

基于 CWM 的元数据管理系统的研究^{*})

雷 琦 宁 洪 李姗姗 汪昌健

(国防科技大学计算机学院 长沙410073)

摘 要 通过对当前各类信息系统的需求分析以及现有元数据管理方案的深入研究,灵活利用 OMG 建模体系提供的各种通用服务,本课题组提出了将建模工具、元数据转换工具、元数据搜索引擎、元数据存储库等基于 CWM 的功能组件集成在一起的元数据管理解决方案 WMMS,试图提供本地和分布式环境下元数据的有效管理,为各类信息系统的快速移植和集成提供支持。本文首先介绍了 CWM 的体系结构,然后说明了 WMMS 的系统组成和主要功能,最后重点描述 WMMS 中建模工具的设计和实现。

关键词 元数据,CWM,WMMS,建模工具,模型检验

Research on Metadata Management System Based on CWM

LEI Qi NING Hong LI Shan-Shan WANG Chang-Jian

(School of Computer, National University of Defense Technology, Changsha 410073)

Abstract With the comprehensive analysis of the requirement of all current information systems and research of some metadata management solutions in existence, a warehouse metadata management solution (WMMS) is proposed. This solution integrates many function components based on CWM such as modeling tool, metadata transformation tools, metadata searching engine and metadata repositories etc., and furthermore, it ingeniously utilizes various common services provided by the OMG. Consequently its aim is to effectively manage metadata in local and distributed environment and to support the rapid porting and integration of all kinds of information systems. In this paper, the architecture of CWM is first introduced and then the systemic structure and main functions of WMMS is detailed, finally what is put stress on is the design and implementation of the modeling tool.

Keywords Metadata, CWM, WMMS, Modeling tool, Model-verifying

1 引言

近年来,数据仓库技术日益受到业界重视,各种数据仓库产品已投入使用,并带来巨大的效益。但是,对于数据仓库中元数据的定义和管理却始终没有一个统一的解决方案。不同开发者的数据仓库工具所基于的元数据结构基本都是私有的,若需对这些工具集成,就会涉及元数据间的复杂映射。为了统一数据仓库开发者元数据管理策略,OMG 组织推出了一个通用数据仓库元模型的标准(Common Warehouse Meta-model, CWM)。CWM 是继元数据联盟 MDC (Meta Data Coalition) 的开放信息模型 OIM (Open Information Model) 以来的一个特定于数据仓库领域的元数据模型集,它独立于具体数据仓库的实现,包含了各类有代表性的数据仓库元数据的有效描述。它的目标是使异构环境下的各种数据仓库工具、数据仓库平台以及数据仓库元数据存储间能方便地交换数据仓库和商务智能的元数据。目前,CWM 已被 IBM、UNISYS、NCR、Hyper ion、Oracle 等多家公司支持并被包含进它们的下一代数据仓库与数据分析产品中。

尽管 CWM 是专门针对数据仓库制定的,但因其考虑了关系型的、面向对象的、面向记录的和 XML 的等各类信息源的元数据描述,因此它同样适用于一般信息系统的元数据的

描述和管理。通过对当前各类信息系统(包括数据库系统、数据仓库系统、数据分析和挖掘系统等)开发、移植和集成等方面的需求分析,以及对现有元数据管理方案的深入研究,本课题组提出了一套基于 CWM 的元数据管理解决方案 WMMS,它灵活利用 OMG 建模体系提供的各种通用服务,适用于分布式环境下的元数据管理。

下面首先简介 CWM 的体系结构,然后说明 WMMS 的系统组成和主要功能,最后重点描述 WMMS 中建模工具的设计和实现。

2 CWM 的体系结构

CWM 使用包机制来组织元模型,这种组织结构使其既易于理解、使用(用户建模),又便于裁减和扩展(工具开发者实现)。CWM 目前的版本中包含了18个包和底层的 Object-Model,每个包尽量涉及一个独立的领域。虽然各个包所描述的领域不尽相同,但它们之间有着紧密的依赖关系。事实上,所有的包按功能和抽象层次组织成四层,同层的包的功能角色类似,每层中的包都为同层或上层的包提供服务,层次越高描述的内容越抽象。CWM 的内容结构图如图1所示。

