数据仓库元数据管理模式的分析与比较

聂　茹, 张　虹
(中国矿业大学 计算机科学与技术学院, 江苏 徐州 221008)

摘　要: 介绍了典型的元数据管理策略,并对三种元数据管理模式进行了详细的分析与比较; 最后提出了一种分布式的元数据管理模式。

关键词: 数据仓库; 元数据; 元数据库
中图法分类号: TP392　　文献标识码: A　　文章编号: 1001-3695(2005)02-0057-02

Comparison Research of Metadata Management Module in Data Warehouse

NIE Ru, ZHANG Hong
(School of Computer Science & Technology, China University of Mining & Technology, Xuzhou Jiangsu 221008, China)

Abstract: Categorizes the major metadata management strategy, makes a detailed analyze and comparison of three kinds of metadata management modules, and presents a distributed metadata management module.

Key words: Data Warehouse; Metadata; Metadata Repository

元数据是数据仓库的一个重要组成部分,是联系数据仓库中各部分的纽带。它作用于数据仓库的创建、维护、管理和使用的各个方面,是内部技术人员开发与维护数据仓库的蓝图,是业务终端用户导航数据仓库以及定位有用信息的路标[1]。而元数据管理是构建、管理、维护和使用数据仓库系统的核心部件[2]。

元数据管理的一个重要方面是元数据的互操作以及与数据访问及分析工具的集成,这一点在元数据管理中非常重要,其最终目的是无缝连接数据仓库环境中的各个工具,实现数据仓库装载、管理、维护和使用的一体化及半自动化[3]。因此,元数据管理模式不仅制约了元数据的使用和普及,甚至可以直接关系到数据仓库项目的成败。

1　元数据管理的两种策略

1.1　元数据仓库

建立一个有较宽专业覆盖范围的元数据标准,让所有行业的元数据均按照此标准进行定义和管理,并通过扩展策略来支持新类型元数据的加入,这是元数据仓库的根本出发点。建立元数据访问和整个元数据生命周期管理的系统——元数据仓库(Metadata Repository),作为元数据访问和聚集的平台[4]。这种元数据管理策略包含一个元数据管理体系结构的框架,这些体系结构框架要求:

(1) 定义通用的元数据标准,描述元数据的结构和语义,这种标准必须是与厂商无关并独立于特定实现技术的。

(2) 对于标准的定义必须有一个标准的建模语言。

(3) 标准的定义必须有一定的逻辑抽象性,能够被多种技术在多种软、硬件平台上实现。

(4) 对于此标准的不同厂商的具体实现系统之间要有一种元数据交换标准,以实现元数据互访和元数据的分布式管理。

这种元数据管理策略以 OMG 的 MOF 和 MDC 的 OIM 为代表[5]。

1.2　元数据交换

元数据交换途径包括:元数据桥和元数据交换标准[6]。

(1) 元数据桥沟通不同工具或应用,提供面向批处理的互操作。现在只有少数几家公司提供这种桥接器,没有明显的市场影响力。

(2) 元数据交换标准允许不同工具用同一种数据格式来共享元数据(标准定义了共享元数据的结构和语义,但没有定义元数据如何在表示层被使用);允许不同领域有不同种类,如地理界的地球空间元数据标准、人文学界的 TEL Header、档案界的 EAD 元数据集等,其中在 IT 业中的以 CDIF, XML 为代表[5]。

2　元数据管理模式的分析与比较

目前,存在三种典型的元数据管理模式,每一种管理模式都对数据仓库的管理和使用等方面具有相应的优势和缺陷,但它们又同时具备一些共同的基本特征和功能。元数据的集成和互操作是元数据管理系统应该具备的基本的和最重要的功能之一。

2.1　集中式元数据管理模式

企业范围内的信息管理要求信息环境中的异构产品能够对元数据进行全局的、高效的访问,因此,企业级的集中式元数据库的数据协调方案应运而生[7]。为了集成企业范围内的不同开发工具和元数据库,共享元数据环境必须提供一致的共享方法,使得元数据能够被一致地存储、管理、集成和全局访问。其参考结构如图 1 所示。

企业级中心元数据库作为元数据集成的平台,一方面有助

收稿日期: 2004-01-11; 修返日期: 2004-05-12

于管理整个企业中数据仓库或数据集市的关键数据;另一方面有助于所有的参与者都能共享通用的数据结构、商业规则定义和企业各系统间的数据定义,而无须提供元数据交换机制。但是,使用统一的元数据模型,并将所有的元数据集中存储在中心元数据库,不利于数据仓库的维护,造价昂贵且实施困难,并且不同的数据仓库工具只能直接访问中心元数据库,而不能自治地局部存储和管理元数据。Microsoft 和 Platinum 开发的 Repository 就是这种面向整个企业的集中式元数据管理模式。

