacks deep-seated information processing and data comprehensive analysis. Based on the data mining technology, through data cleaning, data integration and induction algorithm, the data mining and prediction of library information management were carried out. The practice showed that the method could mine and predict the information of the readers borrowing behaviors, demand for books and reading interest information. It could improve the library management strategy.

Key words: data mining; library information management system; association rule; borrowing service

0 引言

随着网络的普及和电子通信技术的进一步发展,尤其是3G技术的广泛应用,网络博客、手机报、手机上网、电子阅读器等现代化、多元化的信息阅读渠道的扩展,高校大学生成为了新工具、新技术的领跑者和推广者,因此高校图书馆面临信息市场新的挑战. 传统的图书借阅、图书采购等图书信息管理模式已不能满足目前信息化管理的需求,因此,必须改变图书馆的工作模式,完成"以馆藏为中心"向"以用户为中心"的巨大转变[1]. 数字图书馆人性化服务要求及时掌握读者阅读习惯、偏好、特点及用户特定的需求,通过对用户阅读习惯的分析进而主动地向用户提供其可能需要的信息服务. 但

收稿日期:2012-04-17

基金项目:河南省社科联项目(SKL—2011—626)

作者简介:周建华(1971—),女,河南省方城县人,郑州轻工业学院馆员,主要研究方向为数据挖掘.

是目前图书馆对读者信息的获取多停留在比较浅显的层面,缺乏深层次的信息加工和数据的综合分析,尤其是对数据隐含的关联缺少深层次的归纳与挖掘[2]. 本文拟基于数据挖掘技术,利用关联规则实现图书借阅数据的自主分析,对读者的借阅行为、图书需求和阅读兴趣等进行信息挖掘和预测,以便图书馆调整管理策略,进一步提高服务质量和服务水平.

1 数据挖掘基本原理

数据挖掘技术又称数据库中的知识发现[3],是从大量、随机、有噪声的应用数据中,提取隐含和潜在信息的过程. 数据关联是反映事件之间依赖或相互联系的知识,关联规则是通过对数据集的分析处理,寻找各数据项之间有趣的关联或相关联系,从而发现事物内在客观规律的技术. 该技术于 1993 年由 R. Grawal 等人提出[4],基本原理为: 商业领域中,给定一段时间内交易数据集合 $T=\{t_1,t_2,\cdots,t_m\}$,其中 $t_i(1\leqslant i\leqslant m)$ 是每次交易的数据记录. 设 P,Q 是任一交易记录 t_i 中可能出现的数据项. 若在一个交易记录中既含有 P,又含有 Q,即存在着数据项 P 对数据项 Q 的关联. 通常使用支持度 S 和可信度 C 这 2 个技术指标表示关联的强度[5]. 支持度 S 表示关联规则 $P\rightarrow Q$ 在交易集合 T 中出现的频度,即数据项 P 和 Q 的数据记录在整个交易集合 T 中出现的百分比(用 $s\%$ 表示);可信度 C 表示该规则在整个交易集合 T 中出现的必然程度,其值为同时存在数据项 P 和 Q 的记录个数与交易集合 T 中包含数据项 P 的数量比(用 $c\%$ 表示). 实际应用中一般人为设定最小支持度和最小可信度,以排除其他因素的干扰.

数据挖掘系统结构如图 1 所示. 首先对于数据库和数据仓库中的原始数据进行数据清理,去除冗余和错误数据,提取感兴趣的数据,形成用于数据挖掘的数据库,然后在知识库的支持下进行基于关联规则的数据挖掘和模式评估,最终通过用户界面呈现出挖掘的结果.

2 图书馆信息管理中的数据挖掘

2.1 数据仓库的建立

数据仓库和数据挖掘关系紧密,数据仓库是一种数据存储和数据组织,负责提供数据源. 数据仓库是对原有数据库中的数据进一步抽取、加工与集成,随着新数据内容的增加不断进行更新. 本文使用 ETL 工具进行数据仓库后台数据的抽取、清洗、转换和集成. 数据仓库中用粒度表示不同的综合级别,粒度越大表示细节程度越低,综合程度越高[6],因而挖掘效率较高,但是呈现细节数据较困难. 因此本文使用磁盘阵列存放 3 个粒度的数据,分别为当前细节级、轻度综合级和高度综合级.

