统计分析及关联挖掘在高校图书馆流通数据中的应用①

韩存鸽

(武夷学院 数学与计算机系, 武夷山 354300)

摘 要：对武夷学院图书馆提供的连续一年的借阅数据进行了预处理，从统计分析、关联关联挖掘两个方面展开工作，根据分析结果，从馆藏布局、图书采购、图书馆工作人员的人力、工作时间安排上给出相关的建议，帮助图书馆做好图书推荐工作。

关键词：关联挖掘；统计分析；Clementine；Apriori 模型

Statistical Analysis and Association Mining of Application in the University Library Circulation Data

HAN Cun-Ge

(Mathematics and Computer Science Department, Wuyi University, Wuyishan 354300, China)

Abstract：In this paper, Wuyi University Library provides loan for 1 year of data preprocessing, statistical analysis, two aspects associated with association mining to start work, according to the results of the analysis, from the collection of layout, book purchases, library staff, manpower, working time arrangements give relevant recommendations, recommended books to help make the library work.

Key words：association mining; statistical analysis; Clementine; apriori model

随着高校图书馆数据库中数据量的迅速增加，高校图书馆所拥有的文献资源数据正在呈几何指数上涨。但增长过快、过多的数据往往会变成“数据坟墓”，失去其指导意义。如何有效的利用这些数据成为一个问题，利用关联挖掘对读者的借阅日志进行分析，发现读者借阅一类图书时的其他借阅行为，可以在读者下次借阅时推荐其他相关的有价值的相关文献，可以调整馆藏布局。

关联挖掘是数据挖掘的一种重要形式，从提出到目前一直受到数据库界的广泛关注。所谓关联规则就是寻找描述数据库中的数据项（属性、变量）之间存在（潜在）的关联或联系。通过对读者借阅检索数据进行关联分析，从中发现读者在借阅检索文献时的其他借阅行为。通过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分析报告和预测报告，优化信息资源建设或馆藏结构，也可以为研究学科相互渗透现象提供依据。通过对读者借阅记录的挖掘找出读者与图书的频繁项集，从而了解不同读者的兴趣爱好，主动向读者推荐相关资料。

1 关联规则挖掘基本理论

1.1 关联规则的概念

R.Agrawal 等人给出关联规则的基本概念[1]

1）项集：$I=\{I_1,I_2,I_3,\cdots,I_k\}$是项目的集合,I 中项目的个数为 k，则集合 I 称为 k-项集。

2）事务：设 $I=\{I_1,I_2,I_3,\cdots,I_k\}$是由数据库中所有项目构成的集合，一次处理所含项目的集合用 T 表示，使得 $T\subseteq I$，并使每一个 T 都有唯一的标识 TID。那么，我们称二元组<TID,T>为数据库事务，表示为 T。

3）关联规则:一个关联规则是形如 $A\Rightarrow B$ 的蕴含式，其中 $A\subset I$,$B\subset I$,并且 $A\cap B=\varnothing$，A 称为规则前提，B 称为规则结果。

4）支持度和置信度:设 D 中有 S%事务同时包含项

① 基金项目:武夷学院校科研资金(xl201014)
收稿时间:2011-12-09;收到修改稿时间:2012-01-20

集 A 和项集 B，则称 S%为关联规则 A⇒B 的支持度，记为 Support(A⇒B)。若在 D 包含项集 A 的事务中，有 C%的事务同时也包含项集 B，则称 C%为关联规则 A⇒B 的可信度,记为 Confience(A⇒B)。

1.2 关联规则挖掘的过程

关联规则挖掘的过程可以由以下两步的来完成：

(1) 找出频繁项集：Apriori 算法是通过项目集元素数目不断增长来逐步完成频繁项目集发现的。首先扫描数据库，积累每个项的计数，并收集满足最小支持度的项，找出频繁 1 项集的集合 L1,然后由 L1 得到 L2，由 L2 得到 L3，如此下去，直到不能找到频繁 k 项目集。

(2) 由频繁项集产生强关联规则： 关联规则挖掘的基本过程[2]如下图 1 所示，由于步骤(2)不需要到数据库中去读取信息，故它的计算量不大，所以关联规则挖掘的重点句放在了查找所有频繁项集和它的支持度上。

