计算机科学 2005Vol. 32No. 11

基于发布/订阅通信的动态数据集成模型*)

张志伟[1] 郭长国[3] 曹贺锋[2] 王伟球[4] 王 睿[4]

(空军指挥学院科研部 北京 100089)[1] (国防科技大学计算机学院 长沙 410073)[2]

(武警总部通信部 北京 100089)[3] (武警上海总队通信处 上海 200031)[4]

摘 要 提供灵活、动态、一致和开放的数据集成模型对于通过集成异构信息资源,实现大规模企业信息系统至关重要。本文以中介模型为基础,提出了一种基于发布/订阅通信的动态数据集成模型,与相关工作相比,该模型的显著特点在于通过发布/订阅机制支持信息资源的动态可插拔、支持分布的信息检索、提供了动态可配置框架、集成节点具有对等性。

关键词 动态数据集成模型,发布/订阅通信

A Dynamic Data Integration Model Based On Publish/Subscribe Communication

ZHANG Zhi-Wei[1] GUO Chang-Guo[2] CAO He-Feng[3] WANG Wei-Qiu[4] WANG Rui[4]

(Scientific Research, Air-force Commanding College, Beijing 100089)[1]

(Department of Computer, National University of Defense Technology, Changsha 410073)[2]

(Communication Part, Armed Polile Headquaiters, Beijing 10089)[3]

(Communication Part, Armed Polile Headquarster of Shanghai, Shanghai200031)[4]

Abstract Providing flexible, dynamic, uniform and open data integration model is vital to implement enterprise information system by means of integrating heterogeneous information resource. This paper extends the traditional mediator-based integration model and presents a dynamic data integration model. The most distinct characteristics of the model compared with related researches lies in the support of dynamic plug of information resources, distributed information retrieve, dynamic configure framework and symmetry of integration nodes by means of publish/subscribe communication.

Keywords Dynamic data integration model, Publish/subscribe communication

1 引言

随着 Internet 和大规模 Intranet 的出现和飞速发展,软件系统的主要形态、运行方式、生产方式和使用方式发生了巨大的变化。现在的软件系统可能涉及大量的异构数据信息,这些数据分布在具有不同存储介质的系统之中,不同的系统各自进行着自主和异构的演化。如何提供统一的数据集成模型,屏蔽系统的异构性和数据的异构性,实现数据集成和资源整合是实现企业信息共享的关键问题。数据集成模型在诸多领域得到了广泛的研究,出现了许多不同的模型和实现系统,如中介模型(Mediated Model)、联邦数据库模型(FDBM)和数据仓库模型等。

本文以中介集成模型为基础,提出了一种基于发布/订阅(P/S:Publish/Subscribe)通信模式的动态数据集成模型 PSDIM(Publish/Subscribe based Data Integration Model)。PSDIM 以 P/S 通信模型和 XML 技术为基础,能够实现分布、异构和动态的数据集成。与现有工作相比,PSDIM 的显著特点包括如下四点:通过发布/订阅机制支持信息资源的动态可插拔、支持分布的信息检索、提供了动态可配置框架、集成节点具有对等性。

2 相关工作

自 20 世纪 80 年代以来,以企业应用集成为原动力的数据集成模型得到了广泛深入的研究,典型的数据集成模型包括:

(1)中介模型[1] 其通过统一的全局数据模型集成异构的数据库和遗留系统。中介模型一般包括应用层(Application Layer)、中介层(Mediator Layer)、适配层(Adaptor Layer)和数据层(Data Layer)。中介层是中介模型的核心,它负责连接应用系统(应用层)和适配层,向上为数据访问提供统一的数据模式和统一的访问接口,向下通过适配层协调对异构数据源的访问。图 1 给出了中介模型的典型架构。

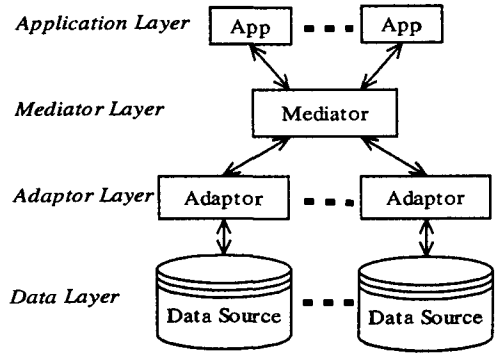

图 1 基于中介的数据集成模型

(2)联邦数据库[2] 其是一种虚拟的数据仓库,它通过联邦数据库管理系统实现对分布异构数据源的统一访问。FDBM 的重要特性之一是各个参与联邦的数据库在参与联邦的同时,又具有自治性。根据耦合度的不同,FDBM 分为紧耦合