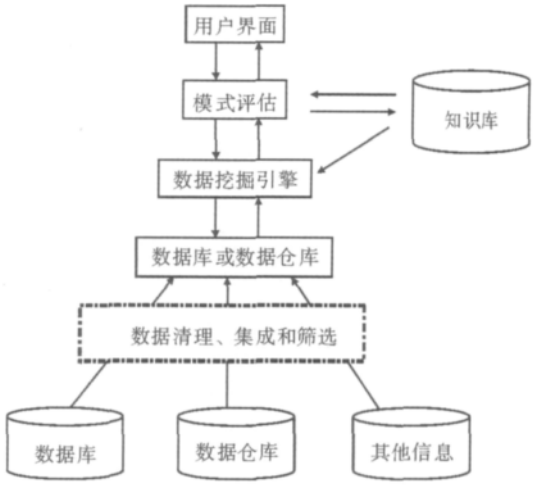

图 1 数据挖掘系统结构

2.2 数据挖掘的关键技术

本文所采用的数据均来自郑州轻工业学院图书馆的数据库中的原始资料,原始数据包括读者库表、书目库表以及流通日志表. 读者库(借书证号、读者级别、级别代码和单位等 24 个属性)有 16 000 余条记录;书目库表(主键码、题名、库键码、索书号等 28 个属性)有大约 180 000 条;流通日志表(主键码、处理时间、读者条码、条形码、登入号、操作类型等 11 属性)记录了图书馆自 1999 年引进管理系统以来的所有借还书记录,共计 1 200 多万条.

2.2.1 数据清理 在读者数据和图书数据处理过程中有一些不完整、不一致和有噪声的数据,因此需要对数据进行清理. 数据清理主要包括缺失值的处理和使用数据光滑技术. 对于“性别”、“续借”等缺失值使用 unknown 替换;对于“年龄”等缺失值使用 ageave 平均值进行替换;对于字段为空的记录,需要手工编写 SQL 脚本来删除;对于部分图书的分类号是中文字符,需要在挖掘程序中进行处理,及时丢弃这些坏数据.

2.2.2　数据整合　导入后的数据分散在各个数据库里,需要整合这些数据得到读者借阅记录的事务数据库. 新建如下数据库:

```
Creat table readertransaction //读者借阅数据库
Class 1 varchar2//年龄层次
Class 2 varchar2//文化层度
Class 3 varchar2//职称
Class 4 varchar2//职业
CallNo varchar//分类号
```

将 B_reader(读者库),B_log(读者流通事务库),B_holding(馆藏书目库),B_bibilos(书目库)中分散的各数据字段导入 readertransaction 中,生成相应的读者借阅事务数据库,可采用以下 SQL 脚本命令:

```
Insert into readertransaction( Class 1,Class 2 Class 3,Class 4,callNo)
Select B_reader. class4,B_reader. class3,B_reader. class1,B_reader. class2,B_biblios. callNo
from B_reader, B_log,B_holding,B_biblios
where B_reader. rdr_RECNO = B_log. data1and B_log. data3 = B_hoding. barNo and B_holding. bib_recoNo = B_biblios. bib_recNo
```

2.2.3　数据规约算法　数据仓库在运行一段时间后,数据仓库中的数据量急剧增加,若不使用规约算法,直接进行挖掘则存在以下 2 个问题: 其一,表的每个字段都占有较大的空间,内存占用率较大,增加了导入内存的时间开销; 其二,大部分单项是汉字字符串,生成候选序列时时间与空间开销会增加. 因此,考虑时间和空间效率,将每个事务记录压缩为长度为 6 的字符串,其中每个字符都是单个的小写字母. 系统在读取 4 个属性配置文件时,读到的属性值按照顺序赋予唯一的字母. 比如程序读到“中专”,把“a”赋给“中专”,读到“本科”,把“b”赋给“本科”,以此类推. 当所有的属性值都被读取,程序便建立了一个属性值到字母字符的一一映射表,随后将读者借阅事务数据库压缩成文件,这样减少了预处理文件的大小,提高了效率. 实际应用中,数据库在未压缩之前为 127 M,压缩后数据的大小为 11.6 M,大大节省了数据占用的内存资源.

3　结果分析

表 1 为郑州轻工业学院计算机与通信工程学院在校本科生在 2011 年度部分借阅图书的信息.

表 1　4 本考研相关书目和借阅人数

序号	书目	总借阅数	计算机与通信工程学院借阅人数
1	《风雨考研路》(桑磊主编)	214	89
2	《2011 考研英语作文周计划》(王长喜主编)	143	52
3	《数据结构习题集》(严蔚敏主编)	98	62
4	《操作系统典型题型精解精练》(杜松主编)	92	67

对表 1 数据使用本文的数据关联的挖掘方法,计算可知 $S(1,2)=0.13$,$S(1,3)=0.26$,$S(1,4)=0.13$,$S(2,3)=0.37$,$S(2,4)=0.35$,$S(3,4)=0.78$,分析可知 $S(3,4)$ 的支持度最高,即借阅《数据结构习题集》的同学同时借阅《操作系统典型题型精解精练》可能性是 78%,因为以上 2 门课程都是计算机专业研究生统考专业课.《风雨考研路》虽然借阅人数最多,但同时借阅其他 3 本书的可能性就比较低. 对该数据进一步分析,该书的借阅者大部分是低年级的同学. 另外,根据借阅不同考研专业图书的挖掘,还可以及时了解当年学生考研人数,以及跨专业考研的人数和专业方向.

