空间数据仓库及其构建策略

李 琦 杨超伟
(北京大学数字地球工作室, 北京 100871)

摘 要 数字地球建设要求广泛共享空间信息,空间数据仓库则为空间信息的有效管理和大众分发提供了有效的工具。文中分析了建设、使用空间数据仓库的背景,探讨了它的建立策略和总体体系结构。然后讨论了空间数据仓库的建设步骤,基于这些理论基础,设计并实现了一个空间数据仓库的原型系统。

关键词 空间数据仓库 策略 体系结构 关键技术 原型 数字地球

0 引 言

(1) 数字地球

自从 1998 年 1 月 31 日被提出以来[1],数字地球已引起了信息产业界、学术界、政界和公共的广泛关注,各国对此给予了极大的重视。在美国,一个由 NASA, NOAA, NSF, DOD 和许多其他组织、政府部门组成的数字地球的机构已经形成。在中国,许多大学[2]和研究所[3]的专家学者围绕着数字地球举行了许多会议,信息产业界对此表现出极大兴趣。政府部门也决心实施数字中国工程。许多其他国家也表示要建设数字国家。但数字地球将以何种方式影响我们的生活呢?

美国副总统戈尔[1]在他的谈话中用一个小女孩漫游数字博物馆的例子展示了数字地球的应用。Ralph Kahn[4]则给出了 15 个例子说明数字地球的未来潜在应用领域。国家科技部副部长徐冠华[5]从国家的角度说明了数字地球可以用于支持可持续发展和国家经济增长及国家安全。正如我们所看到的,这些数字地球的潜在应用例子都要求它的使用者能同时广泛使用数字地球上的空间信息和空间分析显示服务。怎样才能建立一个实现这些功能的复杂系统呢?其中最为重要的步骤之一就是建立一个稳定、快速的空间数据仓库来管理和处理这些空间信息,也就是说,我们需要空间数据仓库[6]。

(2) 数据仓库

专家学者对数据仓库的研究已持续了许多年[7]。很多信息公司,如 Microsoft, Oracle, Informix 等已提出了自己的实施策略。数据仓库之父[8]是这样定义的:"空间数据仓库是一个面向主题,集成的,基于时间段的动态信息集合,这个集合用于支持管理决策。"

事实上,数据仓库是从一种崭新的哲学观念来看待信息管理方法和技术的[9]。

有两种数据仓库:一种叫做数据市场;另一种则直接被称为空间数据仓库。虽然许多人,尤其是系统集成商从商业的角度考虑,认为这两种系统是一样的,但它们彼此存在非常大的差别。这种差别体现在:数据模型,历史数据的存储量、主题相关性、查询访问类型、用户类型和主要体系结构等 6 个方面。

许多成功的数据仓库已存在多年[7],但由于空间关系、空间计算和空间分析的复杂性,特别是高维分析的复杂性,空间数据仓库还停留在理论阶段[6]。

(3) 空间数据库与空间数据仓库

自从 E. F. Codd 博士提出关系代数以来,近几十年数据库已获得了飞速的发展。关系数据库及其访问语言 SQL 已成为信息领域进行信息存储、操作及访问的标准。

最初空间信息的管理是分为两部分:文件系统进行空间数据的管理,属性数据则采用关系数据库来进行管理。然后 DBMS 开始用于空间信息的统一管理来建设空间数据库。空间数据库是由数据库和空间数据引擎(如 SDE 和 SpatialWare 等)构成的,数据库用于对数据进行结构化管理,空间数据引擎用于提供空间信息访问(如空间分析,空间运算,空间查询)功能[10]。

自从地理界的量化革命以来,人们在空间信息的建模、空间决策支持等方面积累了大量的方法、知

收稿日期:1999-11-08

识。信息科学和管理科学的发展要求将这些方法、知识归纳起来与空间数据引擎、数据库管理系统进行集成以构建对于空间决策支持系统的支持[11],也就是建设空间数据仓库[6]。

从这些历史信息,我们不难看出,空间数据仓库是数据仓库与空间处理分析的综合。但它们存在区别,从一个信息科学家的角度来说,空间数据仓库实际是对数据仓库加进了非结构化信息处理。由于存在这种差别,而且数据仓库技术已经成熟,本文将讨论空间数据仓库。