2.2 分散式元数据管理模式

为了解决中心元数据库管理模式存在的缺陷,目前大多数数据仓库系统中采用一种基于交换机制的元数据管理模式。这种管理模式通过建立相应的元数据交换标准,使得不同的数据仓库工具能够使用不同的数据模型、不同的表示方式,而这些工具之间可以通过元数据交换标准进行连接和通信(图 2)。在数据仓库的实施中先建立各部门的数据集市,再在其基础上建立企业级数据仓库,这种管理模式的最大优点是不同的工具可以高度自治地访问局部元数据库,提高了访问速度;但是系统需要提供元数据交换机制来满足不同局部元数据库之间的互操作和连接等问题,相应地增加了系统的负担[8]。另外,这种模式形成了冗余、分离的数据局面,其数据和元数据分散在多个系统中,增大了协调和管理它们的难度,产生了许多问题。OMG 提出的 CWM 规范已经成为这个数据仓库业内共同支持的元数据管理的唯一事实标准。

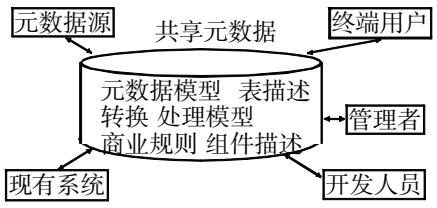

图 1 集中式元数据管理模式

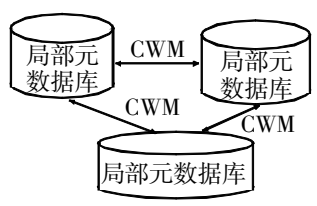

图 2 分散式元数据管理模式

2.3 共享式元数据管理模式

共享式元数据管理模式结合了前两种方法的优点,并且克服了它们的诸多弊端,(图 3)[9]。元数据库采用前述的元数据标准(如 CWM),建立一个从原始元数据到规范元数据的包装器,将分散在不同厂商的数据仓库工具中的元数据转换成统一的元数据语言,并且提供统一的关键数据结构和业务规则,将企业内的各数据集市组合到一起。

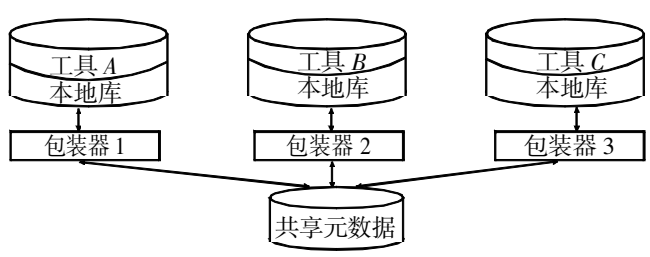

图 3 共享式元数据管理模式

元数据库作为元数据的集成平台,包括用于操作和查询元数据的机制,使整个机构都能方便地访问元数据。这样不但有利于管理所有数据仓库和数据集市中出现的元数据,而且还为各个数据集市提供了一个统一的数据结构和业务规则。

3 分布式元数据管理模式

随着网络技术的飞速发展,传统的数据仓库元数据管理体系已逐渐不能适应数据仓库分布化、异构化的要求,相应的分布式数据仓库元数据管理体系亟待推出并逐步完善。在对传统元数据管理模式进行深入分析和比较研究的基础上,本文提出了基于分布式的元数据管理模型(图 4)。

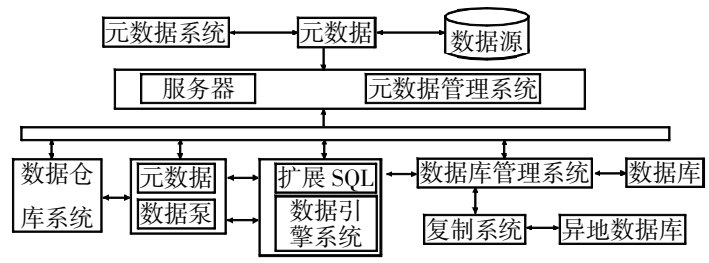

图 4 分布式元数据管理模式

模型中的元数据是分布在各个站点的数据仓库中数据的描述性信息的集合。数据仓库是网络环境下的全数字信息流,其数据来自网络环境下的分布式存储的各类数据库。数据仓库通过元数据系统动态抽取分布式信息来满足用户决策支持的需要。

模型中的元数据管理系统是一个对各站点元数据库、全局共享元数据库及其相互之间的关系进行协调与管理,并提供面向用户的应用接口与应用界面的软件平台,它包括如下子系统:

