广西科学院学报
Journal of Guangxi Academy of Sciences
2006,22(4):275～278
Vol. 22,No. 4 November 2006

基于 SQL Server 分析服务构建数据挖掘方案的研究

A Constructing Data Mining Project Based on SQL Server Analysis Services

贺 杰,郭 慧
HE Jie,GUO Hui

(梧州学院计算机科学系,广西梧州 543002)
(Department of Computer Science, Wuzhou College,Wuzhou,Guangxi,543002,China)

摘要:在分析 SQL Server 分析服务的基本结构的基础上,提出基于 SQL Server 分析服务构建数据挖掘解决方案,并用实例说明这种方案的优点与可行性。

关键词:数据挖掘 SQL Server 分析服务
中图法分类号:TP311 **文献标识码**:A **文章编号**:1002-7378(2006)04-0275-04

Abstract: The basic structure of analysis services of SQL server is discussed. A data mining project based on the analysis services of SQL Server is presented. The advantages and feasibility of the project are explained in an example.

Key words: data mining, SQL Server, analysis services

目前市场上有形形色色的专用和通用数据挖掘软件,但是整体而言,数据挖掘工业仍比较零散,每一个数据挖掘软件的开发商都有自己的数据挖掘包,它们适用的范围单一,软件开发者很难把它们与不同的知识发现工具结合起来应用。大多数商业数据挖掘产品不能在关系数据库里直接进行数据挖掘,使用这类产品从关系数据库抽取数据需要一个中间存储形式,但是数据的导入和导出以及数据转换是非常昂贵的操作,将数据挖掘在传统数据库系统的框架中无缝集成,是数据挖掘面临的最为关键的挑战之一。微软公司的数据库产品 SQL Server 中包含了数据挖掘特性,本文利用它的数据挖掘组件——分析服务,直接从数据库或者数据仓库中进行数据挖掘操作,实现数据挖掘与数据库以及应用程序的紧密耦合,从而大大提高数据挖掘效率。

1 SQL Server 的分析服务

SQL Server 的分析服务采用开放的体系结构[1],是一整套的决策支持引擎和工具,并且无缝集成了多种符合 OLE DB For DM 和预测模型标记语言(PMML)的数据挖掘算法。分析服务的访问是通过图形用户界面工具,并且通过微软管理控制台接口运行[2]。它可以访问两种形式的决策支持机制:数据挖掘和联机分析处理。

分析服务可以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,从而为复杂的分析查询提供快速解答,使用户可以从多维(OLAP)和关系数据源这两者来创建数据挖掘模型,还可以对这两类数据应用数据挖掘模型。分析服务的结构可以划分为服务器部分和客户机部分,其中,服务器部分存储各种功能和服务的引擎,创建和管理 OLAP 多维数据集及数据挖掘模型,并以多维结构、关系数据库或标准化 XML 格式的 PMML 形式存储数据挖掘模型的数据[3],而元数据、用于定义多维数据集的信息、数据挖掘模型和服务器上其它对象由 Meta Data Services 存储在关系数据库中[4]。并通过 PivotTable 服务为客户端提供数据。客户机部分用于提供前端应用软件界面。为访问服务器上的 OLAP 数据和数据挖掘数据,我们通过 PivotTable 服务为客户端服务。客户端应用程序使用 C++ 的 OLE DB 接口或组件对象模型(COM)自动化语言(例如 Microsoft Visual Basic)的 Microsoft ActiveX

收稿日期:2006-07-03
修回日期:2006-08-17
作者简介:贺 杰(1982-),男,助教,主要从事数据库,网络图形学的教学工作。

对象(ADO)模型来连接到 PivotTable 服务。

2 基于 SQL Server 分析服务构建数据挖掘的解决方案

基于 SQL Server 分析服务构建数据挖掘解决方案主要有以下几步:(1)组织数据挖掘源数据;(2)建立数据挖掘立方体;(3)从服务器端或者客户端构建、训练数据挖掘模型;(4)利用数据挖掘模型进行预测查询。接下来按照以上步骤来说明基于 SQL Server 分析服务构建数据挖掘解决方案。

2.1 组织数据挖掘源数据

组织数据挖掘源数据包括以下几个方面:(1)根据挖掘主题,设计数据仓库或数据集市。通常采取的设计模式有两种:星型模式和雪花模式。在选取模式之后,根据需求主题设计事实表和维代码表。(2)为挖掘主题建立关系数据库。该数据库是为数据挖掘主题准备源数据的数据集市。创建数据库后,根据设计的事实表和维代码表创建相应的表格,并且创建相关的索引。(3)提取和加载数据。使用数据转换服务,把分散在不同业务数据库的相关数据经过抽取、转换和装载,汇集、规范到事实表中。

