基于 Apriori 算法的高校招生的关联规则分析 *

赵祖应,丁 勇,潘明波
(云南工商学院,云南 昆明 651701)

摘 要:数据挖掘是适应信息社会从海量的数据库中提取信息的需要而产生的新学科。它是统计学、机器学习、数据库、模式识别、人工智能等学科的交叉。以往的数据挖掘技术的应用大多是在金融领域,而在其他领域里面应用不是很多,如在高校招生中的应用更是如此。数据挖掘技术对招生工作的深层研究与挖掘将会得到各高校的更多重视。以某高校招生数据作为招生信息为依据,对高校招生的关联规则进行分析。从而对关联性规则的应用作进一步的研究。

关键词:关联规则;Apriori 算法;置信度;支持度;建模

中图分类号:TP311.12 **文献标识码**:A **文章编号**:1674-7720(2014)05-0087-03

The analysis for association rules in university enrollment based on Apriori algorithm

Zhao Zuying, Ding Yong, Pan Mingbo
(Yunnan Technology and Business University, Kunming 651701, China)

Abstract: Data mining technology is an emerging discipline in recent years, which meets the requirement to extract information from massive database. It is some sort of inter-discipline combined with statistics, machine-learning, database, pattern recognition, artificial intelligence and others. The data mining technology has been mostly applied for the financial sector, while in other areas there is not used a lot, especially in enrollment in colleges and universities. The further research of data mining technology on enrollment of colleges and universities will achieve more attention. In this paper, we are going to analyze association rules of enrollment based on the college admissions information; therefore, we are to do further research on the application of association rules.

Key words: association rules; apriori algorithm; confidence; support; modeling

1 民办高校招生的现状分析

招生工作一直是民办学校最重要的工作,民办学校在招生上的投入占一年总支出的很大部份,采用的招生方式也在不断的更新,使用新方法,新模式。但同时也会发现,有些方式方法并不能解决招生问题,浪费了有限的资源,得不偿失,主要表现在招生成本高、没有严格的招生机制,宣传模式单一等。归根原因是没有找到适合本校的招生方法与模式,而要能做到这一点,必须要对招生工作做一个详细的研究,根据以往招生的情况,总结分析,找出问题所在点和发光点,为招生工作更好的方式提供有力的依据。

2 Apriori 算法分析

2.1 挖掘关联规则的主要步骤

步骤 1:发现所有的频繁集。项集的频度至少应等于(预先设置的)最小支持度。关联规则的整个性能主要取决于这一步。

步骤 2:根据所获得的频繁项集,产生相应的强关联规则。这些规则必须满足最小置信度阈值。

2.2 Apriori 算法

Apriori 算法是挖掘产生关联规则所需要的频繁项集的基本算法,是数据挖掘领域里面常用的一种关联规则挖掘算法。该算法利用一个层次顺序搜索的循环方法来完成频繁集的挖掘工作。这一循环方法就是利用(k-1)-项集来产生 k-项集,具体的做法是首先找出频繁集 1-项集,记为 L_1;然后利用 L_1 来挖掘产生 L_2,即频繁 2-

* 基金项目:云南省教育厅科研项目(2012Y078)

项集，如此循环往返，直到无法发现更多的频繁 k-项集为止。在每一层挖掘产生 L_k 时，都需要对整个数据库扫描一遍。Apriori 算法利用 L_{k-1} 来生成 L_k。

该算法实现过程包括两个步骤，即连接和剪枝，具体实现过程如下。

连接步骤：设 l_1 和 l_2 为 L_{k-1} 中的的两个项集，符号 L_{ij} 表示 L_i 中的第 j 项，如 $L_{i,k-2}$ 就表示 l_i 中的倒数第二项。若 L_{k-1} 的连接操作记为 $L_{k-1}\oplus L_{k-1}$，它表示若 l_1 和 l_2 中的前 $(k-2)$ 项是相同的，即若有下面关系。

$$(l_{11}=l_{21})\wedge(l_{12}=l_{22})\wedge\cdots\cdots\wedge(l_{1,k-2}=l_{2,k-1})$$

