地理信息系统空间异构数据源集成研究

黄照强 冯学智
(南京大学地理信息系统与遥感实验室,南京 210093)

摘 要 通过对地理信息系统(GIS)中空间异构数据源的访问和集成,对比了目前比较通用的几种集成技术和方法,讨论了空间异构数据集成的关键理论和技术,着重介绍了基于 XML 和中间件技术的集成方法。认为 XML 的最大优点在于它的数据存储格式不受显示格式的制约。随着 RDF 和 VDB 的发展,以 XML 作为集成层的数据描述工具和转换工具,建立具有多数据源集成能力的中间件,能适合空间异构数据源集成的需要。

关键词 空间异构数据 数据集成 XML 中间件
中图法分类号: TP311 **文献标识码**: A **文章编号**: 1006-8961(2004)08-0904-04

The Research of Spatial Heterogeneous Data Source Integration of GIS

Huang Zhao-qiang, Feng Xue-zhi
(*Laboratory of GIS & RS, Nanjing University, Nanjing* 210093)

Abstract The article compares several universal integration techniques and methods via to access and integrate Spatial Heterogeneous Data on GIS. And in the paper, the authors discuss the key theory and technique of Spatial Heterogeneous Data Integration. Last, we give emphasis introducing the Integration method based on XML and Middleware technique. And we consider that the most major of XML is the independence of it's data storage format with display form. Along with the development of RDF and VDB, with data describing tool and transition tool based on XML by way of integration layer, middleware possessing capability of data sources integration is suit to the need of spatial heterogeneous data source integration.

Keywords spatial heterogeneous data, data integration, XML, middleware

1 引 言

随着美国副总统戈尔于 1998 年 1 月 21 日提出数字地球的概念之后,在国家战略体系和国家空间数据基础设施(NSDI, national spatial data infrastructure)建设的驱动下,中国许多城市提出数字城市的构想。国家 1∶100 万、1∶25 万的基础数据库、数字高程模型库、地名数据库已经建立,1∶5 万的数据库正在建设。同时各机关企事业单位正逐步实现计算机化,各类数据库中的数据量剧增。空间数据的应用越来越广,城市规划、建设、市政设施等基础数据不断服务于人们的应用需要,就要求从分布式、异构的数据源中检索和集成信息,从各种异构系统和异构数据库中提取信息,甚至在某些情况下还要保证数据库24×7的高可用性和实时响应,这越来越成为当前研究的一大难题。

2 空间异构数据源的异构特性

目前 GIS 应用系统很多,开发平台多种多样,如 MapInfo、Arc/Info、GeoMedia 等,而且大部分系统都是针对某一类特定的 GIS 数据集及其相关应用而设计开发的。同时随着目前空间信息科学的迅猛发展,空间数据的信息量呈现级数增长,分布式存储和管理已成为空间信息的主要访问形式,而且使用的空间数据库管理系统也不尽相同,数据存储的方式也不尽相同。各种系统没有统一的数据标准,但

基金项目:国家 973 项目(2002CB410810);国家自然科学基金(40301038)
收稿日期:2003-11-05;改回日期:2004-04-26

又有可能交叉，需要进行信息共享。随着信息化应用的发展和数据库技术的发展，以及各类应用要求的差异，也引发了数据源集成问题。空间数据是人们赖以认识自然和改造自然的重要基础数据，空间数据库包括空间数据和非空间数据[1]。空间数据源（同构和异构）的集成给人们用系统科学的方法研究地理环境提供了基础。空间异构数据源（如图 1）的异构特性主要表现在以下两个方面：

（1）系统异构　指的是数据源所依赖的操作系统（如 Windows、Unix 等）、数据库管理系统（如 DB2、Oracle、SQLServer、Sybase、Foxpro、Access 等）以及相应的业务应用系统（如土地、规划、市政设施等）的不同构成了异构数据；

（2）模式异构　指数据源在存储模式上的不同，存储模式主要包括关系模式、对象模式、对象关系模式和文档嵌套模式等几种，同时，即便是同一类存储模式，它们的模式结构可能也存在着差异，例如不同的关系数据管理系统的数据类型等方面并不是完全一致的，如 Oracle 所采用的数据类型与 SQLServer 所采用的数据类型并不是完全一致的。

