

基于 Web 的数据仓库集成环境研究^{*})

段江娇 黄震华 陈 昕

(厦门大学计算机科学系 厦门361005)

摘 要 随着 Web 技术的日臻成熟,基于 Web 的数据仓库技术显示了强大的优势,更好地适应了企业生存发展的需要。本文描述了一个基于 Web 的数据仓库集成环境(WIEDW)及其主要技术,将数据仓库的设计、部署、管理与利用集成在同一个平台中,使得数据仓库具有良好的整体性能,为用户提供一个数据仓库的设计、运行、维护平台环境,提高了企业建立、维护数据仓库应用系统对决策活动的实用性。该集成环境架构在 J2EE 软件体系结构上,利用了 Jdbc、JavaBean 等技术,使系统具有良好的可靠性、可扩展性及安全性。

关键词 数据仓库,集成环境,XML,Web

Research on Web-Based Integrated Environment for Data Warehouse

DUAN Jiang-Jiao HUNAG Zhen-Hua CHEN Xin LIU Xiao-Gang

(Department of Computer Science, Xiamen University, Xiamen 361005)

Abstract With the rapid growth of Web technologies, the Web-based data warehouse has become a dominant trend, because it can be more beneficial to enterprises. This paper describes a Web-based Integrated Environment for Data Warehouse (WIEDW) and the main technologies which integrate the design, running and maintenance data warehouse, and it will improve enterprises to build and maintain data warehousing for decision. The integrated Environment is based on the J2EE software architecture, which uses the techniques of Jdbc and JavaBean, and the reliability, expansibility and security of the system are favorable.

Keywords Data warehouse, Integration, XML, Web

1 引言

数据仓库技术目前已成为企业决策者进行管理分析和决策的主要手段,随着 Internet 的普及和 Web 技术的日臻成熟,人们访问和管理信息资源的方式发生了巨大的变化。WWW 信息资源的数据量每时每刻都在呈指数级增长,但是对其上的数据难以管理和访问,基于 Web 的数据仓库技术显示了强大的优势,更好地适应了企业生存发展的需要^[1,2]。如何有效地使用 Web 技术对数据仓库进行管理和访问,并且根据用户的综合需求提供方便易用的平台是数据仓库环境发展的趋势^[3]。

目前,人们提出了几种数据仓库的数据集成方法。最常见的方法是基于 XML 的集成^[4,5],这是因为 XML 具有很强的自描述能力、良好的跨平台特性,使得 XML 可以很好地表示各种外部数据源,并在异构数据库之间交换数据。但是,由于 XML 是在逻辑层描述并对数据进行集成,因此它的应用范围有一定的限制。文[6]提出了将 XML 表示数据转换成描述概念层的 UML 表示,并实现概念层的集成,使得数据仓库的集成可以提高到概念层的高度,具有更广泛的应用范围。除此之外的集成方法是使用专有的语言,如在 LORE^[7]系统中使用自定义的 Lore Query Language 语言实现外部数据源的集成。但是实际的数据仓库集成环境除了采用某种数据集成方

法外,关键的是要从整个系统出发,将从数据抽取、清洗到查询报表输出整个过程涉及到的数据仓库活动有机集成在一起,为用户提供更灵活有效的系统。因此,本文提出了基于 Web 数据仓库集成环境(WIEDW, Web-based Integrated Environment for Data Warehouse)。

WIEDW 以向导、自定义模板等用户界面元素为核心,以简单、通用、直观的软件工具作为用户的主要操作界面,如 Web 浏览器,更易于非专业用户使用,如使用图表实现多维数据的可视化输出。WIEDW 将数据仓库的设计、部署、管理与利用集成到一起,最终用户可以自行管理、维护已建立的数据仓库应用系统。通过对数据仓库质量的定量描述,计算质量指标的偏差,并反馈给数据抽取模块,使得进入数据仓库的数据是可控制的、有效的。各类企业都可以在该环境上快速建立数据集市与仓库,并可轻松实现复合转换,并对元数据实施强健的管理,同时监控数据的存储等。

