文章编号: 1008-1542(2005)02-0156-04

# 基于超市 OLAP 系统数据挖掘的实现

付瑞雪[1], 杨会志[2], 张　峰[3], 武建章[1]

(1. 石家庄经济学院管理科学与工程学院, 河北石家庄　050031; 2. 河北科技大学经济管理学院, 河北石家庄　050018; 3. 河北科技大学研究生学院, 河北石家庄　050018)

摘　要: 针对SQL Server 2000 Analysis Services 中的数据挖掘功能, 及其在某大型超市的 OLAP 系统中加以应用的情况, 介绍了基于 SQL Server 2000 的数据挖掘模型的建立、训练、浏览模型和进行预测的关键技术, 给出了一种建立在已有的 OLAP 系统基础上的实现数据挖掘功能的具体方法, 并指出运用该方法可以快速地开发数据挖掘的应用, 有效地支持决策。

关键词: 联机分析处理; 数据挖掘; 多维数据; DSO

中图分类号: TP311.11　　文献标识码: A

# Implementation of data mining based on a OLAP system for supermarket

FU Rui-xue[1], YANG Hui-zhi[2], ZHANG Feng[3], WU Jian-zhang[1]

(1. School of Management Science and Engineering, Shijiazhuang University of Economics, Shijiazhuang Hebei 050031, China; 2. College of Economics and Management, Hebei University of Science and Technology, Shijiazhuang Hebei 050018, China; 3. Graduate School, Hebei University of Science and Technology, Shijiazhuang Hebei 050018, China)

**Abstract:** This paper briefly introduces functions of data mining in SQL Server 2000 Analysis Services and their application in a OLAP system for supermarket. It also discusses the implementation techniques of creation, training, browse and prediction of data mining model, and supplies a method of implement ation of data mining based on a OLAP system. This method can rapidly develop application of data mining and support decision efficiently.

**Key words:** OLAP; data mining; multidimensional data; DSO

为了从大量的数据中发现和利用有价值的知识、模式和规则, 支持领导的决策, 并更好地了解顾客以改进客户支持功能, 在数据仓库的基础上采用联机分析处理(OLAP)技术。建立 OLAP 系统的技术数据量非常大, 要识别真正有价值的信息以及找出这些信息之间的关联, 就需要对这些数据进行深层分析, 因此笔者在已有的系统上应用数据挖掘技术, 对原有系统进行了扩展。

## 1　超市原有系统介绍

原有的系统采用客户机/服务器的 3 层体系结构, 包括数据服务层、应用服务层和数据访问层。数据服务层由数据仓库服务、数据集成和转换部分组成。OLTP 产生的信息数据、历史资料数据和外部数据通过 SQL Server 2000 提供的 DTS 服务输入数据仓库; 应用服务器由 OLAP 服务器和 Web 服务器组成; 用户访问层包括使用 Web 浏览器、管理客户端和分析客户端; 使用 Microsoft SQL Server 2000 Analysis Services

收稿日期: 2004-06-20; 责任编辑: 张　军
基金项目: 河北省教育厅基金资助项目(2002259)
作者简介: 付瑞雪(1973-), 男, 河北怀来人, 硕士研究生, 主要从事决策支持系统方面的研究。

作为 OLAP 服务器[1]。体系结构见图 1。

管理客户端是在 Visual Basic 6.0 中使用 DSO(Decision Support Objects)在程序中对 OLAP 各种数据结构进行全面管理;分析客户端是在 Visual Basic 6.0 中使用 ADO MD(ActiveX Data Objects, Multi Dimension)来访问多维数据集,利用 MDX(多维表达式, multidimensional expressions)对检索的数据进行下钻、上卷、切片和切块等操作;网上分析模块采用 MSE(Microsoft Script Editor)为开发工具,利用 ASP(Active Server Pages)技术、ADO MD 技术和 MDX 建立 Web 与分析服务器的接口,实现利用 Web 浏览器对服务器的访问。

