异构数据集成技术的发展和现状

靳强勇　李冠宇　张　俊
(大连海事大学,大连 116026)

摘　要　该文对异构数据集成技术的产生、发展及目前的最新情况进行了介绍。文章首先介绍了这一技术产生的背景和研究目的。然后介绍了相关的概念、技术及开发工具。随后对现有的异构数据集成系统进行了简单的介绍,并对其中很有代表性的 Lore 系统进行了着重介绍。在文章的最后,介绍了笔者所做的研究工作和下一步的工作目标。

关键词　异构数据　数据集成　CORBA　体系结构

文章编号 1002-8331-(2002)11-0112-03　**文献标识码** A　**中图分类号** TP311

Development and Present Situation of Heterogeneous Data Integration Technology

Jin Qiangyong　Li Guanyu　Zhang Jun
(Dalian University of Sea, Dalian 116026)

Abstract: This paper discusses the origination of technology of heterogeneous data integration as well as its development and latest situation.At the first part the background and the object are introduced then followed the related concepts, technologies (such as CORBA and XML) and developing tools (such as Delphi and Java).At the second part some existing heterogeneous data integration systems are presented and emphasis is put on the Lore system.At last the paper briefly introduces what the authors have done and their future work.

Keywords: heterogeneous data, data integration, CORBA, system architecture

1　前言

随着计算机网络的普及,数据资源的共享已经成为一个热门话题[1]。很多企业需要将 DBMS,MIS,OA 数据集成起来,构成企业的管理决策的网络信息平台[2]。传统的数据库集成方法现在已经远远不能适应人们获取数据的需求,因此迫切需要一种新的数据集成系统。这种系统不仅能集成传统的数据库中的结构化数据,而且还可以集成在 web 上应用日益广泛的半结构化数据和非结构化数据。在这种背景下,异构数据集成系统受到越来越多人的重视,这方面的研究也成为当前数据集成研究的一个热点。

异构数据集成系统为企业解决多平台、多结构数据的集成问题提供了一条解决途径。通过这样一个集成系统,可以把企业内部和外部的各种相关数据资源进行整合,为企业的信息资源规划提供了可能,从而搭建起整个企业的信息平台。

在深入探讨异构数据集成系统之前,先介绍其中两个基本概念,这两个重要概念是整个异构数据集成系统的基础。

异构数据　异构数据是一个含义丰富的概念,不仅指不同的数据库系统之间的数据是异构的,如 Oracle 和 SQL Server 数据库;而且还包括不同结构的数据之间的异构,如结构化的 SQL Server 数据库数据和半结构化的 XML 数据。

数据集成　数据集成是对各种异构数据提供统一的表示、存储和管理,这些功能在异构数据集成系统中实现。数据集成屏蔽了各种异构数据间的差异,通过异构数据集成系统进行统一操作。因此集成后的异构数据对用户来说是统一的和无差异的。

2　涉及的技术

异构数据集成系统的研究涉及多种计算机技术,如分布式对象技术、XML、面向对象技术等,下面逐一加以介绍。

2.1　CORBA 技术

目前分布式对象技术主要包括 OMG (Object Management Group)组织制订的 CORBA (Common Object Request Broker Architecture)标准、Microsoft 的 COM/DCOM 标准以及 Sun 公司的 Java RMI (Java Remote Method Invocation)标准。这三种模型各有特点,这些特点已有诸多文献详细介绍,这里不再赘述[3]。

CORBA 应用程序非常类似于其他面向对象的应用程序。所不同的是,当对象在另一台机器上的时候,客户端和服务器端必须分别通过一个特殊的层来管理网络通信,在客户端称为 Stub,在服务器端称为 Skeleton。Skeleton 与 ORB 之间通过 Basic Object Adaptor (BOA)通信。CORBA 应用程序结构如图 1 所示。

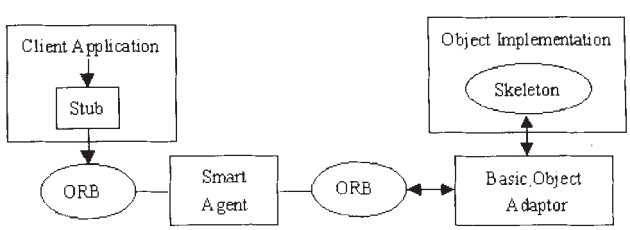

图 1　CORBA 应用程序结构图

客户访问 Stub 的方式类似于访问一个实现某种接口的对象。与一般对象不同的是,Stub 要通过安装在客户端的 ORB 软

基金项目:国家自然科学基金(编号:60172043);辽宁省自然科学基金(编号:002055)