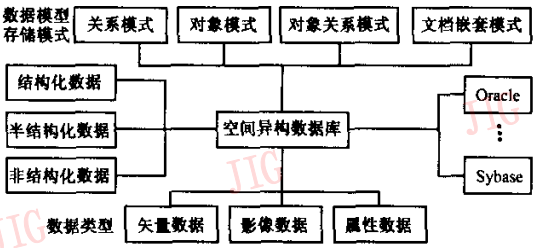

图 1　空间异构数据库

随着城市信息化平台的建设，如何充分利用各种已建立的数据库资源，实现不同数据库之间的连接、数据交换、数据共享、协同工作，解决语义冲突，已经成为一个关键问题。同时在集成后数据要保证集成性、完整性、一致性和访问安全性。

3　空间异构数据集成技术

随着分布式应用越来越广泛，多数据源的集成以及数据访问透明性问题已变得越来越重要。实现异构数据库的集成一般有两种方法。一种是将源数据库中的数据复制到新的数据库管理系统中，为了集成不同类型的数据，必须将一些非传统的数据类型转化成新的数据类型。许多关系数据库供应商提供了类似的功能，如 DTS（data transformation service）提供了数据转换服务。另一种是利用中间件技术来集成异构数据库，该方法并不需要改变源数据的存储和管理方式。中间件位于空间异构数据库系统（数据层）和应用程序（应用层）之间，向下建立与数据层的联系，协调各数据库系统，向上为访问集成数据的应用，提供统一数据模式和数据访问的通用接口。通过在系统的业务逻辑、通用构件服务（如事务、安全、并发控制等）和数据源之间建立一个中间层，对服务层屏蔽数据源的差异。显然，中间件系统模式是实现空间异构数据集成较理想的解决方案。中间件层向服务层提供一致的数据视图，完成从实际数据源到用户数据视图的转换，并在中间充当数据总线的作用。

目前常用的异构数据集成技术方法主要有：通过 ODBC 建立连接；传统的模式集成；通用数据访问结构；基于视图的集成技术；DCOM/CORBA 技术；利用 XML。

（1）ODBC 方法　ODBC 是 Microsoft Windows 开放服务体系结构（WOSA）的主要部分，是一个数据库访问的标准接口。ODBC 体系由 4 部分组成：ODBC API 应用程序、驱动程序管理器、ODBC 驱动程序、数据源。程序执行时调用相应数据库的 DLL 来执行对数据的操作，具有良好的数据独立性，而且使用层次方法来管理数据，在数据库通讯结构的每一层引入一个公共接口。但仍需针对某平台和 DBMS 进行 DLL 文件的装载，执行时调用相应数据库的 DLL 来完成对数据的操作。

（2）传统的模式集成方法　模式集成是指将各个数据库中的信息在逻辑上用同样的概念模式表示以形成一个统一的异构数据库，达到数据共享的目的。分为有全局模式和无全局模式[2]。有全局模式指每个参与集成的数据库有自身的局部概念模式，用户可以通过建立在局部概念模式上的局部外模式访问本地库，在所有局部概念模式的基础上建立全局概念模式，用户通过建立在全局概念模式上的全局模式访问集成系统中的其他数据库。无全局模式指通过一系列定义在某个或某几个局部模式上的外视图访问下层的各个局部数据库系统，但需建立逻辑模式转换器。

（3）通用数据访问结构　通用数据访问结构是 Microsoft 公司继 ODBC 后推出的新一代数据访问组件，实际上是一组软件组件，这些组件之间通过

OLE DB 定义的一组公共的系统级界面进行互操作。此软件组件包括 3 层，下层是数据提供者，储存数据；上层是数据消费者，使用数据；中间层是一系列服务组件，用以对数据进行各种处理。需要提供 OLE DB 的支持。

(4) 基于视图的集成技术　即先建立空的集成视图，然后各个异构的数据库将自身想要参与集成的类(即共享信息)输入到集成视图中。集成系统通过语法、语义的分析解决各输入类之间的冲突，并进行类的派生操作，从而建立适宜于数据共享的集成视图。其最典型的应用是数据仓库。但由于各个数据源中的数据处于不断地变化之中，数据仓库系统要及时刷新数据以反映这种变化，维护效率较高。

