

基于数据挖掘技术的图书馆流通数据的关联分析

陆觉民 马国栋 郑宇

(上海大学图书馆, 上海 200072)

〔摘要〕运用改进的 Apriori 算法, 对具有代表性的与分析任务相关的数据进行样本抽取, 利用 Weka 作为数据挖掘算法开发工具, 产生了一系列强关联规则。根据这些规则, 我们可以解读出一些现象, 它不仅能揭示隐藏在大量数据后的重要关系信息, 同时也为这种关系提供了量化描述手段。这些定性定量的信息不仅能对图书馆的各项工作提供技术上的支持, 还可对学校的教学, 课程的设置, 学科的交叉渗透等提供信息。

〔关键词〕数据挖掘; 关联规则; 图书馆

〔中图分类号〕G250.7 〔文献标识码〕A 〔文章编号〕1008-0821(2009)09-0108-03

The Association Analysis for Library Circulation Data Based on Data Mining Technique

Lu Juemin Ma Guodong Zheng yu

(Library, Shanghai University, Shanghai 200072, China)

〔Abstract〕This paper presented an improved algorithm based on the analysis of the Apriori method, collected typical samples related to our task analysis, used WEKA as Development Tools to discover strong association rules. According to these rules, we can reveal important relations between mass data and quantize the relations. Those quantized information not only provided support for routine work in library, but also for education, curriculum and interpenetration.

〔Key words〕data mining; association rules; library

随着图书馆数字信息化的进展, 信息的种类越来越多, 且变化频繁, 信息资源呈爆炸性的增长。与此同时, 知识的不断更新和科研课题的时间性和阶段性, 使高校读者对信息的需求具有针对性、及时性和新颖性, 并呈多元化和个性化的特征。

然而, 在信息需求多样化、个性化的趋势下, 人们发现要准确、快速地查找自己所需的信息并非容易。从需求内容上, 他们要求提供的信息更具全面性和精确性, 不再仅仅满足获得信息载体方面的信息, 还需要权威性相关信息, 并希望进一步得到经过整合、创新, 能解决问题的知识内容; 从需求时效上, 他们要求个人的信息需求及时得到满足。在这样的背景下, 高校图书馆传统的服务方式受到了严峻的挑战, 高校图书馆不仅需要根据用户明确提出的个性化要求提供信息服务, 而且需要通过认真分析用户个人特征和使用信息的习惯等来发现其潜在需求并主动地向他们提供可能需要的服务。为此, 2008 年上海市图书馆学会将此作为立项课题。

1 研究的内容

用户需求是图书馆工作存在和发展的前提, 只有加强用户需求信息需求行为特点的研究, 才能有针对性地开展工作。就目前数字图书馆个性化信息服务系统普遍比较单一, 个性化智能程度不高的特点, 本文提出利用数字挖掘技术进行图书馆个性化技术的研究, 我们以上海大学图书馆部分流通数据作为研究对象通过用户的历史访问记录, 采用关联规则挖掘技术, 发现用户潜在可能的兴趣, 进行针对性的提炼整合和更高层次的分析。

1.1 运用改进的 Apriori 算法

通过对经典的 Apriori 算法的改进, 采用 JAVA 作为数据挖掘矩阵算法的开发环境, 针对其算法性能瓶颈, 根据频繁项集的性质和二进制逻辑运算的基本思想, 提出基于矩阵的数据挖掘算法。挖掘关联规则的关键问题在于提高算法的效率, 对于类似图书馆这样的信息量大且数据分散的大型数据库系统矛盾更为突出, 采用矩阵的数据挖掘技

