Hans Journal of Data Mining 数据挖掘, 2012, 2, 25-31
http://dx.doi.org/10.12677/hjdm.2012.23006 Published Online July 2012 (http://www.hanspub.org/journal/hjdm)

Application of Data Mining Technology in Library Circulation Services

Lixia Zhang[1], Haiming Gu[2*]

[1]Binzhou Medical University, Yantai
[2]Qingdao University of Science & Technology, Qingdao
Email: bzzlx@163.com, *guhm@ns.qd.sd.cn

Received: Jun. 19th, 2012; revised: Jul. 2nd, 2012; accepted: Jul. 4th, 2012

Abstract: The data mining program based on the classic Apriori algorithm was designed in this paper. It has gone through the pretreatment data mining by using this mining program to mine those data. In order to make the obtained association rules true, reliable, adopting supporting calculation methods on the same set of data used in the mining process, multiplying the runs of different granularity of data refinement combined with detailed Readers information on the association rules and obtaining comprehensive analysis, finding some valuable relationships between them. Using the obtained association relations to guide the library circulation service, constantly opening up the depth and breadth of circulation and service work, gradually increasing readers' satisfaction toward our work.

Keywords: Data Mining; Association Rules; Apriori Algorithm; Circulation Service

数据挖掘技术在图书馆流通服务中的应用

张丽霞[1]，顾海明[2*]

[1]滨州医学院，烟台
[2]青岛科技大学，青岛
Email: bzzlx@163.com, *guhm@ns.qd.sd.cn

收稿日期：2012 年 6 月 19 日；修回日期：2012 年 7 月 2 日；录用日期：2012 年 7 月 4 日

摘　要：本文根据经典 Apriori 算法思想编写了数据挖掘程序，并利用此挖掘程序对经过预处理后的数据进行了挖掘。在挖掘过程中为保证获得的关联规则真实、可靠，对同一组数据采用了不同的支持度计算方法，不同的数据细化粒度多次运行，结合详细的读者借阅信息对获得的关联规则进行综合分析，发现了一些有价值的关联关系。将获得的关联关系用于指导图书馆的流通服务工作，不断开拓流通服务工作的深度和广度，逐步提高读者满意度。

关键词：数据挖掘；关联规则；Apriori 算法；流通服务

1. 引言

数据挖掘[1,2]的概念是上世纪 80 年代被提出的，90 年代我国开始对其研究，现已广泛应用在各领域。数据挖掘(Data Mining)就是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。数据挖掘是一门广义的交叉学科，它把人们对数据的应用从低层次的简单查询提升到从中挖掘有用的信息并为人们提供决策支持。数据挖掘技术不仅在商业界备受青睐，它还延伸到图书馆的建设领域当中，目前，国际上已经将图书馆的信息服务纳入电子商务行为，其广阔的应用前景也备受瞩目[3,4]。但是国内大量的有关数据挖掘技术的研究讨论大都局限在理论上，研究具体实现的较少，其在图书馆服务领域

*通讯作者。

的具体应用也是理论讨论多，具体实践少。数据挖掘技术在图书馆的应用主要体现在信息资源的优化建设、文本数据的自动化处理、信息服务质量的提升、业务范围的拓展等方面[5]，并显示出强大的生命力。

随着数据库技术的迅速发展和数据库管理系统在图书馆的广泛应用，图书馆数据库中积累了大量读者访问图书馆资源的历史记录，这些数据背后隐藏着许多重要的信息。但目前的图书馆管理系统缺乏对数据的深层次处理，通常只提供简单的查询和统计等功能，无法发现这些数据背后潜在的有用的关系和规则，无法利用这些关联规则预测读者的信息需求。数据挖掘技术可以实现对流通数据库的数据进行挖掘，并对挖掘出的结果进行综合分析和评价，为图书采购、读者导读等服务提供信息支持，为优化馆藏结构、馆藏布局等提供决策支持。如何从图书馆管理系统存储的大量读者借阅记录中提取有用的信息用于指导流通服务工作，提高读者服务的深度和广度，逐步提高读者满意度，成为高校图书馆流通服务工作者的一个非常有意义的研究课题。本文第二节讲述了数据预处理的必要性及方法，并据此对待挖掘数据进行了预处理；第三节基于经典的 Apriori 算法思想，编写了数据挖掘程序，利用挖掘程序对待挖掘数据进行处理。为使获得的规则真实有效，在挖掘过程中采用不同的支持度定义来对获得关联规则进行有效性验证。为进一步发现各类图书之间隐含的关系，对待挖掘数据进行了不同粒度的处理；第四节结合详细的读者借阅记录对挖掘结果进行了综合分析，并找出各类图书之间真实有效的联系；最后对该文的工作进行了总结。将获得的真实有效的关联规则具体运用到图书馆的流通服务工作中。

