数据挖掘技术在远程教育教学中的应用

董彩云[1],刘培华[2]
(1.山东广播电视大学 直属学院,山东 济南 250014;
2.济南大学 信息科学与工程学院,山东 济南 250022)

摘 要:数据挖掘是一种新兴的信息处理技术,在信息的利用和提取中发挥着日益重要的作用。简要介绍了数据挖掘技术。给出了一个完整的数据挖掘系统设计与实现过程。它包括数据的准备与选择、数据的预处理、挖掘算法的选择与实现、挖掘结果的描述四个步骤。文中详细介绍了上述各个步骤的处理过程,并通过改进Apriori算法来提高挖掘的效率,为远程教育教学决策提供了科学依据。

关键词:数据挖掘;关联规则;Apriori算法
中图分类号:TP311.13 **文献标识码**:A **文章编号**:1673-629X(2009)02-0179-04

Application of Data Mining Technology in Instance Education

DONG Cai-yun[1], LIU Pei-hua[2]
(1. Open College, Shandong Broadcast TV University, Jinan 250014, China; 2. School of Information Science & Engineering, Jinan University, Jinan 250022, China)

Abstract: Data mining is a promising new technology to transact information, and becoming an increasing role in utilizing and extracting knowledge. Firstly, introduces the data mining technology briefly. An integrated datamining system is presented in this paper. It includes preparation and selection of data, preprocessing, implementation of mining algorithm, description of mining results, etc. All those will be presented in detail. More importantly, It puts forward a new enhanced method to improve the mining efficiency, so provides a scientific basis for instance education and decision making.

Key words: data mining; association rules; Apriori algorithm

0 引 言

远程教育是以学生为中心进行自主学习,采用学分制的教学管理制度。学分制主要以选课为核心,允许学员在一定范围内根据个人特长与爱好选修课程,选择合适自己的学习量和学习年限,使得学员在课程学习的过程中有了极大的自主性。让学员可以发挥主观能动性,使其个体潜力得到充分发展,可以塑造出具有相当专业能力的专门型人才但学分制的实行,也出现了一系列的问题:

一是学员自主选课缺乏有效地指导,使得学习过程缺乏连续性,缺乏专业学习所需的系统性。

二是出现部分学员在尚未学习先修课程的情况下,选择了某一后继课程,除影响学员本身的学习,也对任课教师的教学评价产生不良的影响。

1 数据挖掘技术

数据挖掘[1,2]的研究具有广泛的应用背景和深远的理论意义,是未来人工智能与数据库研究的热点前沿课题之一。DM系统[3]不是多项技术的简单组合,而是一个完整的整体,它还需要其他辅助技术的支持,才能完成数据准备、数据挖掘、结果表述、算法评价这一系列任务。根据功能,整个DM系统可以大致划分为三级结构[4](如图1所示)。将数据挖掘技术应用到远程教育教学中,能客观地反映各课程之间的关系。使用该技术所发现的知识可以方便地用于教学管理人员的决策支持,从而为学员选课提供有价值信息。

2 总体设计

在设计挖掘系统之前首先要确定主题,即从大量的教学管理数据中定位一个有意义的挖掘目标。经过调研,认为学员课程设置是关系到学员学习成绩的一

收稿日期:2008-06-10
基金项目:国家863高技术发展计划项目(2002AA4Z3240);教育部世行贷款——21世纪初高等教育教学改革项目(1283P0843)
作者简介:董彩云(1978-),女,山东东营人,讲师,硕士,研究方向为数据仓库、数据挖掘。

个关键,它的合理性直接影响到学员对课程的学习,进而影响其学习郊果。因此对课程进行相关性分析是一个有价值和有意义的研究课题,挖掘结果将为教学管理人员提供一个有指导意义的参考。主题的名称定为“课程的相关性分析挖掘”。学院有几年的采集数据,数据量非常之大,因此首先选定一个专业(计算机专业)进行课程的相关性挖掘。进而扩展到其他专业进行挖掘。