(1) 元数据管理系统;

(2) 请求响应端管理系统;

(3) 数据仓库系统,根据决策问题及数据仓库的设计方案,驱动数据引擎,抽取分布式存储的信息,构建数据仓库。

4 元数据库的实现

元数据管理的任务主要有两个方面:①负责存储和维护元数据库中的元数据;②负责数据仓库建模工具、数据获取工具、前端工具之间的消息传递,协调各模块与工具之间的工作。元数据库系统结构如图 5 所示。以下我们通过对数据仓库各个主要功能模块的具体分析,全面地了解元数据管理的具体实现。

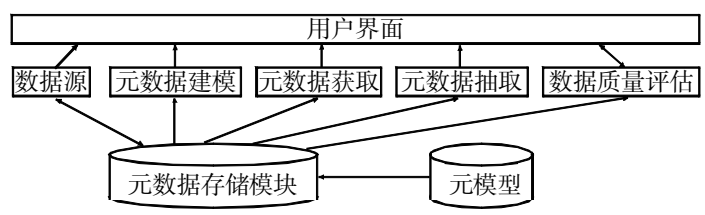

图 5 元数据库系统结构

4.1 数据源

数据仓库的数据来自于企业内外的多个异构的数据源,它们可以是各种类型的数据库、文本文件,甚至是 Web 数据。具体来说,主要包括数据源的元数据、数据模型的元数据、数据源与数据仓库映射的元数据等[2]。

4.2 数据仓库建模

数据仓库建模工具帮助用户通过需求分析创建仓库的数据模式,同时还能够定义模式中各个表的数据来源、数据转换规则、有关聚集操作等信息。具体转换规则有如下几种[8]:

(1) 基本的转换规则有两个:①一对一抽取,即源数据库中某一列直接抽取到数据仓库中的对应列,可分为直接抽取和通过聚集函数抽取两种情况。②多对一抽取,即两个或两个以上的源数据库中的列被抽取到数据仓库的一个列中,通过算术运算符或通过 Union 操作完成。

(2) 复杂的转换规则可以由上面的基本转换规则得到。

(3) 转换规则中不可避免会用到一些函数,当抽取时需要一些特殊的操作,允许用户编程定义一些自定义函数。(下转第 61 页)

然运算过于复杂,因此可以选用 Floyd 算法或是根据著名的旅行商问题(亦称货郎担问题) 的解法求解。在求得最优路径的基础上,再根据现有车辆运行情况可确定车辆调配计划。

(2) 机构设施地理位置的选择

对于供应商、第三物流企业、配送中心、用户而言,需求和供给这两方面总是存在着空间分布上的差异。此外无论是供应商还是销售商其服务范围和销售市场范围也具有一定的空间分布形式,因此机构设施的布局是电子商务下物流管理所必须面临的问题,其合理程度直接影响利润获取的多少。

机构设施地理位置的选择包括位置的评价和优化。评价是对于现有设施的空间位置分布模式的评价,而优化是对于最佳位置的搜寻。地理位置的合理布局实质上就是在距离最小化和利润最大化两者之间寻求平衡点。

现有的针对市场功能区域进行空间分析和模拟的模型很多,如 Peily 的零售重力模型、Batty 的裂点方程、Tobler 的价格场和作用风以及空间线性优化模型。空间线性优化模型是由一系列的边界条件和一个目标函数组成。边界条件即规划目标的约束条件,目标函数代表最大限度可能达到的目标。如某一公司要在某市设立 x 个配送中心,要求分配中心完全覆盖这个城市,而且每个配送中心的顾客数量大致相等。在求解过程中,首先可在城市交通图上标出居民地的空间分布位置,分析已有的供应点和潜在的供应点,按照给定的条件列出需求点和供应点的二元矩阵;根据矩阵约简方法,排除多余供应点,求出满足条件的且最少量的配送中心。详细方法可参考文献[2] 。

(3) 辅助决策分析

物流系统与企业和用户有着最直接的联系,对消费和市场的分析,可以帮助企业制定正确的生产和销售计划。GIS 提供全方位的信息,历史的、现在的、空间的、属性的。通过这些可以获得客户资料以及与企业相关的综合数据,如用户的历史购买力、购买行为、年龄构成、地理分布;所在区域的交通状况、经济发展程度、消费水平等。在空间数据上集成各种信息,并以此为基础,进行如消费趋势分析、销售力量分析、目标市场分析以及潜在客户分析等,为管理者提供决策支持。

