

# 基于 SQL Server 2005 的高校人力资源数据挖掘的研究<sup>①</sup>

Based on SQL Server 2005 university human resources data mining research

宣军英 (嘉兴学院信息工程学院 浙江嘉兴 314001)

**摘要:**介绍了 SQL Server 2005 数据挖掘平台的优势和功能,提出了利用 SQL Server 2005 数据挖掘对高校人力资源数据源中的数据进行挖掘的方法,论述了挖掘模型对高校人力资源管理的支持。

**关键词:**SQL Server 2005 人力资源 数据挖掘

## 1 引言

进入 21 世纪,人才的竞争已经成为各个城市、各个企事业单位竞争的焦点。同样,高校的发展需要大批各个学科、各个层次的人才。目前,高校的人力资源结构基本上由教学科研人员、管理人员、服务人员、离退休人员组成,其中教学科研人员是主体,管理人员是关键,服务人员与离退休人员是补充。随着计算机技术的发展,许多高校相继建立了日益完善的人事信息管理系统,并积累了大量数据。为了充分利用这些数据,在大规模人力资源管理方面做出有价值的探索,可以利用数据挖掘发现有价值的知识,通过数据挖掘技术对高校人力资源数据源中的数据进行分析,寻找其中有价值的关系和规律,对教师聘用、培养、选拔等实际工作能够在一定程度上起到辅助作用,从而为制定人才需求规划、人才招聘和培养提供比较客观的决策支持。本文基于 SQL Server 2005 平台,通过对高校人力资源数据挖掘的研究,希望对高校人力管理起到一定辅助作用。

## 2 SQL Server 2005 数据挖掘平台

数据挖掘又称知识发现(Knowledge Discovery in Database,简称 KDD),是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中抽取有意义的、隐含的、以前未知的并有潜在使用价值的知识的过程,是一个多学科交叉性学科,它涉及统计学、机器学习、数据库、模式识别、可视化以及高性能计算等多个学科。

其目标是在大量的数据中发现令人感兴趣的模式。目前,国外有许多研究机构、公司和学术组织从事数据挖掘工具的研制和开发,自 1989 年出现以来,经过十多年的发展,数据挖掘技术已趋于成熟,并已投入商业应用。世界上比较有影响的典型数据挖掘系统有: SAS 公司的 Enterprise Miner、IBM 公司的 Intelligent Miner、SGL 公司的 SetMiner、SPSS 公司的 Clementine、Sybase 公司的 Warehouse Studio、RuleQuest Research 公司的 See5 等等。但其中很多工具对操作者提出了较高的要求,例如要精通统计学等理论。然而,SQL Server 2005 的出现改变了这一现状。SQL Server 2005 数据挖掘功能具有优于传统数据挖掘应用程序的众多优势,并非一个独立应用程序,而是开发智能应用程序的平台。SQL Server 2005 数据挖掘功能与所有 SQL Server 产品实现了集成,包括 SQL Server、集成服务 SQL Server Integration Services(SSIS)和分析服务 Analysis Services。SSIS 可用于数据预处理阶段,完成数据的转换、清洗加载过程,分析服务用于完成模式发现功能,两部分被有机的统一在 SQL Server 2005 数据挖掘,功能嵌入到整个过程中,可以实时运行,且结果可以发送回集成、分析或报告过程。

### 2.1 SQL Server 2005 数据挖掘功能的优势

(1) 易于使用 通过 SQL Server 2005,Microsoft 努力将数据挖掘从高深的实验室中搬出来,让负责设置和运行数据模型的开发人员和数据库管理员(DBA)、任何分析人员、决策者或可以使用模型输出的其他用

<sup>①</sup> 附:本文为嘉兴学院科研课题成果(编号:70106034,题目:高校人力资源的数据挖掘)

户也可以使用数据挖掘,而不需要具有任何专门知识。

(2) 简单而丰富的 API SQL Server 2005 中的数据挖掘功能具有一个非常强大却甚为简单的 API,这使得创建智能应用程序非常简单。利用该 API,无需了解每个模型的内部细节及其工作原理,就可从客户端应用程序调用预测模型。这使得开发人员可以根据分析的数据调用引擎并选择能提供最佳结果的模型。