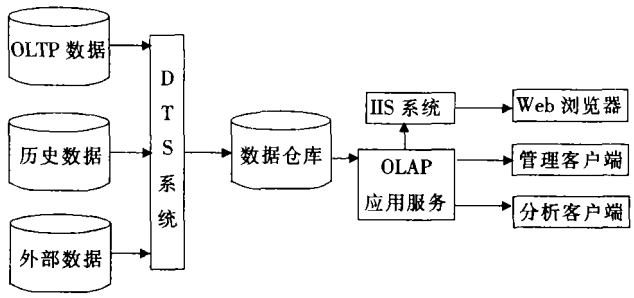


图 1　体系结构

Fig. 1　System structure

## 2　SQL Server 2000 的 Analysis Services 数据挖掘功能简介[2,3]

### 2.1　有关概念介绍

SQL Server 2000 的 Analysis Services 主要由 2 部分构成,即 OLAP 和数据挖掘。OLAP 的目标是满足决策支持或多维环境特定的查询和报表需求;数据挖掘是在大量的数据中通过自动化或半自动化的方法去分析和探索抽取出潜在的,有价值的知识、模型或规则的过程。在 Analysis Services 中的数据挖掘模型可以建立在现有的多维数据集或者直接建立在关系型数据库的数据上。一个数据挖掘模型是由 Analysis Services 创建的结构,它代表一个在多维数据集或关系型数据库中建立的数据聚集。数据挖掘模型是由事例(case)和属性(attributes)构造的,被设计用来分析的数据挖掘模型实体称为一个事例,属性是有关事例的信息。在选择完一个数据挖掘模型的事例和属性后,就可以开始"训练"数据挖掘模型。数据挖掘和 OLAP 都是数据分析的重要技术,自动化或半自动化的知识发现是数据挖掘和 OLAP 的区别之一。数据挖掘综合了人工智能、数据库及统计分析技术,而 OLAP 主要是基于 SQL 和一些聚合技术。数据挖掘和 OLAP 又是互补的,例如:在销售立方体中的顾客维中有很多顾客,要在众多的顾客中找到顾客的购买模式是很困难的,数据挖掘技术可以根据顾客的特征和数量来分析顾客的聚类。

### 2.2　算法简介

SQL Server 2000 包含 2 种数据挖掘算法:Microsoft Decision Trees 算法和 Microsoft Clustering 算法,即微软决策树和微软聚类。决策树算法是被设计用来在大量的数据上建立决策树的,它构造决策树,最重要的信息被定位在树的根部,而比较不重要的信息向树的叶子方向分布。决策树能够处理多维数据,并且发现的规则很容易被理解。微软聚类算法寻找数据中的事例群组,它的典型用途是以现有事例为基础,对新事例做出预测[4]。

### 2.3　建立和训练模型

数据挖掘模型是 OLE DB for DM 中提出的概念。数据挖掘模型在某种程度上可以看成是由许多不同数据类型的列构成的关系表,其中有些列是输入的列,而有些列则是预测列。数据挖掘模型就是容器,但是,数据挖掘模型和关系表是有区别的,它并不存储行数据,而是存储数据挖掘算法在关系表中发现的模式及必须描述与它相关的挖掘算法以及任何已存在的参数列表。以下是基于现有的"Sales"多维数据集手工建立"myModel"的步骤。

在 Analysis Manager 树视图中,右击"Sales"多维数据集,打开挖掘模型向导,在"选择数据挖掘技术"步骤中选择"Microsoft 决策树";在选择"事例"步骤中,在"维度"框中选择"Customer",在"级别"框中选择"name";在"选择被预测实体"步骤中,选择"事例级别的成员属性",然后在"成员属性"框中选择"MemberCard";在"选择训练数据"步骤中,滚动到"Customer"维度,清除"Province"和"City"框(因为不需要在聚集级别上而只需要在单独的客户级别上确定客户模式);在"创建维度和虚拟多维数据集(可选)"步骤中,在"维度名称"框中输入"myModelDim"。然后在"虚拟多维数据集名称"框中输入"myModelCube"。

