第16卷 第1期
2006年1月

计算机技术与发展
COMPUTER TECHNOLOGY AND DEVELOPMENT

Vol. 16 No. 1
Jan. 2006

数据挖掘技术在成才因素分析中的应用研究

黄江涛，刘自伟
（西南科技大学 计算机科学与技术学院，四川 绵阳 621002）

摘 要：如何在高校有效地开展素质教育是一个迫切的问题，而通过对高校学生数据库的挖掘，有可能在这方面获得一些科学依据。文中介绍了数据仓库及数据挖掘基本技术，并将数据挖掘技术应用到成才因素分析之上，通过大量的实际数据测试，获取了感兴趣度较大的规则模式，在高校素质教育的具体应用中起到了促进的作用。最后给出了一个基于MS Analysis Server的多维数据集及数据挖掘结果分析可视化平台。

关键词：数据挖掘；成才因素；素质教育
中图分类号：TP391.77　　**文献标识码**：A　　**文章编号**：1005－3751(2006)01－0165－02

Application Study of Data Mining Technology in Becoming a Useful Person Factor Analysis

HUANG Jiang-tao, LIU Zi-wei
(School of Computer Science, Southwest Univ. of Sci. and Techn., Mianyang 621002, China)

Abstract: The education circle of our country keeps probing into the educational problem of quality, and should obtain some scientific basis in this respect probably through excavating to university's student's database. This paper introduces data warehouse and data mining technology at first, and then describes them used in application of becoming a useful person factor, also offers a great deal of real data for test to find rules which have high interestedness and have been proof useful for college anlagen education, provides finally a multidimensional data and data mining results visual platform based on MS Analysis Server.

Key words: data mining; useful person factor; anlagen education

0 引 言

随着数据库技术的广泛应用，数据库中存储的数据量急剧增大。数据库系统提供了对这些数据的管理和处理功能，人们可以对这些数据进行分析研究。但对如此庞大的数据需要进行较高层次的处理技术，从中找出规律和模式，以帮助人们更好地利用这些数据进行决策和研究。数据挖掘(Data Mining)技术就是在这样一个背景下产生的。它的宗旨就是在数据库中发现有用的知识。

我国教育界一直在探讨素质教育问题，那么什么是素质教育，素质教育包括哪些方面，如何开展素质教育，通过对高校学生数据库的挖掘，有可能在这方面给出一些科学依据，这正是本项目的研究意义所在。天才与勤奋是成功的秘诀，但具有同样天才与勤奋的人，在不同学科领域中取得的成就可能有较大的差异，那么成才究竟有哪些主要因素。数据库中知识发现根据对数据的分析建立对数据的特性以及数据之间关系描述的模式，利用这项技术对成才因素进行深入挖掘极有可能发现某些规律性的存在，对在大学中如何开展素质教育将有着重大指导意义。

收稿日期：2005－04－25
基金项目：四川省教育厅重点项目(01LC07)
作者简介：黄江涛(1979—)，男，广西玉林人，硕士，研究方向为数据仓库、数据挖掘、数据挖掘可视化等；刘自伟，研究员，研究方向为数据仓库、数据挖掘、人工智能等。

1 基本概念概述

1.1 数据仓库

数据仓库是一个面向主题、集成的、非易失的、随时间变化的用来支持管理人员决策的数据集合[1]。数据仓库系统在数据分析和决策方面为用户或“知识工人”提供服务，称为联机分析处理(On－Line Analytical Processing, OLAP)系统。而关系数据库系统的主要任务是执行联机事务处理和查询处理，称为联机事务处理(On－Line Transaction Processing, OLTP)系统[2,3]。

OLAP技术应用中涉及到以下一些基本概念[4]：

(1)维、维的层次、维的成员。维是人们观察数据的特定角度，是考虑问题时的一类属性。属性集合构成一个维，如时间维。维的层次是指人们观察数据的某个特定角度(即某个维)还可以存在细节程度不同的各个描述方面。如时间维可以分为日期、月份、年等层次。维的成员是维的一个取值，是数据项在某个维中位置的描述，如“2004年8月27日”是在时间维上位置的描述。

(2)度量。它是一个数值函数,该函数可以对数据立方体的每一个点求值。

(3)维表、事实表。维表用来存储维的元数据。事实表用来存储事实的度量值和各个维的码值。

1.2 多维数据集构架

多维数据集构架是目前最流行的数据仓库构架,同时,也是 SQL Analysis Services 中所使用的数据仓库构架。在此,主要介绍两种最常用的架构模型:

(1)星型架构(star schema)。星型架构是大多数数据仓库采用的架构模型。它由一个规模很大的无冗余数据的中心表(事实数据表,fact table)和一组规模较小的附属表(维度表,dimension table)所组成。