## 1.2 SQL Server 2005 数据挖掘功能

(1) SSIS 结构及功能 数据预处理是数据挖掘之前的一项准备工作,需要花费较多的时间和精力。SQL Server 2005 中的集成服务器提供了数据抽取、转换、下载功能,大大简化了数据预处理阶段的任务。SSIS 基于工作流环境,提供了控制流(Control Flow)和数据流(Data Flow)两种工作流。一个 SSIS 项目中最基本的部署和执行单位被称为“包”,“包”就是装载 SSIS 工作流的容器。一个 SSIS 项目可以包含多个包,其中一个包只能包含一个控制流,但可以含有一个或多个数据流。

控制流也叫任务流,每个节点就是一个任务,这些任务按照事先定义的顺序执行,可以根据执行结果的不同有不同的分支。这些任务定义“包”需要完成的任务,比如下载数据执行 SQL 语句,激活其他程序等。数据流是专门针对数据操作的工作流,一个数据流又称为一个管道,其中的每个节点被称为一次转换。数据流包含于任务流中,要使用数据流,必须先定义任务流中的数据流任务。

(2) Analysis Services SQL Server Analysis Services 是用于联机分析处理(OLAP)和数据挖掘的中层服务器。Analysis Services 系统包括一个服务器,可以构造用于分析的数据多维数据集,同时 Analysis Services 系统还提供对多维数据集信息的快速客户端访问。Analysis Services 将数据仓库中的数据组织成包含预先计算聚合数据的多维数据集,以便为复杂的分析查询提供快速解答,Analysis Services 可以从多维(OLAP)和关系数据源这两者创建数据挖掘模型。

(3) Analysis Services 中的算法 SQL Server 2005 数据挖掘中包含了多种有效的数据挖掘算法,包括决策树、回归树、时间序列、聚集、序列聚集、关联规则、Naive Bayes、神经网络和文本挖掘。除了这 9 种算法外,用户还可以根据自己的需要加入其他算法。

## 3 SQL Server 2005 数据挖掘的过程

### 3.1 数据挖掘模型构建

OLE DB 是 Microsoft 定义的公用数据访问规范。许多数据存储产品(如电子表格、数据库或其它服务器应用程序)都提供有 OLE DB 提供程序,可供 OLE DB 应用程序在访问数据时使用。使用 OLE DB API 的应用程序可以访问任何有相应 OLE DB 提供程序的数据。OLE DB FOR DM 的核心部分是数据挖掘扩展语言(DMX)的定义,这是一种用于数据挖掘的类 SQL 语言。在数据挖掘模型应用方面,SQL Server 开启了新的篇章。Data Mining Extensions for SQL(DMX)的出现使得开发人员能非常容易地创建与数据挖掘相关的应用程序,使用数据挖掘模型进行预测的语句就像 SQL 查询中的 Join 查询,利用已经熟悉的工具和已具有的知识就能够操作数据挖掘技术。

在 SQL SERVER 2005 数据挖掘平台中,创建关系型挖掘模型的语法如下:

```
CREATE MINING MODEL <MODEL NAME> ( <Column Members> ) USING <Algorithm Name> . 它类似于建立一个关系表,其中包括关键字、输入、预测三类属性,每一个模型与一个数据挖掘算法相关联。
```

### 3.2 模型测试

数据挖掘算法利用输入的数据,分析属性间的关系,发现隐藏在数据背后的规律和模式,方法类似于关系表中的数据插入,语法为: INSERT INTO <model name> ( <Column Names> ) <Data>

## 4 高校人力资源数据挖掘的实现

### 4.1 高校人力资源挖掘的基本过程

(1) 建立高校人力资源数据库 通过整理人力资源信息,确定要挖掘的数据源,进行挖掘数据源的收集,建立 SQL Server 人力资源挖掘库。

(2) 数据预处理 由于人力资源数据库中的数据有许多是不完整、有噪声和不一致的,为了提高数据挖掘质量,利用 SSIS 对数据进行预处理,利用删除语句清理数据库中影响挖掘的数据,为数据挖掘做好准备。

