基于 SQL Server 2005 的数据挖掘技术在石油储运中的应用

张 镇 （后勤工程学院研究生 3 队 重庆 400016）
张振军 （长安大学研究生院 西安 710064）
石 进 （后勤工程学院研究生 3 队 重庆 400016）

摘要：介绍了 SQL Server 2005 数据挖掘平台及其功能，较详细地给出了 SQL Server 2005 数据挖掘过程，论述了基于 SQL Server 2005 的数据挖掘技术在石油储运信息中的实现。通过对石油储运的数据资源挖掘的研究，希望能为石油储运中存在问题的解决给予一定的帮助。

关键词：SQL Server 2005 数据挖掘 石油储运

1 引言

石油在人类社会的发展和进步中起着越来越重要的作用。石油运输和储存所涉及的信息量日益剧增，如何在这一海量信息中及时地发现和提取有用的信息，是目前石油储运工作者面临的亟待解决的问题。石油储运是指从油田开采出来的原油经管道等途径运送到炼油厂，成品油由炼油厂生产出来之后，经过各级油库运送至加油站，供给消费者的整个流通过程，在这个过程中，即包括油品的流动，还包括资金的流动，各种流动的信息量巨大。随着计算机技术的发展，许多石油公司相继建立了日益完善的石油信息管理系统，并积累了大量的数据。为了充分的利用积累的数据，为石油储运信息解决方案做出有益的探索，可以利用 SQL Server 2005 数据挖掘平台来寻找有价值的知识，并通过数据挖掘技术对石油储运的数据源进行分析综合，寻找它们的规律，对石油储运中存在的诸如成本偏高、整体运行效率偏低、某种油品的剩余数量和剩余时间不明确等问题的解决，能够起到一定的辅助作用，从而为降低成本、提高效率等提供具有一定价值的参考和支持。

2 SQL Server 2005 数据挖掘

SQL Server 2005 是一个全面的数据库平台，使用集成的商业智能（BI）工具提供了企业级的数据管理。SQL Server 2005 数据库引擎为关系型数据和结构化数据提供了更安全可靠的存储功能，可以构建和管理用于业务的高可用和高性能的数据应用程序。SQL Server 2005 数据引擎是企业数据管理解决方案的核心。此外 SQL Server 2005 结合了分析、报表、集成和通知功能。这使得企业可以构建和部署经济有效的 BI 解决方案，帮助团队通过记分卡、Dashboard、Web services 和移动设备将数据应用推向业务的各个领域。其与 Microsoft Visual Studio、Microsoft Office System 以及新的开发工具包（包括 Business Intelligence Development Studio）的紧密集成使 SQL Server 2005 与众不同。SQL Server 2005 都可以提供创新的解决方案，帮助您从数据中更多地获益。

2.1 SQL Server 2005 数据挖掘平台

所谓数据挖掘，就是从大量的、不完全的、有噪声的、模糊的、随机的数据集中识别有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程。它是一门涉及面很广的交叉学科，包括机器学习、数理统计、神经网络、数据库、模式识别、粗糙集、模糊数学等相关技术。其目标是在大量的数据中发现有用的信息。一般来讲，数据挖掘工具根据其适用的范围分为两类：专用数据挖掘工具和通用数据挖掘工具。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究和开发，自 1989 年出现以来经过十几年的

发展，数据挖掘技术已趋成熟，很多工具以投入商业应用。在世界上比较著名的开发工具有 IBM 公司的 Intelligent Miner、SAS 公司的 Enterprise Miner、SPSS 公司的 Clementine、Data Mining Technologies 公司的 Nuggets?、NCR 公司的 Teradata Warehouse Miner? 等等。其中 SAS 公司的 Enterprise Miner 在我国的企业中采用的比较多。但是其中好多工具对使用者提出比较高的要求，它要求使用者要精通数理统计、计算机语言等方面的理论。然而，SQL Server 2005 的出现改变了这种状况。SQL Server 2005 数据挖掘功能具有优于传统数据挖掘工具应用程序的众多优势，并非一个独立的应用程序，而是开发智能应用程序的平台。SQL Server 2005 提供了一些工具，可用于创建解决特定业务问题的数据挖掘解决方案。在 Business Intelligence Development Studio 中，使用数据挖掘向导可以轻松创建基于 OLAP 和关系数据源的挖掘结构和挖掘模型。可以使用该向导定义结构和模型，使用特定的数据挖掘技术来分析数据。还可以使用数据挖掘设计器来定义挖掘模型，并浏览和利用模型的结果。SQL Server 2005 Integration Services（SSIS）提供了一些工具，可用于自动完成常见的数据挖掘任务，如处理挖掘模型和创建预测查询等。SQL Server 2005 在数据挖掘时对使用者不要求什么专门的知识，自身具有一个功能非常强大而甚为简单的 API，这使得创建智能应用程序非常简单。