针对文学类图书的借阅情况,在 415 421 条借阅图书的记录中进行统计和分析,其结果如表 2 所示.

表 2 中偏离度的绝对值越小,文理平衡度越高,偏离度的绝对值就越大,对本专业的“忠诚度”也就越高. 从表 2 可以看出,体育系的学生文理平衡度最高; 艺术学院、外语系、政法学院的学生对理工类图书不感兴趣,更忠诚于本专业.

表 2　各教学单位学生借阅图书对比

教学单位	文学类可能性/%	理工类可能性/%	偏离度
计算机学院	2.48	12.56	-10.08
电气学院	3.94	18.21	-14.27
机电学院	1.96	13.41	-11.45
化工学院	0.78	10.84	-10.06
食品生物学院	3.54	9.57	-6.03
数学系	2.34	8.92	-6.58
外语系	20.23	0.78	19.45
物理系	1.82	7.66	-5.84
政法学院	19.56	0.43	19.13
体育系	5.63	5.52	0.11
艺术学院	23.48	0.16	23.32
软件学院	6.85	11.93	-5.08
国际学院	15.30	9.87	5.43

4 结语

数字图书馆作为图书馆未来的发展趋势,将会拥有海量的数字资源. 图书馆的目标就是要充分发挥这些数字资源的作用,而避免信息过量. 本文基于数据挖掘技术,通过数据清理、数据整合,并加上规约算法,对图书馆信息管理的数据进行挖掘和预测,实验结果表明,数据挖掘的结果不仅能为图书馆的业务管理提供数据参考,而且能指导传统图书馆管理员的日常工作. 从而有利于图书馆调整图书管理策略,进一步提高服务质量和管理水平.

参考文献:

[1] Littman M L, Dumais S T, Landauer T K. Automatic Cross-language Information Retrieval Using Latent Semantic Indexing [M]. Belongia: Kluwer Academic Publishers, 1998: 15 - 21.

[2] Vinokourov A, Shawe-Taylor J, Cristianini N. Inferring a Semantic Representation of Text Via Cross-Language Correlation Analysis [M]. Cambridge: MIT Press, 2002: 91 - 98.

[3] 孙权森,曾生根,王平安,等. 典型相关分析的理论及其在特征融合中的应用 [J]. 计算机学报,2005,28(9):1524.

[4] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases [C] // Proc of the 1993 ACM SIGMOD Int Conf on Mana of Data, Washington DC: ACM Press, 1993: 207 - 216.

[5] 郭力平,雷东升,冷永杰,等. 数据库技术与应用 [M]. 北京: 人民邮电出版社,2007.

[6] 李静. 数据仓库中的数据粒度确定原则 [J]. 计算机与现代化,2007(2):57.

(上接第 30 页)

进行数据流程的控制,超时机制和文件描述符表的引入,提高了文件打开的速度; 数据库连接池技术的使用,提高了网络利用效率,缩短了程序运行时间; 内存映射机制的加入,提高了文件访问的效率.

参考文献:

[1] 常玉连,朱保国,任福深,等. 面向油田地面工程系统的数据库接口软件设计 [J]. 油气田地面工程,2003,22(8):62.

[2] 凡哲元,邢绍献,苏映宏,等. 油田开发规划优化决策系统研究 [J]. 油气地质与采收率,2003,10(6):34.

[3] 梁达平. 试析大型制造业 ERP 软件数据库性能优化技巧 [J]. 甘肃科技,2008(10):24.

[4] 陈悦,白杰,王林. 软件项目开发的性能优化 [J]. 微处理机,2009,30(3):99.

[5] 关晓晶,魏立新,杨建军. 基于混合遗传算法的油田注水系统运行方案优化模型 [J]. 石油学报,2005,26(3):114.

[6] 赵改善,孔祥宁,王于静,等. 64 位集群计算平台波动方程叠前深度偏移的性能优化 [J]. 勘探地球物理进展,2005,28(1):57.

[7] 杨存祥,张晓辉,石军. 基于 DSP 和 FPGA 的油田测井系统总线通信接口设计 [J]. 仪表技术与传感器,2010(4):67.