底层的 ObjectModel 是 UML 的子集,CWM 的元模型中的所有包的类与关联都是直接或间接地从 ObjectModel 中的

^{*}) 基金项目:武汉大学软件工程国家重点实验室开放基金项目(SKLSE04—01)。雷 琦 硕士研究生,主要研究方向为数据库、数据仓库和数据挖掘等。宁 洪 教授,主要研究方向为数据库、软件工程和网络安全等。李姗姗 博士研究生,主要研究方向为数据仓库与数据挖掘等。汪昌健 硕士,主要研究方向为数据仓库与软件工程等。

类与关联继承而来, ObjectModel 也作为 Resource 层中的一个包来描述面向对象的数据源。Foundation 层的元模型包含的模型元素主要是代表其它 CWM 包所共享的概念与结构, 如表达式、索引、数据类型、软件配置信息等, 它们比更高层次的包中的模型元素更具有通用性, 放在这一层上为更高层的包提供服务; Resource 层中包含了 OLTP 系统与数据仓库所使用的各种数据资源, 有关系的、层次的、多维的等等, 这些在 Resource 层中都有相应的模型描述; Analysis 层提供了支持逻辑服务的元模型, 这些逻辑服务包括 OLAP、数据挖掘、转

换等, 它们会被映射到由 Resource 层的包所定义的数据存储中去; Management 层元模型通过定义、调度操作任务 (Warehouse Process 包) 和记录数据仓库活动以及相关统计信息 (Warehouse Operation Package) 来支持所有数据仓库的操作。其中每个包都由一系列 UML 表示的类图来描述。这样极大地方便了系统开发者的建模工作, 因为在建模时只取所需的包即可, 比如建模者就可以在 Resource 层中取 Rational 包来描述当前应用非常广泛的关系型数据库的元数据。

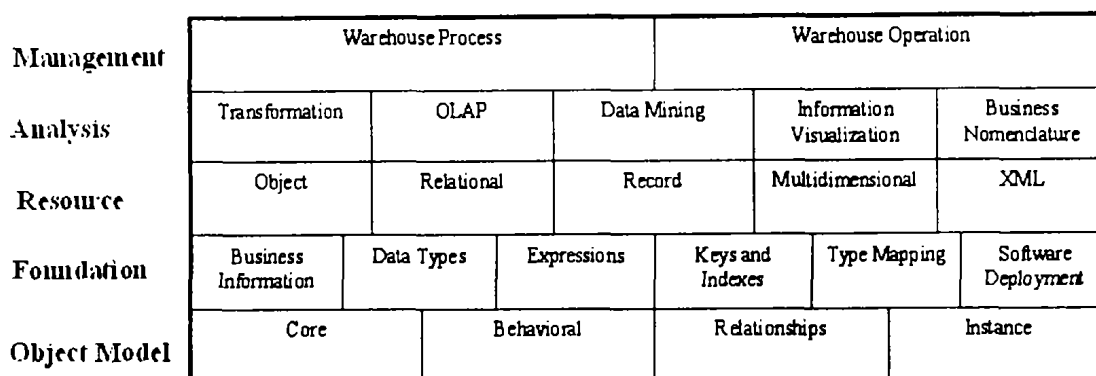


图1 CWM 的内容结构图

此外, CWM 包含于 OMG 的建模体系中, 该体系结合了多种工业标准, 为体系中的所有元模型提供了许多通用的服务。CWM 模型既包括元数据存储, 也包括元数据交换, 它主要采用了以下三个业界规范作为它的底层标准: UML 表示了所有 MOF 描述的元模型, 包括 UML 本身。CWM 的所有类和关联都是直接或间接地继承 UML 的一个子集 Object Model, CWM 元模型也是采用 UML 图形符合描述的; MOF (元对象设施) 定义了一种抽象语言和一个框架描述与技术无关的元模型, 是 OMG 元模型和元数据的存储标准, 提供在异

构环境下对元数据知识库的访问接口; XMI (XML 元数据交换) 是由 OMG 采纳的分布式环境中元数据交换格式的标准, 所有 OMG 建模体系中的元模型都能方便地使用 XMI 进行元数据交换。UML, MOF, XMI 作为 OMG 建模体系中的三个基础规范, 彼此协作, 为模型从建立到交换再到使用等各个环节搭建了整条集成链, 为各种仓库工具的无缝集成奠定了基础。