2. 图书馆流通数据的预处理

数据预处理(data preprocessing)[6,7]是指在对数据进行主要的数据挖掘处理之前，先对原始数据进行必要的清洗、集成、转换和归约等一系列的处理工作，以达到挖掘算法进行知识获取研究所要求的最低规范和标准。

2.1. 数据预处理的必要性

实际采集到的原始数据一般是含噪声的、不完整的或不一致的，我们需要在数据挖掘之前先对原始数据进行预处理，提高数据质量，使之符合挖掘算法的规范和要求。据统计发现：在整个数据挖掘过程中，数据预处理花费整个项目 60%左右的时间，而后的挖掘工作只占整个工作量的 10%左右。经过数据预处理后再进行挖掘，不仅可以节约大量的时间和空间，而且得到的挖掘结果能更好地服务于决策和预测。

2.2. 数据预处理的主要方法

常见的数据预处理方法有：数据清洗、数据集成、数据转换和数据归约。

以上几种方法并不是各自独立的，而是相互关联的。例如，冗余数据的删除既是一种数据清理形式，也是一种数据归约，而往往做完数据集成之后还需要再次进行数据清理工作。

2.3. 数据的选取

数据的选取是进行数据挖掘时很重要的一步，由数据源中选取出合适的数据，并且了解数据的形式，才能更好的处理问题。为了有效的执行数据挖掘，本文选取某医学院校图书馆自 2011 年 9 月至 2011 年 11 月的 15,000 条读者借阅记录明细作为分析的数据源。

2.4. 数据预处理过程

本文主要研究学生读者借阅过的各类图书间的关联关系。通过对图书馆管理系统数据库进行分析，确定需要的主要事务数据库有读者库(表)、流通库(表)、馆藏书目库(表)和馆藏典藏库(表)四个数据库(表)。利用相关字段进行关联，生成新的事物数据库，结构如表 1。在新生成的数据库中去掉借阅记录数为 1 的记录(1314 条)，因为一个借阅者只有一条借阅信息(只借阅过一本图书)，无法形成图书之间的关联关系，在后续的挖掘工作中不能产生关联规则，可以去掉。这种数据形式还不适合挖掘，还需将数据进行转换和规约，将表 1 中同一读者的记录合并成一条记录，如表 2 所示。

3. 数据挖掘算法的实现

3.1. Apriori 算法的基本思想及特点[1,2,8,9]

该算法的基本思想：

Table 1. The new transaction database
表 1. 利用字段关联后生成的新事务数据库

条形码	读者条码	姓名	性别	单位	索书号
2601585	000253	张文文	女	护理学院	R47-42/1
2300019	001150	张月	男	临床医学系	R155/5
2751288	001150	张月	男	临床医学系	R151.4/44

Table 2. Final data for datamining form
表 2. 最终适合挖掘的数据形式

读者条码	借阅书目
030806	T,R,H,I
031356	R,H,Z,J
031457	B,K,I,Z

第一步，统计所有含一个元素项目集出现的频率，并找出一维最大项目集。

第二步，用递推的方法生成所有频集。递推过程是：第 k 步中，利用第 k – 1 步生成的(k – 1)维最大项目集产生 k 维侯选项目集，然后对数据库进行搜索，得到侯选项目集的项集支持度，与最小支持度比较，从而找到 k 维最大项目集。

第三步，使用找到的频集产生期望的规则，保留那些不小于用户给定的最小可信度(支持度)的规则。算法流程图见图 1。

Apriori 算法在理论上比较简单，但它存在以下缺点：

1) 在每一步产生侯选项目集时循环产生的组合过多，没有排除不应该参与组合的元素。

2) 每次计算项集的支持度时，都对数据库 D 中的全部记录进行了一遍扫描比较，如果是一个大型的数据库的话，这种扫描比较会大大增加计算机系统的 I/O 开销，而这种代价是随着数据库的记录的增加呈现出几何级数的增加。

