地理信息数据仓库的技术研究

杨 群 闾国年 陈钟明
(南京师范大学地理科学学院，南京 210097)

摘 要 分析了地理信息系统(GIS)的数据管理，在描述数据仓库技术的实质和特征的基础上，论述地理信息系统数据仓库的功能、特点，给出GIS数据仓库基本体系结构及基于数据仓库的GIS资源共享模式，着重叙述其关键技术。

关键词 地理信息系统(GIS) 数据管理 数据库 数据仓库

0 引 言

我国自70年代开始使用GIS技术。从最初的利用GIS系统来收集和管理空间数据和属性数据，到现在用GIS进行城市管理、交通运输、生产建设、规划、资源勘探、行政管理、旅游、军事指挥等，GIS的数据库技术迅速地发展。进入90年代，完善的数据库管理系统允许高效率地管理和更新大量的数据，使得GIS数据库能得到好的维护和不断更新，从而使用户能方便地利用最新的数据来决策和分析[1]。

1 GIS的数据管理

GIS能将地图对象与数据库的数据连接，实现地图与数据库的双向查询。GIS将数据库中的信息进行直观的可视化分析，发掘隐藏在文本数据之中的有用信息，从而为用户提供一种崭新的决策支持方式，但GIS的数据管理是一个复杂的问题，有如下特点：

(1) 地理信息系统涉及的信息媒体多，格式多变，类型复杂，信息量大，信息源多。因而数据库系统在GIS中处于至关重要的地位。

(2) 由于地理信息系统的数据量大，信息复杂，不同的分析需要从中提取最有用、最关键的信息；同时很多GIS的使用者都是非专业人员，因此GIS应是面向用户、面向主题的数据管理和操作，而且不应该让用户关注更多的技术细节。

(3) 地理信息的本质是分布式的。地理信息分布在不同的行政区域、不同的部门，不同的部门对GIS有不同的需求，因而宜于用分布式的计算机系统进行处理[1]。

(4) 为避免资源的浪费，并考虑到不可能长期保持多个数据库的维护和更新，好的GIS系统应提供计算机资源、技术资源和数据的共享功能。

2 GIS数据仓库(Data Warehouse)

2.1 数据仓库

数据仓库的提出是以关系数据库、并行处理和分布式技术的飞速发展为基础的，其目标是解决在信息技术发展中存在的拥有大量数据资源，但有用信息贫乏的问题。数据仓库概念的创始人W. H. Innmn给数据仓库作出的定义是："数据仓库就是面向主题的、集成的、稳定的、不同时间的数据集合，用以支持经营管理中的决策制定过程"。

可以认为GIS数据仓库是GIS各专用性数据库系统的数据的中央仓库，这些数据采用统一的模式进行集成。

2.2 数据仓库的实质和特征

数据仓库是传统数据库技术的一种新的发展和应用，其实质仍是计算机存储数据的系统，但它存储的数据在量上和质上都与专用性数据库有所不同。数据仓库侧重于综合分析，专用性数据库侧重于一般性数据处理。专用性数据库是数据仓库的基础。

收稿日期：1998-12-21；收到修改稿日期：1999-02-08

数据仓库除了具有传统 DBMS 的共享性、完整性、数据独立性外，还具有以下特征：

(1) 面向主题而集成。传统的数据库是面向应用设计的，数据仓库是面向主题的。主题是在较高层次将数据归类的标准，每一个主题基本对应一个宏观的分析领域[2]。

(2) 历史性和稳定性。数据仓库信息的历史性是指不同时间信息的有用性，由于历史信息的不可改变性，自然地数据仓库的信息具有稳定性。

(3) 时间属性。数据仓库数据是用作趋势分析的，需用较长的时间。因此数据仓库里的数据必须建立含有时间项的码键，以表明该数据的历史时期。由时间维和各个主题域一起可以构成多维数据。

2.3 GIS 数据仓库的特点和功能

(1) 面向主题性和集成性。GIS 数据仓库是面向主题的，它以主题为基础进行分类、加工、变换，从更高层次上进行综合利用，并遵照一个统一的地理信息的分布模型，采用一致的命名规则或编码结构，在分布式计算环境中进行数据集成，因而具有集成性。

