

基于四叉树的海量空间数据无缝组织研究

张莉¹, 唐立文²

(1. 中广电广播电影电视设计研究院, 北京 100045;

2. 装备指挥技术学院 试验指挥系, 北京 101416)

摘要:随着数字地球相关技术的发展,海量数据逐渐成为在数字地球研究中的一个最根本的特征,主要体现在数据种类多、数据量大,访问速度不容易提高。由于原始数据的表达与组织不同,也导致了数据出现不同的缝隙。介绍了数字地球中海量数据的特点,空间数据缝隙的产生,在此基础上,探讨了基于全球四叉树的组织方式,阐述了无缝拼接的方法,并对线和多边形数据进行无缝组织和处理,以使地理数据库中的空间数据在逻辑上达到无缝状态,成为一个整体。

关键词:数字地球;海量数据;无缝组织;四叉树

中图分类号:TP311

文献标识码:A

文章编号:1673-629X(2011)01-0077-04

Study of Organizing Seamless Great Capacity Spatial Data Based on Quadtree

ZHANG Li¹, TANG Li-wen²

(1. Radio, Film and Television Design and Research Institute, Beijing 100045, China;

2. Department of Testing and Command, Academy of Equipment Command & Technology, Beijing 101416, China)

Abstract: Great capacity data has become one of the important characteristics to the study of Digital Earth (DE) with the development of the corresponding techniques. It has all kinds of data, great capacity, slow access speed. Because of the expression and organization of the data, seam accrues. The paper introduced the great capacity characteristics in DE, the birth of seam. And in this way, discussed the structure of global quadtree, studied the methods of seamless joining, organized and handled the spatial data of line and polygon. Thus, the study makes the spatial database become seamless and unitary in logical.

Key words: digital earth; great capacity data; seamless organization; quadtree

0 引言

“数字地球”概念的提出,引起了国际政治、军事和学术界的高度关注,并已成为许多国家竞相发展的国家战略和军队信息化建设的制高点。随着数字地球相关技术的发展,人们逐渐对数字地球的整体轮廓、发展趋势、技术定位等有了较为明确的认识^[1,2]。然而,由于实际的地图数字化导致相邻地图图幅边界的几何位置与属性出现不吻合,割裂了地理实体的完整性,使得空间数据存在着逻辑缝隙和物理缝隙。由此产生了地理空间数据的拼接问题,而海量数据拼接的好坏则直接影响着数字地球技术的应用。文中从海量数据的组织出发,以四叉树的形式对空间海量数据进行分割存储,同时在此基础上提出了对空间数据的无缝组织

与实现方案。

1 海量数据的概念

数字地球的提出导致了海量数据成为GIS界广泛接受的常用术语。随着数字地球的研究、地理信息系统应用领域的拓广、各种问题研究的动态性,以及数字地图数据比例尺的增大,使得整个空间数据量不断增加,人们所面临的需要处理的数据量也越来越大,而这些空间数据目前已达到GB、TB乃至PB数量级以上(1PB=1000TB=1000000GB)。一般认为,海量数据指的是采用不同的存储介质在地理空间数据库中集中或分布存储,至少达GB数量级的空间数据(包括图形、图像数据等)^[3,4]。因此,如何有效对海量数据进行无缝组织和组织,就成为了业界内研究的重点所在。

2 空间数据缝隙的产生

实际的地图数字化是分幅进行的,数字化后的存储和处理也采用分幅的方式,这使得相邻图幅边界的

收稿日期:2010-05-13;修回日期:2010-08-03

基金项目:财政部重大专项课题(40602001)

作者简介:张莉(1974-),女,工程师,研究方向为地图综合、信息融合、广播电视工程等。

几何位置与属性难免会出现不吻合,因而割裂了地理实体的完整性,使得同一地理实体穿越多个图幅。这样,就导致空间数据存在着逻辑缝隙和物理缝隙,由此产生了地理空间数据的拼接问题,而拼接的好坏直接影响着数字地图的质量和使用的。