数据挖掘模型建立以后,是空的容器。在训练阶段,数据挖掘算法分析放入的实例并且把它已发现的模

式存于挖掘模型。为了与 SQL 一致, OLE DB for DM 采用了数据插入查询的语法。训练完后, 数据挖掘算法发现了存于数据挖掘模型之内的模式, 用户能浏览这个挖掘模式, 或者利用这个已训练好的挖掘模型进行预测。

## 3 开发编程

### 3.1 建立模型和训练

如果要在服务器上创建一个由 Analysis Services 来管理的数据挖掘模型, 只有通过 DSO 来实现。如果想在自己的磁盘上创建一个本地的数据挖掘模型, 就要通过 PivotTable Service 来使用 DDL 来实现。使用 PivotTable Service 无需通过服务器端的 Analysis Services 就可让用户查询、创建、修改和删除本地模型。在 Visual Basic 6.0 中使用 DSO 可以对数据挖掘模型进行管理, 控制 OLAP 服务器及其各种数据结构的全面管理工作, 生成数据挖掘模型。使用 DSO 建立数据挖掘模型可以摆脱手工建立及维护数据挖掘模型的方式。在 DSO 对象模型中, 服务器、数据库、立方和分区对象都有自己的集合, 称为 MDStore, 这个集合存在于各个级别。MDStore 集可以包含数据库对象、立方对象、分区对象和聚集对象。从简单的自动维护过程到构造完整的数据库程序 DSO 均可以创建各种自定义解决方案。下面是利用现有的数据库、数据源和多维数据集建立数据挖掘模型的主要代码:

```
Dim myServer As New DSO. Server
Dim myDB As DSO. MDStore
Dim myDmm As DSO. MiningModel
Dim myColumn As DSO. Column
Dim myRole As DSO. Role
myServer. Connect"servername"' 连接到 OLAP 服务器
Set myDB = myServer. MDStores("databasename")' databasename 为数据库名字
Set myDmm = myDB. MiningModels. AddNew("modelname", sbclsOlap)' 增加新模型
Set myRole = myDmm. Roles. AddNew("All Users")' 创建角色
myDmm. MiningAlgorithm = "Microsoft_Decision_Trees"' 使用决策树算法
myDmm. SourceCube = "Sales"' 建立在现有的"Sales"多维数据集上
myDmm. CaseDimension = "Customers"' 选取"Customers"为事例维度
myDmm. TrainingQuery = ""' 让 DSO 进行训练查询
myDmm. Update' 依据"Sales"的结构自动加入列
Set myColumn = myDmm. Columns("Profession") 对"Profession"列进行设定
myColumn. IsInput = True' 处理前使"Profession"列成为输入列
myColumn. IsDisabled = False' 处理前使"Profession"列激活, 同样在处理前使 Name 列激活, Marital Status, Yearly Income 列激活并作为输入列, 同时对 MemberCard 使用 myColumn. IsPredictable = True 使 MemberCard 成为预测列
myDmm. LastUpdated = Now' 设置更新日期为现在
myDmm. Update' 保存模型的元数据
myDmm. LockObject olapLockProcess' 处理模型前进行锁定
myDmm. Process processFull' 处理模型
myDmm. UnlockObject' 解锁
```

### 3.2 浏览挖掘模式

可以使用 mining_model_content 常量为模型选择"架构行集"来达到抽取一个架构行集的目的。架构行集是被设计用来为 ADO 从一个 OLE DB 提供者返回元数据的方法。每一个架构行集是由一个常量指定的, 它包含了实际的挖掘模型的节点, 打开一个架构行集, 可以使用 ADO 的 OpenSchema 方法。OpenSchema 方法有 3 个参数, 第 1 个对于数据挖掘常为 adSchemaProviderSpecific; 第 2 个是一个约束的数组, 它用来限制返回的架构信息; 第 3 个是指定某个架构的常量。例如: 在训练完利用决策树算法的数据挖掘模型