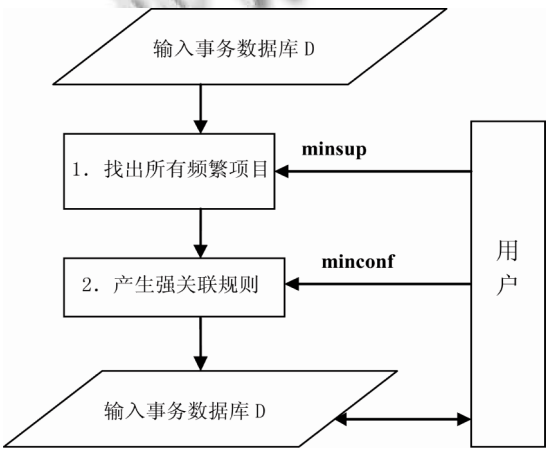

图 1 关联规则挖掘的基本过程

2 图书馆借阅数据的预处理及统计分析

2.1 图书馆借阅数据预处理工作

2.1.1 图书馆数据的采集

从武夷学院图书馆管理系统中导出读者借阅信息，SQL 查询语句如下：

select ltxxb.读者证号,ltxxb.读者姓名,dzxxb.读者单位,借出时刻,wxxxb.索取号;

from ltxxb,dzxxb,wxxxb,tmxxb into table alldata;

where ltxxb.条码=wxxxb.条码 and tmxxb.索取号=wxxxb.索取号 and ltxxb.读者证号=dzxxb.读者证号 and 借出时刻 between {^2006.08.01} and {^2007.07.31 23:00:00}

2.1.2 数据处理

先对图书按索取号进行分类，SQL 语句如下：

sele left(索取号,1) as 分类,year(借出时刻)as 年，month(借出时刻)as 月,day(借出时刻)as 日,hour(借出时刻)as 时间,cdow(借出时刻) as 星期,*

from alldate into table clalldata

group by 读者证号,借出时刻

删除每次只借一本书的记录，最后剩余 34608 条记录。

2.1.3 建立事务数据库

① 建立大类事务数据库：与该课题相关的属性是读者证号、借阅图书在中图法中的分类，将借阅记录数据集中的图书分类号信息转化为二元数据形式[3]。如果有借阅某类图书，就显示.t。否则.f.

② 细类数据的采集及处理（以 tp 类书籍为例）

SQL 语言如下：select 读者证号,借出时间，索取号

from clalldata into table zd

where 分类= ”T” and 索取号 like “TP%”

2.2 对图书借阅信息的统计分析

2.2.1 各大类图书借阅情况的统计分析及建议

在 excel 中统计出全年各类图书借出图如图 2 所示。

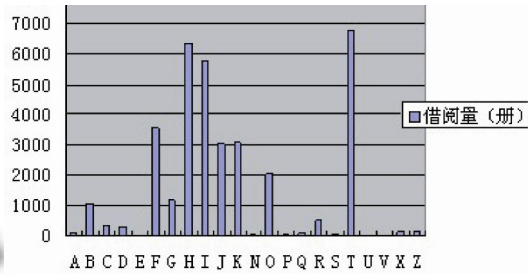

图 2 全年各类图书借阅量统计图

E 类(军事)、N 类(自然科学总论)、V 类(航空、航天)、U 类(交通运输)、S 类(农业科学)、 P 类(天文学、地球科学)书的借阅量非常低。T 类(工业技术)H 类(语言、文字)、I 类(文学)借阅量居前三位。