件来处理接口调用,而 ORB 又通过 Smart Agent (可能运行在局域网中的其他机器上)所提供的目录服务来动态地定位一个可用的服务器,该服务器才真正地实现了接口。

2.2 XML

XML (Extensible Markup Language 可扩展标志语言)是由万维网协会 (W3C)设计,特别为 Web 应用服务的 SGML (Standard General Markup Language 通用标识语言标准)的一个重要分支。XML 是一种中介标示语言 (Meta-markup Language),它作为一种可用来制定具体应用语言的元语言,具有强大的描述能力,可提供描述结构化资料的格式。由于 XML 大大提高了 INTERNET 的接入速度和查询检索 Web 数据的速度,因此网络得到越来越广泛的应用,并将逐步成为数据组织和交换的标准。

由于基于 XML 的数据是自我描述的,数据不需要有内部描述就能被交换和处理。利用 XML,用户可以方便地进行本地计算和处理。XML 可以被利用来分离使用者观看数据的界面,使用简单灵活开放的格式,可以给 Web 创建功能强大的应用软件。

2.3 其它

此外异构数据集成系统还涉及到面向对象技术、数据库技术等。借助面向对象技术可以把异构环境中数据和对数据的操作融为一体,并对不同的数据类型进行包装。数据库技术包括数据模型技术、数据查询分解和优化策略、数据表示和描述等。

3 开发工具

目前可视化的开发工具不少,但不是所有的工具都适合于开发异构数据集成系统。下面介绍的是两种有代表性的开发工具,它们支持一种或多种分布式对象标准,但支持程度上有差异,功能上也不相同。

3.1 Delphi

Delphi 中提供了 COM 和 CORBA 两套组件标准,以及包括 CORBA 在内的多种数据接口。Delphi 通过 Class Name 项来标识 CORBA 对象的类名,采用 Instance-per-client (每个请求创建一个实例)和 Shared instance (一个实例处理所有请求)两种实例方式,提供了两种线程模式 :Single-threaded,Multi-threaded 每个实例是否只能接受一个客户的请求,即每个请求是否有自己的线程)。Delphi 的核心组件库是 VCL (Visual Component Library),其所有对象都有继承和被继承的关系。

3.2 VisiBroker

VisiBroker 作为 CORBA 标准的 ORB(Object Request Broker)产品,完全遵循对象管理组 (OMG)CORBA 规范。它提供了一个完整的 CORBA ORB 运行环境和支持开发环境,用于建立、调配和管理分布式 Java 应用程序。基于 Web 的应用程序可轻易访问由 VisiBroker 建立的对象,通信时采用 OMG 的因特网 ORB 间协议标准 (Internet Inter-ORB Protocol,IIOP)[4]。VisiBroker 的关键特性包括 :VisiBroker 智能代理 (Osagent)体系结构、利用位置服务增强对象查找、对对象实现和对象激活的支持等。

4 现有的异构数据集成系统

目前已经研制出来的异构数据集成系统有很多,这里选取其中有代表性的三个系统,介绍其功能和特点。

Lore :Lore (Lightweight Object Repository)是美国 Stanford 大学设计的一个专门用于管理半结构化信息的数据库管理信息系统,可以对半结构化数据 (OEM)和 XML 数据进行存储和管理。Lore 包含标准数据库的一些特性,如 :多用户支持、日志和恢复功能以及查询和更新语言。同时 Lore 也提供了其他一些操作数据库的工具[5]。下面将详细介绍 Lore 系统。

Ozone :Ozone 是美国 Stanford 大学设计的一个集成模型,该模型对面向对象数据库的 ODMG 模型进行了扩展,可以对结构化和半结构化数据进行集成。Ozone 系统特别适合处理杂合性数据,尤其是广泛应用于 Web 上的数据[6]。

Versatile :Versatile 是由东南大学开发的一个基于 CORBA 的可扩展的异构数据源集成系统原型。该模型在 IONA 公司的 Orbix 产品上,对 SQL Server、Versant、文件系统、超文本数据进行包装和集成。该系统不仅能集成上述数据源,而且能集成随时插入的新数据源中的数据[7]。