GIS 数据仓库在功能上可满足用户管理本部门和共享其它部门数据的要求。

(2) 空间序列的方位数据。自然界是一个立体的空间，任何事物都有自己的空间位置，彼此之间有相互的空间联系，因此任何信息也都应该具有空间标志。一般的数据仓库是没有空间维数的，不能做空间分析，不能反映自然界的空间变化的趋势。进入 GIS 空间数据仓库的空间数据必须具有统一的坐标系和相同的比例尺。

(3) 时间序列的历史数据。自然界是随时间变化的，地理数据库需要随环境的变化而不断更新，在研究、分析问题时可能需要了解过去的数据，数据仓库中的数据包含了数据的时间属性，因而 GIS 能管理不同时间的数据，满足用户数据版本管理的要求。

(4) 由于数据仓库能从多个数据库中提取面向主题的信息，这些数据库可能来自不同行业，可能不全是为 GIS 建立的，比如办公自动化系统、统计系统、建筑系统等，因而，GIS 不必包括所有的功能和数据就能为用户提供进一步开发的便利环境。

(5) 基于数据仓库的 GIS 能将数据仓库中的数据以多种表现形式直观地呈现给用户，为决策人员提供面向主题的分析工具，并不要求用户是 GIS 专业开发人员或计算机专业技术人员。

3 GIS 数据仓库的体系结构及关键技术

3.1 GIS 数据仓库系统的基本体系

GIS 数据仓库系统按照功能可以分为以下几个部分：

(1) 元(Meta)数据。元数据是数据的数据，是关于数据和信息资源的描述信息。它通过对地理空间数据的内容、质量、条件和其他特征进行描述和说明，帮助人们有效地定位、评价、比较、获取和使用地理相关数据。其中，对空间数据某一特性的描述，称为一个空间元数据项。空间元数据是一个由若干复杂而简单的元数据项组成的集合。

(2) 源数据。指分布在不同的地理信息系统的应用系统之中，存储在不同的平台和不同的数据库之中的大量的地理信息，是 GIS 数据仓库的物质基础。

(3) 数据变换工具。为优化空间数据仓库的分析性能，源数据必须经过变换以最合适的方式进入数据仓库。主要的变换包括：数据的提炼、转换、空间变换。

(4) 数据仓库。源数据经过变换进入数据仓库。数据仓库用多维数据库来实现，即以多维方式来组织和显示数据。空间维和时间维是空间数据仓库反映现实世界动态变化的基础，它们的数据组织方式是整个空间数据仓库技术的关键。多维数据库的结构类似超立方体。在实际分析过程中，可以按照需要把任意一维和其他维进行组合，以多维的方式显示数据，让人们从不同的角度来认识世界。

(5) 数据仓库工具。数据仓库系统的目标是提供决策支持，它不仅需要一般的地理信息查询和分析工具，更需要功能强大的分析和挖掘工具，是数据仓库系统的重要组成部分。客户端的数据仓库工具包括查询工具、分析工具和发掘工具。查询工具主要实现对分析结果的查询，如发展趋势或运行模式，而不是对记录级数据的查询，这类查询在数据仓库中是比较少的。数据仓库的查询工具主要为用户提供可视化工具，充分利用人们的视觉能力，从多种不同的角度以各种不同的图表来表示数据，使人们能更方便、清晰地了解综合、分析和挖掘的结果，快速发现数据间的潜在关系，了解数据的复杂性和动态性。分析工具主要实现对数据仓库中的数据进行分析和综合。发掘工具负责从大量的数据中发现数据的关

系,找出可能忽略的信息,预测趋势和行为。

GIS 数据仓库系统如图 1 所示,其中:DB 是关于元数据和源数据的数据库。

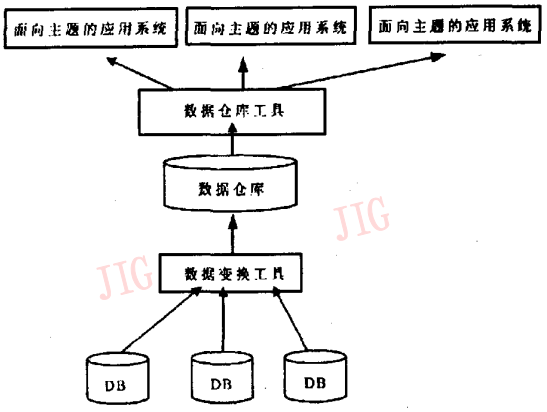

图 1

3.2　基于 GIS 数据仓库的共享模式结构

现代计算机技术和网络技术的发展和应用的需求,促进了数据仓库的发展。GIS 数据仓库提供的空间数据处理功能,帮助用户处理大量的信息,为信息时代提供了空间数据管理分析的功能。基于空间数据仓库,GIS 可以实现对数据、软件和硬件的共享。其结构可采用如图 2 所示模式。