3 WMMS 的系统组成和主要功能

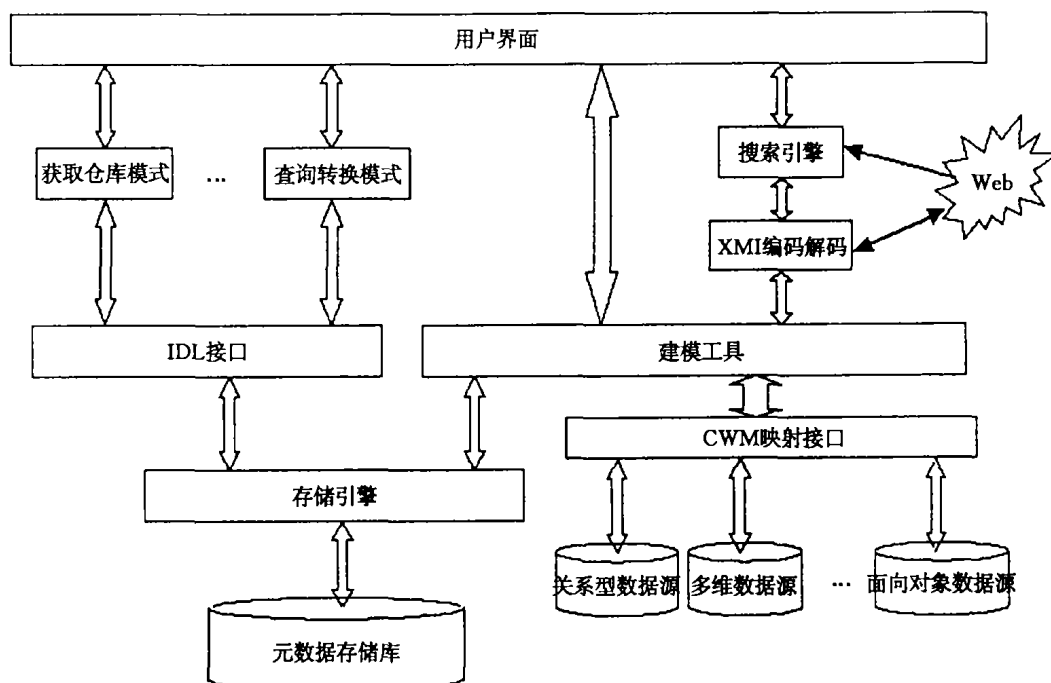


图2 WMMS 系统结构图

元数据管理是企业级数据仓库中的关键组件, 它贯穿于数据仓库设计、开发、运行和维护的全过程。现有元数据管理主要有两种方法: 对于相对简单的环境, 按照通用的元数据管理标准建立一个集中式的元数据知识库; 对于比较复杂的环境,

分别建立各部分的元数据管理系统, 形成分布式元数据知识库, 然后, 通过建立标准的元数据交换格式, 实现元数据的集成管理。

在参考了现有各种元数据管理方案的基础之上, 结合当

前数据仓库的发展需要和一般信息系统移植和集成存在的问题,利用 OMG 建模体系提供的各种通用服务,本课题组提出了一套基于 CWM 的元数据管理解决方案 WMMS。WMMS 的系统组成如图2所示。

WMMS 主要由元数据存储库、建模工具、XMI 编码解码(也称为元数据转换工具)、搜索引擎、IDL 接口等模块组成。WMMS 中元数据管理的主要任务有两个方面:一是负责存储和维护元数据库中的元数据;二是负责数据仓库建模工具、元数据转换工具、前端工具等之间的消息传递,协调各模块之间的工作。