在上述的三种模型中,Lore 系统是存储 XML 数据的一个专门数据库系统,它是异构数据集成系统代表性的例子。该系统的体系结构设计、查询处理和优化方式为后来的异构数据集成系统提供了很好的借鉴。通过对 Lore 系统的分析可深入了解异构数据集成系统,因此有必要介绍一下 Lore 系统的数学模型、查询语言和体系结构。

4.1 OEM 模型

OEM 模型 (Object Exchange Model)是专用于描述半结构化数据的。在这种模型中,数据被表示成带标签的有向图。在这个有向图中,每个节点都是一个对象,每个对象有一个唯一的 id。原子对象只有入边没有出边,它们的值只能取自基本的数据类型,如 integer,real,string 等。其他的对象称为复合对象,可以同时拥有入边和出边。

在 OEM 模型中,没有固定的数据模式。与数据模式相关的信息都存在标签中,因此可以动态地变化。对于一个 OEM 对象 X 和一个标签 l 来说,表达式 X.l 表示对象 X 中所有带标签 l 的子对象的集合[8]。

4.2 Lorel 查询语言

Lorel 查询语言是对 OQL 的一种扩展,它的优势在于 :不必事先知道要查询对象的结构,也不存在类型不匹配的问题。例如,你可以使用这样的查询语句来查询办公室中年龄较大的人群信息 :

```
QUERY
  Select DBGroup.Member.Office
  where DBGroup.Member.Age>30
```

在 Lorel 查询语言中,也支持使用模式符号|、?等及统配符#、%进行查询。例如,你可以使用这样的查询语句来查询房间号信息 :

```
QUERY
  Select DBGroup.Member.Name
  where DBGroup.Member.Office (Room% |.Cubicle)? like "%252"
```

4.3 Lore 系统体系结构

对 Lore 系统的访问是通过各种应用程序或是直接访问 Lore API 实现的。Lore 系统的查询编译层 (Query Compilation layer)包括解析器、预处理器、查询计划产生器和查询优化器。Lore 系统的数据引擎层 (Data Engine layer)包括 OEM 对象管理器、查询操作符、外部数据管理器和其他工具。

当解析器收到一个查询后,生成一棵解析树,并传递给预

处理器。预处理器负责把 Lorel 查询翻译成 OQL 查询，同时生成一个查询计划，传递给查询优化器。经过优化的查询被传递到数据引擎层。在数据引擎层，执行具体的数据操作，例如取一个对象，进行两个对象的比较操作等。

5 笔者的工作

笔者的前期工作包括收集国内外有关的文献资料，分析目前异构数据集成理论、方法以及目前国内外异构数据源集成系统的优缺点，在此基础上初步确定了异构数据集成系统的体系结构，提出了改进的模型以及“异构数据目录”、“信息资源目录”等一些新概念，并在异构数据的表示模型上取得了一定的进展[9,10]。这个改进的模型，对于解决异构数据源主动参与集成有很大的帮助，同时异构数据客户程序和异构数据目录服务等概念的提出，为异构数据集成研究提供了一个新思路。

笔者的研究目标是在综合分析研究目前异构数据集成理论和方法的基础上，提出异构数据集成的表示模型、分类模型和检索模型等集成管理模型，并在此基础上研制出相应的异构数据集成支持工具和系统原型软件。

在异构数据集成系统的体系结构初步建立之后，下一步的研究重点和难点是异构数据的表示模型、分类模型和检索模型以及海量数据查询，多任务调度算法的研究。

6 结束语

文章介绍了异构数据集成产生的背景和发展现状，现在已经研制出来了一些异构数据集成模型，并详细介绍了 Lore 模型的数学模型、查询语言和体系结构。最后，介绍了项目的一些情况和最新进展，以及下一步研究的重点。

（收稿日期：2002 年 1 月）

参考文献

1.王宁，陈滢，俞本权等.一个基于 CORBA 的异构数据源集成系统的设计[J].软件学报，1998，9（6）

2.姜宁，王忠，迟忠先.空间对象模型用于 Web 下数据源集成的研究[J].计算机工程与应用，2001，37（6）：93~95

3.黄为民，陈世福.分布式对象构件及其应用[J].计算机应用研究，2000-10

4.Borland/Inprise Company.VisiBroker for Java Reference & Visibroker for java Programmer's Guide[M].北京：机械工业出版社，2000

