改进的数据挖掘算法在高职院校图书馆中应用

罗晓丽

(福州职业技术学院 计算机系，福州 350108)

摘要：随着高职教育的专业建设和发展，作为直接服务于教育的高职院校图书馆在职业教育中的作用越来越不容忽视。然而目前高职院校的图书馆中大量的数据没有被利用和挖掘并为图书馆的管理服务。因此，本文选取高职院校图书馆系统作为数据仓库并结合学生的基本信息，利用数据挖掘技术，针对高职院校图书馆借阅信息进行数据分析，试图本着方便学生，为学生服务、引导学生钻研专业知识的原则，对专业和图书利用情况进行了相关分析，为图书馆管理者和领导者管理和决策提供有价值的信息.

关键词：数据仓库；数据挖掘；高职教育；借阅数据

中图分类号：TP311.132.3 **文献标志码**：A **文章编号**：1671－1807(2011)07－0121－05

1 应用背景

由于图书馆是服务部门，提高图书馆利用率取决于使用率最大权重者即学生。很多高职院校的图书馆仍以藏书为主，而不尊从以人为本的服务意识，没有针对高职学生的借阅行为、需求、偏好、学生专业进行收藏和管理图书；合理设置图书馆馆藏结构；扭转图书馆的工作作风、完善工作制度等。

2 Apriori 算法改进技术的介绍

Apriori[1]算法每次循环都要扫描一遍数据库，计算候选项集的支持数。随着项集阶数的增加，候选项集的个数逐渐减少，但扫描的事物量并没有改变，即还是原来的数据库。因此，提高效率的一个有效方法就是在循环的过程中逐渐减少扫描的事物量或者减少扫描的次数两个方面。

3 本文算法介绍

具体步骤描述如下：

设 T_1、T_2、T_3、T_4、T_5、T_6共 6 位顾客在某超市内购买了 A、B、C、D、E 共五类商品，A 类商品有四种：A_1、A_2、A_3、A_4；B 类商品有二种是：B_1、B_2；C 类商品有四种是：C_1、C_2、C_3、C_4；D 类商品有二种是：D_1、D_2；E 类商品有三种是 E_1、E_2、E_3；数据库事务记录如表 1，设最小支持度阈值为 30%(共 6 条记录，故最小支持数为 2)。

表 1 数据库记录表

顾客(TID)	A 类商品	B 类商品	C 类商品	D 类商品	E 类商品
T_1	A_1	B_2	C_1	D_1	E_3
T_2	A_2	B_1	C_3	D_2	E_2
T_3	A_3	B_2	C_4	D_1	E_3
T_4	A_4	B_2	C_2	D_1	E_1
T_5	A_1	B_2	C_1	D_1	E_3
T_6	A_1	B_2	C_2	D_1	E_1

1)先扫描数据库，生成候选频繁 1－项目集，删除支持度计数小于 2 的记录。并用 sql 语句拷备生成包含该属性的记录，生成新数据库 D_1(如表 2)，减少事务数据库记录。扫描数据库，计数生成频繁项集的支持度(如表 3)。

表 2 新生成的数据库

顾客(TID)	A 类商品	B 类商品	C 类商品	D 类商品	E 类商品
T_1	A_1	B_2	C_1	D_1	E_3
T_3	A_3	B_2	C_4	D_1	E_3
T_4	A_4	B_2	C_2	D_1	E_1
T_5	A_1	B_2	C_1	D_1	E_3
T_6	A_1	B_2	C_2	D_1	E_1