(5) DCOM/CORBA　DCOM 是组件对象模型(COM)的进一步扩展[3]。客户通过组件对象提供的接口直接访问组件中的方法。CORBA 使得面向对象的软件组件在分布式异构环境中可重用、移植、互操作，同时具有跨平台、语言中立、伸缩性健壮性好等特性。两者都采用远程进程调用的方式和封装的思想，以统一的接口方式向外提供调用，并且二者也都实现了对对象的透明访问。但 DCOM 只适用于 Windows 平台，而 CORBA 与操作系统的交互必须通过中介代理进行，运作效率较低。

(6) 利用 XML 进行集成　XML (extensible markup language)是扩展标示语言，它以一种开放的自我描述方式定义数据结构，具有可扩展性、结构性、平台独立性。XML 使用 DTD 和 Schema 来定义数据的结构，使用 DTD，不同组中的人就能够使用共同的 DTD 来交换数据，并可以验证你接受到的数据是否有效，更重要的是还能够定义数据的类型和数据间的关系。基于 XML 的 RDF[4](resource description framework)提供一种处理元数据的环境。

由以上几种方法的基本原理及特点可以看出，在现今 Internet 时代，利用 XML 进行集成的方法是开放的、平台独立的、可扩展的，并且 RDF 提供了一种处理元数据的环境，对于目前繁杂的、日益增长的空间异构数据来说，它的数据访问能力更加灵活、强大，更适用于目前日益复杂的空间数据异构集成，是未来发展的方向。空间异构数据集成的总体体系结构如图 2 所示，分为 3 个层次：源数据库层、中间控制层(中间件层)、目标应用层(GIS 数据仓库，更高级的集成，应用层)。

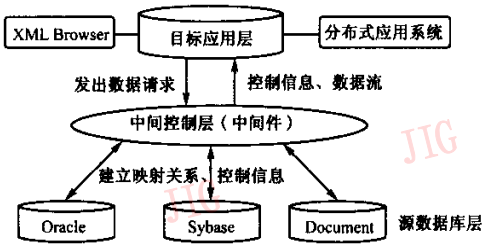

图 2　集成总体体系架构

4　基于 XML 的集成技术实现

XML 是在文本处理系统地发展过程中产生的，是 SGML(standard generalized markup language，标准化通用标记语言)经过优化后的一个子集[5]。它将 SGML 的丰富功能与 HTML 的易用性结合到 Web 的应用中，以一种开放的自我描述方式定义了数据结构，在描述数据内容的同时能突出对结构的描述，从而体现出数据之间的关系。这样所组织的数据对于应用程序和用户都是友好的、可操作的。XML 的最大优点在于它的数据存储格式不受显示格式的制约。XML 能够描述不规则数据[6]，以 XML 作为集成层的数据描述工具和转换工具，建立具有多数据源集成能力的中间件，能适合空间异构数据源集成的需要。XML 中间件层结构如图 3 所示。

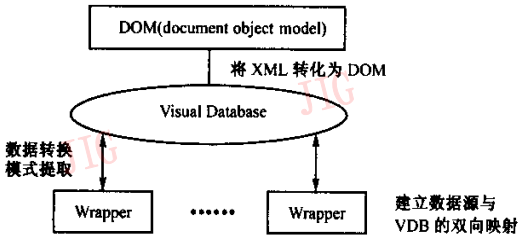

图 3　XML 中间件层结构

在数据管理上，首先建立一个统一的 DTD，所有的站点使用统一的文档类型定义(DTD)。所有数据的存取请求都要通过 XML 中间件层，中间件层存储着所有数据的集成模式。在中间件层中连接每个数据源的是一个适配器(Adapter)或“外套”(Wrapper)，只需用一种统一的可扩展标示语言 XML 和空间数据库引擎(SDE，一种中间件，专门针对空间数据而设计的空间数据库引擎用来访问空间数据库)为各类源