2 WIEDW 的技术基础

2.1 数据抽取、转换等预处理

数据的抽取是数据进入数据仓库的入口,目前数据抽取的方案很多,而且该技术环节相对成熟。数据抽取、转换等在技术上主要涉及数据的互连、复制、增量、转换、调度、刷新和监控^[9]等几个方面。数据抽取实际上是对源数据库中的某些

^{*}) 本课题得到福建省自然科学基金资助(A0310008)和福建省高新技术研究开放计划重点项目资助(2003H041)。段江娇 讲师,研究方向为数据库理论与应用、分布式数据库、数据仓库、数据挖掘、网络技术等。黄震华 硕士研究生,研究方向为数据库理论与应用、分布式数据库、数据仓库数据挖掘等。陈 昕 助工,研究方向为数据库技术、数据仓库等。

数据进行描述、聚集、划分子集、数据变换及运算等过程。在用户操作时,可以选择源数据库中的子集,并利用该子集计算新的数据项,如计数、求和、联接和映射等。源数据可以是多种数据文件或不同文件格式。数据仓库的数据并不要求与联机事务处理系统保持实时的同步,因此数据抽取可以定时进行,但多个抽取操作执行的时间、相互的顺序、成败对数据仓库中信息的有效性则至关重要。

文[9]给出国内比较有代表性的数据仓库工具集 NGDW-1 的抽取机制:

(1)从元数据中读取数据源的连接字符串,连接数据源。

(2)从元数据中读取数据仓库的建模规则,生成数据抽取、数据仓库建模代码,并连接数据仓库数据库,建立数据仓库数据库数据模式。

(3)根据数据抽取原则,通过数据源提供的 ODBC 接口,将数据从数据源中读出来。

(4)根据元数据中定义的数据转换规则,将读出的数据作相关的类型转换或统计处理,以形成符合数据仓库数据库需要的数据。

(5)将处理的数据通过数据仓库数据库所提供的 ODBC 访问接口写入到数据仓库数据库中去。

2.2 元数据管理

元数据管理涉及到数据仓库的构造、运行、维护的整个生命周期,是企业级数据仓库构建过程中重要的一个环节。元数据可分为两类^[10-11],即技术性元数据、商业性元数据。元数据管理的目标是保证数据仓库于企业业务的运作得到同步,一旦需求发生变化,可以通过对元数据的维护使数据仓库的运行作出快速响应。元数据管理具有完全的开放性,通过分析企业的数据库,自动生成元数据并保存在元数据库中。

首先建立一个元数据访问和整个元数据生命周期管理的系统,即元数据仓储,它是元数据访问和集聚的平台。企业内所有的元数据都存放在此仓储中,实现元数据的集成。为了实现数据管理,需要定义一个具有较宽专业覆盖范围的元数据标准及数据定义的扩展策略。

其次,为了支持分布式的元数据仓储应用,需要建立一种元数据交换协议,不同系统中的元数据都通过这种协议来交换数据,从而把分布、异构的系统集成在一起。

2.3 数据的存储和管理

数据存储和管理是数据仓库的关键。数据仓库中数据量非常庞大,其访问方式也非常复杂,因此这些数据必须采用恰当的形式来存储。为了有效利用存储空间和支持用户决策,当前的数据仓库主要采用如下三种数据组织方式:虚拟存储方式、关系存储方式和多维结构存储方式^[12]。

(1)虚拟存储方式。采用这种存储方式的数据仓库没有专门的数据存储,数据仓库“视图”中的数据仍然在数据源中。当用户需要利用数据仓库中的信息时,根据用户的多维需求以及多维视图的定义,通过计算临时在数据源中找出所需的数据来完成多维分析。这种存储方式具有传统数据库“视图”的一切优点:组织方式简单,费用低廉,使用灵活,数据一致性易得到保证。当数据源频繁变化,且数据源的数据组织规范,数据的完备性好,同时又接近多维模型时,采用这种虚拟存储方式是比较恰当的,但是这种系统的运行效率较低,并且在一般的数据源中,数据的组织关系往往非常复杂,同时又存在许多