WMMS 通过对 Rational Rose 二次开发作为建模工具,建模者可以使用它完成基于 CWM 的数据仓库的建模工作,产生的元数据即可以直接存储到元数据存储库中,也可以通过元数据转换工具(即图2中的 XMI 编码解码工具)转换成 XMI Document 的形式用于分布式环境下的元数据交换。对于分布式环境下基于 CWM 的 XMI Document 形式元数据的管理,开发和管理人员可以使用搜索引擎进行分类搜索,从某种程度上说,搜索引擎是属于整个元数据管理系统外延的一个工具,它负责搜寻符合条件的 XML 形式的元数据,交付给元数据管理系统存储和使用。同样对于目前国内企业应用非常广泛的关系型的(如 Oracle, Sybase 等)、多维的、面向对象的等各类数据源,企业数据库开发人员可以通过建模工具的 CWM 映射接口转换成统一的基于 CWM 规范的元数据描述,实现对这些信息系统元数据的交换和集成,从而实现原系统的进一步开发、移植和集成。WMMS 中元数据存储库是基于 MOF 规范的,并且要求元数据以 CORBA 对象的形式存储,元数据存储库主要用于存放建模和操作过程中产生的元数据。WMMS 系统结构图中左面对应的是本地环境下的元数据操作接口,用户对数据仓库元数据的所有操作如查询数据仓库的数据模式或查询某个数据抽取过程等操作都必须通过 CWM IDL 接口进行,该接口可以通过 MOF 的 IDL 映射产生。

目前已有多家比较成熟的元数据管理系统,如微软的 Meta Data Service 以及 Sybase 的 WCC(Warehouse Control Center),它们所基于的元模型分别是 OIM 和 MDIS。MDIS 是 MDC 以前的一套老标准迟早会被 CWM 所替代,而 OIM 在数据仓库建模的能力上也不及 CWM。随着 CWM 标准的逐渐完善,以及 MDC 和 OMG 组织的合并,我们认为 OMG 组织的 CWM 标准将会成为数据仓库元数据领域事实上的标准。我们的元数据管理系统 WMMS 就是基于这个标准建立起来的,因此系统本身具有很好的可扩展性,并且由于增加了元数据转换功能以及搜索引擎功能,它能更好地适用于分布式环境下的元数据管理。

本课题组已就 CWM 及 OMG 建模体系中相关的规范进行了深入的研究,并已经完成基于 XMI 的元数据转换工具和建模工具原型的设计和实现,目前正在研究基于 XML 表示的 CWM 元数据搜索引擎和元数据存储库。由于篇幅的原因下面着重介绍基于 CWM 的建模工具的设计与实现。

4 数据仓库建模工具的设计与实现

数据仓库建模工具是辅助数据仓库开发的 CASE 工具,其能帮助用户和开发者通过需求分析创建数据仓库的数据模式,同时还可以定义模式中各个表的数据来源、数据转换规则、有关的聚集操作等信息。下面将从建模工具的定制、功能

以及各功能模块的设计等三个方面来介绍建模工具的设计和实现。

4.1 建模工具的定制

由于 CWM 的所有类和关联都是直接或间接地继承 UML 的一个子集 ObjectModel,因此在构建基于 CWM 的建模工具时,可以方便地利用现有 UML 建模工具中的图元符号及其内涵。Rational Rose(以下简称 Rose)是目前支持 UML 建模最为通用的工具,其界面友好功能强大,最重要的是它提供了 REI 扩展接口以支持二次开发。于是,我们重新定制它的图元符号,使其支持 CWM 语义,将其改造为我们的建模工具。CWM 建模工具将在两个层次上实现对 Rose 的利用,第一层次就是对 Rose 的图形开发界面的利用,第二层次就是对 Rose 本身已经具有的 UML 内部关系的一种利用,如包和类之间的关系,等等。

4.2 建模工具的功能

目前在课题研究的初期,我们对建模工具的建模能力进行了简化,还只实现了它的一个原型。我们对 CWM 中的所有包进行了分析,但因为考虑到目前国内企业各类信息系统实际的发展需要以及实现的工作量,所以只是完成对如图1中阴影部分所含盖的包进行了深入的研究和设计实现,它们是描述一个基于关系数据库的独立完整的数据仓库过程所需要的最少建模元素。同时它们足以支持一般基于关系数据库的信息系统的建模和集成。该建模工具只需改造 Rose 的类图以支持简单的对象建模功能,但是为了保证模型的正确性和有效性,必须针对 CWM 的语义和特性进行一系列模型检验工作。因此,WMMS 中建模工具的功能主要包括两大模块:模型创建和模型检验,另外还支持简单的日志功能,用于记录用户所有的操作。建模工具的核心用例图如图3所示。

