基于数据挖掘的图书采访系统①

Design of Book Acquisition Based on Data Mining

马金林 (北方民族大学 图书馆 宁夏 银川 750021)

摘 要: 利用挖掘技术对馆藏、借阅及其他信息进行处理，分析出对采访工作有益的数据，为采访人员提供了科学的参考依据。

关键词: 采访 数据挖掘 聚类

1 引言

图书馆藏书是图书馆赖以生存的物质基础，也是评定图书馆文献服务水平的一个重要指标，图书馆的文献保障与建设工作是由采访部门根据采访计划完成的。由于各图书馆的职能和读者的需求不同，导致各个图书馆的文献建设需求不一样，而且这一需求是动态的，不断变化的。各图书馆的采访计划是根据学校的学科建设和发展需要制定的，多为中长期计划，并未考虑到短期的需求，且其更新与调整的周期较长，不能及时准确反映短期内的需求。采访计划不能与现有图书的馆藏、利用率等指标挂钩，导致许多常年无人问津的图书被重复采购，有时甚至加大采购力度，而读者关注和急需的图书却一直得不到采购。如何建立动态、实时调整的采购策略，成为各图书馆做好图书采访、发展文献建设急需解决的问题[1]。建立根据读者需求、馆藏量、利用率、关注率等指标动态调整的采访计划，将能高效地利用购书经费，既能最大限度地满足中长期的馆藏计划，又能快捷地响应读者动态的阅读需求。

数据挖掘(Data Mining)是近年来随着人工智能和数据库技术的发展而出现的一门新兴技术。数据挖掘能从庞杂的信息中提取有用的数据，通过公正客观的统计分析，快速而且正确地得知读者需求信息，找出图书采访方向，准确掌握未来的文献建设动态。可以把数据挖掘定义为一个利用各种分析工具在海量数据中发现模型和数据之间关系的过程，这些模型和关系可以被用来分析馆藏结构、阅读倾向，从而指导文献建设工作。

2 系统设计

本系统通过图书自动化管理系统的书目数据、借阅数据、检索数据以及学科建设和学生阅读需求为数据来源，分析比较现有馆藏、阅读倾向、现有文献的满足率、文献需求信息等数据，产生量化的分析数据。然后将资源建设的需求进行量化后建成资源建设需求表，用它来表示近期文献采访的重点。最终对各类统计信息和需求信息进行数据挖掘，计算出各类图书的采访权重系数，采访人员依据该结果进行图书采访工作。

本系统主要由数据库系统、数据预处理模块、数据挖掘分析模块和结果分析模块构成，其系统结构如图 1 所示。

数据库系统，提供数据的存取与维护，进行相关数据表的创建与更新等工作。数据预处理，根据采访需要，对各数据来源表中的数据进行预处理，如数据的清洗、统计等功能。挖掘分析，通过关联规则、统计分析、聚类分析、分类等分析方法对预处理后的数据进行处理，以求找出读者的阅读习惯与阅读倾向，各专业与该专业相关图书的借阅需求关系以及现有馆藏之间的的关系，找出现有图书的满足率与需求之间的最佳关系等信息。结果分析，将挖掘分析结果进行分析解释，筛选出最有价值的数据以量化的方式形成