空间数据缝隙的产生的主要原因包括数据的获取、表示与处理等^[5]。数据获取方面,由于地图是按图幅的形式划分,人为地将连续的空间地物划分成若干相互关联的子地物,使同一地物的位置和属性分别落在不同的图幅内。再加上纸张伸缩、图数转换误差以及人工操作失误等各种因素,两个相邻图幅边界处就会出现逻辑或物理裂隙。

数据表示和组织方面^[5],以经纬度表示的空间地理坐标系完全可以完整地覆盖整个球面;而使用公里网坐标来描述球面则必然会遇到跨越投影带的裂缝问题。空间数据库组织和管理上,传统的文件型空间数据库、文件与关系混合型空间数据库,均是按照图幅或者一定的区域范围以文件的形式来进行组织与存储空间的几何数据,属性数据则贮存在关系数据库中,因此,不同的图幅或区域之间存在缝隙。全关系型和对象关系型则将属性与几何数据作为一个整体来表示,可以基本解决缝隙问题,但是速度不能达到系统的要求。

数据处理方面^[5],处理过程中的容差或处理过程的不一致也会导致缝隙产生。由于在数字化或节点聚合过程中两点间的距离在容差允许的范围内,因而会导致某些系统将其认为是同一个点,而当数据转到其他一些系统,或者当数据精度要求更高,或者进行空间数据分析处理时,都会导致缝隙的出现。

因此,缝隙产生主要包括:

- (1) 在不同坐标系之间进行数字化相邻地图;
- (2) 空间对象(如公路、河流等)从一个区域跨越到地图的另一个地图区域时,可能导致不连续或不规则出现;
- (3) 数字化的错误也会导致图形的扭曲;
- (4) 相邻图幅之间比例尺的不一致;
- (5) 用于进行矢量化的原始图幅边界处本身就有误差。

3 空间数据的无缝组织

一般数据库系统在处理空间数据时存在着如下不足:

- (1) 完整空间数据在计算机中分块或分段式的分散存储,导致了空间地物在整体的一致性维护上复杂性增加;
- (2) 数据库的检索和存取的效率还有待优化和提

高;

(3) 空间数据查询需要跨越多个图幅,这就要求系统能依靠地物标识,在多个相邻图幅之间进行空间和逻辑上的匹配和关联。

基于上述的讨论,文中研究采用逻辑四叉树的方法对空间数据进行存储。处理的方法是将分散信息处理成连续无缝的数据,并建立相应的四叉树索引机制以及数据模型进行空间数据的无缝管理。

3.1 全球四叉树结构

四叉树结构,就是把 $2m \times 2m$ 个象元组成的阵列按照树的结构进行划分,树的高度为 m 级。树的四个节点分支分别代表了东北、西北、东南、西南四个象限。在四个分支中或者是树叶或者是树叉,树叶表示已划分完成,而树叉则表示可继续划分,最终形成了多层次的数据结构^[6]。

全球海量空间数据,若是存储到空间数据库中,则由于时时刻刻都要对数据库进行检索,因而无法满足空间信息的快速变化需要。因此,本文研究采用全球四叉树的形式存储空间数据:即针对全球范围进行逐层划分,先将全球划分为东西半球两个节点,在每一个半球再采用四叉树的形式向下进行分割,将地图划分为级别、大小不同的正方形,每一方形通过它的上一级较大的方形四等分而产生,如此逐级划分到最终的叶子节点。这种结构是将空间各种不同精度的数据经过分割处理,存储到不同的四叉树节点中,而读取信息时,再直接定位到各个节点,其结构见图1。