2)利用 CA 算法，对此数据库进行编码转换，随着记录数的减少，编码转换的工作量也减少了，这将提高算法运行效率，转换结果如下表 4。

收稿日期：2011－03－20

作者简介：罗晓丽(1971—)，女，黑龙江五常人，福州职业技术学院计算机系，讲师，福州大学软件工程硕士毕业，研究方向：多媒体。

表 3　扫描生成频繁 1 项集及支持数

项	支持数	频繁项	项	支持数	频繁项
$\{A_1\}$	3	$\{A_1\}$	$\{C_3\}$	1	
$\{A_2\}$	1		$\{C_4\}$	1	
$\{A_3\}$	1		$\{D_1\}$	5	$\{D_1\}$
$\{A_4\}$	1		$\{D_2\}$	1	
$\{B_1\}$	1		$\{E_1\}$	2	$\{E_1\}$
$\{B_2\}$	5	$\{B_2\}$	$\{E_2\}$	1	
$\{C_1\}$	2	$\{C_1\}$	$\{E_3\}$	3	$\{E_3\}$
$\{C_2\}$	2	$\{C_2\}$			

表 4　数库 D_1 项编码表

项	TID_1	TID_3	TID_4	TID_5	TID_6
A_1	1	0	0	1	1
B_2	1	1	1	1	1
C_1	1	0	0	1	0
C_2	0	0	1	0	1
D_1	1	1	1	1	1
E_1	0	0	1	0	1
E_3	1	1	0	1	0

3)将频繁 1－项目集进行与运算，得到候选 2－项集，并统计 1 的个数，计算支持度。如表 5。

表 5　频繁 2－项集

项	编码	支持数	频繁项	项	编码	支持数	频繁项
$\{A_1B_2\}$	10011	3	$\{A_1B_2\}$	$\{B_2E_3\}$	11010	3	$\{B_2E_3\}$
$\{A_1C_1\}$	10010	2	$\{A_1C_1\}$	$\{C_1D_1\}$	10010	2	$\{C_1D_1\}$
$\{A_1C_2\}$	00001	1		$\{C_1E_1\}$	00000	0	
$\{A_1D_1\}$	10011	3	$\{A_1D_1\}$	$\{C_2E_1\}$	00101	0	
$\{A_1E_1\}$	00001	1		$\{C_2D_1\}$	00101	2	$\{C_2D_1\}$
$\{A_1E_3\}$	00010	1		$\{C_2E_1\}$	00101	2	$\{C_2E_1\}$
$\{B_2C_1\}$	10010	2	$\{B_2C_1\}$	$\{C_2E_3\}$	00000	0	
$\{B_2C_2\}$	00101	2	$\{B_2C_2\}$	$\{D_1E_1\}$	00101	2	$\{D_1E_1\}$
$\{B_2D_1\}$	11111	5	$\{B_2D_1\}$	$\{D_1E_3\}$	11010	3	$\{D_1E_3\}$
$\{B_2E_1\}$	00101	2	$\{B_2E_1\}$				
频繁 2 项集有：$\{A_1B_2\}\{A_1C_1\}\{A_1D_1\}\{B_2C_1\}\{B_2C_2\}\{B_2D_1\}\{B_2E_1\}\{B_2E_3\}\{C_1D_1\}\{C_2E_1\}\{D_1E_1\}\{D_1E_3\}$							