① 基金项目:北方民族大学研究项目(2007Y037)
收稿时间:2009-03-17

报告，提交给采访人员，以期调整和实施更加科学的采访策略。

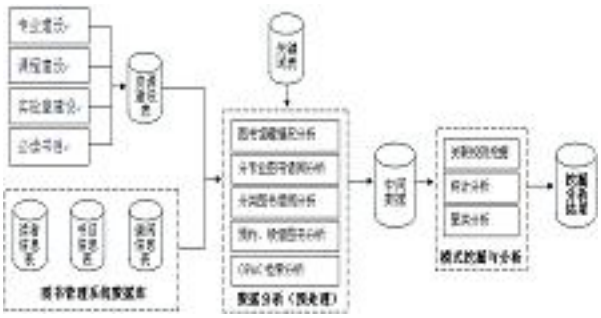

图 1　系统结构图

3　数据库系统设计

为了满足本系统的需求，需要自建若干数据表，如资源建设表、专业课程表、必读书目表、关键词表等。

3.1 资源建设表

用于记录本馆的资源建设情况，反应各类图书的建设需求。通过权重系数来调整和体现当前对该类图书的发展需求，用 0 到 1 的小数来体现其需求的急需程度，1 为急需采购，0 为不需要，采访人员可以根据本馆的资源建设需求来调整这一系数。

该表是对专业建设表、课程建设表、实验室建设表、必读书目表中的数据进行统计和处理后形成的。该表体现了图书馆的中短期图书建设需求。

专业课程表，用于记录本校所开设专业的专业课程情况，该表是数据挖掘和分析过程中重要的数据源。必读书目表，由各专业主讲教师提供，用以指导本馆的专业课程文献建设，是图书馆必须采购的图书。

3.2 关键词表

通过设置关键词，系统按照这些关键词去统计该类特征的图书情况，如书名中的核心词汇、某一学科的学科名称、程序设计语言的名称、操作系统名称、人名、地名、各专业名词等都可以进入关键词表。比如需要统计 Linux 操作系统相关书籍的馆藏情况，系统查找书名中有 Linux 的图书即可。该词表可融于本系统所需的各特定数据表中。

3.3 中间表

将图书管理系统数据库中的数据与资源建设表中的数据进行统计，清洗、统计后形成中间数据，用于后续的数据挖掘与分析。

本系统数据挖掘所需要和产生的表,如表 1 所示：

表 1　数据挖掘系统表

表名	性质	主要字段	功能
读者表	自动化系统表	姓名、学号、读者号、专业	记录学生信息，是数据分析的直接数据源
书目表	自动化系统表	书名、分类号、馆藏量、复本信息	记录书目信息，是数据分析的直接数据源
借阅表	自动化系统表	条码号、分类号、读者号、借阅日期、还书(应还)日期、续借日期	记录读者图书借阅、续借、还书信息
分类表	系统表	类号、学科名、关键词、发展程度	反应图书学科和类号、特征关键词的关系以及每一类图书中各种图书的发展情况（趋势）
专业表	系统表	专业名称、所属院系、专业类别	记录本校各个专业情况
资源建设表	系统表	项目名、项目性质、权重系数	按照本校的学科发展状况和专业建设现状设定的文献建设表，反应和体现现阶段文献建设的重点和方向，以权重系数来表示该类图书建设的重要程度
专业课程表	系统表	专业名、入学年份、开课年份、课程名、学分	各专业的专业课程开设情况和开课情况
OPAC 点击表	系统表	查询表、命中情况、查询次数	记录读者查询情况
必读书目表	系统表	专业名、书名、ISBN、作者	各专业教师提供的必读书目信息

续表 1

续借图书表	中间表	分类、学科、关键词、种数	记录读者图书的续借情况
图书分类表	中间表	分类、学科、关键词、种数	分类、分关键词统计各小类图书的馆藏情况
专业借阅表	中间表	学科、分类、借阅量、专业学生数、是否开课学期、累计节约天数、平均借阅天数	按专业统计各专业相关图书的借阅情况
图书借阅表	中间表	类号、分类、借阅量、累计借阅天数、平均借阅天数	按图书类别统计各类图书的借阅情况
分析结果表	结果表	学科、关键词、馆藏量、借阅率、发展程度、学科性质、学科发展权重、综合指标	按照工作人员的设定，将各个相关表的数据进行数据挖掘后产生的结果信息，工作人员按照这一信息来参考实时图书采访工作

