第 24 卷 第 4 期
2009 年 8 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol.24 No. 4
Aug. 2009

基于关联规则挖掘的课程相关性研究与应用

吴江红
（天津科技大学计算机科学与信息工程学院，天津 300222）

摘 要：应用关联规则挖掘对高校课程相关性进行了研究. 将某高校的毕业生成绩数据库经过预处理之后，采用不设定成绩界限的方法，用改进的 Apriori 算法进行挖掘. 不仅能挖掘出成绩为优时的课程相关规则，还能发现若某些课程成绩差，则其他课程成绩也为差的规则，可以为学分制体系下学生选课和管理者进行决策等提供参考.

关键词：数据挖掘；关联规则；最小支持度；课程相关性
中图分类号：TP311 **文献标志码**：A **文章编号**：1672-6510(2009)04-0073-03

Research and Application on Correlation Between Courses Based on Mining Association Rules

WU Jiang-hong
(College of Computer Science and Information Engineering, Tianjin University of Science & Technology, Tianjin 300222, China)

Abstract: By using mining association rules, correlations between courses were studied. Not using method of achievement limit and using the improved Apriori algorithm, the achievement database of students can provide some interesting rules of correlation between courses of excellent grade after pretreatment. At the same time, the rules that some courses of low grade are the cause of other courses of low grade were discovered. The results can guide the students to select subjects and help the leaders to make decision.

Keywords: data mining; association rules; minimum support; correlation between courses

随着计算机在高校的应用，高校教学管理系统积累了大量的有关教学信息数据. 同时，由于教学规模的扩大，授课教师和相关人员很难再直接通过学生的成绩分布情况找出先修课程与后继课程之间的关系，无法指导教学. 因此需要借助于关联规则挖掘技术，发现数据中潜在的课程相关性，并为决策提供支持. 进而，根据课程之间的相关性得出一些具有价值的规则和信息，以指导学生选课.

目前，课程相关性研究中大部分都集中在挖掘某些课程成绩为优时，其他课程也为优的规则. 在这种情况下，不能发现一些教学和学习中存在的不足和问题[1–3]. 采用的关联规则挖掘算法有 Apriori 算法[1–3]和 AprioriTID 算法[4]，这些挖掘算法挖掘时间长，算法效率较低. 文献[5]对改进的 Apriori 算法进行了研究，在挖掘速度和性能上有所提高.

本文以某高校计算机学院连续三届毕业生的成绩库为数据集，在进行预处理之后，利用文献[5]方法进行课程相关性的挖掘，以期挖掘出一些有意义的规则. 由于采用不设定成绩界限的方法，不但能发现多数已知的结论，同时还能发现某些课程成绩为差，则其他的课程成绩也为差的规则，可用来指导学生的选课和学习.

1 关联规则和改进的 Apriori 算法

关联规则[6–7]是在交易数据等数据中，查找存在于项目集之间的关联等，通过分析数据知道哪些事情将一起发生，此类信息可指导决策者决策.

收稿日期：2008–11–28；**修回日期**：2009–04–21
作者简介：吴江红（1976—），女，甘肃天水人，讲师，wujh@tust.edu.cn.

1.1　问题描述

设 $I=\{i_1,i_2,\cdots,i_n\}$是物品标识的集合；T 是物品的集合，且 $T\subseteq I$；D 为交易 T 的集合，即交易数据库.

关联规则是形如 $X\Rightarrow Y$ 的蕴涵式，其中 $X\subset I$，$Y\subset I$，且 $X\cap Y=\Phi$. 规则 $X\Rightarrow Y$ 在交易数据库 D 中的支持度是交易集中同时包含 X 和 Y 的交易数与所有交易数之比. 可信度是同时包含 X 和 Y 的交易数与只包含 X 的交易数之比.

1.2　关联规则挖掘方法

关联规则挖掘的任务是：在数据库 D 中找出支持度和可信度分别大于等于相应阀值的关联规则，从而指导决策者决策. 该任务可以分解为两个问题：(1)求 D 中满足相应阀值的所有频繁项目集；(2)利用频繁项目集生成满足相应阀值的所有关联规则.