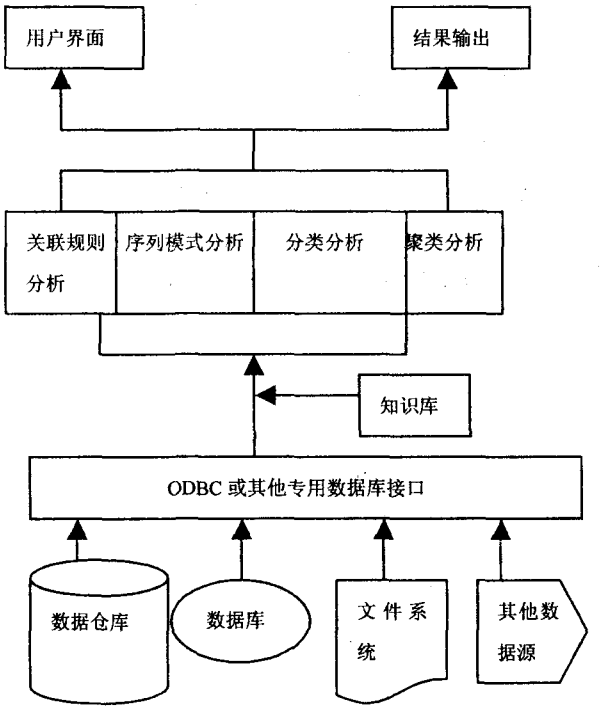


图 1 DM 系统结构图

系统的框架图如图 2 所示。

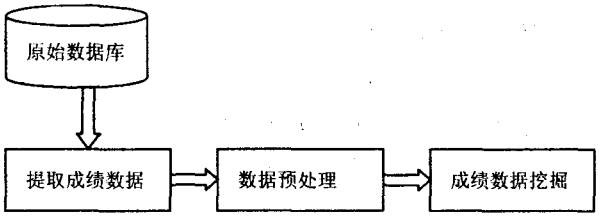


图 2 框架图

3 具体设计与实现

3.1 预处理

一般来说,计算机系统日常产生的数据并不适合直接作为数据挖掘算法的输入,在本系统中主要做了以下工作。

3.1.1 数据选择

根据用户的要求从数据库中提取与挖掘主题相关的数据,数据挖掘算法将主要从这些数据中进行知识提取。选择的原始数据如下:学员学号,各门课程成绩。

3.1.2 数据加工

对选择的数据进行加工处理,检查数据的完整性及一致性,对丢失的数据进行填补。为了保证数据的准确与完整,对下列两种情况进行了适当的处理:第一,学员补考成绩的计算问题。为了保证数据挖掘结果的准确与合理,对于不及格学员的成绩应以他们的原成绩为准,而不按其补考后的成绩计算。第二、没有正常参加考试的学员成绩的计算。对于参加正常考试科目少于应参加考试科目的一半者,该学员所有成绩均不记入挖掘数据中,对于参加正常考试科目多于应参加考试科目的一半者,该学员成绩记入挖掘数据中,并对其没有正常参加考试的科目,参照其参加补考的成绩和学员其它课程的成绩,然后给予一适当的成绩,以维护数据的完整性。

3.1.3 数据变换

表中完全相同的“属性 - 值”很少,如果直接将其作为项进行挖掘,不可能得到理想的结果。因此,做如下的数据变换:

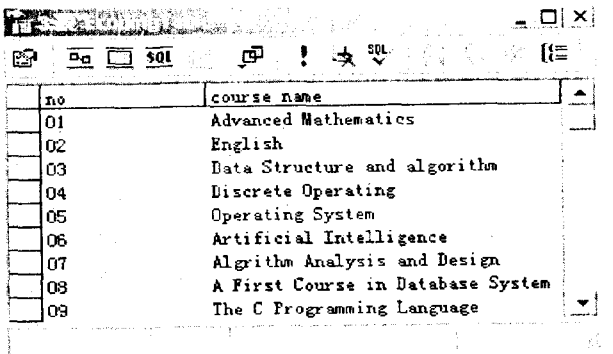
(1)成绩进行离散化。即将成绩数据变换为三个等级数据,即:

