

关联规则在流通库读者数据挖掘中的应用^{*}

金松根 王琳琳 石 丹 付 佳

(牡丹江医学院图书馆 牡丹江 157011)

〔摘要〕 介绍关联规则在图书馆的应用,以不同专业大一读者借阅数据为挖掘对象,从学院不同专业和年级两个角度出发,提取出适合进行数据挖掘的部分记录,从数据准备、数据挖掘两个阶段进行关联规则分析,为深入开展读者服务工作提供技术支撑。

〔关键词〕 阅读行为;关联规则;Weka

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2015.09.018

Application of Association Rules in Data Mining of Circulation Library Readers JIN Song-gen, WANG Lin-lin, SHI Dan, FU Jia, Library of Mudanjiang Medical University, Mudanjiang 157011, China

〔Abstract〕 The paper introduces the application of association rules in the library. With circulation data of freshmen in different majors, in terms of different majors and grades of the college, it extracts partial records suitable for data mining, analyzes phases of data preparation and data mining by use of association rules, and provides technical support for in-depth conduction of reader services.

〔Keywords〕 Reading behavior; Association rules; Weka

1 引言

随着网络化的发展以及数字图书馆的崛起,高校图书馆也开始走入了数字化时代。高校图书馆现有的自动化管理系统中存储着大量的书目数据,读者流通借还数据、书目检索记录、Web 访问记录等,为实现数据挖掘技术在图书馆的应用提供了基础。其中,流通日志挖掘可以帮助图书馆更好地了

解读读者的阅读目的、动机、内容、兴趣特点及层次等阅读行为特征,不仅能为馆藏建设提供参考依据,而且还有助于图书馆积极主动地开展个性化服务,提高读者满意度。目前,国内高校图书馆投入到流通日志数据挖掘工作中,充分利用以往读者借阅服务中产生的海量数据,以关联规则、决策树、聚类分析等多种方法进行了实践研究。本文以金盘图书馆集成管理系统中流通日志文件为原始数据,在 Weka 智能分析环境中利用关联规则方法进行了读者阅读行为挖掘分析研究。

2 关联规则在图书馆的应用

2.1 关联规则简介

关联规则是研究较多的数据挖掘方法,也是在图书馆数据挖掘中应用最为广泛的方法。关联规则的概念由 Agrawal、Imielinski 和 Swami 提出,是一

〔修回日期〕 2015-09-24

〔作者简介〕 金松根,副研究馆员,硕士,发表论文 10 余篇。

〔基金项目〕 黑龙江省高校图工委项目“泛在环境下高校大学生阅读行为研究”(项目编号:2013-A-113);“面向 Y 一代的高校图书馆用户信息行为研究”(项目编号:2013-A-109)。

种简单但很实用的规则。关联规则模式属于描述型模式，发现关联规则的算法属于无监督学习的方法^[1]。关联分析中最著名的算法是 Apriori 算法，其核心是基于两阶段频繁项集思想的递推算法。寻找最大项集合（频繁项集）的基本思想是：算法需要对数据集进行多步处理。第 1 步，简单统计所有含有 1 个元素项集出现的频数，找出那些小于最小支持度的项集，即一维最大项集。从第 2 步开始循环处理，直到再没有最大项集生成。循环过程是：第 k 步中，根据第 $k - 1$ 步生成的 $(k - 1)$ 维最大项集产生 k 维候选项集，然后对数据库进行搜索，得到候选项集的支持度，与最小支持度进行比较，从而找到 k 维最大项集。