2.2 SQL Server 2005 数据挖掘功能

（1）SQL Server 2005 集成服务数据挖掘。在典型的数据挖掘项目中，最消耗资源的步骤是数据准备。它包括数据收集、数据清理和数据转换。我们可以使用 SQL 数据脚本来准备数据，但是用来准备数据的更好工具是 SQL Server 集成服务（SSIS）。SSIS 提出了控制流和数据流。SSIS 中最基本的部署和执行单位是包。SSIS 包是 SSIS 流的容器。它除了包含控制流和数据流，还包含了 SSIS 连接和包的变量。SSIS 项目中可以包含多个包。包只包含一个控制流，而该控制流可以包含一个或多个数据流。

在任务流中，包通常包括多个任务。多个任务按优先权约束的顺序来进行组织的。优先权约束按以下顺序连接两个任务：执行第一个任务的结果决定是否执行第二个任务。可以在工作流中使用优先权约束来创建条件分支。可以将多个优先权约束进行组合，然后将其作为一个约束来求值。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。数据流总是包含于任务流中。有一个特殊的任务，其名称为数据流任务，该任务是用于容纳数据流的容器。

（2）SQL Server 2005 Analysis Services。SQL Server 2005 Analysis Services（SSAS）为商业智能解决方案提供联机分析处理（OLAP）和数据挖掘功能。Analysis Services 通过允许开发人员在一个或多个物理数据源中定义一个称为统一维度模型（UDM）的数据模型，从而很好的组合了传统的基于 OLAP 分析和基于关系报表的各个最佳方面。基于 OLAP、报表以及自定义 BI 应用程序的所有最终用户查询都将通过 UDM（可提供一个此关系数据的业务视图）访问基础数据源中的数据。

（3）SQL Server 2005 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，包括贝叶斯特、决策数、时序算法、聚类算法、序列聚类、关联规则、神经网络、回归树和文本挖掘。除了这 9 种算法外，用户还可以加入自己需要的算法。

2.3 SQL Server 2005 数据挖掘过程

（1）创建数据访问对象。在拥有数据库对象之后，下一步是创建数据源对象和数据源视图（DSV）对象。数据源对象相当简单，只由连接数据库的连接字符串组成，而 DSV 稍微复杂一些。DSV 主要的元素是模式，模式是增加了自定义属性的标准 Dataset 对象。为了将模式加载到 DSV 中，则要为希望加载的每个表创建数据适配器，然后将这些表的模式加入到某个数据集中。然后增加任何必需的关系，最后将数据集加入到 DSV 中（然后，DSV 会被加入到 AMO 数据库中）。

（2）创建数据挖掘结构。挖掘结构描述了数据挖掘引擎可以理解的问题领域。必须创建 Mining Structure Columns，然后指定它们的数据类型、内容类型以及与 DSV 中的源列的数据绑定。

（3）创建数据挖掘模型。OLE DB 是 Microsoft 定义的公用访问规范。许多数据存储产品均提供有 OLE DB 提供应用程序，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,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,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。例如,DMX 查询与以下所示类似。

SELECT TOP 25 t. CustomerID

FROM Customer ChurnModel

NATURAL PREDICTION JOIN

OPENQUERY ('Customer DataSource', 'SELECT * FROM Customers')

ORDER BY Predict Probability ([Churned], True) DESC。

(4) 数据挖掘模型测试。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,其语法为:INSERT INTO < model name > (< Column Names >) < Date >