2.2 建立数据立方体

建立数据立方体时,需要利用分析服务创建数据库,这个数据库本质上是虚拟的,用于存放 OLAP 服务结构的对象,包括 5 个对象:(1)Data Source:用于存放数据库的数据源;(2)Cubes:用于存放立方体对象;(3)Shared Dimensions:存放可以用于所有立方体的维;(4)Mining Models:存放数据挖掘模型;(5)Database Roles:存放数据库中的角色信息。数据立方体的建立过程如下:

(1)指定数据源。在每一个数据库中,都可以指定一个或多个数据源,为立方体、数据挖掘模型等对象提供数据。使用 OLE DB Provider 清单,可以完成指定数据源的要求。

(2)维和立方体。通过使用 Shared Dimensions 对象,可以创建共享维。然后再利用 Cubes 对象,可以选择数据源、事实表、维和度量等来创建一个立方体。

(3)存储立方体。建立后的立方体有三种存储类型可供选择:①MOLAP:数据和聚合都存储在多维结构中。需要花很长时间才能把数据从数据集市或数据仓库传送到多维数据库中,但具有查询快速而有效的特点。②ROLAP:数据和聚合都存储在关系型数据库中。其占用磁盘空间少,但是查询时间较长。③HOLAP:利用 MOLAP 和 ROLAP 的优点,将数据存储在关系数据库中,将聚合存储在多维数据库中,使用分区管理工具来实现灵活的立方体存储。

2.3 构建、训练数据挖掘模型

SQL Server 分析服务提供了服务器端和客户端两种架构,利用其中集成的数据挖掘方法可以分别从服务器端和客户端构建数据挖掘模型。

2.3.1 服务器端构建数据挖掘模型

在服务器端通过应用程序创建和训练数据挖掘模型,只能使用决策支持对象(DSO)。DSO 中公开的对象具体体现了 Analysis Service 中对象的内部结构,这样可以比较容易的通过编程的手段来对 Analysis Service 进行管理控制,DSO 对象模型如图 1 所示。

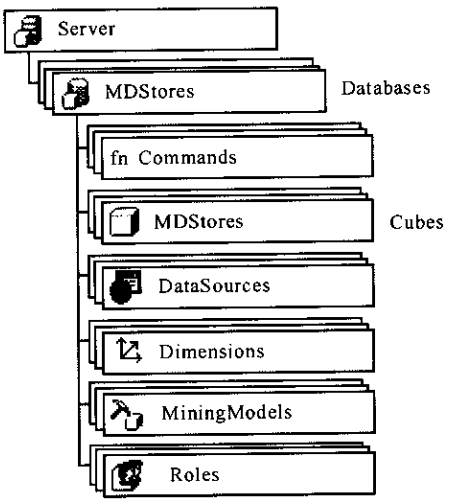

图 1 DSO 对象模型

Server 对象是 DSO 层次结构模型树中的根对象,可以通过它来进行与 Analysis 服务器相关的一些操作。Server 对象使用 Server 接口,Server 对象用于:与分析服务器的连接或断开操作;启动,暂停或者停止 Analysis 服务器服务(MSSQLServer OLAPService)引擎;提供 Analysis 服务器的诸如版本和编辑等详细信息;创建其它 Service 对象,如数据库,数据源,命令集,维,立方体,数据挖掘模型及角色等;管理 Analysis Services 中锁定的对象,控制多用户状态下对象的读写操作。

Database 对象代表着 Analysis Services 中的一个数据库。数据库里含有立方体和数据挖掘模型,二者处在同一层次上。数据库对象只能通过 DSO 的 Server 对象中的 MDStores 集合来进行操作,而数据库对象中只有立方体这个对象能够被该接口访问。Database 用于创建、编辑及删除一个分析服务中可用的命令集、数据源、立方体、维及数据挖掘模型等,

同时 Database 用于数据库对象的事务管理。

利用 DSO 可以在服务器端创建一个关系型或者 OLAP 型数据挖掘模型,然后在应用程序中可以利用各种高级语言操作 DSO 对象模型来完成数据源连接、数据库的创建等工作,应用程序中创建一个新的数据挖掘模型,需要以下几步:(1)连接到一个分析服务器并得到一个指向希望创建一个新的数据挖掘模型的数据库的对象指针。(2)创建新的挖掘模型对象。(3)为挖掘模型的第一列创建一个 Column 对象。(4)为 Column 对象设置属性。(5)使用对象模型的更新方法保存挖掘模型。(6)使用对象模型的处理方法训练数据挖掘模型。