后,这个模式行集中就包含了树结点信息,基于这个模式行集的内容,用户的应用程序能够显示它的树型图。以下是使用 TreeView 控件浏览模型内容的主要代码:

```
Dim myCon As New ADODB.Connection
Dim myRes As New ADODB.Recordset
Dim varrestrict As Variant    ' 下面是架构常量
Const mining_model_content As String = "{3add8a76- d8b9- 11d2- 8d2a- 00e029154fde}"
strCon = "provider= msolap; datasource= servername; initial catalog= databasename"
myCon.ConnectionString = strCon    ' strcon 是连接字符串
myCon.Open    ' 进行连接
varrestrict = Array(Empty, Empty, "modelname")    ' 约束数组
Set myRes = myCon.OpenSchema(adSchemaProviderSpecific, varrestrict, mining_model_content)
```

' 打开架构行集,此时可以利用 TreeView 控件对 myRes 记录集内的内容进行显示,此时记录集的字段的名字(name)是"node_unique_name","parent_unique_name","node_caption","node_probability","node_rule"等,利用 TreeView1.Nodes.Add 的方法显示字段的值(value)。显示的结果见图 2。

### 3.3　预　测

进行预测需要一个已训练了的数据挖掘模型和一组新的实例,预测的结果是一个包含预测列值的新的记录集,这个过程总体来说非常类似于关系连接。不同于连接 2 个表的是,预测连接的是一个数据挖掘模型和一个输入表,叫预测连接。通常,预测是基于一个实例集的,同时也能在基于单个实例的基础上做预测,称这些预测查询为单一查询。语法如下:

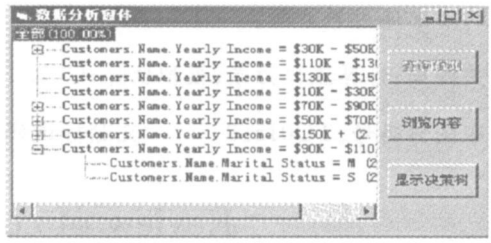


图 2　数据挖掘模型的浏览结果

Fig. 2　Result of browse of data mining model

```
SELECT[ FLATTENED] < SELECT- expressions>  FROM < 数据挖掘模型>
PREDICTION JOIN < 被预测的新事例>  ON < join condition>
[ WHERE < WHERE- expression> ]
```

表达式中 Flattened 表示返回一个 2 维的记录集,可以使用 ADO,而不是 ADO MD; ON 子句规定了数据挖掘模型中的列如何匹配从 OpenRowset 语句返回的列;使用 OLAP 数据挖掘模型作为预测的基础,需要在 OpenRowset 语句中使用 Microsoft Shape Provider 来创建 1 个整理过的行集,使它和数据挖掘模型的源数据具有相同的结构, OpenRowset 是 Microsoft 到 SQL 的扩展,可从 OLE DB 数据源创建一个行集(虚拟表);OLE DB for DM 同时也定义了一系列的预测函数,它能够在预测语句中的 SELECT 子句中被调用,这些函数将返回预期值的概率,以及相关概率、最高期望聚类 ID 等。

在应用程序或者 ASP 页中定义查询语句 strQuery 字符串后,就可以使用 ADO 执行这个预测查询。例如: Set myRes= myCon.Execute(strQuery)。

## 4　结束语

笔者采取在原有的 OLAP 系统的基础上实现数据挖掘的功能,既减少了开发的时间和成本,又可以利用已有的 OLAP 系统来对数据挖掘维度和关于它的多维数据集进行分析,具有很好的支持决策效果。但是设计和实现一个支持数据挖掘功能的系统是一项十分复杂的工程,还需要增加一些数据挖掘算法,并经过反复的调查研究和设计。

参考文献:

[1]　WILLIAM C A. SQL SERVER OLAP 开发指南[M]. 北京: 电子工业出版社, 2000.
[2]　王　珊. 数据仓库技术与联机分析处理[M]. 北京: 科学出版社, 1999.
[3]　武延军. 精通 ASP 网络编程[M]. 北京: 人民邮电出版社, 2001.
[4]　[美]希德曼(Seidman C). SQL Server 2000 数据挖掘技术指南[M]. 刘　艺译. 北京: 机械工业出版社, 2002.