建议：i、采购部门少采购或近期不采购 E、N、V、U、S、P 类图书。

ii、图书馆调整书架，可以把借阅量非常低的 E、N、V、U、S、P 类图书放在比较偏僻的位置，而把居于借阅量前三位图书放在读者容易找到的位置。

2.2.2 各月份借阅情况的统计分析及建议

SQL 语言如下：select 月,count(索取号) from

clalldata group by 月

在 excel 中最终形成各月份的借书量如图 3 所示

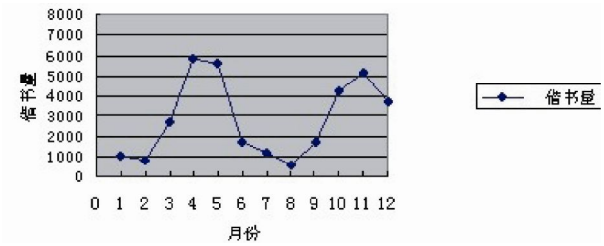

图 3　各月份的借书量

建议：i、居于借阅量前三位的是四、五、十一月份，图书馆应当增加人力支援这三个月份。

ii、三、六及九月份借阅量都比较低，图书馆可以适当的减少人力安排。当然图书馆可以在这三个月中多做图书宣传，鼓励学生充分利用图书馆资源。

2.2.3 一周内每天借阅情况统计分析及建议

select 星期,count(索取号) from clalldata group by 星期经过统计在 excel 中得到每天的借阅量，如图 4 所示：

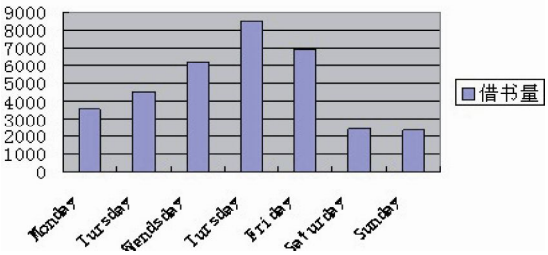

图 4　一周内每天的借阅量

建议：i、星期四借阅量最高，是因为周四下午全校没教学安排，建议图书馆工作人员的例会，尽量不要安排在周四，这天还要增加图书馆人员的安排。

ii、除去周六、周天，周一的借阅量应该是比较低的，可以考虑把图书整理的时间放在周一。

3　高校图书馆数据的关联挖掘及分析

目前市场上的数据挖掘工具主要有 IBM Intelligent Miner、SAS Enterprise Miner、SPSS Clementine 等，由于数据挖掘本身需要考虑的因素很多，很难按照原则给工具排一个优劣次序。最重要的还是用户的需要，通过对图书馆流通数据的分析，我选择了 SPSS 公司的 Clementine[4,5]作为本次挖掘工具，通过挖掘分析以找出图书之间的关联规则，以及读者背景信息与借阅行为之间的关联规则。

3.1 Clementine 中实现大类之间关联规则的挖掘

在 Clementine 中关联挖掘有“GRI 模型”、“Apriori 模型”、“Carma 模型”以及“有序模型”，选择经典的“Apriori”算法建立模型。设置最小支持度为 15%，最低置信度为 65%，得到如图 5 的七条关联规则。

图 5　生成的关联规则

生成的七条规则分别为:

rules1：T→F　support =18.971%，confidence =71.731%

rules2：O→F　support =16.7%，　confidence =71.25%

rules3：T→H　support =20.3%，　confidence =70.6%

rules4：H→I　support =35.616%，confidence =69.029%

rules5：J→I　support =17.221%，confidence =67.6%

rules6：B→K　support =15.439%，confidence =65.7%

rules7：G→B　support =17.307%，confidence =65.428%

分析：规则 5 的提升度小于 1，该规则可信度低，不予讨论，其余 6 条规则的前项与后项是正相关的，这些规则可信程度应该较高，规则 4　H→I 不是有趣规则，也不予分析。所以后面分析的是其余 5 条规则。

建议：根据以上挖掘出的五条关联规则，图书馆可以根据读者需求，优化馆藏布局，比如可以把可能被同时借阅的 T 类和 F 类、H 类放在一起，O 类和 H 类放在一起，G 类和 B 类放在一起，J 类 K 类放在一起。这样既方便师生借阅，又可以在一定程度上辅助教师的教学。图书采购部门可以进行相应的策略调整，如在购进某类新书的时候，同时购入与其相关联的其他类图书，增加这些图书的借阅率，以更好地服务读者。