1　建设空间数据仓库需采用的策略

空间数据仓库内部或它所使用的数据是分布在不同地点的,我们需要解决这种数据分布的问题。一个公司的空间数据仓库将为公司中的大部分人服务,这些人在公司中处于不同的岗位,具有不一样的重要性,他们对空间数据仓库的需求也是不一样的。因此应对这些用户提供多级存储策略以满足他们的需要。对不同地点的不同数据源访问和安全性控制则要求一个用于管理和交换空间信息的数据交换中心。

1.1　空间数据分布

空间数据由数据本身和对他们进行描述的元数据组成,元数据是关于数据的数据。有 3 种方法实施空间数据存储。第一种是将所有数据存储在数据中心;另一种是在不同地点存储不同数据;第三种是在任一使用地点存储所有的数据。

1.1.1　数据中心策略

在这种策略中,所有数据存储在数据中心,任何一个用户在任一时刻访问数据都访问数据中心。数据可以被直接访问,但是数据中心服务器的负载非常重。处理过程可描述如下(图 1):

① 用户将请求通过 GUI 提交给客户端;

② 客户计算机分析请求并将请求发往数据中心;

③ 数据中心接受数据请求并在数据中心对其进行处理;

④ 处理结果被传回客户计算机;

⑤ 结果被显示给用户。

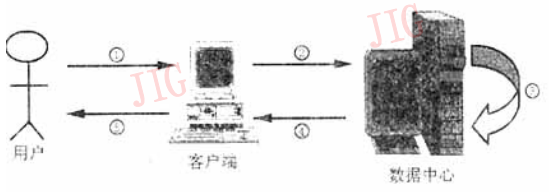

图 1　数据中心策略图示

这种策略用于人数不多,而公司里的许多人却都要求使用空间数据仓库的一些简单功能。数据市场就是这种方式的一个典型应用。

1.1.2　分布式策略

在这种策略中,数据分布于不同的地点,不同数据及其相关元数据存储在不同地点,这种存储策略用于中、大型公司,而且大多数用户对数据仓库的实时性要求不高的情况[16,17]。

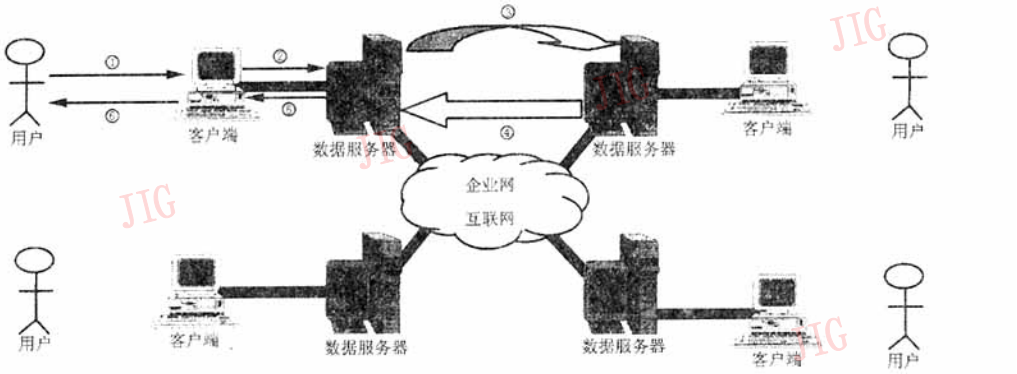

图 2　分布式策略

处理过程如下(图 2):

① 用户通过 GUI 将请求发往客户计算机;

② 客户计算机分析请求并将数据请求发往本地数据中心;

③ 本地数据中心接受数据请求,分析请求内容并按数据定位将数据请求发往其它的服务器;

④ 其他数据服务器处理请求并将结果传回本地服务器;

⑤ 本地服务器接受并处理数据且将结果传回客户端计算机;