问题(2)较易解决，而问题(1)的解决是整个关联规则挖掘的核心部分. 最经典的关联规则挖掘算法是 Apriori 算法，但是其挖掘效率比较低，因此本文采用改进的 Apriori 算法来挖掘频繁项和关联规则.

1.3　改进的 Apriori 算法

Apriori 算法效率较低的主要原因是：需要对整个数据库进行多次扫描；并且在修剪(Prune)中需进行多次的模式匹配. 针对这两个问题对 Apriori 算法进行了改进. 主要是变换了数据结构，即将普遍采用的水平结构变换为项目事务垂直对应关系的数据结构. 采用的垂直结构由项目及包含该项目的二进制事务集构成. 采用二进制一方面不用多次扫描数据库，从而降低了对存储空间的要求. 此外，二进制的运算速度要远远快于字符串的运算速度.

2　关联规则挖掘在课程相关性中的应用

挖掘课程相关性的研究目标是发现课程之间的相关性，得到课程的先修课和后继课的信息. 因此，挖掘结果可以作为对教务管理人员有指导意义的参考，为决策提供重要依据. 在学生自主选课的情况下，可以为学生提供一个科学的选课指导.

2.1　数据预处理技术

现实世界中数据多数都是不完整、不一致的“脏数据”，不能直接用于数据挖掘，为了提高数据挖掘的质量，产生了数据预处理技术.

数据预处理技术有：数据清理、数据集成、数据归约等. 在数据挖掘之前使用这些数据处理技术，可以提高数据挖掘模式的质量，降低实际挖掘所需要的时间. 在本学生成绩数据库中，用到的数据预处理方法主要包括数据清理、数据集成和数据归约.

2.1.1　*数据清理*

数据清理用于处理数据中的遗漏和清洗脏数据，主要处理的是空缺值，平滑噪声，识别、删除孤立点. 在学生成绩数据库中，某些成绩为零，可能因为缺考，也可能就考了零分. 为了挖掘的准确性，需要删除这样的数据. 此外，为了使挖掘结果更准确，对于有补考成绩的不采用补考成绩，仍然以原始成绩为处理数据.

2.1.2　*数据集成*

在数据挖掘之前需要对数据进行集成，即将多个不同数据源中的数据合并、存放在一个统一的数据存储(如数据库)中，数据源可以是多个数据库、数据立方体等. 例如将 Access 和 SQL Server 等数据源中的数据存放在一个 Oracle 数据库中.

2.1.3　*数据归约*

数据归约用来得到数据集的归约表示，它接近于保持原始数据的完整性，但数据量比原始数据小. 如果在归约之后的数据上进行挖掘，所需要的资源和时间都更少，挖掘结果更有效，并能产生相同或很相近的分析结果. 学生成绩数据库中的成绩为 0 至 100 之间的任意一个数. 由于成绩的取值较多，此时相当于寻找无意义的小概率事件. 为了使挖掘结果有意义，需要将其进行离散化和归约，按照成绩段离散化为 4 类：85 ~ 100 分为 A 类；70 ~ 84 分为 B 类；60 ~ 70 分为 C 类；60 分以下为 D 类.

在改进的算法中，事务数据库使用横向数据结构，如学生成绩数据库中的每个学生为一个事务，该事务包含此学生的所有课程成绩数据. 但是，此结构不符合通常的管理系统结构，通常在成绩数据库中采用如表 1 所示的纵向结构. 表 1 中只列出了部分数据和属性，在表 1 中每个学生可以有多条记录，每条记录包含有相对较多的信息.

表 1　部分成绩表

Tab.1　Partial achievement table

学号	学院编号	专业编号	学期	课程编号	成绩	…
04101101	10	1	06-07-1	K1	89	…
04101101	10	1	06-07-1	K2	92	…
04101101	10	1	06-07-2	K3	83	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮

要应用改进的 Apriori 算法在上述数据中挖掘关联规则，需要进行数据转换. 表 2 是经过转换后的部分数据. 表中每一条记录代表一个事务，学生的学号可看成事务的 ID 号，课程及成绩属性的内容为该事

务的项目集，即该学生所有的课程和成绩，其中每一个项目为课程编号（K1）和所取得的成绩的类别. 这样转换完之后，数据量少了，并且可以直接用改进的 Apriori 算法进行频繁项的挖掘.