因 Apriori 算法存在上述缺点，所以人们寻求了一些能减少这种系统 1/O 开销的更为快捷的算法，即 Apriori 的改进算法。例如 AprioriTid 和 AprioriPro 就是 Apriori 的两种改进算法。这些算法的思路基本上与 Apriori 算法保持一致，大都采用同 Apriori 同样的产生候选集的思想。两者的不同之处在于改进算法的侧重点主要有：减少扫描数据的次数；减少产生的候选集数目；减少候选频繁项集的计算时间。因本文挖掘程序处理的数据量不大，所以仍基于 Apriori 算法来实现。

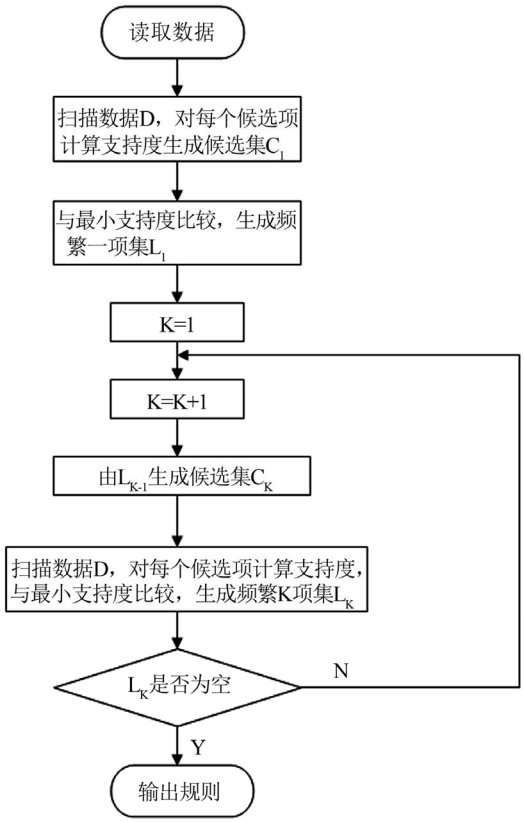

Figure 1. Apriori algorithm flowchart
图 1. Apriori 算法流程图

3.2. 算法举例

现有含有四条读者借阅记录的图书借阅记录集，如表 3 所示。“读者条码”为借阅者的唯一标识，“借阅书目”为读者在一段时间内的图书借阅情况。

在 Apriori 算法中每一次循环先找出该次的侯选集，并统计每个侯选项目集的支持度，再和预先定义的最小支持度比较，找出该次循环的频繁项目集。算法执行示意图如图 2 所示。