⑥ 结果由客户端计算机通过 GUI 显示给用户。

由于在构建数据仓库时已存在许多的数据库系统,这种策略广泛应用于数据仓库中。

1.1.3 分布式重复数据策略

在这种策略中,公司的所有数据存储在任意一台服务器上,数据将被定期更新。这种策略的网络结构与分布式策略一样,但处理过程则与数据中心策略所采用的一致。这种策略一般用于大公司或重要部门中对空间数据仓库的实时性要求较高的情况。这种策略的一个非常复杂的问题是数据复制和完备性、完整性的维护。

1.2 空间信息的存储策略

空间数据仓库一般是由一个组织、机构或公司建立的。在这样一个实体内部,几乎每个人都需要访问它。这些用户来自于不同的背景,对空间数据仓库也有着不同的需求。这种应用的逻辑和层次结构非常明显。为满足不同级别用户请求和快速响应要求,采用了空间信息的多级存储策略[6]。

存在 3 种级别的存储(图 3):数据市场,部门空间数据仓库和全局空间数据仓库。数据市场是低级的查询结果数据集,它可用于许多普通用户;部门空间数据仓库用于为特定部门的领导服务,它是基于该部门的主题构建的;全局空间数据仓库是为整个公司的高层决策构建的,它的目的是为空间决策支持系统提供信息基础。

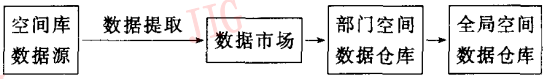

图 3 空间数据仓库的存储策略

1.3 数据交换中心

图 2 所示的数据请求与数据传输是直接在本地服务器和远程服务器之间直接进行的。这种方法存在着许多问题,例如,对于数据交换的管理将变得非常复杂,为管理空间数据仓库中的数据,元数据将不得不存储在每台计算机上。为实现对元数据的统一管理和解决一系列问题,需要采用空间数据交换中心。

空间数据交换中心用于管理数据交换,并存储所有系统元数据,从某种意义上说,它是一个目录服务器。在空间数据交换中心维护着一个有关所有数据的目录。

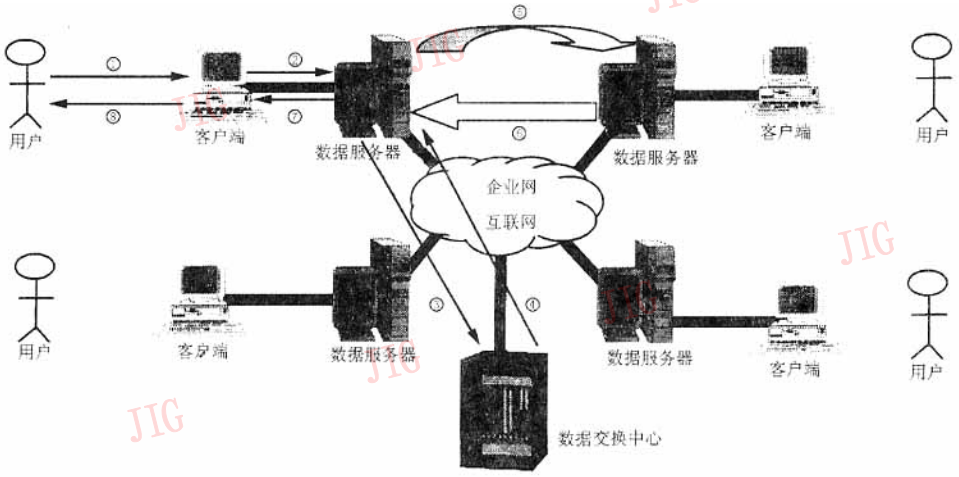

图 4 数据交换中心的作用

用户可以如下步骤访问空间数据仓库中的数据(图 4):

① 客户向客户端发送数据请求;

② 客户计算机分析请求并将请求发往本地服务器;

③ 本地服务器分析请求并向数据交换中心发送请求;

④ 数据交换中心找到数据的地址并将地址传回本地服务器;

⑤ 本地服务器按地址到远程服务器进行数据请求;

⑥ 远程服务器查到数据并将数据传回本地服务器;

⑦ 本地服务器接受并处理数据将结果传回客户端计算机;

⑧ 结果在客户端显示给用户。

这些策略涵盖了在建立空间数据仓库之前必须解决或进行的一些策略性问题。但一个空间数据仓库的总体是怎样的呢?