表 2　预处理后的部分成绩表

Tab.2　A part of achievement table after pretreatment

学号	课程及成绩
04101101	K1A,K2A,K3B,K5A,…
04101102	K1B,K2A,K5B,K6B,…
04101103	K1C,K2C,K3C,K5B,…
04101104	K1B,K2A,K4A,K5A,…
04101105	K1B,K2B,K6B,K7B,…
⋮	⋮

2.2　挖掘结果及分析

设置支持度为 10%，可信度为 60%，用改进的 Apriori 算法进行挖掘. 部分挖掘结果如图 1 所示.

图 1　部分挖掘结果

Fig.1　Partial mining results

为了便于理解，用课程名称来代替课程编号. 通过挖掘得到部分关联规则如下：

规则 1：高等数学 A、C 语言 A→算法分析 A，可信度为 69%. “高等数学的成绩为 A 类”、“C 语言的成绩为 A 类”和“算法分析的成绩为 A 类”3 个事件同时发生的可能性较大，因此可以通过加强先修课程高等数学和 C 语言的教学质量，来提高算法分析的教学效果，同时指导学生选课.

规则 2：C 语言 A、数据结构 A→操作系统 A，可信度为 67%. 说明 C 语言和数据结构是学习操作系统的先修课程，并且对操作系统的课程学习有直接的影响. 以下规则的含义相似，不再赘述.

规则 3：C 语言 A、操作系统 A→UNIX 操作系统 A，可信度为 62%.

规则 4：高等数学 B、大学物理 B→数字信号处理 B，可信度为 62%.

规则 5：高等数学 D、C 语言 D→数值分析 D，可信度为 80%. 可见，高等数学和 C 语言学不好，那么利用计算机技术解决数学问题的数值分析也学不好. 同时还发现了一些无法解释的规则，将其删除.

从以上规则中可以看出，C 语言和高等数学对于计算机专业的学生是相当重要的. 通过以上分析也验证了计算机专业课程设置的合理性. 发现的有用规则为计算机专业课程间的相关性. 规则中的条件都是该专业的基础课程，必须在低年级开设，以保证学生学习的连续性.

根据得到的规则可以为学生选课提供必要的指导. 根据规则，学生可以知道某门课程的先修课程有哪几门，还可以通过认真学习该门课的先修课程来提高此课成绩. 管理者也可以通过增加其先修课程的课时，减少本课课时的方法来达到提高成绩的目的.

3　结　语

数据挖掘在商业等盈利机构中应用的较多，也很成功，但是在高校、政府等一些非盈利性机构中应用的很少，本文对关联规则在课程相关性的应用做了研究，使用改进的 Apriori 挖掘算法进行课程相关性分析，得出了有效的规则.

本文只对计算机科学与技术专业的学生成绩库进行了挖掘，下一步可以应用其对其他专业进行挖掘，以指导课程设置和学生选课.

参考文献：

[1] 李雪婵. 关联规则在课程相关性中研究与应用[J]. 计算机与数字工程，2006，34(9)：173–176.

[2] 王员根，李爱凤. 关联规则在课程相关性模式中的研究与应用[J]. 现代计算机，2007(2)：79–82.

[3] 陈启买，彭利宁，刘海，等. 基于关联挖掘的课程相关性模式研究[J]. 华南师范大学学报：自然科学版，2008(1)：52–59.

[4] 贾文，臧明相，周鸿. 基于数据挖掘的课程相关性研究与分析[J]. 计算机技术与发展，2006，16(12)：178–180.

[5] 吴江红，周长英，于秀丽. 一种改进的关联规则挖掘算法[J]. 天津科技大学学报，2005，20(2)：57–60.

[6] Agrawal R，Imielinski T，Swami A. Mining association rules between sets of items in large database[C]// Proceedings of ACM-SIGMOD Conference on Management of Data，Washington D C：ACM Press，1993：207–216.

[7] Agrawal R，Srikant R. Fast algorithms for mining association rules[C]//Proceedings of the 20th International Conference on Very Large Database，Santiago：Morgan Kaufmann，1994：487–499.