冗余和冲突,数据模型和多维模型相距甚远,因此数据仓库的多维语义层很难定义,难以建立实际的有效的决策支持系统。

(2)关系存储方式。数据仓库中的数据以2维表的形式,存储在关系数据库表结构中,通过 RDBMS 对这些数据的管理和操作来完成数据仓库的功能。

(3)多维结构存储方式。这是一种直接面向 OLAP 的数据组织形式,当前多数数据仓库产品采用这种存储方式。多维数据库中,数据以多维数组的形式来组织,同时为了方便操作,为这些数据构建了相应的维索引及数据管理文件。

2.4 支持多维分析的查询模式

用户在使用数据仓库时的访问方式与传统关系数据库有很大不同。对于数据仓库的访问往往不是简单的表和记录的查询,而是基于用户业务的分析模式,即联机分析。联机分析处理 OLAP 是对数据的一系列交互查询过程,这些查询过程要求对数据进行多层次、多阶段的分析处理,在 OLAP 中,多维数据集^[13](cube)是分析的基本单位。一个多维数据集代表了一个特殊的可查询区域,是由一定数目的维度和度量结合而成的整体存储单位。多维分析对多维形式组织起来的数据采取切片、切块、旋转等各种分析动作^[13],以求剖析数据,最终用户能从多个角度、多侧面地观察数据库中的数据。

3 WIEDW 的系统结构

图1所示为基于 Web 的数据仓库集成环境 WIEDW 的系统结构图。从系统的硬件组成来看,除了底层计算机网络设备外,整个系统由客户端浏览器、Web 服务器、应用服务器和数据仓库服务器组成。其中,应用服务器是 WIEDW 的关键部件,它负责数据的抽取、清洗、数据仓库建模以及 OLAP、数据挖掘和分析报表等,完成了整个数据仓库的设计、部署、管理与利用整个生命周期过程。应用服务器所处理的数据存储在数据仓库服务器中,逻辑上包含了元数据、数据集市和数据仓库。而 Web 服务器则主要为 WIEDW 的各类用户提供相关的操作界面及简单的处理逻辑。基于 Web 的数据仓库集成环境与传统的 C/S 的数据仓库环境的区别在于多了 Web 服务层和应用服务器。在系统软件结构设计上,应用服务器上的各个关键部件是以服务(Service)的形式封装了各种数据处理功能,通过 Web 服务器上的各种界面程序调用 Service 的形式实现相应的功能。WIEDW 中应用服务器采用基于 J2EE 结构的 Weblogic Server 6.1^[14]作为系统的运行平台,数据处理过程为各种 EJB 并封装成 Service,这样系统易于集成并具有良好的扩展性。

从系统使用人员角度来看,分为数据仓库构建维护人员、数据分析人员、决策人员和网络管理员。数据仓库构建维护人员负责数据仓库设计,完成数据从联机事务系统和工程数据库向数据仓库的迁移,进行必要的数据库清洗和数据格式转换,以及数据仓库的维护工作。数据分析人员运用 OLAP、数据挖掘等数据分析工具,对数据进行分析。网络管理员负责 Web 服务器的日常操作维护,并和网络维护人员一起共同负责整个计算机网络稳定运行。决策人员通过 Web 服务器请求访问由数据分析人员对数据仓库进行分析得到的结果,或是调用由数据分析人员在应用服务器上建立的应用程序模块,以动态网页格式返回请求结果,根据该信息进行企业决策。此外,决策人员、数据分析人员、数据仓库维护人员之间需要经常进