2.3.2　客户端构建数据挖掘模型

许多情况下,本地机器足以处理本地使用的数据挖掘模型,在与网络断开时也需要执行数据挖掘任务。这时就要用到 PivotTable Service 来创建客户端应用程序,该程序可以用来创建本地数据挖掘模型,而且无需通过集中数据到远程服务器就能对其进行查询。客户端构建数据挖掘模型时,首先要使 PivotTable Service 连接到 Analysis Services 中的数据挖掘模型,这种连接可以通过本地的 OLE DB Provider 实现,如 ADO,也可以通过 HTTP 在互联网上实现。当使用 URL 作为连接的数据源特性,可以使用适当的端口,通过 Internet 穿过防火墙直接与分析服务连接,这种连接可以通过使用一种特殊的 Active Server Pages 网页 Msolap.asp 来完成。也可以编写代码利用 URL 作为数据源,以 ADO 的形式来连接到默认的分析服务器。

2.4　利用数据挖掘模型进行预测查询

使用 SQL Server 分析服务中的预测连接(PREDICTION JOIN)完成预测查询,一个预测连接的完成需要一个数据挖掘模型并指定它与新数据的关系,然后利用该模型对新数据进行预测[5]。同时 OLE DB For DM 中制定了大量的预测函数,利用这些预测函数可以得到大量的附加信息,例如预测精确度,最大可能性的行的统计偏差等信息。

3　软件开发实例

在 Visual Basic 环境下,构建一个顾客工资水平与其性别、教育水平以及婚姻情况之间的关系为例,说明基于 SQL Server 分析服务构建数据挖掘解决方案。我们设计的数据库中存有一个大型超市的顾客情况表,该表中详细记录了该超市的会员的情况,包括年龄、性别、婚姻、居住地址、收入水平等,以此来分析这些顾客教育水平以及婚姻情况和他们的收入之间的关系。顾客的资料已经预先存储到了相应的数据库,所以不需要再组织数据挖掘源数据,而直接将数据通过 DTS 存储到 SQL Server 数据库。

3.1　指定系统数据源

在使用分析服务之前首先要指定系统数据源,在 Windows 操作系统中,可以在控制面板中打开"数据源"(ODBC)管理器,如图 2 所示,通过在系统 DSN 选项卡中单击添加项来指定 FoodMart 数据库作为系统数据源。

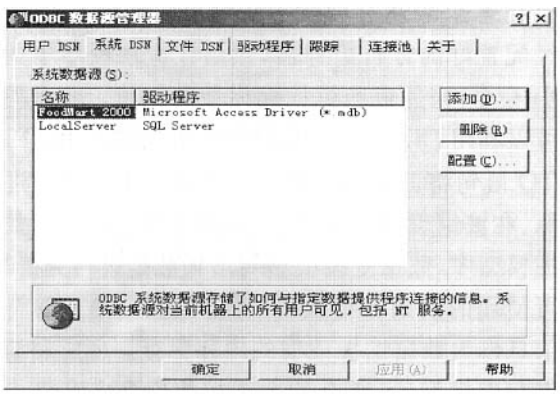

图 2　指定系统数据源

3.2　指定立方体的数据源

启动分析服务器,在分析服务器中构建数据立方体,首先要在分析服务器中创建一个数据库,该数据库是一个虚拟的对象,但是它包含了数据挖掘模型对象。分析服务器中建立虚拟数据库以后,需要它指定数据源,通过分析服务其中的数据源选项卡可以指定立方体的数据源,如图 3 所示。

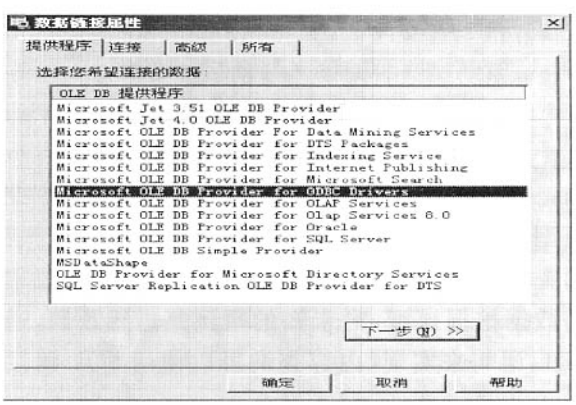

图 3　指定立方体数据源

3.3　构建基于分析服务的数据挖掘模型

在应用程序中构建基于分析服务的数据方法如下:(1)如果应用程序运行在服务器端,则可以使用 DSO 来构建数据挖掘模型,首先要定义 DSO 对象模型,并将 DSO 连接到分析服务器并得到

