文章编号:1671-8860(2010)04-0379-05　　**文献标志码**:A

浅论自发地理信息的数据管理

李德仁[1]　钱新林[1]

(1　武汉大学测绘遥感信息工程国家重点实验室,武汉市珞喻路 129 号,430079)

摘　要:分析了自发地理信息(volunteered geographic information,VGI)数据的来源、分类、特点与管理要求,探讨了 VGI 数据清理与质量控制,研究了以高效处理绘图查询与数据更新为目标的 VGI 图形数据管理问题,提出了动态线综合二叉树与缩放四叉树的设计思想,以解决 VGI 图形数据管理中的难点问题。

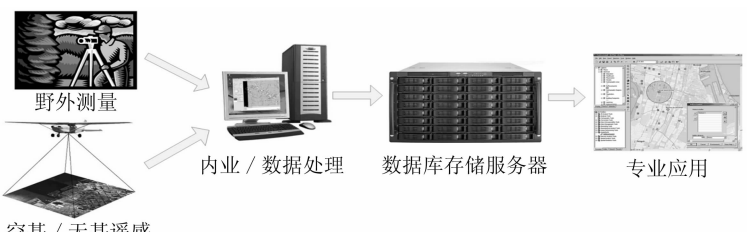
关键词:自发地理信息;数据管理;绘图查询;动态线综合二叉树;缩放四叉树

中图法分类号:P208

自发地理信息是随着地球信息科学在新地理信息时代[1]中发展出现的新概念[2],认为地理信息的创建、维护、应用可由大众完成。狭义的自发地理信息是由大量非专业用户利用 3S 技术自发创建的地理信息;广义的自发地理信息是与狭义的自发地理信息相关的概念、模式、方法和技术。

自发地理信息产生的技术条件是新一代互联网和无线网技术的发展。云计算技术的提出,使 Web Service 平台可方便地实现与包括智能手机在内的传感器网相连,Web 2.0 的标注和上传功能使大众用户成为义务的信息提供者[2]。自发地理信息产生的社会条件是公众存在着对地理信息传播与共享的需求,同时,地理信息相关知识与技能逐渐被公众认识、了解、掌握,成为社会常识。

在线 VGI 应用系统的基本模式是以高分辨率遥感影像为底图,用户判读地物、创建矢量化的几何对象并添加属性资料,积累形成开放共享的地理信息数据库(图 1)。目前,VGI 应用有 OpenStreetMap 和 Google Map Maker 等。VGI 数据的意义在于补充地理框架数据的不足,提供丰富的细节和准实时更新;VGI 数据是协作劳动和集体智慧的产物,它集中了普通大众的地理空



(a) 专业地理数据生产流程与应用模式

(b) 专业地理信息数据创建流程与应用

图 1　专业测绘数据与 VGI 数据的生产流程

Fig. 1　Productions of Survey Data and VGI Data

收稿日期:2010-01-20。

项目来源:国家 863 计划资助项目(2009AA121404);国家自然科学基金资助项目(10978003);武汉大学测绘遥感信息工程国家重点实验室专项科研经费资助项目;IPV6 重大应用示范资助项目(126)。

间知识并为传播与共享创建了基础平台。

目前,学术界对VGI的研究集中在其对人类社会以及地理信息科学的意义[2]、应用实例与应用方法的探讨[3]、VGI数据的质量与可靠性[4]、数据的处理与提炼[5-6]等方面,在VGI数据管理、存储与索引结构等方面的研究较为缺乏。本文从分析VGI数据的特点以及数据的应用需求出发,讨论了VGI数据管理的方案。

1 自发地理信息的数据特点与应用要求

VGI数据来源多样,包括GPS终端记录的兴趣点与轨迹,带GPS的智能手机上传的具有时空位置的图像、视频和语音记录,地物描述信息,用户勾绘的点、线、面等几何对象。VGI数据根据其性质可分为以下两类:① 兴趣点、轨迹、几何对象属于图形信息;② 描述、地名、多媒体属于属性信息。相比于专业测绘数据,VGI数据中图形信息有以下4个特点:① 数据质量的不可预测,数据可能存在偏差、重复、错误;② 几何对象复杂,代表大范围自然地物的几何对象随着不断的编辑会变得更加详细且复杂,可包含10万数量级的点,类似于GSHHS数据;③ 几何对象分布不均匀,数据创建是自发的、无规范的,因而数据条目分布不均匀且疏密不一致;④ 更新连续发生,几何对象随时会被编辑,数据更新操作频繁。