4　数据预处理模块设计

分析数据来源于多方面，主要有以下预处理操作。

数据的清洗，由于很多数据并非真实体现读者实际需求的信息，如读者在仓促中选中的图书，在认真阅读时却发现并非其真正所需图书，他们一般都要在当天或 3 天内将该书归还。那么我们就认为这条记录是无效记录，把它视为脏数据。将系统各数据表中的无效记录进行清洗，可以较好地控制挖掘的效果，挖掘出真实存在的规则与模式。

4.1 读者借阅信息预处理

分别按不同属性统计出各个属性下图书的借阅信息，如按专业、年级、学院、生源等的借阅情况，甚至统计读者对图书的续借情况。需要用到读者信息库、读者借书库、读者还书库、续借库等数据源。

4.2 图书信息预处理

统计当前馆藏图书的情况，按不同的属性统计出各种图书的馆藏。如按中图法标识统计、按关键词统计、按学科统计、按出版社统计、按图书出版年份统计等。需要用到书目数据库、中图法表、关键词表、出版社表、专业课程表等数据源。

4.3 图书流通情况预处理

统计图书的流通情况，以反应各类图书的流通状况，体现当前这类图书的使用情况。如借阅率、预约率、续借率、流通周期、周转率等数据。需要用到书目数据库、借书库、还书库、续借表、预约表等数据源。

5　模式设计与分析模块设计

数据预处理后，数据的模式挖掘工作主要完成各数据模式发现工作。本系统采用统计分析和关联规则来进行模式发现。

5.1 统计分析

统计方法是从图书馆自动化管理系统中提取有用信息的一种有效技术。通过对读者借阅信息的分析，可以对读者的阅读倾向、借阅意图、学习动态、研究方向等信息进行统计分析，也可以对现有馆藏的藏书结构、藏书数量等数据进行准确体现。

5.2 关联规则

关联规则主要用于发现读者之间、图书之间以及图书借阅与读者专业、读者兴趣、研究方向之间存在的潜在关系。关联规则的步骤是：迭代识别所有的频繁项目集，要求频繁项目集的支持率不低于用户设定的最小支持度，从频繁项目集中构造可信度不低于用户设定的最小置信度。其中支持度与置信度的定义如下：

支持度指交易集 T 中包含 X 和 Y 的交易数与交易数据库 D 中所有的交易数之比：

$$\text{support}(X \Rightarrow Y) = \left|\{T : X \cup Y \subseteq T, T \in D\}\right| / |D|$$

置信度指交易集 T 中包含 X 和 Y 的交易数与包含 X 的交易数之比：

$$confidence\ (X \Rightarrow Y) = \left|\{T : X \cup Y \subseteq T, T \in D\}\right| / \left|\{T : X \subseteq T, T \in D\}\right|$$

5.3 聚 类

聚类是将数据点集合分成若干类或簇(cluster)，使得每个簇中的数据点之间最大程度地相似，而不同簇中的数据点最大程度地不同，从而发现数据集中有效的、新颖的、可以理解的数据模式分布[2]。在本系统中，聚类技术是对符合某一访问规律特征的读者进行读者特征挖掘。通常可以将借阅图书的总和视为数据空间，构造一个 BOOK_Use2rID 关联矩阵 Mm×n，如式所示。

$$M_{m\times n} = \begin{bmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,j} & \cdots & h_{1,n} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,j} & \cdots & h_{2,n} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ h_{i,1} & h_{i,2} & \cdots & h_{i,j} & \cdots & h_{i,n} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ h_{m,1} & h_{m,2} & \cdots & h_{m,j} & \cdots & h_{m,n} \end{bmatrix}$$