FoodMart 数据库的指针，然后调用 AddNew 方法来制定挖掘模型对象是基于关系型数据还是 OLAP 数据，再就是对各个属性列的重复定义和数据挖掘算法的选择这样就完成了微软决策树数据挖掘模型的创建。挖掘模型构建好后还是一个空的容器，现在需要保存挖掘模型，并且训练挖掘模型。通过利用 DSO 对象模型在应用程序中直接利用关系数据库中的数据完成了数据挖掘模型对象的构建和训练，可以在分析服务中查看所构建的数据挖掘模型，如图 4 所示。(2)如果应用程序运行在客户端，那么首先要在客户端通过使用 PivotTable Service 连接到已经配置好的分析服务器，然后在应用程序中运行 DDL 命令，来完成数据挖掘模型的创建。客户端同样可以利用 DDL 训练挖掘模型，可以使用 INSERT INTO 语句将训练数据插入到模型中来对模型进行训练。在该进程中，数据被插入到未经过培训的数据挖掘模型中，数据挖掘模型对培训数据进行分析，找出可使用的规则和模式，以确定预测列的值并将统计信息作为数据挖掘模型的内容保存。生成和培训数据挖掘模型后，所有的数据挖掘模型内容将作为数据挖掘模型节点存储。

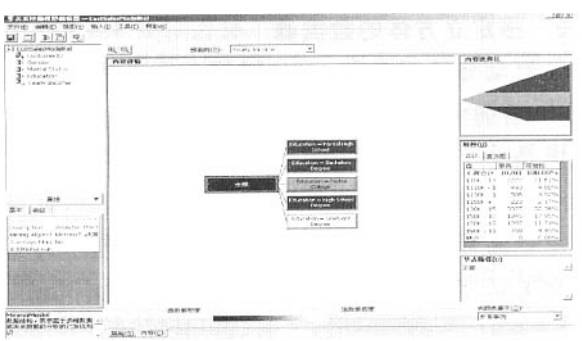

图 4 CustomerModel 挖掘模型

3.4 利用数据挖掘模型进行查询

经过培训的数据挖掘模型节点可以深入分析数据，从而提供有价值的信息。为获得已培训数据挖掘模型的信息，SQL Server 提供了两种查询方式：预测查询和内容查询。预测查询可以使用事先已培训的数据挖掘模型，对未知的事例集进行预测，通过 SQL 语句来实现，内容查询可以使用 SQL 语法查看挖掘模型培训后的信息。通过数据挖掘模型浏览器，也可以实现对模型内容的查看。数据挖掘模型浏览器中显示的信息，代表数据挖掘模型通过检查培训数据而学到的趋势统计模型，其图形化的表示有利于进一步理解培训数据所表示的一般模式和规则，并能够精确调整这些模式和规则以更好地适合培训数据。

4 结束语

本文分析了 SQL Server 的服务器端和客户端架构，在这两种结构体系中，分别使用 DSO 和 OLE DB For DM 中指定的 DDL 语言完成了数据挖掘模型的创建、训练以及做出了相应的预测，而且利用 SQL Server 自带数据挖掘算法来进行数据挖掘操作，使应用程序和关系数据库以及数据挖掘无缝集成。这种方式大大简化了数据挖掘操作的复杂性，使普通的程序员也可以使用类 SQL 语言来完成数据挖掘操作，由此为数据挖掘的广泛应用提供了一个切实可行的方法和工具。

参考文献：

[1] Microsoft corporation. SQL Server 2000 Books Online [DB/OL]. [2006-07-03]. http://www.microsoft.com.downloads/details.aspx? FamilyID: 469793eo-db8c-469e-aac4-271dfb74ea70&Displaylang=en.

[2] 东方人华. SQL Server 2000 与 Visual Basic.NET 数据库入门与提高[M]. 北京：清华大学出版社，2002：5-13.

[3] FABIO ARCINIEGAS. XML Developer's Guide[M]. New York：McGraw-Hill Companies Inc，2001：97-121.

[4] MIKE GUNDERLOY，TIM SNEATH. SQL Server Developer's Guide to OLAP with Analysis Services [M]. Alameda：SYBEX Inc，1999：336-339.

[5] DORIAN PYLE. 业务建模与数据挖掘[M]. 杨冬青，马秀莉，唐世渭，等，译. 北京：机械工业出版社，2005：185-191.

(责任编辑：凌汉恩 邓大玉)