A:对应分数 80～100

B:对应分数 60～79

C:对应分数 0～59

(2)课程编码。将各课程分别以 01,02…进行编码后的部分课程编码如图 3 所示。



no	course name
01	Advanced Mathematics
02	English
03	Data Structure and algorithm
04	Discrete Operating
05	Operating System
06	Artificial Intelligence
07	Algrithm Analysis and Design
08	A First Course in Database System
09	The C Programming Language

图 3 课程编码表

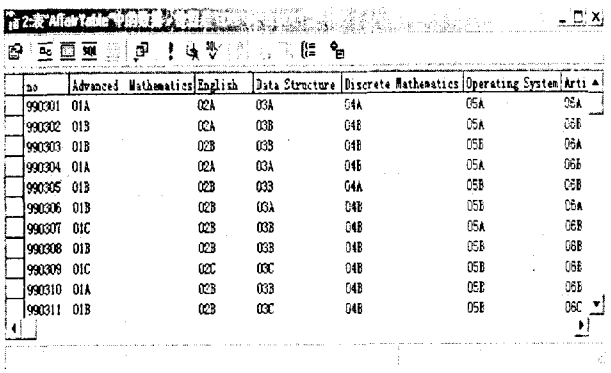
(3)项的表示和事务数据库构成。将一个学员的信息看作一个事务,事务标识号就是学员的“学号”,项采用由“属性 - 值”变换成的“课程编码等级”的形式,如:01A。部分事务数据库如图 4 所示。

3.2 改进的关联规则挖掘算法

关联规则挖掘是在数据或信息知识库的项目集或者对象集中寻找关联、相关或者有因果关系的结构。

使用传统的 Apriori 算法进行关联规则挖掘[5],可以比较有效地产生关联规则,但也存在着以下两种缺

陷:(1)整个数据库进行多次访问;(2)识别频繁项目集时算法采用模式匹配,执行效率低。



no	Advanced Mathematics	English	Data Structure	Discrete Mathematics	Operating System	Arti
990301	01A	02A	03A	04A	05A	06A
990302	01B	02A	03B	04B	05A	06B
990303	01B	02B	03B	04B	05B	06A
990304	01A	02A	03A	04B	05A	06B
990305	01B	02B	033	04A	05B	06B
990306	01B	02B	03A	04B	05B	06A
990307	01C	02B	03B	04B	05A	06B
990308	01B	02B	03B	04B	05B	06B
990309	01C	02C	03C	04B	05B	06B
990310	01A	02B	033	04B	05B	06B
990311	01B	02B	03C	04B	05B	06C

图 4　事务数据库

在本系统中,针对这些问题对算法作了改进。主要思想是将设定的支持度转换为支持度计数,只在开始时要对整个数据库进行访问,当 $k>1$ 时,只需对频繁 $(k-1)$ - 项目集 L_{k-1} 进行访问即可。而且随着 k 的增大,频繁 $(k-1)$ - 项目集也在不断减小,这样算法访问的数据量不断减小。例如计算由频繁项目集 A 和 B 组成的候选项目集的支持度,方法就是计算项目集 A 和 B 的事务列表中相同事务的数目,这一数目即为包含候选项目集的事务数。当候选项目集的支持度计数大于最小支持度计数时,它就为频繁项目集,同时还生成该项目集的事务列表,以便今后计算由该频繁项目集组成的候选项目集的支持度计数。

从上述描述可以知道,只在计算频繁 1 - 项目集时需要对整个数据库进行访问,之后在计算候选 k - 项目集($k>1$)的支持度时,仅需要数据库中频繁 $(k-1)$ - 项目集的信息即可,而随着 k 的增大,频繁 $(k-1)$ - 项目集的数目不断减小。因此,需要访问的数据量也在不断减小。另外,在计算候选项目集的支持度时,避免了模式匹配,这使得改进算法的速度提高了。