(2)雪花架构(snowflake schema)。雪花架构是对星型架构的拓展。它对星型架构的维度表进一步层次化,使其中某些维度表被进一步规范化表示为主维度表和附属维度表,只将主维度表与事实数据表联接,其它维度表联接到主维度表,其架构形式类似于雪花的形状[5]。

1.3 数据挖掘

数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的数据中,提取隐含在其中、潜在有用的信息和知识的过程。

数据挖掘通过对观测到的数据集(经常是很庞大的)进行分析,发现未知的关系和以数据拥有者可以理解并对其有价值的新颖方式来总结数据[6]。它出现于 20 世纪 80 年代后期,是数据库研究中一个很有应用价值的新领域,是一个多学科领域,从多个学科汲取营养。这些学科包括数据库技术、人工智能、机器学习、神经网络、统计学、模式识别、知识库系统、知识获取、信息检索、高性能计算和数据可视化等。

2 成才因素分析数据仓库实现

本课题从西南科技大学学生处与教务处获取了多个年级的学生信息以及成绩信息。在这些操作型源数据中存在着缺失数据、噪音数据、冗余数据等,如某个学生因重修存在多条成绩信息或人为导致其他错误重复信息等。在进行成才因素挖掘分析前,通过以下步骤将操作型数据源清洗、转换、装载到分析型数据仓库中:

(1)由事务型数据库作为源系统组成数据仓库与数据集市。

(2)从操作型数据库中抽取数据。

(3)将抽取获得的数据进行清洗,主要是对源数据中存在的一些脏数据进行修改。对离散字段使用一个全局的常数对缺失值进行填充;对连续字段使用一个计算值对缺失值以及存在冲突的值进行填充或者修改。

(4)对操作型数据源中的一些字段进行概念树提升,如"学生来源"字段将学生来源城市提升到省份再提升到区域。

(5)将进行过以上处理的数据进行转换、集成、装载到一个分析型数据仓库中。

(6)对其中一些典型的规律包括从两个或多个字段汇总数据创建汇总表或总量字段。

在建立数据仓库时,必须先确定采用哪种模型。在多维数据集架构中,星型架构是使用最频繁的架构模型,而在成才因素分析中,主要以学生入学信息表 student 为核心,事实表为学生成绩表 score,因为学校 99 年以前学生数据的信息所采用的学号与后期教务处采用网络办公平台所采用的学生登陆号之间存在着冲突,而学生成绩表采用的是学生登陆号,因此在两表之间必须建立一个转换维度。因此,学生成才因素分析的多维数据集架构采用雪花架构。其架构如图 1 所示。该架构模型的数据挖掘查询语言(DMQL)定义如下:

```
Define cube student [face,gender,grade,major,nation,comefrom]:
    Total_Score = Sum(project_score),
    Total_Score_Point = Sum(score_point),
    Count_here = Count(project_score),
     Averager_Score = [Measures].[Total_score]/[Measure].
[Count_here]
Define dimension face As (face)
Define dimension gender As (gender)
Define dimension grade As (grade)
Define dimension major As (major, department, kebie)
Define dimension nation As (nation)
Define dimension comefrom As (city, province, layer)
```

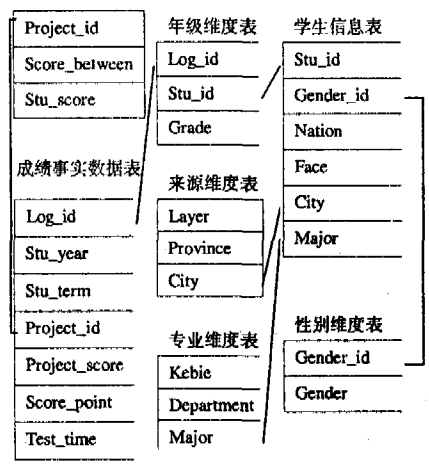

图 1 成才因素分析数据仓库的雪花架构

3 成才因素数据挖掘

在上面建立起来的数据仓库模型之上,采用 MS Analysis Services 中提供的两种数据挖掘算法:Microsoft Decision Tree 和 Microsoft Clustering,进行设置实现成才因素挖掘分析。

在 Analysis Services 中,数据挖掘模型(DMM)是一个虚拟结构,表示关系和多维数据的分组和预测分析。数据挖掘模型的结构主要由一组数据挖掘列和数据挖掘算法来进行定义。经过设置、生成和培训,可得挖掘结果模型。

(下转第 169 页)

表中自动生成与接收单位相对应的 N 个记录,每个记录对应一个单位。第二,接收单位签收通知时,就将其签收收据填入签收数据表,这样,发送单位通过查看签收数据表就可以了解通知签收情况。