VGI数据相对专业数据在较多方面具有劣势,但其现势性强、成本低。随着3G、4G乃至5G无线通信技术的发展,当无线上网达到10 MB/s~100 MB/s后,上亿个手机将成为自发地理信息的重要来源。

VGI数据与专业测绘数据的特点差异使传统数据管理系统难以处理VGI数据,其数据特点和应用需求需进行VGI数据管理。

非专业性是VGI数据的基本特征,初始VGI的数据质量难以保证。虽然VGI通过用户协作以改善其数据质量,但为了减少对人工操作的依赖,VGI数据管理系统要求具备数据清理与质量控制的功能。

绘图是地理信息表达与可视化的重要步骤。在VGI模式下用户浏览、编辑几何对象,客户端需要实时绘制图形而不能采用预渲染成图像的方法,因此,与绘图相关的查询是VGI数据管理中最常用的查询,其处理性能是衡量数据管理效率的关键指标。VGI应用将传统的数据"查询-浏览"模式转变为"创建-浏览-更新"的循环,需要灵活的数据结构以表达和存储高度动态化的地理数据。

综上所述,VGI数据管理需要解决数据清理与质量控制、绘图相关查询的高性能处理以及频繁更新等问题。

2 自发地理信息的数据清理与质量控制

由于VGI数据具有自发性、无序性、非规范性等特点,VGI数据必须经过数据清理和质量检查以保证数据的形式有效和内容合法。质量检查包括数据的有效性验证、运用约束条件对数据进行纠正等。数据清理包括恶意内容的发现与清除、重复数据的检测与合并、涉及秘密与隐私等内容的处理。

VGI数据有效性包括几何与属性信息的有效以及数据约束条件的满足。充分利用几何约束条件对数据进行纠正才能提高质量,如建筑物拐角的直角化纠正与圆弧轮廓的规范等。

处理垃圾信息与恶意内容是进行数据质量控制的关键。除了通过上传信息的实名注册登记外,还须采用自动化鉴别与社区审核相结合的方式来鉴别恶意内容。自动化鉴别根据关键词等文本模式排查非法信息、广告等。社区审核是将用户依据其贡献与水平进行分级,积极提供高质量内容的熟练用户能赢得较高级别,成为社区管理员并负责审核其他用户创建、编辑的数据。恶意删除、污染数据等非法操作可通过分析和挖掘用户行为模式进行初步判定,经人工确认后处理,必要时终止此类用户的编辑权限。

数据重复和冗余是Web 2.0产生的数据存在的普遍问题,空间数据中对重复和冗余内容的检测已有较多的方法,如将数据记录映射到高维空间并进行聚簇,利用其在特征空间的邻近性发现重复记录。

秘密与隐私数据的处理也是VGI数据管理中的重要问题。为保障公共安全和国家利益,有些信息必须严格保密,不得随意传播,更不允许公开,涉及秘密的信息严禁创建。涉及个人隐私的数据,例如房屋的住户等,不应当作为VGI数据的传播和共享,隐私类型数据的判定一般较为困难,需人工参与。

3 VGI图形数据的管理与组织

以遥感影像为底图由用户勾绘的几何对象和

添加的描述信息所形成的 VGI 数据是本文讨论的管理对象,大几何对象的表达存储和分布不均匀的海量几何对象集的组织管理是讨论重点。

已有的研究表明,线综合二叉树(BLG-tree)[7]和条带树(Strip Tree)[8]适用于大型复杂线对象的多比例尺表达,但其截窗查询效率较低且为静态结构。本文提出的动态线综合二叉树是线综合二叉树的扩展,能较好地执行截窗查询与简化查询,支持动态更新,适合表达和存储大几何对象。

目前的地理数据库能够利用空间索引高性能地处理海量几何数据集的窗口查询[9],但缺乏对查询结果集进行综合与简化等细粒度控制。本文提出了缩放四叉树结构建立空间索引,采用顶点直方图选择算法控制查询结果集并生成绘图查询的实化视图金字塔[10]。

动态线综合二叉树与缩放四叉树形成了 VGI 数据管理的框架,前者用于单个几何对象的表达与存储,处理截窗查询与简化查询;后者用于几何对象集的索引与组织,处理绘图查询。