收稿日期: 2009-03-16

作者简介: 陆觉民 (1957-), 女, 高级工程师, 研究方向: 计算机管理工程、数据库建设研究, 发表论文 20 余篇。

术较好避免了 Apriori 系列算法固有的缺陷, 算法占用内存小, I/O 操作少, 执行速度快, 系统效率大大提高。

1.2 数据的预处理

数据预处理的质量直接影响后续工作, 高质量的数据预处理, 不仅能节约系统资源, 而且能提高数据挖掘过程的精度和性能, 提高系统效率。

对具有代表性的与分析任务相关的数据进行样本抽取, 读者的借阅习惯与其所从事的专业有很大的联系, 因此需

要从图书馆系统的数据库中根据读者专业属性提取借阅数据, 将相关数据库转换整合, 数据归约, 把用户空间分成若干相似用户聚类群, 实现与数据挖掘矩阵算法的对接。

我们着重跟踪上海大学机电工程与自动化学院及知识产权学院 2005 级大一及大三学生借阅 O- 数理学科和化学类, H31- 英语类, D- 政治法律类, I- 文学类, TP- 自动化及计算机技术类书籍的数据, 总计 12 747 条记录, 分类统计见表 1。

表 1 参与挖掘的数据分类统计

项 目	总记录数 12 747	英语类 借阅记录数	文学类 借阅记录数	自动化及计算机 技术借阅记录数	数理学科和化学类 借阅记录数	政治法律类 借阅记录数
自动化 05 大 1	4 692	514	2 751	326	1 007	94
自动化 05 大 3	2 771	266	667	1 698	93	47
知识产权 05 大 1	2 605	282	1 615	21	64	623
知识产权 05 大 3	2 679	175	1 079	100	6	1 319

表 2 为经过处理以后的借阅人数统计及相关的支持度, 由此为关联挖掘矩阵算法做好数据准备。

表 2 处理后的借阅人数统计及相关的支持度

项 目	借阅人数	英语类 借阅人数 (支持度)	文学类 借阅人数 (支持度)	自动化及计算机 技术借阅人数 (支持度)	数理学科和化 学类借阅人数 (支持度)	政治法律类 借阅人数 (支持度)
自动化 05 大 1	N = 315	114 (36%)	268 (85%)	88 (28%)	167 (53%)	55 (17%)
自动化 05 大 3	N = 293	74 (25%)	145 (49%)	239 (82%)	35 (12%)	21 (7.2%)
知识产权 05 大 1	N = 156	53 (34%)	133 (85%)	10 (6.4%)	22 (14%)	110 (71%)
知识产权 05 大 3	N = 172	43 (25%)	127 (74%)	23 (13%)	4 (2.3%)	140 (81%)

1.3 用户隐私安全与保护问题

为了更好地开展个性化服务, 用户的个人信息是不可缺少的, 这就涉及到了用户的隐私问题。由于个性化信息服务需要对用户的基本信息和查询行为进行基本的分析, 因此有关用户日常行为日志、个人信息、注册信息等都在用户个性化特征分析之中。个性化信息服务应该使用户相信其个人信息不会被滥用, 而是用于有效满足用户的需求。同时应该在用户中树立良好的信誉感, 制定出较为完善的隐私保护政策, 保证用户个人信息不被第三方使用。

2 关联挖掘结果及评估

经统计 05 级自动化学院、知识产权学院和文学院参与关联分析的 5 类书籍借阅人数占总借阅人数的比例都在 90% 以上, 样本选取合理, 可信度高。05 级自动化学院、知识产权学院文学院大一、大三学生借阅率变化如图 1。工科类的借阅率呈下降, 文科类的借阅率呈上升。文科大一大三的借阅率都高于工科。

根据统计, 学校的文理科都有这个变化趋势。我们分析主要原因是当今社会科技发展日新月异, 工科专业类的图书更新相对落后于需要, 上网查资料成了学生解决问题的重要途径。而文科则不同, 随着学习的深入, 需要的是