2　体系结构与信息流程

空间数据仓库将对一个企业或公司内的所有信息和数据进行管理以提供对空间决策支持系统的支持。它是怎么实现的呢?让我们从它的体系结构来看看它是如何实现这些功能的。

一个空间数据仓库的结构包括它的数据信息,硬软件和相关的人力资源。如图 5 所示,它的信息流程是:数据源、数据库、空间数据仓库统计信息、分析信息、决策知识[6,12]。

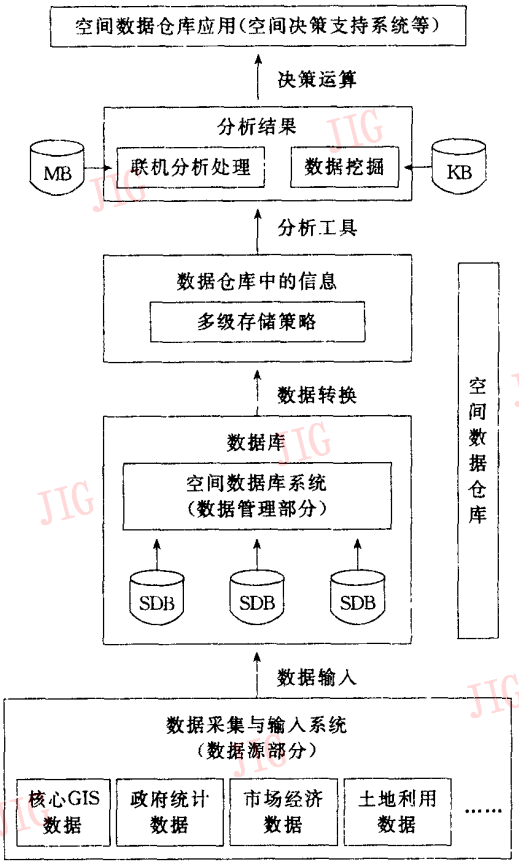

图5　空间数据仓库的体系结构

处理过程是数据采集输入、数据转换、数据分析、决策知识运算。

2.1　数据源

数据源是空间数据仓库的信息源,它包括两部分:GIS 核心数据(在 NSII 和 NSDI 中也称框架数据)和附加的特定应用数据。GIS 核心数据包括大地测量控制,数字正射影像,高程,交通,水文,行政边界,地籍信息。附加特定应用数据是为某一特定应用领域增加的应用空间数据。例如支持城市可持续发展决策支持系统的人口分布数据。

框架数据集可以通过测量或直接用 GIS 软件(如 ERDAS,ARC/INFO 等)从影像上提取获得。特定应用数据可以从应用部门或相应部门的数据中心获得。随着数据采集工具智能化水平的不断提高,框架数据的获取自动化将会很快变为现实。

在将数据输入数据库系统之前,数据必须被投影到特定的坐标系统,并进行地理编码、格式化等一系列工作。在这个过程中,会引入许多不同的异构性。如数据格式异构性、主题异构性、语义异构性、编码异构性、投影参考坐标系异构性、精度与数据处理异构性等一系列异构性[7]。

为解决这些问题,在数据转换精化之前需进行以下工作:

(1) 建立元数据;

(2) 对空间对象的命名、投影、坐标系采用统一标准;

(3) 采用统一地理编码、数据格式和精度标准;

(4) 选择空间数据仓库中使用的数据;

(5) 记录数据的时间信息。

2.2　数据库

所有空间数据仓库中的数据将由数据库统一管理。开始空间数据是存在于文件之中的,然后它被存放于关系数据库的 BOLB 字段中,一个空间数据引擎被加于其上以提供空间访问功能,它们一起构成了空间数据库,下一代用于管理空间数据的数据库将是 OODB (Object-Oriented DataBase)或 ORDB (Object-oriented and Relational DataBase)。

关系数据库访问语言的新版本-SQL3 已经被扩充为具有空间信息访问能力的语言。在不久的将来,我们可以象关系数据库中访问结构化数据一样对空间数据进行访问。

有许多成熟的数据库产品可以用于空间数据库,如微软的 SQL Server,Oracle,Informix 等。今天的许多成熟的空间数据库就是在这些已有的数据库基础上增加一个空间数据引擎构成的,如 ESRI 公司的 SDE,MapInfo 公司的 SpatialWare 等。