3.1　VGI 数据中大几何对象的表达与更新

动态线综合二叉树(dynamic BLG-tree)用于表达大几何对象,其性质如下。

1) 平衡二叉树,一棵树代表一条多义线 L,子树代表子线 L'。

2) 中间结点表示线内顶点,存储点坐标、误差值、顶点编号、子树最小外接矩形等数据项,该顶点将线 L 分为两条子线,对应于该结点的左右子树。叶结点表示相邻顶点构成的线段。除叶结点,所有结点有两个子结点。

3) 中间结点 n_i 所表示的顶点 v_i 是其子树所对应子线 L' 中最重要的顶点,DP 算法中即为离多义线的首尾顶点所构成线段最远的顶点。

4) 中间结点存储的顶点编号采用差值,如子线 L 其起始顶点为 v_m,终止顶点为 v_n,在代表 v_i 的结点内其顶点编号即为 i-m。

5) 顶点的插入、移动、删除会使性质 3)在特定子树中不再成立,此子树为无效子树,需进行重建。

动态线综合二叉树能处理简化查询、截窗查询与同时带有两者条件的截窗简化查询。动态线综合二叉树的"滞后更新"算法,为在"即时更新"与"代价最低"这一对矛盾的目标中达到平衡,为了摊薄更新操作所耗代价,使用无效子树大小 n 与其包含的顶点更新次数 c 之间的比值 $R_u = n/c$,称作平均重建代价,决定何时重建无效子树。顶点更新的实例过程如图 2 所示。

图 2 中的重建代价阈值设为 1,图 2(a)是初始线,图 2(e) 是其对应的动态线综合二叉树。图 2(b)～2(d)分别表示顶点的更新、插入、删除,图 2(f)～2(h)分别代表更新操作后的树结构,在删除顶点 p_2 的操作之后,由于更新次数 3 与无效子树大小之比值即平均重建代价为 1,无效子树被重建,图 2 中的无效子树用浅灰色表示。

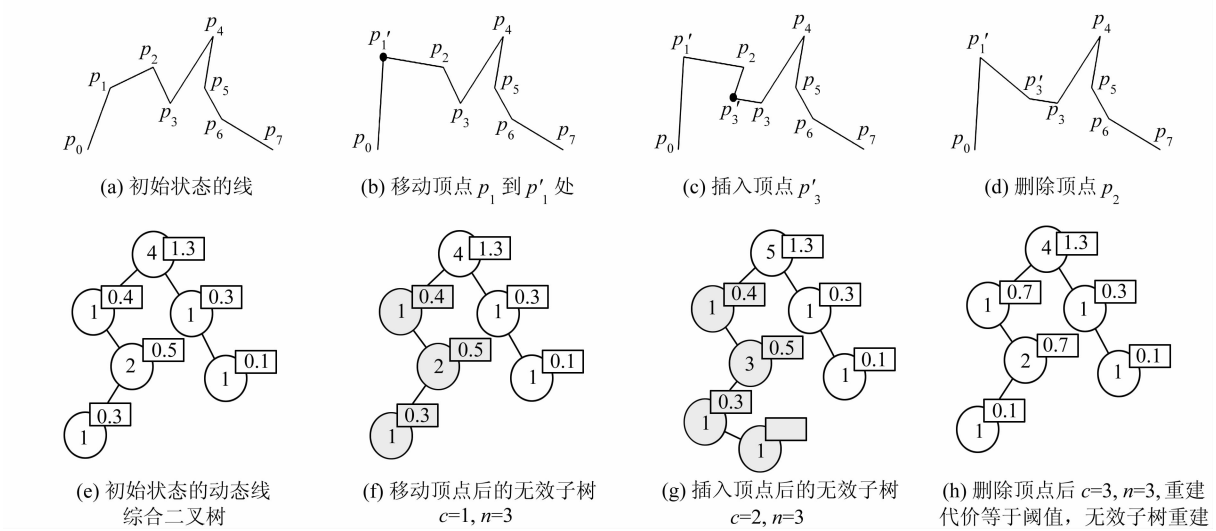


图 2　动态线综合二叉树的顶点更新过程

Fig. 2　Vertices Updating of the Dynamic BLG-tree

3.2　几何对象集的管理与组织

3.2.1　绘图查询

VGI 数据的绘图需要通过窗口查询取出与目标范围相关的所有空间对象,由于图面表达和网络传输的需要,须控制结果集的大小。这种包括窗口查询条件 w,又包含依据分辨率 e 的制图综合操作和结果集大小上限值条件 u 的查询,称为 VGI 数据的绘图查询。其处理流程如图 3 所示。