3.2 在 clementine 中实现细类关联规则的挖掘

在 clementine 中首先应该导入事务数据 zd.txt，继续选择“Apriori”模型。设定最小支持度为 6%，最小

置信度为 62%，得到如图 6 中十一条规则：

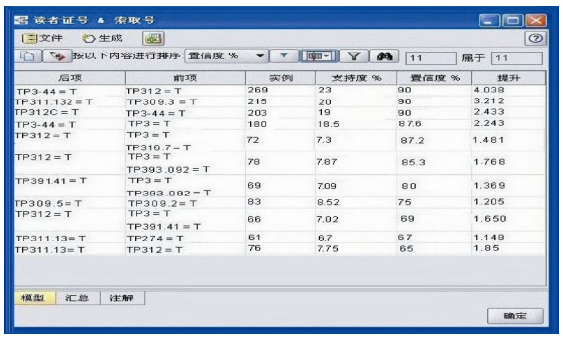

图 6 细类实现关联挖掘后得到的规则

得到关联规则共十一条分别如下：

rules1:TP312→TP3-44 support =23%，confidence =90%

rules2:TP309.3→TP311.132 support =20%，
confidence =90%

rules3:TP3-44→TP312C support =19%，
confidence =90%

rules4:TP3→TP3-44 support =18.5%，
confidence =87.6%

rules5:TP3, TP316.7→TP312 support =7.3%，
confidence =87.2%

rules6:TP3， TP393.092 → TP312 support =8.87%，
confidence =85.3%

rules7:TP3，TP393.092→TP391.41 support =7.09%
confidence =80%

rules8:TP309.2→TP309.5 support =8.52%,
confidence =75%

rules9:TP3， TP391.41 → TP312 support =7.02%
confidence =69%

rules10:TP274→TP311.13 support =6.7%
confidence =67%

rules11:TP312→TP311.13 support =7.75%，
confidence =65%

分析：由 rule1 和 rule4 可以知道，借阅程序设计语言类图书、计算机技术类图书后同时借阅 TP3-44 的可能性分别为 90%和 87.6%,图书馆采购部门在增加计算机等级考试类图书的同时,补充一定数量的程序设计语言类图书和计算机技术类图书，以满足读者需求。

由 rule3 知道，在借阅的计算机等级考试类图书中 C 语言类图书居多,这也反映了知识发展的方向。

rule2 从馆藏来看符合我们挖掘的规则及图书馆的图书推荐工作，读者想借阅数据库系统类图书可以向其推荐数据库备份与恢复类书籍。老师在进行数据库系统讲解时可以适当的增加数据库恢复与备份知识。

rule5、rule6、rule9 分别列出了借阅了计算机技术、Windows 操作系统类图书、网络浏览器类图书、图像识别及其装置类图书后同时借阅程序设计语言类图书可能性分别为 87.2%、85.3%、69%，图书馆采购部门在计划增加程序设计语言类图书时候，需要补充一定数量的计算机技术类、Windows 操作系统类图书、网络浏览器类图书、图像识别及其装置类图书，以满足读者的需求。管藏模式也应该根据该要求进行改进。

按照 rule7 图书馆采购部门可以进行相关的采购。

rules8:是计算机安全类图书与病毒与防治类图书之间的借阅关系，该规则不是有趣规则，不作分析。

rule9、rule10 列出了数据处理类图书与程序设计类图书和数据库类图书之间的关系，根据该关系可以对馆藏适当调整，把数据处理书籍放在数据库类书籍的附近，提供个性化服务。教师在向学生介绍数据库或数据仓库类内容时，适当增加数据处理方面的内容。

4 结语

本文对武夷学院图书馆提供的借阅数据进行了预处理，从大类、月份、星期三个方面在 Excel 中进行统计，根据统计结果从馆藏布局、图书采购、图书馆工作人员的人力、工作时间安排上给出相关的建议。

在 Clementine 中采用 Apriori 模型分别从大类、细类两方面进行了关联挖掘，根据挖掘结果，找出书籍之间大类与大类、细类与细类之间的借阅关系。以了解读者的阅读兴趣，根据兴趣调整馆藏布局，给图书采购部门相关建议，帮助图书馆做好图书推荐工作。

参考文献

1 Han JW, Kamber Mi.范明,孟小峰,译.数据挖掘概念与技术.北京:机械工业出版社,2006.

2 高明.关联规则挖掘算法的研究及其应用[硕士学位论文].济南:山东师范大学,2006.

3 鲍静.关联规则挖掘及其在图书流通数据中的应用研究.合肥:合肥工业大学,2007.

4 岳小婷.数据挖掘工具 Clementine 应用.牡丹江大学学报,2007,(4):103–105.