其中: $h_{i,j}$是 j 读者在一段时间内借阅第 i 类图书的次数；每一行向量 M[1,jj]表示所有读者对图书类“1”的访问情况；每一列向量 M[i,1]表示读者“1”对该图书馆中所有图书的借阅情况。因此可以这样认为：行向量既代表了藏书结构，又蕴涵有读者共同的借阅模式；而列向量既反映了读者类型，也勾勒出了读者的个性化借阅子图。再使用一些度量方法(如 Hamming 距离)分别度量行向量和列向量的相似性，就可以得到两种类型的聚类，即读者聚类和图书聚类。读者聚类主要是把具有相似特性的读者聚集在一组，这类知识对日后为读者提供个性化的服务特别有用；图书聚类可以找出具有相关内容的图书群，这对推出智能 OPAC 检索和提供荐书有很大的帮助。

6 系统实现

系统使用 Windows 2003 上安装的 Oracle 数据库 9i，数据库选用我馆实际运行的自动化管理系统数据库，系统由数据预处理和数据挖掘两部分组成。数据挖掘采用 PL/SQL API 与 Oracle Data Mining Java API[3]两个 API 来实现。

实验使用“决策树分类”算法创建、计算和应用预测模型分类读者，采用 Oracle JDeveloper 进行系统开发，结合采用 Oracle Discoverer 进行输出和图形可视化处理。

系统封装多个 Java 类，其中 Java 类 OracleModelSettings 用于维护模型设置表：

```
public OracleModelSettings(String modelSettingsName,
    Connection databaseConnection,
    String[] keyToValueStringMap)
    throws SQLException
```

Java 类 OracleMiningModel 用于创建和维护数据挖据模型

```
public OracleMiningModel(String modelName,
    OracleModelSettings oms,
    String[] keyToValueMappings,
    // keyToValueMappings 确定算法并指定必须的挖掘模型
    boolean recreate)
    throws SQLException
```

7 结果分析

下图为挖掘后的部分结果，该表给定了系统分析后的综合指标，该指标越大表示该类书的需求越大，可加大该类书的采购力度。

分类	学科	关键词	馆藏量	借阅率	发展程度	学科性质	学科发展权重	综合指标
TP312	程序设计	C,C++	683	16%	0.3	普通学科	0.6	0.35
TP312	程序设计	java	102	87%	0.9	重点学科	0.9	0.89
TP309,TP312,TP311	信息系统	信息系统	37	93%	0.9	重点学科	0.93	0.92

图 2　挖掘分析结果

下图为聚类分析后的部分结果，显示某类图书的主要阅读人群。

图书分类	关键词	主要借阅人群
TP312	Visual C++	2007网络工程,2007信息系统,2007应用数学,2007软件工程,2006软件工程,2006计算机
TP391.41	Photoshop	2005艺术设计,2005广告学,2005新闻学,2006计算机科学与计算
TP391.72,TH12	AutoCAD	2005艺术设计,2005过程装备,2006计算机科学与计算,2006测控技术与仪器,2005艺术设计

图 3　某类图书借阅人群分析

下图为聚类分析后的部分结果，显示某读者的阅读倾向。

读者	主要借阅图书	关键词
JS03020	K281,K820	回族,回回,回话,习俗,风俗,民族
JS12006	K879	岩画,壁画,文字,墓葬
JS02001	H15	修辞,王希杰

图 4　读者借阅分析

通过分析，采访人员可以掌握近期的采购重点，根据聚类分析的结构对特定读者推荐与其相同属性人群所关心的图书，对特定读者发布与其相关的新书通报。

8 结语

数据挖掘是图书馆信息化决策系统的重要组成部分，如何充分利用图书馆现有条件，构建高效的挖掘系统是一个值得研究的课题。利用挖掘技术为图书采访提供参考决策，可以有效提高采访质量，有助于采访部门买来学校和读者最需要的图书。

参考文献

1 李志明,胡森树.数据挖掘及其在现代化图书馆中的应用.图书馆学研究, 2006,(6):39－41.
2 毛国君,段立娟.数据挖据原理与算法.第 2 版,北京:清华大学出版社, 2007.80－85.
3 Using the Oracle Data Mining API.http://www. oracle.com