行沟通,数据分析人员要依据决策人员的要求提供相关信息,数据仓库维护人员则要依据数据分析人员的要求提供相应数据。

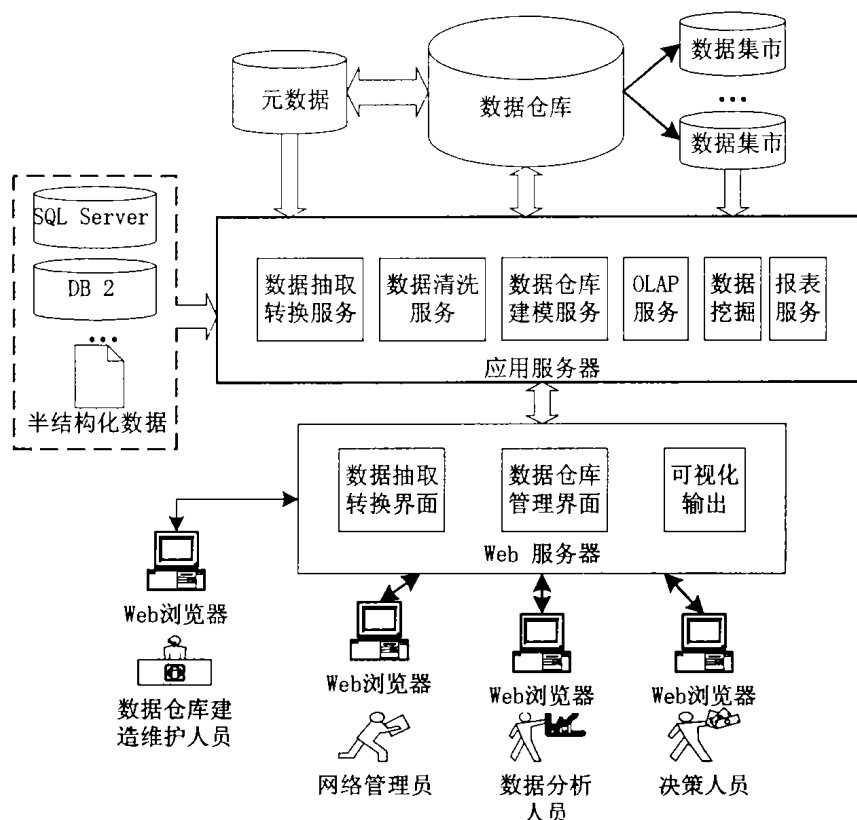


图1 基于 Web 的数据仓库集成环境的系统结构

4 WIEDW 的功能及实现

4.1 数据抽取与刷新

数据仓库从多个数据源获取数据,包括大型关系数据库系统、桌面数据库、文件型数据库等。数据抽取采用基于 XML 的数据交互技术,通过 JDBC 的数据访问技术来实现。基于 XML 技术的数据抽取的基本原理是这样的:按照元数据先定义好一些相应的 XML 元素以及这些元素的属性,然后据此制定一份 XML 接口规范,再按预定义的规范,组织相应的 XML 文档实现对数据库或其它数据文档进行抽取。

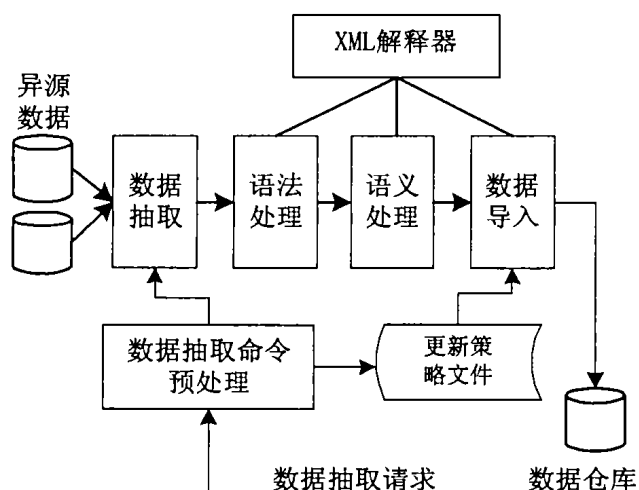


图2 基于 XML 的数据抽取

在这个基本原理的指引下,WIEDW 系统的数据抽取与更新模块的实现原理如图2所示。首先,由“数据抽取命令预处理”接收用户或系统发送来的数据抽取请求,并对请求进行解释生成抽取命令的执行计划,确定本次抽取的更新策略,如完