2.3　存储结构

空间数据仓库的大量数据管理和分布特性要求

它提供有效的信息存储策略以获得较好的性能。我们不可能在接受到请求后,再到数据库中进行数据访问,但我们可以建立一个多级存储印象(如图 3 所示)。

空间数据仓库可能会访问互联网上的任何数据,我们可以以特定时间间隔的方式来对全球数据进行挖掘。通过这种方式,我们可以提取出需用的信息,进而构造领域空间数据仓库。

2.4 空间数据市场/仓库

在空间数据仓库或数据市场中,信息是基于主题组织的,主题又可能由许多因素组成,如果将这样的一个 1NF 因素作为一个维,例如空间三维,时空四维,时空人口分布为五维。这样,空间数据仓库中的维将是高维的。

空间数据仓库使用多维技术来组织大量数据,建立立方体或超立方体数据模型。维数是由空间查询要求决定的。普通的地理查询可以归结为回答什么时候,什么地点,什么事情怎样发生? 我们可以依据空间三维、时间一维、主题多维来组织数据,也就是人们观看世界的本来方式。主题维可以视具体的不同主题定义。

维是按不同粒度、不同层次组织的,粒度直接与数据源的抽象和聚集方法相关,粒度越小,信息量越大,粒度越大,信息量越小。

在空间信息集成的处理过程中,我们可以按这种方式处理,这正是在数字地球中所说的 4.5 维或 5 维模型。事实上,应是多维的,空间三维,时间一维,属性多维。

建立空间数据仓库或数据集市的过程也就是一个空间信息融合的过程,地图综合、图形边界处理、DEM 或三维集成在这个过程中都有使用。时间集成就是按时间段对信息进行综合,属性集成就是按特定属性对信息进行综合,这种多维集成方式有利于空间数据仓库的分析工具开展分析工作。

2.5 分析结果/知识

空间数据仓库的建立目的是支持空间决策支持系统。无论我们开始进行何种工作,最后我们必须获得空间决策支持系统所需的知识。在分析过程中,我们需要用到知识库和方法库。

3 关键技术

空间数据仓库的构建是一个大而复杂的系统工程,它的实现有赖于一些关键技术的成熟,这些关键技术包括:快速计算、大容量存储、高速网络、空间数据库的无缝连接、元数据、数据挖掘、空间数据联机分析处理、服务与互操作等关键技术,详细介绍请看文献[6,12—14]。

4 建设空间数据仓库

正如 Gary Dodge 和 Tim Gorman 所说[9],数据仓库并不是一个工程,它是一个不断进化的过程,是一个基于现有成熟的数据库系统的不断完善改进的创新过程。

空间数据仓库也与此类似,为建立一个空间数据仓库,也需要一个不断进化的过程,虽然它不是一个工程所能概括的,它却需要按一个工程的方法来进行。也就是按软件工程的方法、步骤来建设。在这一过程中,存在一些特定的问题需要讨论,它本身则是一个全工程化的进程模式。

4.1 建设空间数据仓库的过程

空间数据仓库的建设逻辑如图 6 所示[15],空间数据仓库的数据流如图中箭头所示,所有这些数据是用它们的元数据进行维护的,数据的处理模块也是以这些元数据作为依据的。这些功能部件是由空间数据仓库管理系统进行管理的。

建设空间数据仓库首先要讨论的问题是分析主题,然后分析提取系统需求。根据这些系统需求就可以进行系统设计并实现系统。在建立完一个空间数据仓库之后,还应进行人员培训和系统维护升级,这一过程可概括为:明确主题、分析系统需求、系统设计、系统实现、人员培训和系统维护升级。