4　结论

地理信息系统与电子商务和物流历史上是完全独立,且分开发展的不同系统,但是在这样一个信息化、网络化的时代,几种技术的整合是难以避免的。无论从特点上、体系结构上、操作的可行性上来讲,它们的结合都是切实可行的,而且是有价值的。将 GIS 技术引入电子商务下的物流管理,一方面能够开拓 GIS 的应用领域,促进其自身的发展;另一方面,也可以完善电子商务下物流管理体系,更好地满足物流管理信息化的要求,提高物流分析技术,帮助解决电子商务下物流配送的瓶颈问题。

GIS 技术和现代物流技术都是在 20 世纪 80 年代初传入我国,发展历史都不长,理论和实践上都还有待进一步的提高。而将 GIS 应用于物流分析和物流研究中,还处于起步阶段,在应用中,基于 GIS 的物流管理信息系统的计算机体系的建立、数据库的建设,以及物流分析模型都还有待于更深入的研究。

参考文献:

[1] 张铎. 电子商务与物流[M] . 北京: 清华大学出版社, 2000.
[2] 陈述彭, 鲁学军, 周成虎. 地理信息系统导论[M] . 北京: 科学教育出版社, 2000.
[3] 郭仁忠. 空间分析[M] . 武汉: 武汉测绘科技大学出版社, 1997.
[4] 王大力, 陈联. 基于 GeoSurf 建立绿色食品交易中心网站的可行性研究[J] . 测绘软科学研究, 2002, 8(2) : 37-40.
[5] 张广军, 荣朝和. 物流分析系统 GIS 仿真[J] . 中国流通经济, 1999, (4) : 19-22.

作者简介:

杨瑾(1973-) , 女, 重庆人, 长安大学讲师, 博士研究生, 主要研究方向为地理信息系统的教学与研究; 陈晏辉(1973-) , 男, 宁夏银川人, 工程师, 硕士, 主要研究方向为电子商务及物流管理。

(上接第 58 页)

4.3　数据抽取与转换

从源系统的数据到数据仓库中的目标数据的转移是一项复杂的工作,主要涉及到两个问题: ①抽取工作之间的复杂关系; ②源数据与目标数据之间的映射,即源表与目标表之间的一种复杂的多对多的关系。

4.4　元数据的浏览与导航

元数据浏览是分门别类地组织和显示各种元数据,供数据仓库管理员或最终用户根据需要浏览或查看他所关心的元数据。当然,不同的用户要通过用户权限使用自己访问级别内的数据。

4.5　数据质量评估

数据仓库把数据从源事务系统移到数据仓库中的目的是用于决策支持,这对数据质量提出了更高的要求。保证质量的第一步是建立支持商业目标的数据质量期望标准以及达不到该标准所冒的风险有多大。衡量数据质量的公共参数包括: 准确性、完整性、一致性、相关性、时间性、唯一性和有效性。

5　结束语

本文讨论了数据仓库实现过程中非常重要的一环,元数据的管理。元数据就像一座桥梁,将数据仓库中的数据与用户有机地结合起来,它不仅在整个数据仓库系统,而且在整个决策支持系统中,都起着非常重要的作用。本文分析、比较了元数据管理的几种模式,并针对目前元数据管理存在的问题,提出了分布式元数据管理模式。随着分布式系统的广泛应用,分布式元数据管理模式必将得到广泛应用。

参考文献:

[1] illiam A Giovinazzo. 面向对象的数据仓库设计[M] . 北京: 人民邮电出版社, 2000. 23- 25.
[2] Inmon W H. Building the Data Warehouse[M] . New York: John Wiley & Sons Inc., 1996. 67- 71.
[3] W H Inmon, Ken Rudin. 数据仓库管理[M] . 北京: 电子工业出版社, 2000. 58- 61.
[4] 王红兵. 数据仓库中的元数据[J] . 微机发展, 1999, 5: 109- 113.
[5] 戴超凡. 数据仓库中元数据技术研究[J] . 计算机工程与应用, 2001, 37(14) : 106-110.
[6] 戴超凡. 数据仓库中的元数据管理[J] . 计算机工程与科学, 2003, (4) : 65- 69.
[7] 胡颖峰, 卢美莲, 程时端. 数据仓库中元数据互通的研究[J] . 计算机应用研究, 2002, 19(11) : 30-31, 64.
[8] Anca Vaduva, K R Dittrich. Metadata Management for Data Warehousing: Between Vision and Reality[C] . 2001 Int1 Database Engineering & Amp; Applications Symp, 2001. 129-138.

作者简介:

聂茹(1976-) , 女, 江苏徐州人, 硕士研究生, 主要研究方向为数据仓库与分布式计算; 张虹(1941-) , 教授, 博士生导师, 主要研究方向为数据仓库与多媒体图像压缩技术等。