Table 3. Readers record of borrowing
表 3. 读者借阅记录

读者条码	借阅书目
031089	A, H, R
031092	B, C, H
031100	C, I, H
031105	B, C, I, H

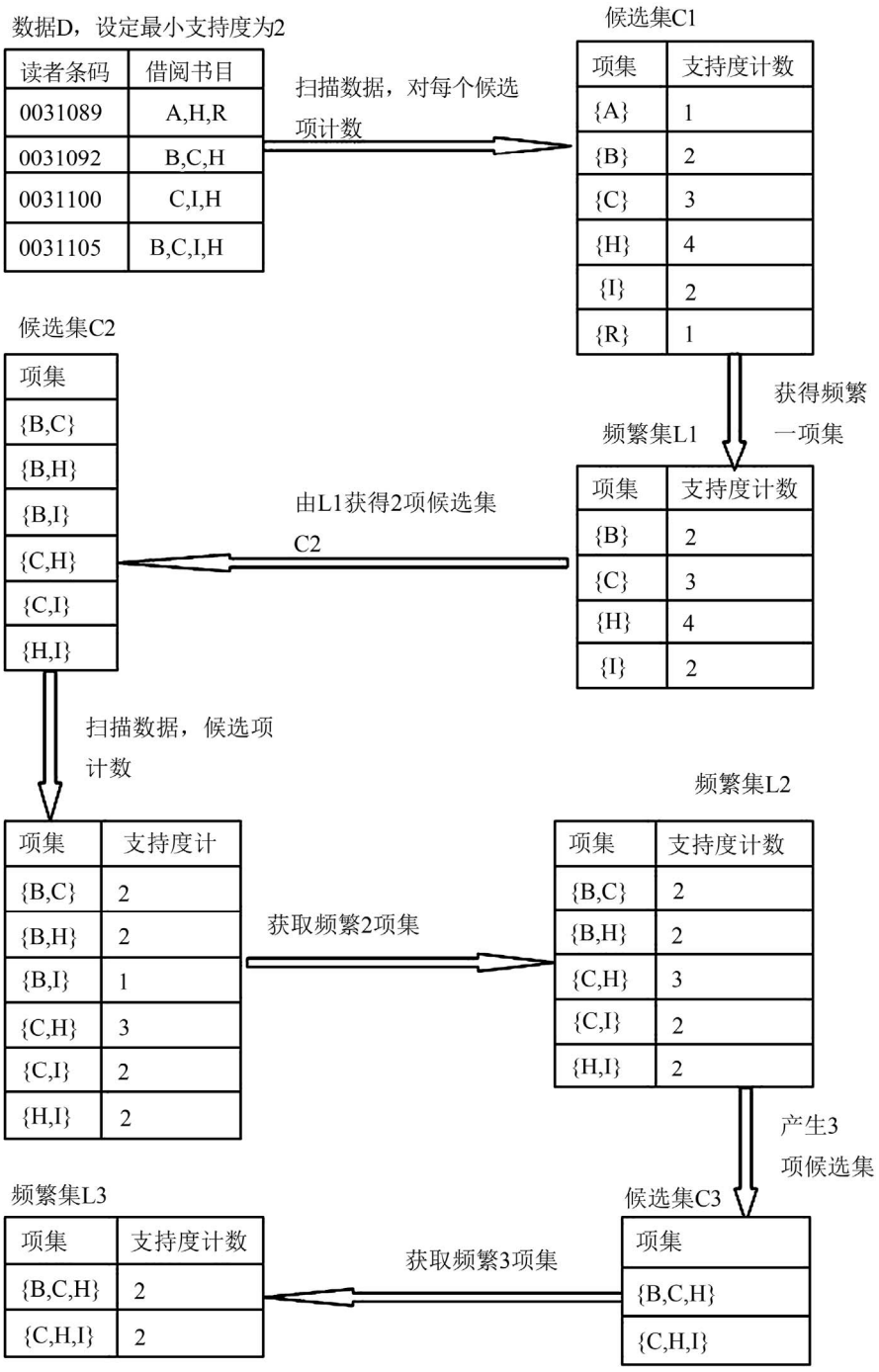

Figure 2. Sketch of algorithm implementation
图 2. 算法执行示意图

4. 实验结果与分析

设定最小支持度为 4，最小置信度 0.6，运行挖掘程序得到如下 5 条关联规则。

规则 1：{Q 生物科学}→{R 医药、卫生}(置信度 = 74.2%)。

规则 2：{I 文学、Q 生物科学}→{R 医药、卫生}(置信度 = 66.67%)。

规则 3：{B 哲学、宗教、O 数理科学和化学}→{H 语言、文字}(置信度 = 66.67%)。

规则 4：{T 工业技术、Q 生物科学}→{R 医药、卫生}(置信度 = 68%)。

规则 5：{J 艺术、Q 生物科学}→{R 医药、卫生}(置信度 = 100%)。

这些规则是否真实的反映了这种借阅关系，还需结合实际进行综合分析。有些关联规则可信度虽然很高，但支持度却很低，说明该关联规则实用的机会很小；相反，如果支持度很高，可信度很低，则说明该规则不可靠。因此，还需要根据实际情况来对获得的关联规则进行具体分析。

规则 1：{Q 生物科学}→{R 医药、卫生}，因为医学专业中生物学和医药、卫生学是紧密联系的，生物学和医药、卫生学都是医学相关专业学生的基础课程，结合具体的流通借阅记录进行分析得出该规则是有效的。

规则 2：{I 文学、Q 生物科学}→{R 医药、卫生}，I 类图书为文学类图书，包括世界文学、中国文学(包括小说、诗歌、散文等)及世界各国文学。该规则反映出了学生在借阅专业课书籍的同时，也会伴随文学类图书的借阅。

规则 3：{B 哲学、宗教、O 数理科学和化学}→{H 语言、文字}，在选取的整个读者借阅数据中，B 哲学、宗教类图书的借阅记录有六百余条，此规则在整个图书借阅事务中所占的比例较小，属小概率事件，当调高了最小支持度后，该规则可能就不再出现。

规则 4：{T 工业技术、Q 生物科学}→{R 医药、卫生}，结合具体的学生借阅记录发现在该条规则中 T 工业技术类图书主要是 TP3 类(包括计算机应用类、计算机软件类)、TS9 类(主要包含 TS94 服装工业、制鞋工业，TS97 生活服务技术类)。这也从一个侧面说明了在计算机、互联网等信息技术飞速发展的情况下，现在的大学生在学习专业课程的同时也会积极的学习掌握一些信息技术，同时对饮食、衣着、美容护肤等也有一定的关注度。进一步分析发现借阅 TP3 类图书的学生中多数是男生，而借阅 TS9 类图书的学生中多数是女生。