3 石油储运数据挖掘的实现

3.1 石油储运数据挖掘基础

(1) 建立石油储运数据库。对于石油储运数据库的建立应该用系统的观点去看待这个问题。也就是说,要全面的、系统的分析,综合的考虑各方面的因素,通过整理石油储运信息,确定要挖掘的数据源,进行挖掘数据源的收集,建立 SQL Server 石油储运挖掘库。由于石油储运领域信息如此广博,在建立石油储运数据库之初,要想做到面面俱到是不可能的,我们要抓住主要环节,先从大的范围构筑框架,收集数据。

(2) 数据预处理。通过填写空缺值,平滑噪声数据,识别、删除孤立点,并解决"不一致"来"清理"数据;将多个数据源合并成一致的数据存储来集成数据;通过 SSIS 对数据进行预处理包括数据清理、数据集成、数据转换、数据规约四步。这样可以解决数据的不完整、有噪声和不一致的问题,有助于提高挖掘质量,节省挖掘时间。

(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,选择适当的挖掘算法,利用分析服务器进行模式发现。

(4) 利用数据挖掘模型进行预测查询。所有生成一个数据挖掘模型的努力都是围绕它的查询能力,并利用这种能力从测试实例中预测未知的值。但是,需要注意的是并不是每次挖掘的结果都是有价值的,有的时候可能有误导性的结果,需要重新分析挖掘的数据和挖掘算法。

3.2 创建和测试石油储运挖掘模型

(1) 创建石油储运挖掘模型。石油储运数据源于石油储运公司信息管理中心。例如有数据表名为石油储运信息 sycyxx,其中字段名主要有油品编号 YPBH、油品名称 YPMC、入库时间 RKSJ、入库数量 RKSL、油品等级 YPDJ、出库时间 CKSJ、出库数量 CKSL、剩余数量 SYSL、剩余储存时间 SYSJ 等。数据存储在 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数据挖掘的数据。

利用 DMX 语法结构构建模型:CREATE MINING SYCY_DM (YPBH varchar(30) LONG KEY, YPMC char(12) discrete, YPDJ char (4) discrete, RKSJ, CKSJ datetime discret, RKSL, CKSJ varchar (30) LONG KEY, SYSJ datetime predict, SYSL datetime predict) USING Microsoft_Decision_trees

在此数据挖掘模型中,属性值包括油品的等级、入库时间、入库数量、出库时间、出库数量、编号等,采用了决策树算法。

(2) 石油储运挖掘模型测试 上例:INSERT INTO SYCY_DM(YPBH, RKSJ, RKSL, CKSJ, CKSL, SYSJ)

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

'select YPBH, RKSJ, RKSL, CKSJ, CKSL, SYSJ, SYSL from sycyxx')

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 SQL OLE 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。通过分析石油储运信息的 YPBH、RKSJ、RKSL、CKSJ 和 CKSL 属性,预测其 SYSL,SYSJ。

4 结论

基于数据仓库和数据集市的数据挖掘是一个新兴

(下转第 97 页)

(上接第 93 页)

而有实用价值的研究领域,目前该技术已经广泛应用于社会生活的各个领域。本文分析了 SQL Server 2005 数据挖掘平台及其数据挖掘功能,利用数据挖掘技术可对石油储运数据进行更深层次的分析。SQL Server 2005 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具有针对性的算法,但是数据挖掘的任何一种算法都不是万能的,针对不同的问题应采用不同的方法去解决。数据挖掘的实现是一个复杂的过程,数据挖掘工具能够嵌入整个过程中,可以实时运行,且结果可以传递给集成、分析或报告过程,可以通过简单灵活的方式向组织中的任何人提供分析及预测结果。

参考文献

1 Rick Dobson, Beginning SQL Server 2005 Express Database Applications, http://www.apress.com, 2006.

2 刘翔,数据仓库与数据挖掘技术[M],上海:上海交通大学出版社,2005 28-33.

3 朱建平,数据挖掘的统计方法与实践[M],北京:中国统计出版社,2005.6 1-9.

4 张水平,SQL server 数据库应用技术[M],西安:西北工业大学出版社,2005.4 1-3.

5 刘芝怡、常睿,数据挖掘在 SQL Server 2005 中的应用[J],电脑知识与技术,2006(6)156-157.