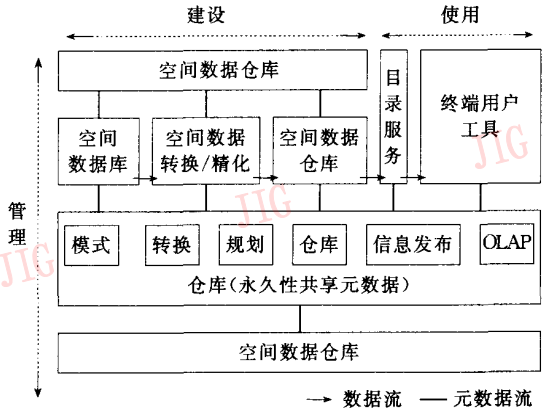

图6 空间数据仓库的信息流和环境

4.2 明确主题

在建设一个空间数据仓库过程中首先需要解决的问题是空间数据仓库的主题是什么?这个过程是与公司的高层领导进行交流,了解公司的运转业务、范围和相关的一些信息。这个过程需要解决以下问题。

(1) 空间数据仓库建立的主题,如产品销售还是交通运输管理;

(2) 相关领域,领域是否与产品、人员相关;

(3) 用户种类,例如,决策者、管理者、财会人员等;

(4) 每类人员的领导者,如总裁、部门经理等。

4.3 系统需求

第二步是了解空间数据仓库的系统需求,这对于系统的成功是非常重要的。在这个阶段,需要与用户群体进行广泛的接触。这些工作基于上一个步骤,在这一步中,需要解决以下问题。

(1) 空间数据仓库每类用户的需求是什么?

(2) 每类用户的工作流程是怎样的?

(3) 用户希望他们将来的用户终端界面是怎样的?

(4) 关于空间数据仓库,已经有些什么工作?

4.4 系统设计

第三步是设计实现空间数据仓库的硬软件。在这一步,需要涉及许多种人员。如用户、硬软件工程师、系统分析员、系统维护人员等等。这个过程中,应定义清楚以下事项:

(1) 空间数据仓库的逻辑结构;

(2) 空间数据仓库的物理结构;

(3) 空间数据仓库的模型;

(4) 已有系统的融合策略;

(5) 系统软件环境及所需设备清单;

(6) 硬件环境及所需软件清单;

(7) 系统信息流程与相应软件结构;

(8) 系统功能模块和实现方法。

4.5 系统实现

在系统设计完成后,下一步是具体实现系统,这一步最重要的是建立一个高效协作的团体队伍。这个阶段应提交以下东西。

(1) 空间数据仓库运行系统及相关文档;

(2) 功能模块及相关文档;

(3) 系统测试报告和文档。

4.6 人员培训

人员培训应在系统移交给用户之前进行。需要培训的人员分为 3 类:高层决策人员,他们只需了解系统运行的界面即可;一般使用人员,需熟练使用自己的专用界面和一些相关的维护技术;系统管理员,需要对系统进行全面培训,最好在系统建设之初就让他们参与进来。

4.7 系统升级

一个公司的发展是非常快的,随着这种飞速发展,公司的决策支持也将变得越来越复杂和频繁。因此空间数据仓库需要不断更新,增加新的功能。

所有这些步骤对于建设空间数据仓库的过程是非常重要的,但以下一些事情在建设过程中必须给予足够重视。

(1) 明确的用户群体和用户的积极参与;

(2) 决策者的支持;

(3) 高效的项目队伍;

(4) 快速稳定的步骤;

(5) 良好的体系环境。

4.8 云南省生态评价系统

为保护中国云南省的生态系统。政府决定建立一个生态评价系统,采用实时更新数据对生态环境进行综合评价。

系统由云南省的 1∶1,000,000 比例尺的政区图和 5 个相关的专题数据库构成:水土流失、土壤、人口、沙化和森林覆盖率。本系统采用了数据中心策略,数据库采用 SQL Server 7.0,空间数据引擎是 WebGIS, 如图 7 所示。

只要用户可以访问互联网,他就可以在任何地点通过 http://www.cybergis.net.cn/yn/yn_main.htm 访问系统。GUI 由 3 部分组成(图 8):地图浏览器,主题图层与结果区,用户可以在地图上任意选择一个县,而后点击“查询”按钮就可以得到结果,它们显示在结果区[12]。