(4)报文管理数据表:报文的处理流程虽然比较繁杂,但是与通知处理不同的是报文单位与审批部门或领导之间基本上是一对一的关系。虽然,偶尔也会有一对多的关系,但是少量的数据冗余在设计数据表时也是允许的。所以,在处理报文数据时只设计了一个数据表。数据表的字段包括:报文的标题、正文、报告单位、报告时间、审批单位、审批内容、审批人姓名、审批时间、批转标志、审批校领导姓名、校领导审批内容、校领导审批时间等,当然还包括一个关键字段,设计方法同上。在处理一个报告时,在数据表中只是对应一个记录,对报告的审批、转批、查看等操作都在这个记录中完成。

(5)文件管理数据表:因为文件管理的工作流程与通知管理类似,所以其数据表的设计也与通知管理相同。但是,由于文件的正文是采用扫描的方式,所以两者正文字段的属性不同:通知管理数据表的正文是"text"属性;文件管理数据表的正文是"image"属性。一份文件可能由多页组成,设计时要设多个正文字段,一个字段对应一页。

3　网络办公系统的安全保障和扩展

办公系统一个重要特征就是安全保证[4],整个网络办公系统的设计采用了三级安全保障。首先是整个系统的构架采用C/S模式,避免Internet上的恶意攻击。其次,对操作人员包括一般工作人员及领导登录系统时都要进行身份认证,因而在一定程度上保证了办公系统的保密性。报文管理模块中领导的批复和意见反馈是非常重要的,而且具有较高的保密程度,因此,还设计了相关的数字签名系统[5],接受报文并有权利批复的领导都有各自的私钥,批复后通过私钥签名,能有效地防止假冒。

以上系统设计都是建立在局域网内,但是如果涉及到多个局域网的办公系统可以通过构建分布式数据库的方式来保证系统可靠地运行。在每个局域网内建立一个数据库服务器,由于办公系统的实时性要求不高,信息、通知、报文等允许有适当的延迟,因此数据库之间通过异步复制的方式来交换数据,局域网内的客户端只需访问本地的数据库就可以完成正常的办公事务。

4　办公自动化系统的应用环境

网络办公化系统以校园局域网为基础,以Windows 2000作为网络操作系统,采用SQL Server数据库平台。由于在客户端不需进行数据处理,所以对客户机的要求不高,机器可以是Pentium系列的任一机型;操作系统可以采用Windows系列的任一版本。由于本系统的模块之间依赖性很少,因此,有较好的扩展性,可以根据用户的需求增加其它新的功能。

参考文献:

[1] 汤　丽,周传玉.办公自动化技术探讨[J].山东电子,2002(3):30-32.

[2] 张亚玲,李　微,张毅坤.办公自动化系统的网络规划与实现[J].现代电子技术,2001(2):50-52.

[3] 刘阶萍,杨长水,刘世军,等.深探SQL Server 7.0与电子商务开发应用[M].北京:机械工业出版社,2000.

[4] 张　青,柏永斌.办公自动化系统网络安全设计策略[J].微机发展,2003,13(6):67-68.

[5] 杨利英,陈基禄,李春祥,等.数字签名技术与校园网办公自动化系统[J].华北电力大学学报,2001(1):63-67.

(上接第166页)

因为MS Analysis Server的数据挖掘组件主要是面向数据库管理员的,所以对于数据挖掘以及多维数据集来说并没有提供完善的可视化图形界面,这对于决策层管理人员清晰快捷地从挖掘结果中获取感兴趣规则信息存在着一定的难度。鉴于此,结合平行坐标多维可视化等可视化技术设计开发了多维数据集及数据挖掘可视化平台MD&DM Player。

4　结束语

通过数据挖掘技术在成才因素分析中的应用研究,可以获取培养专才或高素质人才的一些隐藏规则。对高校的学科设置以及高校的本科教学质量控制将起到促进的作用。同时,目前学生的数据信息仍然不够完善,对学生个人的基本属性如性格、爱好、成长环境等无法获取,而这些信息的获取对学生素质教育分析将起到更大的作用。

参考文献:

[1] Inmon W H. Building the Data Warehouse[M]. New York: John Wiley & Sons Inc, 1993.

[2] Kalnis P, Papadias D. Multi-query optimization for on-line analytical processing[J]. Information Systems, 2003,28:457-473.

[3] Gal A, Eckstein J, Stoumbos Z. Scheduling of data transcription in periodically connected databases[J]. Stochastic Analysis and Applications,2001,21(5):1021-1058.

[4] 鲍　钰,黄国兴,张　召. 基于OLAP的上海社区服务网后台数据仓库的设计与实现[J]. 计算机应用研究, 2003(4):144-146.

[5] Levene M, Loizou G. Why is the snowflake schema a good data warehouse design[J]. Information Systems, 2003, 28(3):225-240.

[6] Hand D, Mannila H, Smyth P. Principles of Data Mining[M]. Cambridge: Massachusetts Institute of Technology, 2001.