4)将频繁 2－项目集进行与运算，得到候选 3－项集，并统计 1 的个数，计算支持度。如表 6。

表 6　频繁 3－项集

项	编码	支持数	频繁项	项	编码	支持数	频繁项
$\{A_1B_2C_1\}$	10010	2	$\{A_1B_2C_1\}$	$\{B_2C_1E_3\}$	10010	2	$\{B_2C_1E_3\}$
$\{A_1B_2C_2\}$	00001	1		$\{B_2C_2D_1\}$	00101	2	$\{B_2C_2D_1\}$
$\{A_1B_2D_1\}$	10011	3	$\{A_1B_2D_1\}$	$\{B_2C_2E_1\}$	00101	2	$\{B_2C_2E_1\}$
$\{A_1B_2E_1\}$	00001	1		$\{B_2C_2E_3\}$	00000	0	
$\{A_1B_2E_3\}$	10010	2	$\{A_1B_2E_3\}$	$\{B_2D_1E_1\}$	00101	2	$\{B_2D_1E_1\}$
$\{A_1C_1D_1\}$	10010	2	$\{A_1C_1D_1\}$	$\{B_2D_1E_3\}$	11010	3	$\{B_2D_1E_3\}$
$\{A_1C_1E_1\}$	00000	0		$\{C_1D_1E_1\}$	00000	0	
$\{A_1C_1E_3\}$	10010	2	$\{A_1C_1E_3\}$	$\{C_1D_1E_3\}$	10010	2	$\{C_1D_1E_3\}$
$\{B_2C_1D_1\}$	10010	2	$\{B_2C_1D_1\}$	$\{C_2D_1E_1\}$	00101	2	$\{C_2D_1E_1\}$
$\{B_2C_1E_1\}$	00000	0		$\{C_2D_1E_3\}$	00000	0	
L3:$\{A_1B_2C_1\}\{A_1B_2D_1\}\{A_1B_2E_3\{A_1C_1D_1\}\}\{A_1C_1E_3\}\{B_2C_1D_1\}\{B_2C_1E_3\}\{B_2C_2D_1\}\{B_2C_2E_1\}\{B_2D_1E_1\}\{C_1D_1E_3\}\{C_2D_1E_1\}$							

5)将频繁 3－项目集进行与运算，得到候选 4－项集，并统计 1 的个数，计算支持数。如表 7。

6)将频繁 4－项目集进行与运算，得到候选 5－项集，并统计 1 的个数，计算支持数。得出最大频繁项集为$\{A_1B_2C_1D_1E_3\}$。如表 8。

4　本文算法的不同之处

通过上述介绍，可以看到该算法的思路是基于 Apriori 算法，开始进行一次数据库的扫描，得到频繁 1 项集，对数据库处理，生成缩小包含所有频繁 1 项集的新数据库，再利用 CA 算法进行编码的转换，但不同于 CA 算法即在一开始就对原来的数据库进行编码转换，而是在对原来的数据库进行消减后生成的新数据库进行编码转换及运算，极大地减少了处理的时间，因此，本文算法在提高挖掘的效率方面，有极大

的优势。

表 7　频繁 4－项集

项	编码	支持数	频繁项
$\{A_1B_2C_1D_1\}$	10010	2	$\{A_1B_2C_1D_1\}$
$\{A_1B_2C_1E_1\}$	00000	0	
$\{A_1B_2C_1E_3\}$	10010	2	$\{A_1B_2C_1E_3\}$
$\{B_2C_1D_1E_1\}$	00000	0	
$\{B_2C_1D_1E_3\}$	10010	2	$\{B_2C_1D_1E_3\}$
$\{A_1C_1D_1E_1\}$	00000	0	
$\{A_1B_2C_2D_1\}$*			
$\{A_1B_2C_2E_1\}$*			
$\{A_1B_2C_2E_3\}$*			
$\{B_2C_2D_1E_1\}$*			
$\{B_2C_2D_1E_3\}$*			
$\{A_1C_2D_1E_1\}$*			
$\{A_1C_2D_1E_3\}$*			

注：* 号数据表示该频繁项的子集中含有非频繁项集而删除的项集。

表 8　候选 5－项集

项	编码	支持数	频繁项
$\{A_1B_2C_1D_1E_1\}$	00000	0	
$\{A_1B_2C_1D_1E_3\}$	10010	2	$\{A_1B_2C_1D_1E_3\}$

5　本文算法所用到的伪代码

下面本文算法的描述(伪代码)[2-4]：

输入：事务数据库 D；最小支持度阈值 min_sup

输出：D 中的频繁项集 L

{C1＝candidate1－Itemsets}；//生成候选频繁 1－项目集，一般情况下，令 $T_1=\{\{I_1\},\{I_2\},\{I_3\}\cdots\{Im\}\}$，即 T_1 为所有项目的集合。

$L_1=\{p\in T|Sup\ (p)\geqslant min\_sup\}$；//扫描事务数据库 D 一次，生成频繁 1－项集。

Copy * to D 2 for c $\in L_1$//生成缩小的新的数库，并对该数库进行编码转换。

for(i＝2；Li－l≠；i＋＋)