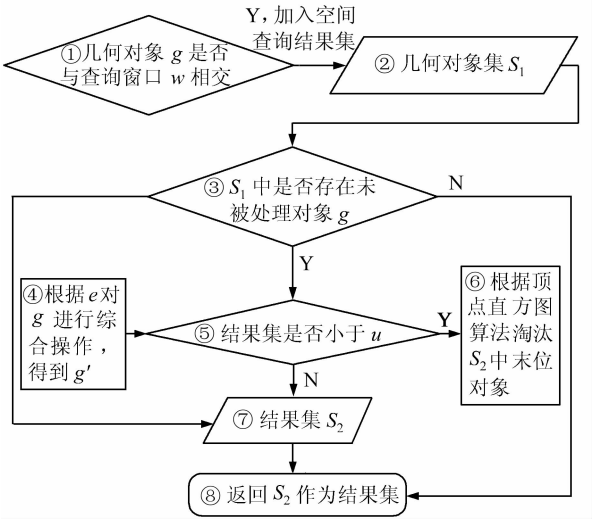


图 3 绘图查询处理流程

Fig. 3 Procedure of Plotting Query

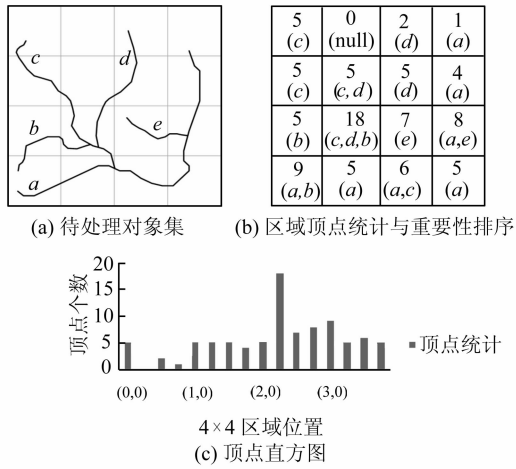


图 4 顶点直方图选择算法示例

Fig. 4 Object Selection Algorithm Based on Vertices Histogram

为了在结果集中选择出重要的几何对象，同时控制结果集大小与对象分布的均匀性，顶点直方图选择算法处理过程如下。

1) 将查询窗口 w 等分为 $m\times n$ 个区域，大小为 $m\times n$ 的一维数组 N 记录几何对象落在各区域内的顶点数之和，每个区域设置一个栈 $T_{i,j}$ 存放对象的引用。将待选集 S 中的几何对象按重要性降序排列，重要性的计算由经验公式或人工赋予。

2) 取出 S 中的对象，计算其落在各区域内顶点的个数，累加到数组 N，将对象压入相交区域的栈中并加入选择集 S_2。

3) 判断 S_2 大小是否超过 u，若未超过则转到步骤 2)，若超过且数组 N 的偏差不大于某阈值，则算法完成，否则，顶点数量最多的区域栈 $T_{i,j}$ 将对象出栈，从 S_2 中删除，并转到步骤 2)。

图 4(a)表示待选集的 5 个对象，按重要性降序排列为 a、c、d、b、e，窗口范围被划分为 4×4 个区域，图 4(b)表示每个区域内顶点的统计个数以及对象栈，图 4(c)表示区域的顶点直方图，其中区域(2,1)包含 18 个顶点，若选择集大小超过 u，则应当在区域(2,1)的对象栈(c,d,b)中将 b 出栈，剔除出选择集。

3.2.2 缩放四叉树

缩放四叉树由索引树与视图树两棵四叉树组成，一棵基于 MX-CIF 四叉树[11]作空间索引；另一棵按照四叉树金字塔结构组织绘图查询的实化视图。

改进的 MX-CIF Quadtree 的性质如下。

1) 子结点规则等分父结点空间。

2) 几何对象只属于某个确定的结点。根据几何对象 MBR 优势边长确定其在树中的层次，根据 MBR 中心点确定其在某层的水平位置。

3) 树的查询与更新算法类似于 MX-CIF Quadtree。

金字塔结构的实化视图四叉树性质如下。

1) 子结点在每一维度上按规则等分父结点空间。

2) 结点包含一个绘图查询结果集，但不包含 MBR 优势边大于结点边长的几何对象，此绘图查询的窗口 w 是结点所对应的区域，上界 u 为经验值，分辨率参数 e 为将 w 绘制到 2 048×2 048 分辨率点距 0.25 mm 的虚拟屏幕时，每个像素所对应的实际地面长度。