在此模式中,系统的客户端是一台连接 Internet 的计算机。服务器端由 5 部分构成,分别是 WebGIS 服务器、数据库服务器、数据仓库服务器、地理信息系统软件库(包括各种 GIS 构件和模块)和应用模型库。其中 WebGIS 服务器提供 2 大功能:①作为不同地理信息系统软件数据的融合和转换器,基于空间数据转换标准对不同格式的数据进行转换。②作为客户请求的解译器,提供对用户请求的解译——语义转换,通过可视化的导航语言将用户引导到数据的提取、分析处理。地理信息软件库主要提供各种空间分析工具。各种地理信息系统软件所提供的空间分析模型有所差异,即使是同一种分析模型的实现方法、运行效率也不尽相同,因此通过对用户语义的解译得到系统实施流程,再采用对称式、嵌入式、动态连接库和构件方法的集成技术来共享各种地理信息系统的软件。对专业性应用系统,如海洋动力模拟、河流演变模拟等同样采用各种集成方式实现模型的共享。

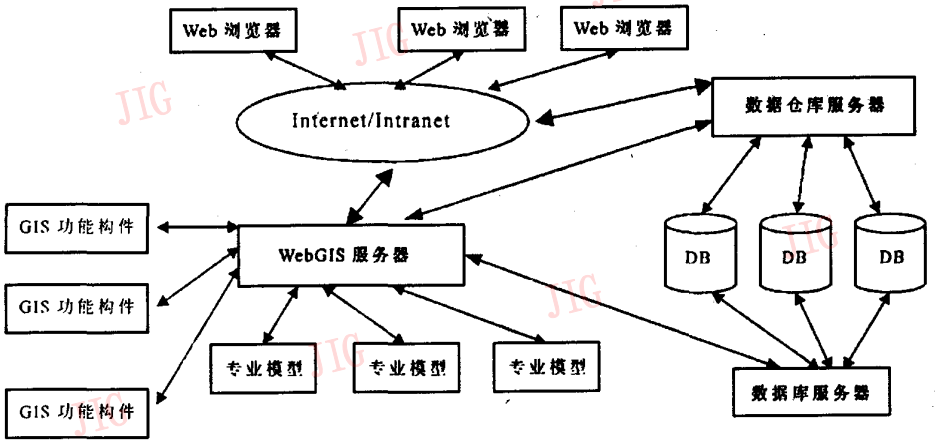

图 2

3.3　GIS 数据仓库的关键技术

目前 GIS 应用系统很多,其开发的平台非常多,如 Mapinfo、Arc/Info、ArcView 等,使用的数据库系统也不尽相同。各种系统没有统一的数据标准,但又有可能交叉,构筑 GIS 数据仓库,可以集成不同运用系统的数据,进行统一的存储与管理,以共享利用。

由于空间信息数据的特殊性,构筑 GIS 空间数据仓库需解决如下关键问题:

(1) 语义异构性

GIS 空间数据仓库的数据来自不同的应用平台,如何将它们高效集成,解决语义异构性的问题是整个空间数据仓库成功的关键之一。不同的系统,对于同一数据的意义在用法上存在不同的解释。如 ESRI 和 Intergraph 公司的产品的数据结构、存储