*)基金项目:863 课题(2003AA115410),自然科学基金(No. 90104020)。张志伟 博士,主要研究方向为分布计算与软件工程技术;郭长国博士,研究方向为分布和实时系统。

和松耦合两种形态。紧耦合 FDBM 能够提供统一的数据访问模式，但是增加数据源比较困难；松耦合 FDBM 不提供统一的接口，但是可以通过统一的语言访问数据源。

(3)*数据仓库* 其是另外一种广泛应用于管理决策的数据集成模型，其典型特点是面向主题、相对稳定和能够反映数据的历史变化轨迹。数据仓库的关键技术包括数据抽取、数据存储和管理、数据表现和数据仓库设计等。

上述三种数据集成模型都在一定程度上解决了应用之间的数据共享和互通问题，但是它们之间存在差异。基于中介的集成模型相对易于实现，成为主流的数据集成模型，引起了研究人员的广泛关注，涌现出了许多基于中介的数据集成系统。文[3]对传统的中介架构进行扩展，提出了一种基于中介机制的动态数据集成框架，他们的工作专注于资源描述和变更描述上。文[4]研究了数据集成系统的三层体系结构和查询改进算法。文[5]研究了基于消息队列的分布数据集成问题。

与文[3~5]相比，PSDIM 是一种具有灵活性和开放性的架构，不仅支持集中式结构，也支持分布式结构，支持信息资源动态可插拔，提供动态可配置框架，集成节点具有对等性。

3 PSDIM

3.1 PSDIM 模型

PSDIM 模型系统结构如图 2 所示。为了便于讨论，下面首先给出 PSDIM 的定义，然后讨论具体的模型和算法。

定义 $PSDIM=\langle PSS, DM, DAA, MD\rangle$

其中 PSS 为发布/订阅服务器(PSS：Publish/Subscribe Server)，它作为 PSDIM 的基础设施，PSDIM 中可以有多个 PSS 实例，例如在集群环境下，多个 PSS 构成一个集群。数据中介(DM：Data Mediator)是 PSDIM 的核心，它主要包括代理(Broker)、适配工作器(AdaptorWorker)和适配器(Adaptor)等三个核心构件。其中代理作为 PSS 的客户存在，它根据元数据(MD：Meta Data)订阅数据访问主题(图 2 中为 Request 主题，可在 MD 中动态配置)。代理从数据访问主题接收数据访问请求，一旦接收到请求之后就启动适配工作器。适配工作器负责启动具体的适配器执行数据访问动作。适配器负责封装底层不同的数据源，向上提供统一的数据访问接口，它根据数据源元数据配置访问具体的数据源，并将结果发布给代理，由代理将结果发布到 PSS 中的数据发布主题(图 2 中为 Reply 主题，可在 MD 中动态配置)。

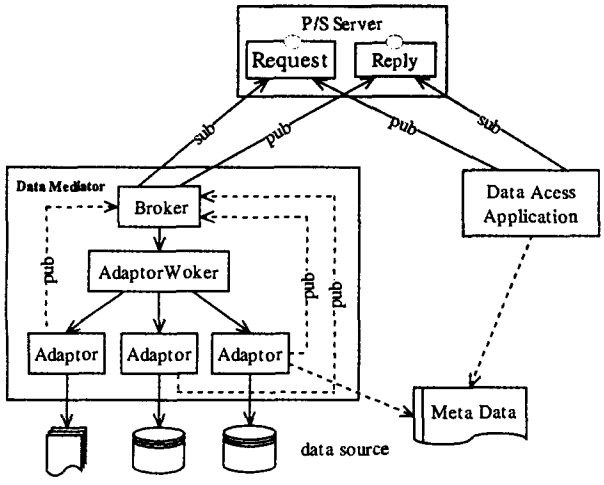

图 2 PSDIM 模型系统结构图

数据访问应用(DAA：Data Access Application)是数据源的访问者，它利用 PSS 中的数据访问主题发布自己的数据访问请求，并订阅到数据发布主题接收数据访问结果。