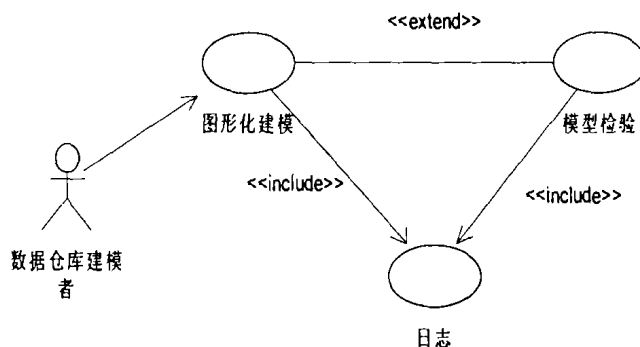


图3 建模工具的核心用例图

在核心用例图的基础上,可以确定建模工具主要包含的类,分别有:界面管理类,建模元素管理类,日志类,画对象图类,建模元素类,检查类。界面管理类控制整个建模界面,触发用户进入 CWM 建模环境,关闭 Rose 时退出 CWM 建模环境等;建模元素管理类管理 CWM 中所有可以直接产生元数据的类;日志类记录用户所做的所有操作;画对象图类根据用户选择的建模元素在图形编辑界面中画相应的对象或关联;CWM 中每个可以直接产生元数据的类都对应一个建模元素类,每个类绑定了该类相关的所有关联信息用于模型检验;检查类中定义了一系列模型检验机制,通过与建模元素类绑定的信息以及建模过程中记录的信息来检验模型。

4.3 建模工具各功能模块的设计

建模工具的主要功能包括模型创建和模型检验。建模的环境分为三个部分:建模元素选择栏、图形编辑界面和日志、

(下转第145页)

它不仅考虑各个节点的向上路径的相似关系而且也考虑它们向下路径的相似关系,这样很好地支持了带分支的路径表达式查询;同时它也可以根据查询情况的变化来动态地改变索引结构,使索引适合当前的查询要求,从而获得较高的查询效率。把基于路径的索引方法和基于元素的索引方法结合起来去处理更复杂条件的查询,这是正在进行的科研工作。

参考文献

- 1 Abiteboul S. Querying semi-structured data. ICDT, 1997. 1~18
- 2 Buneman P, Kaushik R, UnQL D S. A query language and algebra for semistructured data based on structural recursion. VLDB Journal, 2000, 9(1): 76~110
- 3 Clark J, Derose S. XML Path Language (Xpath). W3C Recommendation, 1999. <http://www.w3.org/TR/Xpath>
- 4 Chamberlin D, Florescu D, Robie J, Simeon J, Stefanescu M.

XQuery: A query language for XML. W3C Working Draft, 2002. <http://www.w3.org/TR/xquery>

- 5 Goldman R, Widom J. DataGuide. Enable query formulation and optimization in semistructured databases. VLDB, 1997. 436~445
- 6 Milo D, Suciu D. Index structure for path expression. ICDT, 1999. 277~295
- 7 Kaushik R, Shenoy P, Bohannon P, Gudes E. Exploiting local similarity for efficient indexing of paths in graph structured data. ICDE, 2002. 129~140
- 8 Wu H, Wang Q, Xu J, Zhou A, Zhou S. UD(k,l)-index: An Efficient Approximate Index for XML Data. WAIM, 2003
- 9 Qun C, Lim A, Win K. D(k)-Index: An Adaptive Structural Summary for Graph-Structured Data. SIGMOD, 2003. 134~144
- 10 Xmark. The xml benchmark project. <http://monetb.cwi.nl/xml/index.html>

(上接第113页)

模型检验器。

建模元素选择栏是一个独立的矩形窗口,内含一个树形控件,其中包含 CWM 的所有建模元素。系统针对用户所选择的建模元素,可以在图形编辑界面中画相应的对象和关联,同时后台记录用于模型检验和元数据转换的信息。