更经典, 更具有积淀的资料, 这些专业信息, 图书馆的藏书更多于网上能提供的资源。总的来说网络是影响借阅率的主要因素之一。

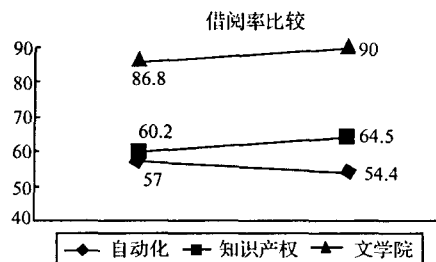


图 1 工科和文科类借阅率比较

课题利用 Weka 作为数据挖掘算法开发工具, Weka 的全名是怀卡托智能分析环境, 是一款免费的、非商业化的, 基于 JAVA 环境下开源的机器学习以及数据挖掘软件。它和它的源代码可在其官方网站下载。WEKA 能承担对数据进行预处理, 分类、回归、聚类、关联规则以及在新的交互式界面上的可视化。而开发者则可使用 Java 语言, 利用 Weka 的架构上开发出更多的数据挖掘算法。频繁项目集 $L = \{O, D, H31, I, TP\}$, 取最小置信度 $\min_Confidence$ 为 0.66。关联挖掘的结果如下:

05 级大一自动化, 时间: 2005 年 9 月 - 2006 年 7 月

表3 自动化05大一强关联规则

关联规则	置信度	期望置信度	作用度
$H31 = > I$	93/114 = 82%	114/315 = 0.36	2.27
$O, H31 = > I$	67/85 = 79%	85/315 = 0.27	2.92
$O = > I$	127/167 = 76%	167/315 = 0.53	1.43
$H31 = > O$	85/114 = 75%	114/315 = 0.36	2.08
$H31, I = > O$	67/93 = 72%	93/315 = 0.30	2.40

读者数 $N = 315$, 最小支持度为 0.2, 得到频繁三项集 $L3 = \{O, I, H31\}$ 。自动化05级大一强关联规则见表3。

05级大三自动化, 时间: 2007年9月 - 2008年7月

表4 自动化05大三强关联规则

关联规则	置信度	期望置信度	作用度
$H31, I = > TP$	37/43 = 86%	43/293 = 0.15	5.73
$H31 = > TP$	62/74 = 84%	74/293 = 0.25	3.36
$I = > TP$	101/145 = 70%	145/293 = 0.49	1.42

读者数 $N = 293$, 最小支持度为 0.1, 得到频繁三项集 $L3 = \{TP, I, H31\}$ 。自动化05级大三强关联规则见表4。

05级大一知识产权, 时间: 2005年9月 - 2006年7月

表5 知识产权05大一强关联规则

关联规则	置信度	期望置信度	作用度
$H31, D = > I$	37/41 = 90%	41/156 = 0.26	3.46
$H31 = > I$	47/53 = 89%	53/156 = 0.34	2.61
$D = > I$	91/110 = 83%	110/156 = 0.71	1.17
$H31, I = > D$	37/47 = 79%	47/156 = 0.30	2.63
$H31 = > D$	41/53 = 77%	53/156 = 0.34	2.26
$H31 = > D, I$	37/53 = 70%	53/156 = 0.34	2.05
$I = > D$	91/133 = 68%	133/156 = 0.85	0.8

读者数 $N = 156$, 最小支持度为 0.2, 得到频繁三项集 $L3 = \{D, I, H31\}$ 。知识产权05级大一强关联规则见表5。

05级大三知识产权, 时间: 2007年9月 - 2008年7月

表6 知识产权05大三强关联规则

关联规则	置信度	期望置信度	作用度
$H31, D = > I$	29/39 = 74%	39/172 = 0.23	3.21
$H31 = > I$	31/43 = 72%	43/172 = 0.25	2.88
$D = > I$	97/140 = 69%	140/172 = 0.81	0.85
$H31, I = > D$	29/31 = 94%	31/172 = 0.18	5.19
$H31 = > D$	39/43 = 91%	43/172 = 0.25	3.64
$H31 = > D, I$	29/43 = 67%	43/172 = 0.25	2.68
$I = > D$	97/127 = 76%	127/172 = 0.74	1.03