2.2 已有研究情况

从中国知网等数据库检索相关文献，发现大多数关联规则挖掘研究选取图书馆自动化系统里的数据，主要提取读者借阅日志数据，按照专业、年级、借阅时间、读者层次等开展挖掘任务。如储文静等在“基于 Weka 读者借阅行为分析”一文中，通过关联挖掘，挖掘出读者和各个院系之间的关系，主题文献与院系或单位间的关联度、文献借阅相似度^[2]；陆觉民等在“基于数据挖掘技术的图书馆流通数据的关联分析”一文中，运用改进的 Apriori 算法，不同置信度设置下的频繁三项集，挖掘出两个专业大一和大三学生借阅数据中强关联规则并进行了量化描述^[3]；彭仪普等在“关联挖掘在文献借阅历史数据分析中的应用”一文中，描述了文献 A 到文献 B 的关联规则发现^[4]；王磊等在“关联规则挖掘在图书馆信息服务中的应用”一文中，对确定恰当的项集、数据层次和产生最佳规则进行了理论描述，对引入提升度和增益度指标的必要性进行了解释^[5]，但未进行实际指标对照，无法了解最佳规则产生情况。针对医学生的阅读特点，韦耘耕^[6]、何卫华等^[7]、毛景珠^[8]进行了阅读倾向和阅读图书分类调查分析；冯研等^[9]、孟凡红等^[10]主要关注点集中在数据挖掘专题，利用文献计量学等方法对医学生的信息行为进行系统研究。

3 读者阅读行为关联分析

3.1 概述

本文利用 Weka 3.7 作为数据挖掘开发工具，Weka 的全名是怀卡托智能分析环境，是一款免费、非商业化的基于 Java 环境的开源机器学习以及数据挖掘软件。Weka3.7 实现了多种关联规则挖掘算法，包括 FP - Growth 算法、Filtered Association 算法及典型的 Apriori 算法^[11-12]。从数据挖掘流程来看，读者阅读行为关联规则大致要经历数据准备阶段、数据挖掘阶段和关联规则分析阶段，其流程见图 1。在具体挖掘实践工作中，根据预期挖掘目标，读者、专业、图书分类号的属性选择和参数优化上可反复调整，直到得到有效关联规则。

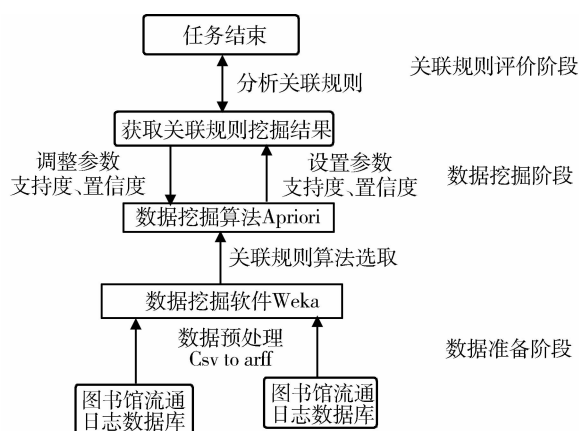


图 1 流通日志中挖掘关联规则流程

3.2 数据准备阶段

在进行读者阅读行为关联分析前，要进行数据挖掘之前的准备工作，将金盘图书馆集成管理系统流通日志中操作代码为“借书”的记录，下载到本地形成 Excel 文件。原始的流通记录中含有读者条码、读者级别、专业、借阅日期、图书条码、索书号、作者项等字段，其中读者条码、读者专业、索书号字段可作为属性参与关联规则挖掘运算，其余可删除或屏蔽。接下来的操作包括：（1）索书号到图书分类号。流通日志索书号字段

中，提取图书分类号时以简表为基准表，凡是包含简表分类号的统一按简表分类号属性处理；若涉及复分表的图书分类号则独立作为一个属性出现，如 E-49、E-53 等。（2）文件格式转换。格式转换涉及两个方面：一是 Excel 文件转换为 csv 格式，需要在 Excel2007 中选择另存为“csv”格式；二是 csv 格式转换成 Arff 格式，需要在 Weka 3.7 中选择 Arffviewer 功能进行转换。（3）分类号离散化。利用 Ultraedit 等文本编辑工具，需要

把形成的 Arff 文件中分类号属性的数据类型从字符串型（string）或数字型（numeric）替换为为关联模式可理解的二元属性形式 { 1 }，完成分类号属性离散化。

数据预处理是数据挖掘的重要步骤，数据挖掘的大部分时间和精力都要花在预处理阶段。经过以上 3 个数据预处理操作后，获取了多个适合在 Weka 平台下进行关联规则挖掘的数据集，如图 2 所示为“不同专业大一读者.csv.arff”。