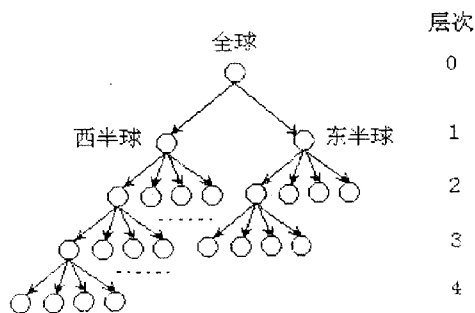


图1 基于四叉树的海量空间数据组织

3.2 空间数据的无缝组织结构

与地图图幅不同,无缝地理数据库中没有图幅的概念,地物入库时,必须进行物理和逻辑接边(即图形数据和属性数据的合并、统一),同一地物在数据库中合成为一个完整的实体对象,该对象记录地物的完整空间信息和属性数据。同一地物在相邻图幅(或四叉树结点)拼接时,在各自的图幅(或四叉树结点)中,系统分配给它一个唯一的标识号,按照事先建立的分级分类体系,用户又分配给它一个用户标识(ID),这样在进行物理和逻辑接边时,两个图幅(或四叉树结点)

的系统标识号进行更改形成整个数据库中系统标识号唯一的完整无缝地物。

由于实际的空间数据一般均以分幅的形式获得,因此,空间数据的无缝组织就涉及到相邻图幅的拼接处理,在系统数据库中则表现为相邻四叉树结点数据的拼接。具体的功能主要有两方面:一方面是逻辑接边,建立相邻图幅(或四叉树结点)之间的空间数据在逻辑上的链接;另一方面是物理接边,即将不同图幅或区域(或四叉树结点)的空间数据合并成更大的区域,再进行分割存储到四叉树结点。具体采用哪种接边方式,则视所采用的空间无缝数据模型而定。其拼接方法见下面一节。

同时,在所研究的地理数据库内部采用经纬度的地理坐标系,这样,在数据的表示和组织方面,可以达到数据的连续无缝。而空间数据在入库之后,对于用户来说,并不需要了解空间数据是怎样分割的、也不需要知道数据的真实物理地址,整个海量空间数据对用户来说应该总是连续的、无缝的数据集,即实现了对数字地图进行逻辑和物理无缝的拼接。

3.3 空间数据的无缝拼接方法

在数据拼接时,点矢量保持不变,线矢量应根据其位置以及前面讨论的情况判断采用何种方法。多边形矢量的边界由线矢量组成,其主要方法可参考线矢量。下面针对图2中的线矢量介绍几种拼接方法^[7-9]。

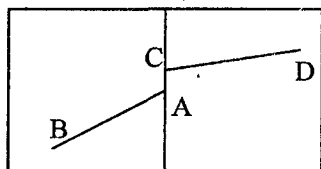


图2 线矢量拼接图

(1)平均法。就是将图幅边界处的待接点坐标均值作为拼接后点的坐标,这种方法操作起来比较简单易行,适用于拼接误差在允许的精度范围内的各种线矢量的拼接处理,而且容易实现全自动批量处理。采用这种方法拼接时,取AC的中点作为拼接点即可。

(2)强制法。强制法就是把某条待拼接边线的待接点强行接到另一条待拼接边线的待接点上,这种方法主要是用于能明显知道一条待接边比另一条待接边更加准确、可靠,适合交互式接边处理情况。即把A或C点作为拼接后的点进行存储。

(3)优化法。无论采用上述哪种方法,在边线处都会产生一个拐点,如果在数字化中进行提取要素时没有错误或错误很少,则此拐角会很小。但是如果待接边明显是一条直线边的话(如直线道路、规则建筑物等),很小的拐角也会影响整个图形的效果,因此,拼接不仅仅要在图幅边线上将两个点连在一起,而且

还要考虑接边后的两段线既要共点,也要共线,此法称为优化法。优化法就是找出A、C两个待接点以及它们的前序点B、D,用连接B、D与边界线的交点坐标代替原待接点坐标。

以上各个拼接方法各有优略,在实际情况中,还要根据图形整体和接边位置等具体情况而定。

多边形数据的拼接方式与线不同,其拼接是在边界处进行的^[10,11]。在存储数据之前,先要得到当前需要入库数据的最大经纬度范围,并存储每一个多边形的经纬度范围。由于存储到数据库时是一个矢量处理一次,因此,可根据其存储的经纬度范围判断是否要与当前已入库数据是否相交。若是,则找出当前多边形矢量与边界的交点(可能只对经度方向或纬度方向判断),然后找到对应的四叉树节点,搜索其可拼接多边形;否则的话,直接存储入库。