(3) 数据挖掘 有了适合数据挖掘的数据,根据实际问题的需要,选择合适的数据挖掘算法,利用分析服务器进行模式发现。

(4) 挖掘结果分析 可以通过现有的模式分析,对人力资源的管理起一定的辅助作用,但要注意并不是每次挖掘的结果都是有价值的,有时候可能有误导性结果,需要重新分析挖掘的数据和选择挖掘算法。

#### 4.2 创建和测试人力资源挖掘模型

(1) 创建人力资源挖掘模型 高校人力资源数据源自学校人事处教职工数据管理中心。例如有数据表名为 RSGI,其中的字段名主要有 ZGBH(职工编号),GZXM(职工姓名),CSNY(出生年月),ZC(职称),XB(性别),XL(最高学历)等。数据存储于 sql server 数据库中,利用 SSIS 对数据进行预处理,Execute SQL Task 用于数据清理,通过删除语句清理数据库中影响挖掘的数据。

利用 DMX 语法构建模型:CREATE MINING rlzy\_dm (zgbh varchar(20) LONG KEY ,csny datetime discrete ,xb char(2) discrete,xl varchar(30) discrete, zc varchar(30) discrete PREDICT) USING Microsoft\_Ddecision\_Trees

在此数据挖掘模型中,属性值包括教师的年龄,性别,学历等,采用了决策树算法。

(2) 人力资源挖掘模型测试 上例为:INSERT INTO rlzy\_dm (zgbh,csny,xb,xl,zc)

Openrowset ('sqldb','servername','loginname','password',

'Select zgbh,csny,xb,xl,zc from rsgl')

使用 OPENROWSET 命令,将一个 SELECT 语句查询传给 SQL OLE 数据库服务器。'sqldb'为 OLE DB 提供者的名字,'servername'为要连接的数据库的名字,'loginname','password'为登陆数据库的用户名和密码。通过分析教职工的 csny,xb,xl 属性,预测其 ZC。

### 5 挖掘模型对高校人力资源管理的支持

通过对现有模式的分析,数据挖掘对高校人力资源管理的支持,主要有以下几个方面:

#### 5.1 合理制定教师培训、招聘计划

利用分类算法对教师的职称进行分类,对不同年龄、学历、职称级别的教师提供进修、培训的机会。确定哪些专业的人才可以通过内部培训产生,哪些专业人才浅缺而且急需,需要通过引进解决,从而达到平衡

内部人才和外部招聘人才的关系。为调整师资队伍结构,专业教师分布均衡,以此制定教师进修、培训、招聘计划。

#### 5.2 有利于人才的选拔

高校的发展需要年轻有为、德才兼备的干部,从高校人力资源数据库中提取数据,对现有人才进行聚类分析,从而发现各种人才的类型及特征,有利于决策者制定组织人才发展战略,进行人才的选拔。

#### 5.3 制定有竞争力的薪金福利体系

综合这几年离职的教职工的数据,通过聚类算法分析他们离开单位的种种原因,加上分析现有教职工的薪金结构,制定合理的薪金福利体系、奖惩规定,达到吸引、留住杰出人才的目的。

### 6 结束语

SQL SERVER 2005 数据挖掘平台提供了简单而实用的数据挖掘方法,所用算法除了分析服务中自带的多种算法外,还能够根据挖掘的情况输入更有针对性的算法,并且实现了数据挖掘和报告的集成,可以通过简单灵活且可伸缩的方式向组织中的任何人提供预测结果,特别是可以将挖掘结果直接反馈给前端的应用程序。因此,SQLSERVER 2005 数据挖掘具有较大的优势。

#### 参考文献

- 1 Jiawei Han ,Micheline Kamber. 数据挖掘概念与技术[M],北京:机械工业出版社,2001.
- 2 ZhaoHuiTang, JamieMaclennan. Data Mining With SQLSERVER 2005 [M]. USA:Wiley ,2005.
- 3 Yidao . Sql Server 2005 data mining 简介 [EB/OL]. <http://blog.csdn.net/yidao/archive/2006/09/10/1204970.aspx>. 2006. 9
- 4 王媛媛、钟永恒,基于 SQLSERVER2005 的 Web 日志挖掘系统构建[J],现代图书情报技术 2006,(5): 58? 61.
- 5 安淑芝,数据仓库与数据挖掘[M],北京:清华大学出版社,2005.
- 6 SQLSERVER 2000 Books on line product Documentation