不同专业大一读者.csv.arff														
Relation: 不同专业大一读者														
No.	1: duzhetiaoma Nominal	2: zhuanYe Nominal	3: A1 Nominal	4: A2 Nominal	5: A3 Nominal	6: A4 Nominal	7: A43 Nominal	8: A48 Nominal	9: A49 Nominal	10: A5 Nominal	11: A7 Nominal	12: A8 Nominal	13: B-43 Nominal	14: B- Nominal
1	J090001	卫生信息...			1									
2	J090002	卫生信息...												
3	J090003	检验医学...										1		
4	J090004	生物工程...												
5	J090005	物流管理...			1									
6	J090006	药物制剂...						1				1		
7	J090009	物流管理...			1									
8	J090011	康复治疗...	1											
9	J090012	检验医学...			1									
10	J090013	康复治疗...												
11	J090014	口腔专业					1							
12	J090015	麻醉专业					1				1			
13	J090016	影像技术...												
14	J090017	预防医学...												
15	J090018	卫生检验...	1											
16	J090019	卫生信息...										1		
17	J090020	预防医学...					1							
18	J090021	检验医学...												
19	J090023	口腔专业												
20	J090024	卫生信息...	1									1		
21	J090025	物流管理...												
22	J090026	口腔专业												
23	J090027	护理专业												
24	J090028	检验医学...												
25	J090029	卫生检验...					1							
26	J090031	公共事业...												
27	J090032	康复治疗...									1			
28	J090033	预防医学...												

图 2 不同专业大一读者.csv.arff

3.3 数据挖掘阶段

关联规则反映一个对象与其他对象之间的相互依赖性，如果多个对象之间存在一定的关联关系，那么其中一个对象就能够通过其他对象进行预测。应用到图书馆流通日志中读者借阅行为关联规则挖掘，通过分析每个读者所借图书的不同分类号，发现所借图书之间的关系，从而分析读者的借阅习惯。

3.3.1 不同专业同年级读者 从图书馆集成管理系统中提取了 2009-2014 年不同专业同年级读者流

通数据发现：一是不同专业读者数存在非常大的差异，特别是 2012 年后新增专业读者数据样本偏少，半数以上的图书分类没有借阅记录；二是大四以后读者借书册数明显下滑，提取出适合进行关联规则挖掘的有效样本难度较大。本研究选取样本较为稳定的 2009-2011 年共计 3 年不同专业大一读者借阅数据为挖掘对象，经过数据预处理后得到 6 478 实例、519 属性的数据集。在 Weka 系统中导入试验数据之后，选择关联规则挖掘 Apriori 算法，最小置信度为 0.9，最小支持度为 0.1，其他参数设置保留系统默认^[12]，见表 1。

Scheme: weka. associations. Apriori -N 10 -T 0 -C 0.9
-D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation: 不同专业大一读者
Instances: 6478

Attributes: 519

经过挖掘数据集试验，不同专业大一读者借阅数据集中，按参数设置挖掘出关联规则 10 条，见表 2。

表 1 不同专业大一读者关联规则 Apriori 挖掘参数设置

参数名称	参数设置	执行含义
car	False	挖掘全局关联规则
classindex	-1	最后的属性被当做类属性
delta	0.05	迭代递减单位
lowerBoundMinSupport	0.1	最小支持度下界
metricType	Confidence	设置对规则进行排序的度量，这里选择默认置信度
minMtric	0.9	度量的最小值
numRules	10	要挖掘的关联规则数
outputItemSets	False	结果中不输出项集
removeAllMissingCols	False	不移除全部为缺省值的列
significanceLevel	-1.0	重要性测试
upperBoundMinSupport	1.0	最小支持度上界
verbose	False	算法不进行冗余模式运行