利用项目事务对应关系数据库进行挖掘的过程:首先对数据库中的每一个项目搜索其事务列表,以确定其支持度计数,这样就得到频繁 1 - 项目集 L_1,不满足最低支持的项目的数据将不参加下一循环的计算;由 L_1 可得到候选 2 - 项目集 C_2,在计算 C_2 中候选项目集的支持度计数时,不需对整个数据库进行访问,只需对频繁 1 - 项目集 L_1 进行访问即可,显然每一循环之后,访问的数据库都在减小。另外,在识别频繁项目集时,也不需要对剩下的整个数据库进行访问,而只需对数据库中该项目集的两个子集的事务列表数据进行访问即可。

改进的算法:

算法:Apriori - new;

输入:事务数据库 D;最小支持度计数 supx;

输出:D 中的频繁项集 L。

```
begin
find_frepuent_1 - itemsets(supx)
k=1
do while(L_k>1)
  k++
  find_frepuent_itemsets(supx)
  move record from D' where itemnum=k-1
  loop
end
```

其中 find_frepuent_1 - itemsets(supx)为产生一个频繁项目集的函数,它通过数据库查询操作得到所有项目的支持度,将满足条件的项目加入频繁项目集中,将不满足条件的项目的所有信息从数据库中删除。

算法:find_frepuent_1 - itemsets(supx)

输入:事务数据库 D,用户最低支持度 supx;

输出:所有的 1 - 频繁项目集,项目事务对应的关系数据库。

```
begin
 for each t∈D do
 for all l_item∈t.Itemset do
   D'=D'∪(l_item,l_trans,1)
 For all Item l_item do
  Let l_c be the number of record in D'
  itemset=l_item
   If l_c>=supx then
    L1=L1∪(l_items,l_c,)
   Else
    Move record from D' where itemset=l_item End
```

产生一频繁项目集后,通过迭代逐渐产生更长的频繁项目集,这一步是在函数 find_frepuent_itemsets(supx)中产生的。

算法:find_frepuent_itemsets(supx)

输入:项目对应的关系数据库,最低支持度计数 supx;

输出:所有的 $(k+1)$ - 频繁项目集。

```
Begin
L=0
For each itemset l1∈L(k-1){
 L++
 For itemset l2∈L(k-1)
  if(l1[1]=l2[1])∧(l1[2]=l2[2])∧…∧(l1[k-2]=l2[k
-2])∧(l1[k-1]=l2[k-1]) then
    c=l1 join l2;
let l_c be the number of records of c
if l_c>=supx then
 for each l_t∈c do
    add c to D'
```

```
end if
}}
end;
```

在算法的实现中，不是先生成所有的候选项目集再确定频繁项目集，而是生成一个候选项目集后就计算其频繁度，减少内存占用，提高算法的效率。

3.3 结果描述

数据挖掘将获取的信息以便于用户理解和观察的方式反映给用户，这时可以利用可视化工具。对于DM系统的挖掘结果，可以用自然语言、图形、表格等多种方式进行表示。在本系统中采用自然语言的方式表示。规则 A=>B 解释的形式为：加强对课程 A 的学习，有助于课程 B 的学习。其中前件 A 和后件 B 均可以包含任意多个属性。并且系统向用户推荐一个课程先后学习的序列。

由频繁项集 L_4 输出的一些关联规则：

加强对《计算机数学基础》《电子技术》《计算机组成原理》的学习，有助于课程《计算机体系结构》的学习。

加强对《计算机数学基础》《电子技术》的学习，有助于课程《计算机组成原理》《计算机体系结构》的学习。

加强对《计算机数学基础》的学习，有助于课程《电子技术》《计算机组成原理》《计算机体系结构》的学习。

课程推荐序列：《计算机数学基础》=>《电子技术》=>《计算机组成原理》=>《计算机体系结构》可由系统给出几种挖掘所得的规则形式及推荐课程设置的序列，给用户提供参考，由决策者决定所采用的规则及序列，做出相应的决策，指导学员选课，以帮助学员更好地完成各门课程的学习。