全更新或部分更新,并将该策略保存起来。其次,将执行计划送给数据抽取执行机构,由它根据整个系统的当前状态,如系统忙闲状态等,在合适的时间内完成抽取任务。抽取得到的数据是 XML 格式的,经过语法判断、语义判断及处理,如果有不合法的记录出现,则将该放弃记录的导入,并保存到相应的日志文件中或提示用户。最后,“数据导入”将数据正确的 XML 数据记录根据更新策略接顶写入数据仓库的方式,并写入到相应的表中。

4.2 元数据管理

对于数据仓库中涉及到的数据表定义,数据建模提供了实体定义、数据字段定义、索引字段定义和实体之间的关系定义。另一方面,数据建模产生的实体定义是作为元数据写入到元数据库表中。

元数据管理实现了元数据的插入、修改、删除等功能,在 Web 页面上,元数据库以树的形式展示给用户,包括:服务器,数据库,数据表及列记录。该树由 JavaBean 实现,体现了面向对象的思想。

元数据管理中的主要问题是,有部分元数据在创建之后是不能随意修改的,否则容易产生数据仓库数据的不一致性或不完整性。因此,在 WIEDW 系统中,对元数据进行分类,不同类别的元数据具有不同等级的可修改性。如异源数据库的连接参数、数据抽取频次等对于一般的管理员或系统本身是可修改的;而数据仓库的定义、维表、事实表、概要表等的定义只对高级管理是可修改的。

4.3 数据仓库建模

在设计数据仓库时,一般使用星型模型和雪花模型来表示事实、维和粒度的关系。在 WIEDW 中主要通过原始数据表、事实表、维表、概要表、元数据表、模板表等完成数据仓库的建模。其中,原始数据表是数据抽取与更新模块写入或更新

的数据表,允许仓库管理人员建立多个原始数据表来存储来自不同数据源的数据;事实表是一个数据仓库的核心,允许它从原始数据表选择所需要的字段,并设定数据选择条件;概要表根据不同的需求可以按照时间以及其他维度建立不同粒度的概要表,因此,允许系统建立多个不同的概要表。概要表的建立有助于提高数据仓库系统的响应时间,但增加了数据的存储,因此,在设计概要表的粒度时,是根据“25%”的原则,即:如果可以让事实表的数据减少25%,则允许建立概要表。

因此,在这些基本表的基础上,用户可以通过该集成环境建立自己的数据仓库系统。

数据仓库的建模就是通过 WIEDW 提供的工具对上述几个基本表进行更新或写入的过程。这是一项复杂工作,在没有较好的工具支持的情况下,往往需要计算机专业人员去完成,而在 WIEDW 中则改变了这种状态。系统中以“所见即所得”效果来建模,如图3所示,在建立销售情况事实表与客户、产品、供应商、时间维度表时,用户可以直接用鼠标拖动相应的表即可实现,这样用户采用更直观、更易使用、操作的界面来进行。