表 2 不同专业大一读者关联规则挖掘结果

序号	关联规则	度量指标
1	B84 =1 I24 =1 R3 =11 720 = = > H31 =11 679	< conf: (0.98) > lift: (1.15) lev: (0.03) [225] conv: (6.34)
2	I5 =1 R3 =1 1718 = = > H31 =11 669	< conf: (0.97) > lift: (1.15) lev: (0.03) [217] conv: (5.32)
3	I26 =1 R3 =11 782 = = > H31 =11 725	< conf: (0.97) > lift: (1.15) lev: (0.03) [218] conv: (4.76)
4	B82 =1 R3 =11 725 = = > H31 =11 669	< conf: (0.97) > lift: (1.14) lev: (0.03) [211] conv: (4.69)
5	I24 =1 R4 =11 904 = = > H31 =11 840	< conf: (0.97) > lift: (1.14) lev: (0.04) [230] conv: (4.54)
6	R3 =1 R4 =11 869 = = > H31 =11 804	< conf: (0.97) > lift: (1.14) lev: (0.03) [224] conv: (4.38)
7	B84 =1 R3 =11 959 = = > H31 =11 886	< conf: (0.96) > lift: (1.14) lev: (0.04) [230] conv: (4.1)
8	B84 =1 I24 =12 061 = = > H31 =11 982	< conf: (0.96) > lift: (1.14) lev: (0.04) [240] conv: (3.99)
9	I24 =1 I5 =11 866 = = > H31 =11 789	< conf: (0.96) > lift: (1.13) lev: (0.03) [211] conv: (3.7)
10	B82 =1 I24 =11 823 = = > H31 =11 747	< conf: (0.96) > lift: (1.13) lev: (0.03) [206] conv: (3.67)

从表 2 中发现，大一读者所借图书中：（1）同时借阅 B84、I24、R3 类图书的读者，有 98% 的读者会借阅 H31 类图书。（2）同时借阅 I5、R3 类图书的读者，有 97% 的读者会借阅 H31 类图书。（3）同时借阅 I26、R3 类图书的读者，有 97% 的读者会借阅 H31 类图书。（4）同时借阅 B82、R3 类图书的读者，有 97% 的读者会借阅 H31 类图书。（5）同时借阅 I24、R4 类图书的读者，有 97% 的读者会借阅 H31 类图书。（6）同时借阅 R3、R4 类图书的读者，有 97% 的读者会借阅 H31 类图书。（7）同时借阅 B84、R3 类图书的读者，有 96% 的读者会借阅 H31 类图书。（8）同时借阅 B84、I24 类图书的读者，有 96% 的读者会借阅 H31 类图书。（9）

同时借阅 I24、I5 类图书的读者，有 96% 的读者会借阅 H31 类图书。（10）同时借阅 B82、I24 类图书的读者，有 96% 的读者会借阅 H31 类图书。分析得出，关联规则前项集中在 B82、B84、I24、I26、I5、R3、R4 等图书分类，后项均指向 H31 类。医学院校读者出于专业学习需要 R3、R4 等基础医学图书借阅频率较高；大一新生又处在价值观形成的关键时期，出自提高修养等目的 B82、B84 等伦理学、心理学图书借阅频率保持较高水平；从拓展知识、休闲的需求出发，大一读者中 I24、I26、I5 等中外文学类图书也颇受欢迎；再者由于大学期间在校 生都要面临四、六级英语考试，过级考试大多又集中在大二、大三期间，因此大一读者中 H31 类图

书始终保持较高的借阅频次。最为有趣的发现是喜欢阅读文学类图书的读者群中, I24、I5 类图书和 I26 类图书出现了明显的分界线, 支持度 10% 的条件下, 没有出现共存关系; 再如读者所借图书当中 B84 类和 B82 类之间, 也存在相同的分界现象。为了发现更多读者借阅行为, 在挖掘参数保持不变条件下, 本文依次利用 Weka 系统中的“移除 (Remove)”功能移除 H31、I24、B84、R3 属性后, 进行了多次关联规则挖掘, 除了移除 R3 图书分类时, 置信度降为 0.6 以外, 移除其他属性时所挖掘出的前 10 项最佳关联规则顺序接近统一, 置信度较高, 均保持在 0.9 以上。