当处理的多边形数据位于当前处理的数据边界处时,虽然在边界处多边形有可能会跨越多个四叉树节点,但是拼接过程仍然是在单一的四叉树节点内进行,详细情况如下。

首先,求出两个待拼接多边形在某一四叉树节点内待拼接处的所有交点,并按照距四叉树最左端或最下端点的距离标准进行排序。找出其距离最远和最近的两个交点P0、P1,然后,根据多边形的走向判断需要从哪一个多边形开始搜索,设为P0。当搜索到交点P1时,转向另一个多边形,如此一直搜索到开始距离最远的那个交点P0,此时得到的多边形是其拼接后的外环。如果交点有两个以上,则说明原始空间数据有岛。先将P0、P1设置为不参与搜索标志,再从第二个距离最远的交点开始,继续上述搜索过程。如果原来两个多边形本身就含岛,则不需拼接,直接存储。

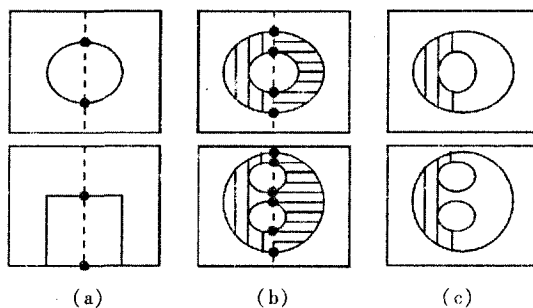


图3 多边形拼接示意图

图3中,(a)是只有两个拼接点的情况,两待拼接多边形拼接后成为一个多边形;(b)是有超过两个交点的情况,此种情况下两个待拼接多边形拼接后的多边形包含岛的信息;(c)中则是原始待拼接多边形含岛信息,而计算的拼接点也超过两个,此时拼接后的多边形既包含拼接时产生的岛,也包括两待拼接多边形原来所含的岛。

3.4 空间数据的无缝实现

矢量拼接过程中,点矢量的信息基本不变;对于线矢量,只对头尾点判断是否需要接边处理;而对于多边形,则需要对每一个点进行比较,若超过一个点与边界相交,则判断为应与其相邻边界要素拼接。而接边时,需考虑四个方向(东南西北),每个方向向上下或左右延伸若干四叉树节点。同时还要记录其与哪个边邻接,然后再判断接边时需要处理一个边方向的三个节点还是两个边方向的五个节点。

拼接处理时,需要判断两个矢量在相邻处是否需要连接,其前提之一是它们均具有相同的要素特征码,如属性码、国标码等;二是接边处两要素在坐标位置上要符合所给定的限差。在确定两个矢量是同一矢量时,将二者的所有信息统一,包括ID、属性等内容。图4是线矢量拼接的示意图。在图4中,竖线表示相邻两幅图的边界处,其中“1”点位置相邻两矢量由于国标码不一致,因此二者虽然位置上满足限差要求,也不进行拼接操作。而在“2”点位置,只有国标码为21012的洋河那条矢量才具有拼接的可能,所以根据前多边形的算法将二者合并为一个矢量。合并后,再根据其在四叉树中的位置,分割存储到相应的四叉树结点,各个四叉树节点存储其相邻边界的信息。