规则 5：{J 艺术、Q 生物科学}→{R 医药、卫生}，从图 2 可以看出该规则的置信度很高。结合学生详细的借阅记录发现在选取的数据源中有六十余条 J 类图书的借阅记录，借阅的艺术类图书主要包括书法类、绘画类、摄影类、音乐类等，读者分布在不同的学院、不同的专业，这说明部分读者有这些方面的业余爱好或兴趣，该规则是符合常理的。但是在整个借阅记录中艺术类图书的借阅是一个低概率事件，若在运行时选定的最小支持度大一些，就不会得到该规则，但在图书推荐或读者导读时有意识的向读者推荐一下艺术类图书，也可能起到意想不到的效果。

4.1. 挖掘条件的修正

支持度的两种定义形式

1) 含 X 的交易的个数为 X 在 D 中的支持度。

2) 假定 X 是一个项目集，D 是一个交易集合或交易数据库，称 D 中包含 X 的交易的个数与 D 中总的交易个数之比为 X 在 D 中的支持度。

这两个不同的概念使用在不同的场合，但其内在含义是一致的。而后一个定义使用得更广泛，因为它是一个规格化的概念，保证了支持度的范围在 0 到 1 之间。上面程序运行时输入的最小支持度为 4，这是一个绝对的数值，相对于最终挖掘的事物记录数量来说这个值是很低的，其缺点是容易得到一些实际上出现几率很小的规则。下面设最小支持度为 0.01，最小置信度为 0.5，对相同的数据进行处理，得到如下 5 条关联规则

规则 1：{H 语言、文字}→{R 医药、卫生}(置信度 = 56.35%)。

规则 2：{Q 生物科学}→{R 医药、卫生}(置信度 = 74.2%)。

规则 3：{H 语言、文字、Q 生物科学}→{R 医药、卫生}(置信度 = 53.26%)。

规则 4：{I 文学、Q 生物科学}→{R 医药、卫生}(置信度 = 66.67%)。

规则 5：{T 工业技术、Q 生物科学}→{R 医药、卫生}(置信度 = 68%)。

规则 2、规则 4、规则 5 与前面获得的规则相同。对规则 1 结合学生的借阅记录看出，这里的 H 类图书主要是 H31 英语类图书，包括英文小说、英语四六级考试书籍、英文口语、考研英语等。该规则说明学生在图书馆借阅英语类书籍和专业相关书籍的相互关联性较大。经综合分析得知该学校的实际情况也确实如此，全校从上到下都很重视英语课的学习，所以该规则是切实可信的。规则 3 包含规则 1 和规则 2，说明英语类、医学生物类和医疗卫生类图书同时借阅的

可能性大。

4.2. 挖掘数据的细化

前述挖掘程序处理的数据，是选取的图书分类号的第一位即中图分类法的一级类目，得到的关联规则体现的也是一级类目之间的借阅关联关系。为了进一步挖掘各类图书借阅的关联关系，现选取图书分类号的二级类目生成待挖掘数据。设定最小支持度为 1%，最小置信度为 50%，修改程序，计算并显示每条规则的支持度，共生成 8 条关联规则：

规则 1：{I3 各国文学}→{I2 中国文学}(支持度 = 3.43% 置信度 = 56.35%)。

规则 2：{R9 药学}→{R3 基础医学}(支持度 = 10.82% 置信度 = 74.2%)。

规则 3：{R6 外科学}→{R5 内科学}(支持度 = 6.94% 置信度 = 53.26%)。

规则 4：{Q5 生物化学}→{R3 基础医学}(支持度 = 11.97% 置信度 = 66.67%)。

规则 5：{I1 世界文学}→{I2 中国文学}(支持度 = 2.71% 置信度 = 68%)。

规则 6：{R9 药学、H3 常用外国语}→{R3 基础医学} (支持度 = 3.29% 置信度 = 53.26%)。

规则 7：{H3 常用外国语、R6 外科学}→{R5 内科学} (支持度 = 1.6% 置信度 = 66.67%)。

规则 8：{H3 常用外国语、Q5 生物化学}→{R3 基础医学} (支持度 = 3.43% 置信度 = 68%)。

从规则 1、规则 5 可以看出 I3 各国文学类图书和 I2 中国文学类图书之间存在关联，I1 世界文学类图书和 I2 中国文学类图书之间也存在关联，这说明读者的阅读兴趣比较广泛。从规则 2、规则 3、规则 6 和规则 7 可以看出，R3、R5、R6 和 R9 类图书间存在着明显的借阅联系，同时可以看到在借阅专业类图书的同时外语类图书(主要是英语类)的借阅也伴随发生。从规则 4，规则 8 可以看出生物科学和医学联系较为紧密，尤其是在学生学习基础课程阶段基础医学类图书和生物科学类图书同时借阅的可能性较大。