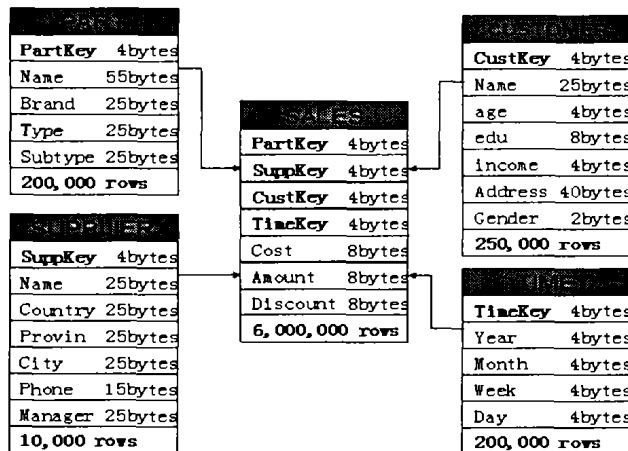


图3 星型模型的可视化建模

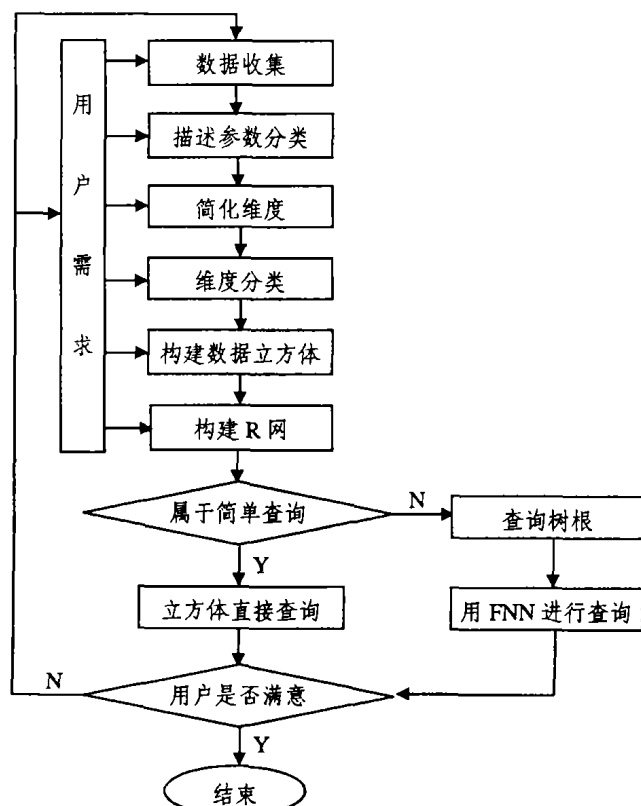


图4 多维分析的查询模式

另一方面,为了使数据仓库具有更广泛的用户,系统引入了计算机操作中常见的“模板”概念,将企业内部一些常见的统计分析行为事先建立好,并保存起来,这样,企业内部不同部门在需要时只需调用该模板,并设置输入数据及格式等,就可以完成部门的数据集市设置。

4.4 OLAP 与展示

在 Web 页面上,事实表和维度表以树的形式显示,方便用户选择需要分析的因素。图4给出多维分析的查询模式。我们采用一个 JAVA 的图表引擎 JFreeChart 用来产生基于 Web 的图表。这些图表包括:表格、柱状图、饼图、点线图。图5和图6分别用平面化和图表的形式显示了分析的结果。

Time	Name			
Quarter 3				
	Andrew Fuller	Anne Bodsworth	Janet Leverling	Laura Callahan
Category Name	Total	Total	Total	Total
Beverages	703.5	3759.75	1152.5	1444.8
Condiments	3748.45	341.25	878.175	1377.855
Confections	680.2		1429.7425	3427.5
Dairy Products	2775	741.75	3523	1001.5
Grains/Cereals	3411	23.8	2009	2315
Meat/Poultry	1056	4875.4		
Produce	1007		69.75	488.25
Seafood	3928	504	1407.3	745.5
总计*	17309.15	10245.95	10469.4675	10800.405

图5 平面化表格显示分析结果

4.5 动态加载算法

从数据仓库中发现各种规律可以使用数据挖掘算法来分析,但是各种数据挖掘算法都有各自的适用范围,即对数据具有一定的敏感性。因此,实现数据挖掘算法、分析算法的动态加载可以针对不同的数据使用不同的挖掘算法,有利于提高数据分析的有效性。

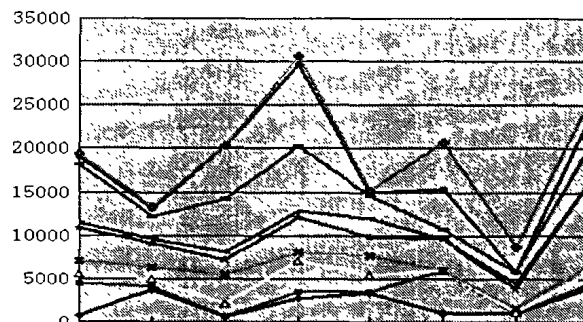


图6 图表显示分析结果

在 WIEDW 中,由挖掘综合器、Web 数据挖掘算法库、方法选择知识库构成一个完整的算法动态加载机制。挖掘综合器是一个挖掘驱动引擎。它根据挖掘要求和挖掘方法的知识库到 Web 数据挖掘算法库中去选择合适的挖掘算法,并使用该算法来完成挖掘任务。Web 数据挖掘算法库是以组件形式组织算法,能动态加载算法,使算法库可扩充性好,并方便使用。