5.Andre Bergholz.Lore Tutorial.http://www-db.standford.edu/lore

6.T Lahiri，S Abiteboul，J Widom.Ozone：Integrating Structured and Semistructured Data.http://www-db.stanford.edu/pub/papers/ozone.ps

7.王宁，王能斌.异构数据源集成系统查询分解和优化的实现[J].软件学报，2000，11（1）

8.McHugh J，Abiteboul S，Goldman R et al.Lore：a database management system for semistructured data[J].ACM SIGMOD，1997，26（1）：54~66

9.李冠宇，靳强勇，张俊.一个改进的基于 CORBA 的异构数据集成系统体系结构[J].交通与计算机，2001-04

10.李冠宇，张俊，靳强勇.The Research of Heterogeneous Data Integration in Information Systems[C].ICMSE 2001/Harbin2001 会议论文集，2001

（止接 37 页）

adaptation in immune system as an evolution[C].In：Proceedings of the IEEE Conference on Evolutionary Computation，1996：150~153

11.Y Ishida，N Adachi.Immune algorithm for multiagent：application to adaptive noise neutralization[C].In：IEEE International Conference on Intelligent Robots and Systems，1996：1739~1746

12.R Deaton，R C Murphy，M Garzon et al.DNA based artificial immune system for self-nonself discrimination[C].In：Proceedings of the IEEE International Conference on Systems，Man and Cybernetics，1997：862~865

13.曹先彬，刘克胜，王煦法.基于免疫遗传算法的装箱问题求解[J].小型微型计算机系统，2000，21（4）：361~363

（止接 49 页）

法访问。有了图像数据，后续处理（如数字滤波、解相位、由物相关系求高度等[4]），就可以直接针对缓存中的数据来进行。

3.6 实验结果

笔者采用的是某公司开发的 MPE1000N 图像卡，由于服务商不再提供该卡 MFC 版本的驱动程序，使得在源驱动程序（API 版本）的基础上进行再开发变得很困难。用户只需自行开发图像后续处理进程，然后采用进程通讯的办法，就可避免对源驱动程序做大的修改。这样既可以使用原有驱动程序的功能，又可以发挥 MFC 的高效性。该方法已成功地应用于光栅投影三维轮廓测量中了，图 3 为石膏像的三维数字轮廓测量结果。

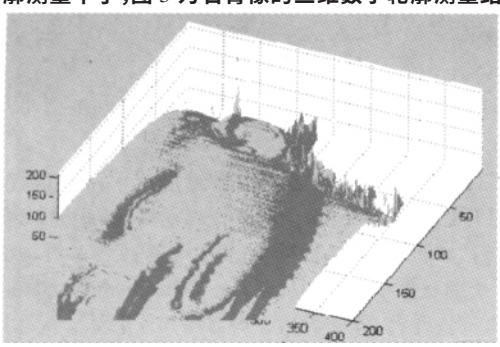

图 3 测量结果

4 结束语

通过在三维数字轮廓术中的图像采集进程和三维图像处理进程间实现数据通讯，说明了基于用户自定义消息和文件映射的通讯方法，给出了具体过程。通过自定义消息和文件映射实现进程间的数据传递，可以方便、灵活、快速地满足用户对数据的不同要求，对于工程软件的再开发有着很强的实用性和高效性。（收稿日期：2002 年 3 月）

参考文献

1.Takada M，Mutoh K.Fourier transform profilometry for the automatic measurement of 3-D shapes[J].Applied Optics，1983，22（14）：3977~3982

2.Sansoni G，Biancardi L，Minoni U et al.A novel，adaptive system for 3-D optical profilometry using a liquid crystal light projector[J].IEEE Transactions on Instrumentation and Measurement，1994，43（4）：558~566

3.John E Swanke.Visual C++ MFC 扩展编程实例[M].北京：机械工业出版社，2000

4.Pierre Soille.Morphological phase unwrapping[J].Optics and lasers in engineering，2000，32：339~352

5.David J Kruglinski.Visual C++技术内幕[M].第四版，北京：清华大学出版社，1995

6.Jeffrey Richter.Programming Applications for Microsoft Windows[M].4th edition，Microsoft Press，1999