格式各不相同,要将它们集成在一个应用环境中,有3种方法:①将数据统一在一个平台上,采用直接转换-关联表的方式,将数据格式进行转换。该方式进行的数据转换是指在两个系统之间通过关联表直接将输入数据转换成输出数据。这种方法是记录间的转换,只对较小的转换工作量才有意义,而且由于是对单个记录进行转换,没有存储功能,不能保证转换过程中语义的正确性;②直接转换-转换器方法。转换器是一个内部模型,通过对输入数据类型及转换规则进行转换,得到指定的数据类型和值。与使用关联表相比,这种方法具有更详细的语义转换功能,也具有一定的存储功能。上述两种方法简单,但通用性较差,移植起来比较困难,效率也比较低;③基于统一的数据标准进行的转换,它仅仅是两个系统之间达成协议,两个系统之间都有一个转换模型,为不同用户、不同系统之间在空间数据仓库环境下空间信息的表示、查询、处理、分析、共享、管理和传输方法提供标准。这种方法具有极强的通用性和可移植性,但是通用标准的建立是一个极其艰苦的过程,需要较长时间的努力。目前较为适用的方法是空间数据仓库提供数据的变换工具,即将源数据经过变换,以最合适的方式进入空间数据仓库。这种数据的变换包括:数据的提炼、转换、空间变换等。数据提炼是指数据项的重构,删去不需要的运行信息,字段值的编码与译码,补充遗漏的信息,检查数据的完整性和相容性;数据的转换主要是指统一数据编码和结构,给数据加上时间标志,根据需要对数据进行各种运算等;空间变换主要是指空间坐标与比例尺的统一,赋予一般数据空间属性。数据转换为空间数据库及空间数据仓库之间架起了一座桥梁,使得源数据得到了统一和增值,最大限度地满足了空间数据仓库高层决策分析的需要。

(2) 数据库系统

空间数据仓库的物理表现为一个多维数据库。空间维数据库是整个数据库的关键。由于空间数据量无法统计及动态增长等非结构化特点,目前的大多数GIS产品,空间数据存储和管理基于文件系统,而空间数据的描述信息即属性数据由关系数据库进行管理,两者之间用关键字或指针联系。空间数据和属性数据的分离管理造成:数据完整性和一致性不能保证;缺乏数据动态增长能力和数据优化管理;系统查询能力和分析效率低下;数据共享和并行处理无保证[3,4]。空间数据仓库不仅管理大量数据,对超大数据库进行动态管理,还要实现多进程、多线程、内存缓冲、快速索引、数据完整型和一致性保证、并发控制、安全和恢复机制及分布式处理机制,这些都要求空间数据与属性数据统一在一个数据库管理系统下。数据库系统必须引入新的技术来满足对空间数据管理的要求。管理GIS数据的数据库系统除了具有普通数据库所具有的功能外,还应具有如下特征:①支持空间数据操作;②查询语言具有可扩充性;③空间数据的有效存储及组织[5]。目前,主要的做法有3种,①数据库管理系统与空间操作的实现模块分开。空间操作由GIS操作模块实现,数据的查询和存取采用双库结构,由通用的DBMS和空间数据管理软件包实现[6]。②以当前的关系数据库技术为基础,按要求进行扩充和完善,用统一的数据库管理系统管理空间数据库和属性数据库,而且支持部分GIS空间操作,如Oracle公司的Oracle 7 Spatial Data Opition是运行在内核上管理空间数据的产品,它通过引入新的数据类型和动态分割技术,完成了属性数据和空间数据的一体化管理[3]。其他如:Informix公司的Data Blade插件,ESRI公司的SDE、MapInfo公司的MapInfo MapXtreme和SpatialWare;③对现有数据库技术做根本性的改造,引入面向对象技术,建立对象关系数据库或纯对象数据库,将对象与其低层表示完全分离,空间属性和非空间属性地位平等,对象的结构与行为封装大大方便了用户定义操作的实现,如最近IBM推出的DB2、Oracle的Oracle 8等。

(3) 数据挖掘

数据挖掘是数据仓库最重要的应用之一。数据挖掘是从大型数据库或数据仓库中发现并提取隐藏在其中的模式和关系的过程,目的是帮助分析人员寻找数据间潜在的关联,发现被忽略的要素。数据挖掘利用数据挖掘工具在数据中查找模型,这个搜寻过程可以由系统自动执行,自底向上搜寻原始事实以发现它们之间的某种联系,也可以加入用户交互过程,由分析人员主动发问,从上到下地找寻以验证假定的正确性。

知识发现处理过程描述了保证获得有意义结果所必须采取的步骤,是成功地实现数据挖掘的基础。GIS隐含的大量知识需要综合使用多种方法进行提取。这些方法包括:归纳与演绎、统计方法、模糊方法、数学公式、空间分析方法等。归纳与演绎是从数据库中发现知识的基本方法,无论是空间数据还是属性数据,在进行抽象和概括时都可用此方法,如从图形数据库中,可以方便地获得关于某一类对象的