和前面的结果比较发现少了以 T 工业技术类为前项的关联规则，分析后发现 T 工业技术类图书，在分类号数据细化后主要转化为 TP 和 TS 两部分，各自的支持度都小于最小支持度，被过滤掉了。

在实际应用中我们还可以根据特定的需求将某段范围内的数据进一步细化，发现更多隐含的规律。例如：对上面 TS 类图书的借阅信息挖掘后发现主要是 TS941(服装工业)类图书和 TS974.1(美容)类图书，并且绝大多数的借阅者为女生。

5. 总结

通过挖掘程序获得的关联规则反映了图书借阅之间的表面联系，但不一定表示它们之间存在必然因果关系。规则是僵化的，规则的使用者是灵活的，需结合实际情况分析判断是否合理。根据挖掘获得的规则，结合具体的借阅记录进行综合分析，获得了以下几点有益的结论用于改进、提高图书馆的流通服务工作。

1) 合理配置文献购置费用

根据获得的关联规则使有限的图书购置经费在各学科间合理分配。对有关联关系的图书进行关联采购，既能提高读者的满意度又能提高图书的流通率。

2) 在读者导读中的应用

图书馆读者服务的重要内容之一就是读者导读，若将挖掘结果正确地运用到读者导读工作中，可以大大提高读者的满意度。例如：利用前面挖掘得到的关联规则{R9 药学}→{R3 基础医学}、{R6 外科学}→{R5 内科学}，当读者借阅药学类图书时为他推荐基础医学类图书；当读者借阅外科学类图书时为其推荐内科学类图书。也可以在书库内安装查询终端，使读者可以方便的查询自己感兴趣的图书，并同时主动给出和该图书相关联的其它类图书信息，方便学生借阅。

3) 以上面提到的图书馆为例，目前此图书馆是以分类号为依据划分了四个书库，分别是医学书库、外文与自然科学书库、文学书库和社科书库。这种书库划分方式分类清晰，方便管理，但是从便利读者的角度来看，未必是一种最好的书库划分方式。从以上获得的关联规则可以看出，读者在借阅专业类(医学类)图书的同时，往往伴随有文学类图书、外语类图书的借阅。以目前的书库划分，读者至少需要到三个书库中去查找才能借到自己需要的图书。若在书库排架时不是按照图书类别顺序排放，而是将借阅相关性大的图书排放在相邻位置，比如英文类、文学类图书可以和医药类图书相邻，这样就能大大节省读者的时间，

更加方便读者，同时又能提高图书的流通率。但是也应看到，此种排架方式虽然方便了读者，但对管理者来说是一个挑战。

参考文献 (References)

[1] 韩家炜[加], 堪博(Kamber. M), 著, 范明, 孟晓峰, 译. 数据挖掘: 概念与技术[M]. 北京: 机械工业出版社, 2007: 3-18.

[2] R. Agrawal, T. Imielinski and A. N. Swanmi. Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 1993: 207-216.

[3] 朱晓华. 浅析数据挖掘技术在图书馆自动化中的应用[J]. 图书馆学研究, 2002, 5: 41-42.

[4] 鲍翠梅, 王尊新等. 数据挖掘技术及其在图书馆中的应用[J]. 情报杂志, 2004, 9: 49-51.

[5] 李玮平. 基于数据挖掘的图书馆读者需求分析[J]. 图书馆论坛, 2004, 24(3): 86-88.

[6] 方洪鹰. 数据挖掘中数据预处理的方法研究[D]. 西南大学, 2009.

[7] 鲍静, 范生万. 基于数据挖掘的图书数据预处理[J]. 图书情报学刊, 2008, 26(2): 31-33.

[8] 蔡伟杰. 关联规则挖掘综述[J]. 计算机工程, 2007, 21(5): 31-33.

[9] G. Li, H. J. Hamilton. Basic association rules. Proceedings 2004 SIAM International Conference on Data Mining (SDM'04), 2004: 166-177.