算法的动态加载的实现原理:将挖掘算法封装成 Javabeans 或 EJB 组件,由挖掘综合器判断当前数据集的模式,并决定使用哪个组件。算法的输入是一个数据集,但是其输出是不同的,因此,需要着重解决的问题是算法输出的规范化表示问题。

4.6 数据仓库质量指标

在 WIEDW 中引入了数据仓库质量(QoD:Quality of Da-

(下转第154页)

- col. RFC2401. Nov. 1998
- 2 Kent S, Atkinson R. IP Authentication Header. RFC2402. Nov. 1998
- 3 Atkinson R. IP Encapsulating Security Payload (ESP). RFC2406. Nov. 1998
- 4 Harkins D, Carrel D. The Internet Key Exchange (IKE). RFC2409. Nov. 1998
- 5 Chan K, Seligson J, Durham D, McCloghrte K. COPS Usage for Policy Provisioning (COPS-PR). RFC 3084. Mar. 2001
- 6 Hodges J, Morgan R. Lightweight Directory Access Protocol (v3): Technical Specification. RFC3377. Sep. 2002
- 7 Yang L, Dantu R, Anderson T, Gopal R. Forwarding and Control Element Separation (ForCES) Framework. draft-ietf-forces-framework-13. txt. Jan. 2004
- 8 Yang L, Halpern J, Gopal R, DeKok A, Haraszi Z, Blake S, Deleganes E. ForCES Forwarding Element Model. draft-ietf-forces-model-02. txt. Feb. 2004
- 9 Putzolu D. ForCES Protocol Evaluation Draft. draft-ietf-forces-evaluation-00. txt. Dec. 2003
- 10 汤隽, 赵荣彩, 宋成杰. 安全策略系统的研究及其在 IPsec 中的应用. www.Linuxipsecvpn.cosoft.org.cn
- 11 Farrell S, Housley R. An Internet Attribute Certificate Profile for Authorization. RFC3281. Apr. 2002
- 12 Blaze M, Feigenbaum J, Ioannidis J, Keromytis A. The KeyNote Trust Management System Version 2. RFC 2704. Sep. 1999
- 13 Blaze M, Ioannidis J, Keromytis A. Trust Management for IPsec. In: Proc. of the Internet Society Symposium on Network and Distributed Systems Security. SNDSS 2001. 139~151

(上接第117页)

ta Warehouse)指标的概念。它的集成性是由数据仓库的 QoD 来实现的, QoD 将各个模块连接起来的, 并形成一种反馈机制, 从而使得整个数据仓库系统朝着质量好的方向演变。

为了实现这个目的, 必须定义数据仓库的质量指标, 并对这些质量指标进行定量计算。在 WIEDW 中, 整个数据仓库的综合质量指标是由若干个分质量指标组成的。分质量指标是指数据仓库的实际输出与理想状态下的数据输出在准确性方面的接近程度, 它的一般定义可以用下面的公式来表示:

$$DW_i = (QI - QF) / QI \quad (1)$$

式中, QI 表示理想状态下数据仓库的某个质量指标, QF 表示实际环境下的相应的质量指标。因此, 整个数据仓库的综合质量指标, 定义如下:

$$QoD = a_1 * DW_1 + a_2 * DW_2 + \dots + a_n * DW_n$$

其中, $a_1 + a_2 + \dots + a_n = 1$

$a_1, a_2, \dots, a_n$ 表示每个质量指标所具有的影响值, 可以通过人为的方式指定。

可以把数据仓库的分质量指标 DW_i 分成三类:

(1) 及时性 这是指数据仓库系统在接收到用户的请求之后将查询结果返回给用户时所花费的时间与理想状态下的值的接近程度, 可以用式(1)来定量表示。

及时性由数据仓库本身的响应时间、数据仓库系统的响应时间组成。前者与数据仓库定义时所指定的数据的粒度、维度的定义方式有关; 后者与数据仓库系统的并发用户、数据仓库实现模式有关。

(2) 完整性 是指数据仓库返回给用户的查询结果是完整的。它的主要影响因素有: 各个数据源数据抽取的频率、元数据的定义、数据仓库的更新算法、数据记录概括算法等。

(3) 一致性 是指在不同用户的相同查询或同一个用户的多次查询请求中, 所得到的数据应该是相同的。它包括数据结构的一致性、数据语义的一致性, 如同一个字段的描述、产量在某一段时间内的增加与减少等。

它的主要影响因素有: 数据抽取时设定的过滤规则、元数据的表达及维护是否合适、数据仓库的更新算法、数据记录概括算法等。

上述分析中, 所涉及到的影响质量指标的因素可以用来作为数据抽取模块、OLAP 展示组件、元数据管理等模块做自我调整的参考。

结束语 基于 Web 的数据仓库集成环境的研究具有以下重要意义: 1) 改变了目前数据仓库构建过程中各个环节相互脱离的现象, 为用户提供一个集成化的设计、运行、维护平台, 提高了企业建立、维护数据仓库应用系统所需要的人力、物力投资, 又大大缩短了企业建立数据仓库的开发设计周期; 2) 实现了最优算法、行业定制、企业定制算法的动态加载, 使数据仓库的分析数据真正对企业的决策活动具有较好的指导意义; 3) 使用户能在任何地方借助互联网通过浏览器方便使用。

参考文献

- 1 Ladley J. Data Warehousing Trends, Data Adviser, Jan. 1998. 35~37
- 2 Xin Tan, Yen D C, Xiang Fang. Web warehousing: Web technology meets data warehousing. Technology in Society, Jan. 2003, 25 (1): 131~148
- 3 Shim J P, et al. Past, present, and future of decision support technology. Decision Support Systems, June 2002, 33(2): 111~126
- 4 Li Xiu, Wang Xuerui, Liu Wenhuan, Liao Lin. The research of web-based data warehouse using XML. Info-tech and Info-net, 2001. In: Proc. ICII 2001 - Beijing. 2001 Intl. Conf. on, 2001, 5: 42~47
- 5 Liao Lin, Wang Ligang, Liu Wenhuan. Building the Web-Based Data Warehouse with XML. The Joint International Computer Conference, Nov. 2000
- 6 Mikael R J, Thomas H M, Torben B P. Converting XML DTDs to UML diagrams for conceptual data integration. Data and Knowledge Engineering, March 2003, 44(3): 323~346
- 7 McHugh J, Abiteboul S, Goldman R, Dallan Q, Windom J. Lore: A Database Management System for Semistructured Data.
- 8 NCSA. Introduction to NCSA Mosaic for X. Notional Center for Supercomputing Application, University of Illinois at Urbana-Champaign.
- 9 唐蕾, 张剡, 等. 数据仓库工具集 NGDW-1 的设计与实现. 见: 第 19 届中国数据库学术会议. 郑州, 2002.
- 10 曹鲍光, 王申康. 元数据管理策略的比较研究. 计算机应用, 2001, 21(2): 3~5
- 11 Hurwiz Group. Enterprise Metadata Management [R]. Hurwiz Group BalancedView Report, 1998
- 12 宋擒豹, 杨向荣, 沈钧毅. 数据仓库技术研究. 计算机工程, 2002, 28(1): 125~127
- 13 王珊, 等. 数据仓库技术与联机分析处理. 北京: 科学出版社, 1998
- 14 Michael G, Woollen R, Emerson L. J2EE 应用与 BEA Weblogic Server. 邢国庆等译. 北京: 电子工业出版社, 2002