3.3.2 同专业不同年级读者 学科服务和学科文献建设中, 图书馆常以二级学院为单位进行划分研究, 因此针对同一专业不同年级读者阅读行为研究, 以二级学院为大专业概念。以护理学院为例, 利用 Weka 分析软件, 设置置信度为 0.6, 支持度为 0.1, 得到了护理学院不同年级读者的前 10 最佳关联规则。研究发现, 大一时护理学院读者借阅量不大, 所借图书主要集中在 I24 和 R3 两个大类; 大二时借阅量逐渐增加, 但所借图书分类变化不大; 大三开始读者借书关联规则发生了较大变化, H31 和 R4 类图书为主要借阅对象; 大四时读者所借医学图书份量正在增长, 分类趋于分散, 除了 R3、R4 类图书, R71、R9、R6 等多个医学类图书也有了较高频率外借, 而 90% 以上最佳规则都指向了 H31 类。从挖掘结果可以分析到护理学院不同年级读者的阅读路径, 大一、大二时期集中在基础医学和心理学方面, 而大三以后对外语的需求逐年增加, 医学类图书借书范围也扩展到了内科学和其他临床医学, 兴趣趋于分散。其他二级学院的读者借阅规则挖掘方法与护理学院相同, 这里不再赘述。

4 结语

阅读动机、兴趣、情绪、能力等阅读意向活动是推动读者阅读的一种内部动力, 直接影响读者的阅读倾向和效果。此次研究从学院不同专业和不同

年级两个角度, 在海量的数据库系统中提取适合进行数据挖掘的部分记录, 经过索书号转换分类号、分类号离散化、数据记录格式转换等步骤完成数据预处理, 利用关联规则 Apriori 算法进行了不同专业同一年级、不同年级同一专业挖掘工作, 预置支持度和置信度属性参数得到了与读者借阅行为相关的最佳关联规则。在关联规则分析阶段, 从增减属性和调整支持度和置信度两个方面, 对得出的关联规则进行了挖掘结果分析。本次研究旨在对用户阅读行为数据进行隐性知识的分析与整理, 对深入开展图书馆读者服务工作提供技术支撑。

参考文献

- 1 史忠植. 知识发现. 第 2 版 [M]. 北京: 清华大学出版社, 2011: 140 - 148.
- 2 储文静, 奉国和. 基于 Weka 读者借阅行为分析 [J]. 情报科学, 2010, 28 (3): 424 - 429.
- 3 陆觉民, 马国栋, 郑宇. 基于数据挖掘技术的图书馆流通数据的关联分析 [J]. 现代情报, 2009, 29 (9): 108 - 110.
- 4 彭议普, 熊拥军. 关联挖掘在文献借阅历史数据分析中的应用 [J]. 情报杂志, 2005, (8): 40 - 41, 44.
- 5 王磊, 刘东苏. 关联规则挖掘在图书馆信息服务中的应用 [J]. 情报杂志, 2008, (2): 154 - 155, 158.
- 6 韦耘耕. 从图书馆流通数据分析读者阅读倾向及馆藏结构 [J]. 医学信息学杂志, 2010, 31 (2): 75 - 78.
- 7 何卫华, 陈志才, 邱广鸣, 等. 医学生阅读心理及阅读倾向调查分析 [J]. 医学信息学杂志, 2010, (3): 71 - 74.
- 8 毛景珠. 医学院校学生阅读图书种类分析 [J]. 医学信息学杂志, 2007, 28 (4): 397 - 399.
- 9 冯研, 刘薇薇, 张兵兵, 等. 国内图书馆数据挖掘研究及应用的文献计量分析 [J]. 医学信息学杂志, 2011, 32 (6): 57 - 60.
- 10 孟凡红, 樊红雨, 侯西娟. 基于数据挖掘技术的用户信息需求研究 [J]. 医学信息学杂志, 2011, 32 (9): 43 - 46.
- 11 马国栋. 基于 WEKA 的高校图书馆流通数据的关联分析 [J]. 图书馆工作与研究, 2010, (12): 42 - 45.
- 12 袁梅宇. 数据挖掘与机器学习——WEKA 应用技术与实践 [M]. 北京: 清华大学出版社, 2014: 102 - 112.