3) 结点内简化对象可能对应多个原始对象，原始对象更新时，简化对象也会更新。

金字塔结构的实化视图四叉树运用多级地理网格[12]思想将地理数据进行分块存放，利用局部性进行数据聚簇，提高查询性能。

缩放四叉树处理绘图查询的过程如下：① 根据窗口 w 的优势边长，在实化视图树中找到相应目标层，将完全覆盖 w 的 4 个相邻结点加入查询结果集；② 根据 w，在索引树中从根结点开始下降至目标层，依次访问与 w 相交的结点，找到与 w 相交的几何对象并进行截窗简化查询后加入查询结果集。

参 考 文 献

[1] 李德仁，邵振峰. 论新地理信息时代[J]. 中国科学(F 辑)，2009，39(6)：579-587

[2] Goodchild M F. Citizens as Voluntary Sensors:

Spatial Data Infrastructure in the World of Web 2.0 [J]. International Journal of Spatial Data Infrastructures Research, 2007(2):24-32

[3] Goodchild M F. NeoGeography and the Nature of Geographic Expertise[J]. Journal of Location Based Services, 2009,3(2):82-96

[4] Haklay M, Weber P. OpenStreetMap: User-Generated Street Maps[J]. Pervasive Computing, 2008, 7(4):12-18

[5] Mummidi L N, Krumm J. Discovering points of Interest from Users' Map Annotations[J]. GeoJournal, 2008, 72(3/4):215-227

[6] Qian Xinlin, Di Liping, Li Deren, et al. Data Cleaning Approaches in Web2.0 VGI Application [C]. The 17th International Conference on Geoinformatics, Fairfax, 2009

[7] Van Oosterom P, Van Den Bos J. An Object-Oriented Approach to the Design of Geographic Information Systems[C]. The First Symposium on Design and Implementation of Large Spatial Databases, Santa Barbara, 1990

[8] Ballard D H. Strip Trees: a Hierarchical Representation for Curves[J]. Communications of the ACM, 1981, 24(5):310 321

[9] Gaede V, Gunther O. Multidimensional access methods[J]. ACM Comput. Surveys 1998, 30(2): 170 231

[10] Qian Xinlin, Zhu Xinyan. Multi-representation Geographic Data Organization Method Dedicated for Vector-based WebGIS[C]. The 34th Congress of ISPRS, Beijing, 2008

[11] Samet, H. The Quadtree and Related Hierarchical Data Structure [J]. ACM Computing Surveys, 1984, 16 (2):187-260

[12] 李德仁.论广义空间信息网格和狭义空间信息网格[J].遥感学报,2005(2):513-520

第一作者简介:李德仁,教授,博士生导师,中国科学院院士,中国工程院院士,国际欧亚科学院院士。主要从事以 RS、GPS 和 GIS 为代表的空间信息科学与多媒体通讯技术的科研和教学工作。近年来提出空间信息多级网格和空间数据挖掘与知识发现理论、广义空间信息网格和狭义空间信息网格,并致力于地球空间信息科学技术的研究与应用工作。

E-mail:drli@whu.edu.cn

A Brief Introduction of Data Management for Volunteered Geographic Information

LI Deren[1]　*QIAN Xinlin*[1]

(1　State Key Laboratory of Information Engineering of Surveying, Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

Abstract: Volunteered Geographic Information (VGI) is a new concept emerging with the development of geographic information science. This paper briefly analyses the sources, categories and characteristics of VGI data, presents the requirement of VGI data management, discusses the issues of data cleaning and quality control, focuses on the research of the VGI data management whose target is high performance processing of plotting queries and data updates. Two kinds of spatial data structures are presented. Dynamic binary line generalization tree, which is an extension to binary line generalization tree (BLG-tree), is designed to process the queries and updates of a single spatial object. Zoom quadtree, which includes two quadtree that function as spatial indexes and materialized views, is designed to index and organize the voluminous spatial objects.

Key words: VGI; data management; plotting query; dynamic binary line generalization tree; zoom quadtree

About the first author: LI Deren, professor, Ph. D supervisor, Academician of Chinese Academy of Sciences, Academician of Chinese Academy of Engineering, Academician of Euro-Asia International Academy of Sciences. He is concentrated on the research and education in multi-media communication, spatial information science and technology represented by RS, GPS and GIS. His recent majors are the theories and methods for spatial information multi-grid, data mining and knowledge discovery , theories and applications of generalized and specialized spatial information grid, etc.

E-mail: drli@whu.edu.cn