{

Ci＝

Li＝

for each J1 Ci－1 do

for each J2≠J1∈Ci－1 do

… …

if count(J1 J2 KJi)≥Vmin—sup//某个项的编码或某几个项做完与运算后的编码中'1'的个数大于或等于最小支持度时

Li＝Li(J1J2 LJi)；

Ci＝CjJ1J2 L ji；

endif

endfor

endfor

endfor

6　算法实验数据比较

为直观地表现本文算法的性能，将其与 CA、Apriori－Tid 算法进行比较，实验如下：

实验环境：在同样的实验条件下(服务器：P42.0GHz、4G 内存、SQL Server2000；客户机：P42.0 GHz、1M 内存。

实验数据：选用了一个真实数据库(本校 2007 级计算机专业学生成绩为原始数据，得到原始数据库。该数据库共有学生 600 名，21 门课程，61 113 条记录。)

使用数据库中前 30 000 条记录在不同最小支持度阈值条件下进行关联规则挖掘。最小支持度阈值 min_sup 分别设置为 5%、10%、15%、20%、25%。分别用 Apriori-Tid 算法(简称 Tid 算法)、CA 算法和本文算法进行数据挖掘。结果如表 9。

表 9　不同支持度三种算法数据挖掘结果比较

支持度	L_1 频繁项集	L_2 频繁项集	L_3 频繁项集	L_4 频繁项集	L_5 频繁项集	关联规则	运行时间		
							TID 算法	CA 算法	本文算法
5	349	1 250	30	21	10	58	70.1	62.5	54.6
10	349	457	25	17	3	35	67.2	57.8	49.7
15	332	256	23	10	2	27	63.4	49.6	47.2
20	210	178	12	3	ϕ	13	42.2	37.6	29.8
25	113	33	5	ϕ	ϕ	6	31.7	30.1	27.5

从表 9 结果，可以看出当最小支持度阈值较小时，本文算法和 CA 算法执行时间比 Tid 算法的执行时间少很多。随着支持度阈值的增大，三种算法的运行时间明显减少，时间差距缩小，但本文算法的时间上仍然占较大优势。

7　该挖掘算法应用实例

利用该论文算法、CA 和 Apriori 算法分别对福州职业技术学院图书馆的数据进行数据挖掘，然后对

挖掘结果进行比较：

图 1　数据源

对于图 1 数据源，分别运用 Apriori－Tid 算法，Coding－based Apriori 算法，本文算法进行数据挖掘，挖掘出支持率大于 30％的挖掘结果，如图 2、图 3、如图 4。

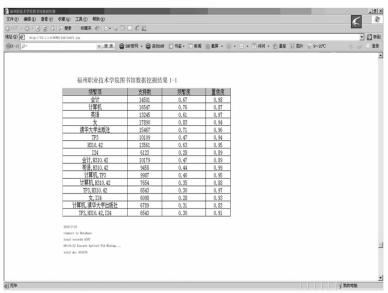

图 2　Apriori 算法数据挖掘结果 1－1

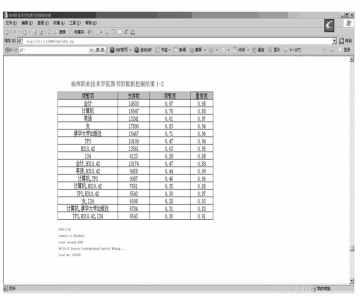

图 3　CA 算法数据挖掘结果 1－2

比较上述三种算法结果，可见挖掘的结果基本相同，只是在挖掘时间上有所不同，本文算法又比其他两种算法在运行效率上有了一定的提高。本文算法优化理论通过实验得到证实。

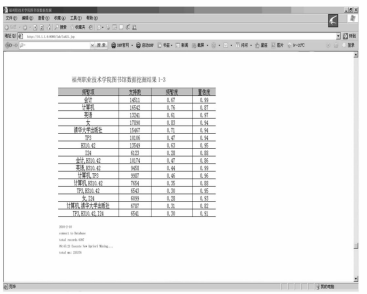

图 4　该论文算法数据挖掘结果 1－3

分析上述结果，得出以下结论：

1）借阅率最高的是计算机、英语、会计专业的学生，分析原因：原来的整合前的主要专业就是这三个专业，而其他新设专业，相应的图书馆藏量不足，应根据学校所设专业调整图书馆藏量。