MD 是 PSDIM 中的元数据配置库，主要包括三类：首先，数据访问主题和数据发布主题的相关元数据；其次，数据源配置的元数据；最后，安全认证元数据。MD 使 PSDIM 具有良好的灵活性和开放性，例如可以根据不同的应用配置不同的数据访问和数据发布主题，可以在 DAA 不感知的情况下灵活地加入和替换数据源。

3.2 PSDIM 算法

为了方便建模，本文定义如下：

①MD 中配置的数据访问主题记为 T_{DA}，数据发布主题记为 T_{DP}；

②PSS 支持的操作包括 connect(建立连接，同时进行安全认证)、publish(向主题发布消息)、subscribe(订阅主题)和 disconnect(断开连接)；

③DAA 对每一个数据访问请求 Req 产生一个唯一标识 ID^{Req}。ID^{Req} 和 Req 一起被封装为一个 XML 消息传递至 DM；

④适配器将数据访问应答 Rly 与 ID^{Req} 一起封装为 XML 消息返回给 DAA。

下面分别讨论 DAA 处理算法和 DM 处理算法。

• DAA 处理算法：

```
Algorithm DAA
Input: Reqs
Output: Rlys
{
 ①con←PSS. connect;
 ②con. subscribe(TDP, this);
 ③construct Req and unique IDReq;
   wrap into xml message xml(req,IDReq);
 ④con. publish(TDA, xml(req,IDReq));
 ⑤OnMessage (Msg){
   Decode Msg , distinguish and associate Rly with IDReq
}}
```

• DM 处理算法：

```
Algorithm DM
Input: Reqs
Output: Rlys
{
Broker;
① con←PSS. connect;
② con. subscribe(TDA, this);
③ onMessage(Msg ) {
   decode Req and IDReq from Msg;
   Start AdaptorWorker with Msg;
}
AdaptorWoker(Msg){
④ Locate Adaptor;
⑤ Get Meta Info mti about data source and topic from MD;
⑥ Rly←Adaptor. invoke(Msg, mti);
⑦ Wrap Rly and IDReq into xml message xml(rly,IDReq);
⑧ con. publish(TDP, xml(rly,IDReq));
}}
```

算法分析：

算法 DAA 和算法 DM 协同工作，共同完成数据集成的任务。DAA 在建立连接(①)、订阅 T_{DP} 主题(②)之后，构造数据访问请求并将其封装为 XML 消息(③)，然后将 XML 消息发布到 T_{DA} 主题。一旦 DM 通过⑧将访问结果发布到 T_{DP} 主题，PSS 就将数据访问结果推送至 DAA(⑤)。

DM 中的代理作为普通的 PSS 客户订阅到 T_{DA}(②)，接收到 DAA 发布的数据访问请求之后(③)，DM 就将数据访问请求交付给适配工作器。适配工作器负责定位适配器(④)，获取元信息(⑤)，通过适配器激活数据访问请求(⑥)，最后将结果和请求唯一 ID 一起封装为 XML 消息(⑦)并发布到数

据发布主题(⑧)。

4 基于 PSDIM 的分布数据集成

第 3 节讨论了单一 PSS 环境下的 PSDIM 模型,PSIM 模型的一个良好性质就是它在集群环境下同样能够很好地工作。这种性质使得通过 PSDIM 模型可以方便地实现分布式数据集成系统。图 3 给出了一种基于 PSDIM 模型的分布数据集成系统,图中三个 PSS 构成一个集群,DAA 需要访问异地远程数据源。对于提供数据访问的节点,只需部署 DM 和配置必要的元数据 MD。DAA 的数据访问请求首先发布到数据访问主题,然后被传播到集群中的其它包含数据源的集成节点,这些节点通过 DM 访问数据源,将数据访问结果发布到数据发布主题,订阅到数据发布主题的 DAA 接收数据访问结果,进行数据综合和抽取,将结果交付至上层应用。DAA 的数据访问请求可以是标准的 SQL 语句,也可以是应用层的业务逻辑表示,后一种情况需要在 MD 中建立相应的映射。