则 L_{k-1} 中的 l_1 和 l_2 的内容就可以连接到一起。

剪枝步骤：C_k 是 L_k 的一个超集，其中由项集组成的各元素不一定都是频繁项集，但是所有的频繁 k-项集一定都在里面，即有 $L\subseteq C_k$。对数据库进行扫描就可以确定 C_k 中各候选项集的支持频度，并由此获得 Lk 中的各个元素，即频度 k-项集。所有频度不小于最小支持频度的候选集就是 Lk 的频繁集。

3 Apriori 算法对民办高校招生分析

3.1 数据预处理

从某高校招生的收集数据中抽出 1 000 条数据进行数据预处理，并对其进行数据筛选，处理结果如表 1 和表 2 所示。

表 1 2012 年招生数据预处理

序号	时间	姓名	地区	考分	性别	是否报到
1	2012 年	杨波	A	415	0	yes
2	2012 年	和增祺	B	453	1	no
3	2012 年	杨珍	C	398	0	yes
4	2012 年	和铭哲	D	346	1	yes
5	2012 年	周彬	E	359	1	no
6	2012 年	郭培杰	D	361	0	yes
7	2012 年	吴翠莲	C	339	0	yes
8	2012 年	凌静轩	B	402	1	no
9	2012 年	马雪雁	A	453	0	yes
10	2012 年	杨阳	E	448	1	yes
11	2012 年	王焕妮	A	468	0	no
12	2012 年	丁聪	A	419	1	yes
13	2012 年	和闪闪	B	267	1	yes
14	2012 年	邱龙	C	373	1	no
15	2012 年	梁景程	D	322	1	yes
16	2012 年	苏敏	E	344	1	yes
17	2012 年	杜瑞康	D	448	1	no
18	2012 年	袁红仁	C	392	1	yes
19	2012 年	唐卓亚	B	403	0	yes
20	2012 年	胡超	A	415	1	no
21	2012 年	和荣昌	E	452	1	yes
22	2012 年	刘鑫凯	A	377	1	yes
23	2012 年	陈烨	A	355	1	no
24	2012 年	刘宁	B	316	1	yes
25	2012 年	和翔	C	429	1	yes
26	2012 年	和钊	D	315	1	no
27	2012 年	和荣	E	353	1	yes

备注：A：表示云南昆明；B：表示云南大理；C：表示云南曲靖；D：表示云南昭通；E 表示丽江；0：表示女；1：表示男；yes 表示报到；no：表示未报到

表 2 2013 年招生数据预处理

时间	姓名	地区	考分	性别	是否报到
2013 年	刘敏	A	287	0	no
2013 年	李娜	A	389	0	yes
2013 年	唐俊	A	270	0	no
2013 年	高晓红	B	467	0	yes
2013 年	徐晓翔	D	435	1	yes
2013 年	彭烨	C	298	1	no
2013 年	付超	E	312	1	yes
2013 年	洪开端	A	349	1	no
2013 年	吴利林	A	358	1	yes
2013 年	刘映涛	A	431	1	yes
2013 年	周进奇	B	345	1	no
2013 年	沈彬	B	408	1	yes
2013 年	顾品柱	B	278	0	no
2013 年	张雪娟	B	333	0	yes
2013 年	吴瑾	B	340	0	yes
2013 年	杨梦芹	C	324	0	no
2013 年	吴继梅	C	444	0	yes
2013 年	郭林霞	C	350	0	no
2013 年	田孟杰	D	385	0	yes
2013 年	尹雪娇	D	379	0	yes
2013 年	李俊杰	E	338	1	no
2013 年	吴利斌	A	357	1	yes
2013 年	黄凯	A	328	1	no
2013 年	沈灿	A	282	1	yes
2013 年	杨进	B	376	1	yes
2013 年	周红姣	D	291	0	no
2013 年	杨梦	C	422	0	yes

3.2 利用 spss Clemention 建模

利用 spss Clemention 工具建立模型，本例用 2012 年入学数据与 2013 年入学数据进行比较，得出两年的地区与是否报到的关联性分析，建模如图 3 所示。