4 结束语

该系统实现了一个完整的数据挖掘过程，用户只要提供必要的数据，系统就可以自动地对选定的数据库进行分析，并且返回用户需要的信息，帮助用户做出决策，从而帮助学员选课及进一步学习，起到了很好的促进作用，具有一定的实用价值。同时数据挖掘技术已经在许多领域取得好的应用，随着数据的不断增长，把数据挖掘技术应用到远程教育教学中，能够较客观实时地反映问题，这一研究也对远程教育教学管理提出了很好的建议。

参考文献：

[1] Agrawal R, Imielinski T, Swami A. Mining Association Rules between Sets of Items in Large Database[C]//In SIGMOD' 93. Washington, DC: [s. n.], 1993: 207 - 216.

[2] Han Jiawei, Kamber M. 数据挖掘概念与技术[M]. 范 明，孟小峰等译. 北京：机械工业出版社，2003: 150 - 221.

[3] 朱 明. 数据挖掘[M]. 合肥：中国科学技术大学出版社，2002: 129 - 140.

[4] 程涛远. 基于园区网络的数据仓库相关技术的研究[D]. 济南：济南大学，2002: 36 - 40.

[5] 贾彩燕，倪现君. 关联规则挖掘研究述评[J]. 计算机科学，2003, 30(4): 145 - 148.

(上接第 175 页)

versity of Cambridge, 1996.

[13] 林 榕. 基于图像的信息隐藏技术综述[J]. 装备制造技术，2007(7): 91 - 93.

[14] 罗建禄. 图像数字水印技术综述[J]. 重庆工商大学学报：自然科学版，2007, 24(1): 10 - 11.

[15] Kutter M, Bhattacharjee S K, Ebrahimi T. Towards second generation watermarking schemes[C]// in Proc. IEEE Int. Conf. Images Processing: vol. 1. Kobe, Japan: [s. n.], 1999: 320 - 323.

[16] Moulin P, O'Sullivan J A. Information - Theoretic Analysis of Information Hiding[J]. IEEE Transaction on Information Theory, 2003, 49(3): 563 - 593.

(上接第 178 页)

Using Key Graphs[J]. IEEE/ACM Transaction on Networking, 2000(8): 16 - 30.

[2] Moyer M, Rao J, Rohatgi P. A survery of security issues in multicast communicastions[J]. IEEE Network, 1999, 13: 12 - 23.

[3] Wallner D, Harder E, Agee R. Key management for multicast: Issues and architecture[S]. IETF RFC 2627, 1999.

[4] Kwak D W, Lee Seung Joo, Kim Jong Won, et al. An Efficient, LKH Tree, Balancing Algorithm for Group Key Management[J]. IEEE Communication, 2006, 10(3): 222 - 224.

[5] Son Ju - Hyung, Lee Jun - Sik, Seo Seung - Woo. Energy Efficient Group Key Management Scheme for Wireless Sensor Networks[M]//IEEE, Invited Paper. [s. l.]: [s. n.], 2007.

[6] 康巧燕，孟相如，王建峰，等. 基于逻辑层次树的动态组播密钥管理改进方案[J]. 计算机工程，2007(8): 123 - 125.

[7] 徐明伟，董晓虎，徐 恪. 组播密钥管理的研究进展[J]. 软件学报，2004, 15(1): 141 - 149.

[8] 周 杰，张金焕，王全迪，基于 RSA 的组播加密方法及其密钥管理方案[J]. 大连理工学报，2005, 45(10): 122 - 125.

[9] 李新国. 数字内容保护系统中的认证和密钥管理技术研究[D]. 西安：西安电子科技大学，2006.