基于 PSDIM 的分布数据集成具有如下特点:首先,体系结构具有灵活性和开放性,加入新的节点只需部署 DM 和必要的 MD 即可,这在 EAI 信息集成和资源整合中具有重要意义;其次,用户接口简单、灵活,易于实现、部署和维护;对等性,任意节点可以是 DAA,普通的提供数据访问的节点,或者既是 DAA 又是提供数据访问的节点。

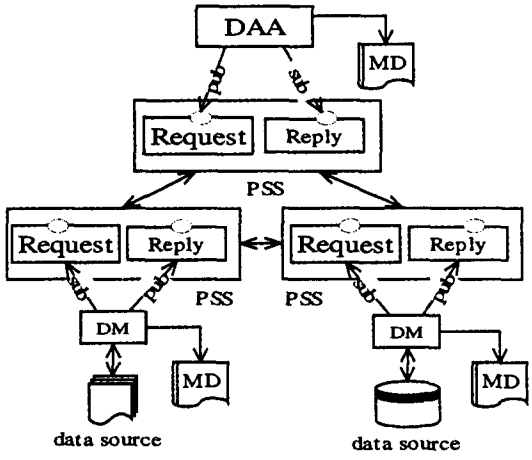

图 3 基于 PSDIM 的分布数据集成

5 实现

我们基于本文提出的 PSDIM 模型实现了一个数据集成系统,主要实现了数据中介 DM。目前的数据中介实现支持通过 JDBC 访问异构数据库。图 4 给出了一个具体的数据中介元数据配置例子,主要包括四部分信息:ID,用于标识一个特定的数据源,一个 DM 可以同时连接多个数据源;安全认证信息,即 DM 作为 PSS 客户连接到 PSS 时的认证信息;数据订阅主题和数据发布主题元信息;数据源元数据,例子以 SQL Server 为例。

```
〈DataMediator xmlmns="http://www.w3.org/2001/XMLSchema-instance"
    xsi:noNamespaceSchemaLocation="DataMediator.xsd"〉
 〈MetaData〉
   〈ID〉SQL〈/ID〉
   〈PSUser〉some〈/PSUser〉
   〈PSPasswd〉tiger〈/PSPasswd〉
   〈SubTopic〉Request〈/SubTopic〉
   〈PubTopic〉Reply〈/PubTopic〉
   〈DataSource〉
    〈DricerClass〉com.microsoft.jdbc.sqlserver.SQLServerDriver〈/DriverClass〉
    〈URL〉jdbc:microsoft:sqlserver://localhost:1043;DatabaseName=db=〈/URL〉
    〈User〉sa〈/User〉
    〈Passwd〉123〈/Passwd〉
     〈ConnectionPool〉
     〈Size〉5〈/lSize〉
     〈BusyTime〉90000〈/BusyTime〉
    〈ConnectionPool〉
   〈/DataSource〉
 〈/MetaData〉
〈/DataMediator〉
```

图 4 PSDIM 中 MD 配置例子

结论 数据集成在大规模企业信息系统中具有重要地位,传统的数据集成技术在处理具有高度自治性和异构性的数据资源整合时还存在缺陷。本文提出了一个基于发布/订阅通信的动态数据集成模型 PSDIM,该模型因为具有支持信息资源动态可插拔、支持分布的信息检索、支持动态可配置和集成节点具有对等性等优点而能够较好地支持具有自治性和异构性的系统之间的动态数据资源整合。PSDIM 模型已经在国税数据大集中系统中得到了初步的应用,应用结果表明了 PSDIM 的有效性。进一步的工作包括查询性能优化和 DM 管理问题。

参 考 文 献

1 Wiederhold G. Mediators in the architecture of future information systems. IEEE Computer, 1992, 25(3):38~49

2 Sheta A P, Larson J A. Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases. ACM Computing Surveys, 1990,25(3)

3 方俊,虎崇林,韩燕波.一个基于中介机制的动态数据集成框架.计算机科学, 2003,30(10 A):259~262

4 谢丽聪,白清源,余建家.数据集成的三层体系结构及其查询改写算法的改进.计算机科学,2003,30(10 A):255~258

5 张磊,陈荦,吴秋云,李军.基于消息队列的分布式信息查询技术的研究与实现.计算机科学,2003,30(10 B):4~6