2）图书借阅者多为女生，经过调查，男生一般上网查阅资料，原因是在图书馆查阅资料比较麻烦，且经常查到过时的信息，降低了查阅者的兴趣。而女生一般比较有耐心。图书馆管理者应该引导借阅者，提供最佳服务。

3）学生较喜欢清华大学出版社出版的图书，大家一致认为该出版社的图书的编排比较好。这类图书可放置在明显的地方。

4）计算机考证、英语等级的图书的借阅率比较高，主要原因是高职院校的毕业是和这两种证书挂钩的，没有相应的证书不予毕业。

5）女生的偏好是看小说，因此小说的借阅率比较高。

参考文献

[1]赵松．APriori 算法的改进及应用[D]．哈尔滨：哈尔滨理工大学，2008.

[2]单明辉．改进的关联规则算法在采购数据挖掘中的应用[D]．上海：上海交通大学 2008.

[3]叶晓波．一种基于二进制编码的频繁项集查找算法[J]．楚雄师范学院学报，2009(3).

[4]尹国琦．改进的 Apriori 算法在大学生素质拓展中的应用[D]．大连：大连交通大学，2006.

Application Research on High Vocational College Library Based on Data Mining

LUO Xiao-li

(Computer Department, Fuzhou Technical and Vocational College, Fuzhou 350108, China)

Abstract: With the development of higher vocational college, the library of which is not negligible in professional education. But now too much data of library hasn't been mined and used to serve for administration. Therefore, this dissertation selects library borrowed data as data warehouse, using data mining to put up data analysis making for supplying valuable information for administrator and leader, aiming server for student conveniently and inducing students.

Key words: data warehouse; data mining; higher vocational college; borrowed data

(上接第 100 页)

Empirical Analysis of Business Performance Based on SME-board Companies

——Based on factor analysis and cluster analysis

LI Xi

(School of Business, Anhui University, Hefei 230039, China)

Abstract: This paper constructs an evaluation index system of business performance, based on SME-board companies, from four aspects, which are profitability, operation ability, solvency, company development ability. By using multivariate statistical analysis methods, it conducts a classified research and a comprehensive evaluation of business performance to 37 groups of effective sample data. First is factor analysis. This paper extracts three common factors from 10 indexes, which could reflect abilities of the four aspects. Second, it makes a Hierarchical-cluster analysis to the sample companies. The analyzed results are aggregated to five types with socio-economic statistical software SPSS. By analyzing the average scores on each factor about various companies, the paper comprehensively evaluates the business performance based on their advantage factors and disadvantage factors. Last, from company management and SME-board investor, suggestions are respectively presented.

Key words: SME-board; listed companies; business performance; factor analysis; cluster analysis

(上接第 120 页)

[3]余治国,江雨燕. 生活中的博弈论[M]. 北京:世界图书出版公司北京公司,2006.

[4]李娟. 我国社会医疗保险中的风险及其规避[D]. 湘潭:湘潭大学,2005.

Game of Four Players in Health Insurance

MA Ji-wei

(School of Management, Ocean University of China, Qingdao Shandong 266100, China)

Abstract: Health insurance because of its severity of asymmetric information led to the establishment of the insurance relationship emerged from the beginning to the end when the number of questions. Through the establishment of the insurer, the insured, insurance agents and medical institutions in the insurance relationship building process the Quartet in the game model that produced the existing balance of game due to the adverse effects on the parties. And the solutions the plight of these games should be taken.

Key words: health insurance; information asymmetric; moral hazard; game