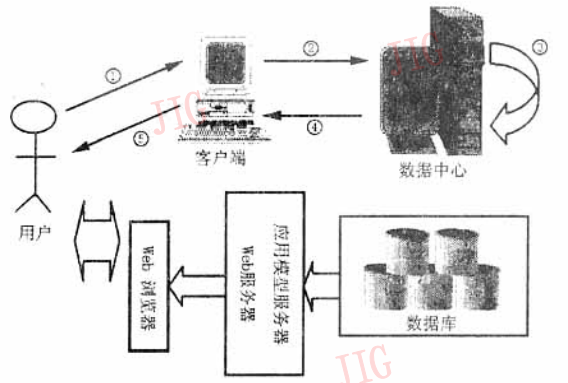

图 7 云南省生态评价系统结构图

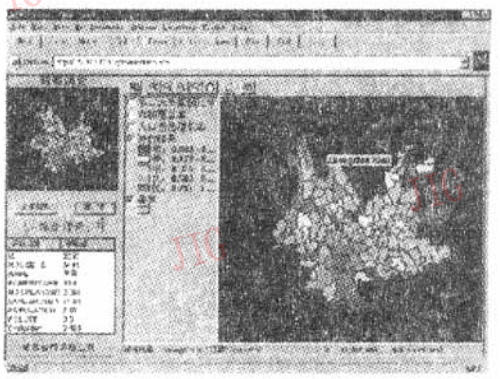

图 8 云南省生态评价系统的运行界面

5 结 论

本文讨论了空间数据仓库的环境和它的体系结构及建立策略,并介绍了如何构建一个空间数据仓库,最后介绍了一个例子。

参 考 文 献

1 Al Gore. Digital Earth:Understanding our planet in the 21st century. 1998, http://www.digtialearth. gov/speech. html

2 http://www.cybergis.net.cn/

3 http://www.digitalearth.net.cn/

4 Ralph Kahn *et al*. 1999, http://digitalearth. gsfc. nasa. gov/Scenarios199902. htm

5 徐冠华. 全社会要高度关注数字地球. 科学新闻周刊,1999,(1).

6 杨超伟,李 琦,李浩川,毛新生. 数字地球中的空间数据仓库研究. 中国图象图形学报,1999(增刊):54～59.

7 Delvin Barry. Data Warehousing:From Architecture to Implementation. Addison Wesley Longman, Inc, 1997.

8 Inmon W H. Buiding the Data Warehouse, 2nd. John Wiley. 1996.

9 Gary Dodge, Gorman T, Oracle8™ data warehousing. Wiley Computer Publishing, 1998.

10 李 琦,杨超伟,陈爱军. WebGIS 中地理空间数据库模型研究. 中国图象图形学报,1999.(待发表)

11 Yee Leung. Intelligent Spatial Decision Support Systems. Springer, 1997.

12 李浩川. 基于网络环境下的地理空间数据组织、管理及其在可持续发展决策支持系统中的应用[学位论文]. 北京大学,1998.

13 Andrej Vckovski. Special issue:Interoperability in GIS. International Journal of Geographical Information Science, 1998,12(4).

14 赵永平. 基于国家空间信息基础设施的 Metadata 研究及其共享示范体系的建立[学位论文]. 北京大学,1998.

15 Microsoft SQL Server 7.0 Data Warehousing Framework. Microsoft White Paper, 1998.

16 杨超伟,李 琦,承继成等. 遥感影像的 Web 发布研究与实现. 遥感学报,1999.(待发表)

17 杨超伟,李 琦,王京傲. 空间对象关系运算的分布式研究. 中国图象图形学报,1999,4(4):331～335.

Spatial Data Warehouse and Construction Strategy

Li Qi, Yang Chaowei

(*The CyberGIS Studio, Peking University, Beijing* 100871)

Abstract Digital Earth calls for the sharing of spatial information, spatial data warehouse (SDW) is a convenient way for managing and distributing information among public. In this paper, the society and institutional context of building spatial data warehouse is given, that is followed by the discussing of strategies used when constructing SDW and architecture of SDW. Then related construction steps of SDW are introduced in detail. Based on these theories discussed, a prototype of SDW is designed and implemented to display how to build a SDW.

Keywords Spatial data warehouse, Strategy, Architecture, Key technique, Prototype, Digital Earth