由于 CWM 只定义了静态属性,因此图形编辑环境只需定制 Rose 的类图。类图及其中的模型元素在 CWM 建模环境下赋予了新的含义:类图成为了对象图,包维持原来的含义,类变成了对象,原来属性的 Initial Value 现在代表对象的属性值,关联在此表示关联的实例。用例图和序列图在新的建模环境仍然可以重用,其它的图在 CWM 建模环境中暂时不需要使用。Rose 各种视图的界面大家很熟悉,这里就不再附图说明。

模型检验器是建模工具中最重要的部分。它必须具备对操作的合法性进行实时检查的能力,并立刻将信息反馈给建模者,同时最大程度保证模型能客观地描述实际信息系统并且符合 CWM 的语义范畴。

由于对象都是从建模元素选择栏中选定的,因此不存在对象超出 CWM 描述范畴的情况。而且在模型维护的时候, Rose 提供了修改对象的级联作用,即当建模者删除一个对象时,系统会自动将与该对象连接的所有关联实例删除,当然我们需要在后台数据结构中记录下这一信息用于后面的模型检验和元数据转换。然而由于 Rose 具有广义的建模能力,因此对关联的创建和修改不作任何的检查,但是这对于我们针对数据库领域的建模需求是远远不够的。因此,模型检验器主要还是针对关联的创建和修改进行检查,根据 CWM 的特点,关联主要有以下三方面的检查:

(1)合法性检查:判断两个对象之间的关联在 CWM 中是否有描述;

(2)导航性检查:针对有向关联(包括有向聚合),用于判断箭头的方向与 CWM 中的导航性约束是否一致;

(3)多样性检查:判断对象之间的关联数目是否满足 CWM 中的多样性约束。

在介绍核心用例图的时候就已经提到,为了方便实现模型检验,每个会产生元数据的类都对应了一个建模元素类,新建一个对象的同时也产生建模元素类的一个实例,其中记录

所有用于模型检验的信息。在实现模型检验的时候,我们为这些信息定义了一个统一的数据结构,同时还设计了几个高效的检验算法(由于篇幅的原因在此就不一一介绍了)。

结束语 本文介绍了一个基于 CWM 的元数据管理系统 WMMS,该系统具有强大的元数据管理能力,并且更适用于分布式环境下的元数据管理。目前本课题组已基本完成对 WMMS 中建模工具和元数据转换工具的设计和实现。接下来将进一步完善和扩展建模工具的功能,其中包括在建模工具中实现把各种数据源的元数据如各类厂商的关系模型的元数据、分布式环境下搜索 XMI 格式的元数据等转换成建模工具可识别的模型信息,以及对建模工具产生的元数据进行格式转换以实现在 WMMS 中基于 MOF 的 CORBA 对象的形式存储。同时展开对专门针对基于 CWM 的、以 XMI 形式的元数据智能搜索引擎的研究。至此我们就可以实现一个在分布式环境下的应用开发的闭环。对于本地环境下的元数据操作接口技术中,需要实现将基于 MOF 的元模型到相应 IDL 接口的映射。所以要实现 WMMS 系统强大的元数据管理功能还需要开展很多研究工作。

参考文献

- 1 OMG. Common Warehouse Metamodel Specification Version 1.1. March 2003
- 2 Poole J, Chang D, Tolbert D, Mellor D. Common Warehouse Metamodel. John Wiley & Sons, Inc. 2002
- 3 Rational the e-development companyTM. Using the Rose Extensibility Interface Version; 2001A. 04. 00
- 4 Inmon W H. Building the Data Warehouse Second Edition. 机械工业出版社, 2000
- 5 Poole J 等著. 公共仓库元模型—数据仓库集成标准导论. 彭蓉, 何璐璐等译. 机械工业出版社, 2004, 3
- 6 李姗姗, 宁洪, 陈波, 等. 通用数据仓库元数据模型(CWM)的研究. 计算机科学与工程
- 7 李姗姗, 宁洪. 基于 CWM 的元数据管理系统中数据转换工具的设计与实现. 计算机工程与应用
- 8 数据仓库与元数据管理. <http://www.dwway.com/file/warehouse&metadata.doc>
- 9 李姗姗, 宁洪. CWM 扩展机制的研究. 计算机科学, 2003, 30(10): 169~171