位置、形状、大小及结构等几何特征,通过推演和归纳即可得出关于该类地物对象(诸如街区、麦地、果园、湖泊等)的一般性(或规律性)的几何信息。统计推理用于 GIS 属性和空间数据的分类。模糊方法则可用于模糊的空间知识的获取,如从姓名、性别、年龄、文化程度、年收入、地址可以推导出某个居委会(或单位)各年龄段的人所占比例。利用数学公式可以从一组存放在数据库的空间数据得出其空间关系,如:从存放点坐标、多边形信息的数据库中运用数学公式可推导出多边形相邻两边的角度。上述方法得出的结论往往还需结合空间分析的方法,空间分析可以从空间关系中发现知识。从 GIS 的图形和属性数据库中可以发现对象与对象之间的相连、相邻与共生(如湖与湖中岛的关系)关系的知识;将 GIS 的图形数据与属性数据对应起来,可以发现对象的几何位置与属性的对应关系。其次,利用空间分析方法,可以找到属性形成与空间分布的对应关系,已知某一对象的属性则可知道其相应的空间分布;若已知其空间分布位置,根据其对应关系亦可知其属性。神经元网络、决策树方法等也可以用来从 GIS 中提取知识。

4　结束语

90 年代,GIS 已走向产业化。数据库技术的发展,GIS 应用的发展,促进了 GIS 数据仓库的发展。本文针对 GIS 中数据管理的特点,分析了 GIS 数据仓库的特点、功能及 GIS 数据仓库系统的 5 个组成部分,即元数据、源数据、数据变换工具、数据仓库和数据仓库工具,重点分析了构建 GIS 数据仓库的关键技术,并给出了一个基于 GIS 数据仓库的 GIS 资源共享模式。GIS 数据仓库的建立,是一项有意义的工作,但也是一个艰苦而漫长的过程,其中有许多问题要考虑和研究,如数据仓库要提供数据的变换工具,与元数据和源数据的组织管理、功能的实现有直接的关系,但目前国内关于元数据的规范、标准、理论和实现方法手段的研究尚不够[7];数据库技术,特别是基于面向对象技术的 DBMS 中面向对象模型和面向对象数据库在理论上及实现中还有很多问题。数据仓库的数据模型的设计和数据源的组织和获取、将各已有的专用性 GIS 数据库数据集成到数据仓库中数据标准的制定等方面也都需要人们付出艰辛的劳动,做有益的探索,并付诸实践。

参 考 文 献

1　陈子坦. 企业化地理信息系统的应用. 软件世界,1996,(8).

2　王　珊,刘　方. 创建数据仓库的方法、模型与步骤. 计算机世界报,1996.7.15.

3　何心远,邱名卿. 空间数据的表示与动态分割技术. 计算机系统应用,1997,(6).

4　Emmanuel Stefanakis, Timos Sellis. Enhancing operations with spatial access methods in a database management system for GIS. Cartography and Geographic Information Systems, 1998, (1).

5　田增平,周傲英,施伯乐. 地理信息系统中的数据库技术. 计算机科学,1995,(6).

6　Medeiros C B. Geographic information system. Tutorial, 20th VLDB, 1994.

7　李　军等. 地球科学数据的元数据研究. 地理研究,1996,(1).

杨　群　1971 年生,本科毕业于北方交通大学计算机系,现为南京师范大学地图与地理信息系统专业研究生。

闾国年　1962 年生,地理信息系统博士,南京师范大学地理系教授。主要从事地理信息科学基础理论研究以及地理信息系统软件应用。著作有《中国自然灾害学》等。

陈钟朝　1969 年生,地理信息系统硕士,南京师范大学地理系讲师。主要从事地理信息软件的开发、研究。

Studies on the Technology of Geographical Information Data Warehouse

Yang Qun, Lv Guonian and Chen Zhongming

(*College of Geographical Science, Nanjing Normal University, Nanjing* 210097)

Abstract This paper makes an analysis on the data management of Geographical Information System (GIS). Besides of generally describing the nature and features of Data Warehouse, it explores the functionalities and features of Data Warehouse specified in Geographical Information System, and also delivers the infrastructure of Geographical Information System and addresses the pattern of shared resources based on Geographical Information System data warehouse, expecially discussing the key technologies throughout the process of it's implementation.

Keywords Geographical Information System (GIS), Data management, Database, Data warehouse