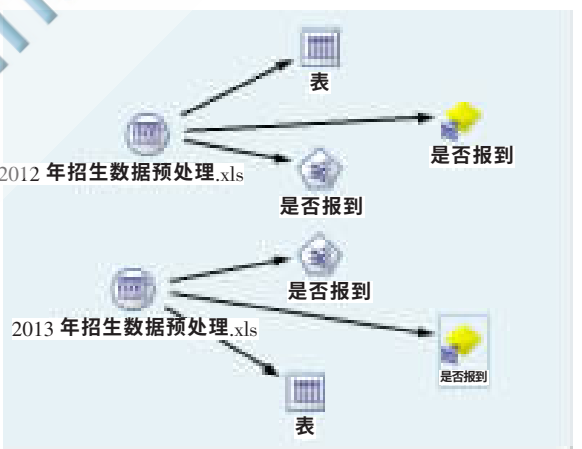

图 3 数据建模

3.3 设置最低条件支持度，最小规则置信度，最大前项数

在 2012 年的数据中，设置最低条件支持度为 8.0，最小规则置信度 60.0%，最大前项数为 5，得到的数据分析结果如图 4 所示。

如果把 2013 的规则支持度和置信度设置和 2012 相同，结果如图 5 所示。

3.4 地区与是否报到关联规则结果分析

根据图 4 和图 5 进行比较，B（云南大理）和 D（云南

图4 2012年执行结果

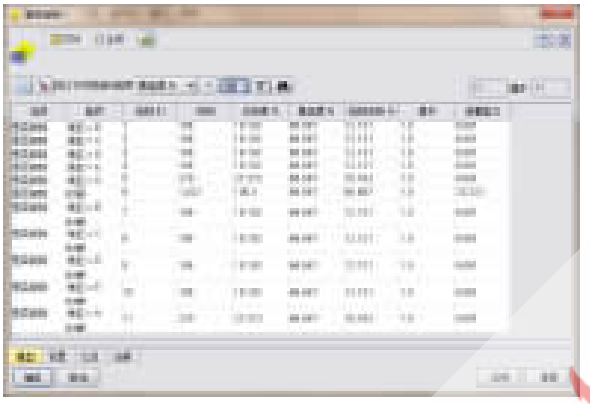

图5 2013年执行结果

昭通)地区的学生报到是趋于正常的发展,在2013年招生中,A(云南昆明)、C(云南曲靖)和E(云南丽江)加大了招生宣传,取得了非常明显的效果,那么在2014年的招生宣传中,还需要在A、C、E地区保持一定的宣传投入,在B和D地区可以适当减少招生投入。

一个学校生源的多少决定了它规模及发展。特别是在民办高校,“招生就是一切”,招生中不仅要数量、质量也是发展的关键。民办院校在不同的发展时期会有不同的发展策略,在不同的历史时期院校也就有不同的招生策略及队伍建设适应发展的需求。因此,只有在清楚制定了院校发展战略规划后,才能顺理成章地制定出院校人力资源需求、发展、策略、培训、扩建和储备计划。充分把数据挖掘技术利用在招生工作中,将对个高校的招生工作提供决策支持,对高校的招生成本的整合具有深远的意义。

参考文献

[1] 赵祖应,丁勇.基于Apriori算法的购物篮关联规则分析[J].江西科学,2012(1).

[2] 王嵩岩.基于数据挖掘的关联规则研究[J].吉林省经济管理干部学院学报,2008,22(1):80-82.

[3] 朱建平,谢邦昌.数据挖掘中关联规则的提升及其应用[J].统计研究,2004(12):34-39.

[4] 姚俊.浅谈关联规则挖掘[J].信息技术,2005(6).

[5] 刘柱文,李丽琳.关联规则技术在数据挖掘中的应用[J].科学技术与工程,2008(6).

[6] 谭建豪,章兢.数据挖掘技术[M].北京:中国水利水电出版社,2009.

[7] 刘世平.数据挖掘技术与应用[M].北京:高等教育出版社,2010.

(收稿日期:2013-12-06)

作者简介:

赵祖应,男,1979年生,副教授,硕士,主要研究方向:数据挖掘与网络安全。

丁勇,男,1975年生,讲师,硕士,主要研究方向:软件工程与敏捷开发。

潘明波,男,1982年生,讲师,主要研究方向:电子商务应用。