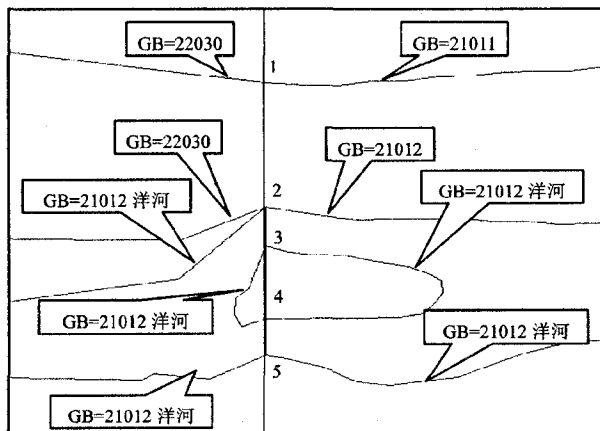


图4 矢量拼接效果图

利用VC++^[12],文中对某一含岛多边形进行拼接研究,其拼接前后的状态见图5。左图为未拼接情况,在两多边形相邻处有一段标识。而右图则是沿着该线段在相应四个四叉树节点内进行拼接处理后的结果示意图。设拼接时按从下往上的顺序在四叉树节点进行拼接(若横向,则从左向右),首先在最下面的四叉树节点内进行拼接,拼接完成后将多边形信息存储到当前四叉树节点内。此时两待拼接多边形的属性信息也要进行合并,以保证最终多边形矢量的一致性。由于在每一个四叉树节点内拼接点都是两个,因此拼接后成为一个完整的多边形矢量,除了保留原来多边形1所含岛信息外,不含其它岛信息。

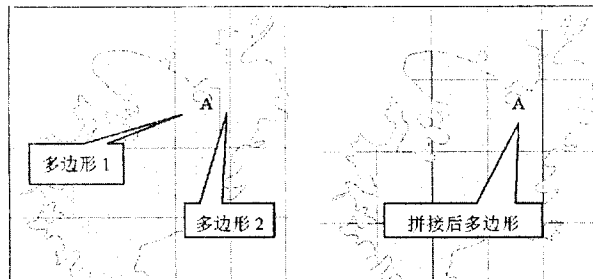


图5 多边形拼接实例

4 结束语

“数字地球”研究的发展,使得海量空间数据日益称为人们研究中的重中之重。由于海量数据的表达、组织,导致了空间数据缝隙的产生。文中探讨了海量数据的内容,空间数据缝隙的产生,并在此基础上,研究以四叉树的逻辑组织方式,对海量空间数据进行无缝组织和处理,以使地理数据库中的空间数据达到无缝状态,整个空间成为一个整体。

参考文献:

- [1] Dodge M, Moberly M, Turner M. Geographical Visualization: Concepts, Tools and Application[M]. Hoboken: John Wiley & Sons, 2008: 1-10.
- [2] Gore A. Digital Earth: Understanding our Earth in the 21st Century[EB/OL]. [2008-11-06]. http://www.isde5.org/al_gore_speech.htm.
- [3] 李爱勤, 龚健雅, 李德仁. 大型GIS地理数据库的无缝组织[J]. 武汉测绘科技大学学报, 1998, 23(1): 57-61.
- [4] 马荣华. 大型GIS海量数据的无缝组织初步研究[J]. 遥感信息, 2003(3): 44-48.
- [5] 朱欣焰, 张建超, 李德仁, 等. 无缝空间数据库的概念、实现与问题研究[J]. 武汉大学学报: 信息科学版, 2002, 27(4): 382-386.
- [6] 唐立文, 廖学军, 汪荣峰. 基于四叉树的海量空间数据模型研究[J]. 装备指挥技术学院学报, 2007, 18(2): 70-74.
- [7] 李德仁, 龚健雅, 张桥平. 论地图数据合并技术[J]. 测绘科学, 2004, 29(1): 1-5.
- [8] 吴立新, 史文中. 地理信息系统原理与算法[M]. 北京: 科学出版社, 2003.
- [9] 文玉树, 郑普红. 数字地图无缝连接及漫游的基本方法[J]. 四川测绘, 2003, 26(3): 104-106.
- [10] 唐立文, 汪荣峰, 廖学军. 基于四叉树的海量空间矢量多边形处理技术[J]. 装备指挥技术学院学报, 2007, 18(3): 104-108.
- [11] Zalik B. Two Efficient Algorithms for Determining Intersection Points Between Simple Polygons[J]. Computer & Geoscience, 2000, 26(2): 137-151.
- [12] 王华, 叶爱亮, 祁立学, 等. Visual C++ 6.0 编程实例与技巧[M]. 北京: 机械工业出版社, 1999.