数据库设计一个“外套”,就可构成一个“虚拟的源数据库服务器”,为系统提供服务。上层针对虚拟的源数据库服务器进行访问,将 XML 转换成一个 DOM 文档对象模型,为上层提供服务。DOM 另外又是一个接口,一个与语言无关的接口,应用通过这个接口来和 XML 或 HTML 内的数据打交道。DOM 由核心(core)、HTML 和 XML 3 部分组成。核心部分是结构化文档比较低层对象的集合,这一部分所定义的对象已经完全可以表达出任何 HTML 和 XML 文档中的数据了。HTML 接口和 XML 接口则是专为操作具体的 HTML 文档和 XML 文档所提供的高级接口,使对这两类文件的操作更加方便。

在地理信息系统中空间异构数据源涉及的面较广,由于应用过程中,所使用的 GIS 平台和操作系统、数据库系统不同,存储模式也千差万别,因此需要一种开放的,具有可扩展性、结构性、平台独立性的集成方法。正因为 XML 及其 RDF 对空间数据、属性数据、文档、多媒体等数据的表达能力,使得能够实现城市信息化平台的建设,充分发挥信息共享的效率。

下面的例子说明了在中间层上 XML 文档可以包括来自多个独立的数据库中的数据,通过连接两个不同数据源的数据库生成一个 XML 文档,然后 VDB 使用包装器和提取器的组合把异构数据源(非结构的、半结构的、结构的)转换成数据库形式,最后转换成一个 DOM,为上层应用提供服务。其过程为

```
<%@ LANGUAGE=VBScript%>
<? xml version="1.0">
<AUCTIONBLOCK>
<%
'The connection to the City data source is made
Set Coon=Server.CreateObject("ADODB.Connection")
Conn.open "City","City","City"
Set ItemRS=Conn.Execute("select * from units")
Do While Not ItemRS.EOF
%>
<ITEM>
  <TITLE><ItemRS("Name")%></TITLE>
  <ADDRESS><%=ItenRS("Address")%></ADDRESS>
  <POSTAL><%=ItemRS("Postal")%></POSTAL>
  </ITEM>
<%
ItemRS.MoveNext
Loop
%>
<%
'The connection to the City2 data source is made
Set Coon=Server.CreateObject("ADODB.Connection")
Conn.open "City2","City2","City2"
Set ItemRS=Conn.Execute("select * from units")
Do While Not ItemRS.EOF
%>
<ITEM>
<TITLE><%=ItemRS("Name")%></TITLE>
<ADDRESS><%=ItenRS("Street")%></ADDRESS>
  <POSTAL><%=ItemRS("PostalID")%></POSTAL>
</TITLE>
<%
ItemRS.MoveNext
Loop
%>
</AUCTIONBLOCK>
```

5　结　论

目前空间异构数据集成已成为信息技术发展的一个重要发展方向。基于 XML 和中间件的技术是目前较实用的集成方法之一,但却没有形成一个面向任何数据的集成方法。相信随着科技的发展和应用的迫切需要,特别是 RDF(资源描述框架)和 VDB(虚拟数据库)的理论和技术的发展,将促进空间异构数据集成方法的发展。

参 考 文 献

1　李德仁,王树良,史文中等.论空间数据挖掘和知识发现[J].武汉大学学报(信息科学版),2001,**26**(6):491～499.

2　谢鸿强,董逸生.异构数据源的集成技术[J].工业控制计算机,2001,**14**(6):1～6.

3　Richard C. Leinecker. COM+技术大全[M].北京:机械工业出版社,2001.

4　RDF Model and Syntax Specification[EB/OL]. http://www.w3.org/TR/REC-rdf-syntax/.

5　Charles F Goldfarb, Paul Prescod. XML 用户手册[M].北京:人民邮电出版社,2000.

6　李军怀,周明全,耿国华等.XML 在异构数据集成中的应用研究[J].计算机应用,2002,**22**(9):10～12.

黄照强　1973 年生。南京大学城市与资源学系博士生。主要研究方向为遥感与地理信息系统。

E-mail:hzhaoq@sina.com

冯学智　1953 年生。教授、博士生导师。主要研究方向为遥感与地理信息系统。已在国内外发表论文 120 余篇,获国家科技进步奖课题 7 项。