读者数 $N = 172$, 最小支持度为 0.1, 得到频繁三项集 $L3 = \{D, I, H31\}$ 。知识产权05级大三强关联规则见表6。

关联挖掘得出的结果与图书馆实际工作及读者调查相比较, 结果是很相近的。现选择典型的加以说明。

(1) 根据所选取的关联规则最小支持度, 机电工程与自动化学院1, 3年级得到的频繁三项集分别为 $L3 = \{O, I,$

$H31\}$ 和 $L3 = \{TP, I, H31\}$ 。从实际情况来看: 自动化专业的学生在整个大学的学习过程中一般很少有借阅政治法律类书籍的需要, 大一的学生对数学等基础课程的图书借阅比较集中而对自动化及计算机技术类书籍的借阅量相对少, 到了大三随着基础课程的结束和专业课的开设学生对数理学科和化学类书籍的借阅急剧减少, 对自动化及计算机技术类书籍的借阅量却大大增加了。

(2) 知识产权学院1, 3年级所得到的频繁三项集没有变化 $L3 = \{D, I, H31\}$, O (数理学科和化学类) 和 TP (自动化及计算机技术类) 不参与各年级的关联规则的运算。这个结果也是很显然的。

(3) 表3~表6中, $H31$ 英语类书籍出现在较多的强关联规则里, 从宏观上来说英语是学校工科、文科各年级的主要借阅书籍。从另一个角度看, 整个大学期间学生在外语上花费了大量的时间和精力。

(4) 表4中, $H31 = > TP$, $I = > TP$ 的作用度分别是 3.32 和 1.41, 表明自动化大三三年级期间, 相对于文学书籍而言, 外语类书籍与专业书籍相关性更高些。在表6中, 因为 $I = > D$ 的作用度小于 1, 所以文学类书籍与法律类书籍的关联是无效的。

(5) 用图表分析后, 表5和表6中的 $D = > I$, $I = > D$ 作用度的变化, 我们可以解释为由于大三专业课的增加, 知识产权学院的学生相对于大一借阅法律书籍数量大大增多, 而借阅文学书籍的学生稍有减少。

3 结束语

数字图书馆的流通信息为我们提供的是最基础的原始的数据, 通过对流通数据的关联挖掘, 不仅能揭示隐藏在大量数据后的重要关系信息, 同时也为这种关系提供了量化描述手段。这些定性定量的信息不仅能对图书馆的各项提供技术上的支持, 还可对学校的教学, 课程的设置, 学科的交叉渗透等提供信息。从表3~表6中我们得到了许多强关联规则, 数据挖掘工具能够发现满足条件的关联规则, 但它不能判定关联规则的实际意义。对关联规则的理解需要熟悉业务背景, 丰富的业务经验对数据有足够的理解, 也可以通过筛选技术排除虚假规则, 只有这样才能去其糟粕, 取其精华, 充分发挥关联规则的价值。

参考文献

- [1] 陆觉民, 郑宇. 基于矩阵的数据挖掘技术在数字化图书馆中的应用 [J]. 现代情报 2007, 27 (12): 92-93, 98.
- [2] 魏育辉, 潘洁. 图书流通数据的关联挖掘量化分析方法 [J]. 现代情报, 2005, (11): 108-110.
- [3] 鲍静, 范生万. 基于数据挖掘的图书数据预处理大学 [J]. 图书情报学刊, 2008, 26 (2): 31-33.
- [4] 王伟, 张征芳, 王明海. 基于数据挖掘的图书馆读者行为分析 [J]. 现代图书情报技术, 2006, (11): 51-54.
- [5] 李虹. 面向用户的数字图书馆信息服务模式研究 [J]. 情报杂志, 2007, (8): 134-136.