现代 IT 与第四代 GIS 软件

方 裕[1),2)] 田国良[3)] 史忠植[4)] 周成虎[1)]

[1)](北京大学计算机科学技术系,北京 100871) [2)](中国科学院遥感应用研究所,北京 100101)

[3)](中国科学院计算技术研究所,北京 100080)

[4)](中国科学院资源环境信息系统国家重点实验室,北京 100101)

摘 要 就现代计算机软件技术发展的若干方面进行了回顾和分析,结合 GIS 软件在这些方面的实现进行了讨论. 特别就 GIS 软件在空间数据的分布式计算,空间数据、属性数据的一体化查询与操作,空间数据挖掘等方面进行了比较深入的分析. 指出目前 GIS 实现技术存在的不足及其原因,主要表现在空间数据模型的组织和存储技术不完善、缺乏完整空间关系描述框架、以及空间数据与其他数据联系不够紧密等方面. 提出了解决这些问题的技术突破方向,是要改变以图层为基础的空间数据存储和操作模式,建立空间同步的数据操作机制,实现空间数据修改的 UNIX 语义;要研究空间数据与属性数据一体化查询语言,提高空间数据的操纵能力,等等.

关键词 RPC 同步 空间数据 空间关系 分布式计算 SQL 空间数据挖掘 数据仓库

中图法分类号: P208 TP311.133.1 **文献标识码**: A **文章编号**: 1006-8961(2001)09-0824-06

Modern IT and 4th GIS software

FANG Yu[1),2)], TIAN Guo-liang[3)], SHI Zhong-zhi[4)], ZHOU Cheng-hu[1)]

[1)](*Department of Computer Sci. & Tech., Peking Univ., Beijing* 100871)

[2)](*Institute of Remote Sensing Application, Chinese Academy of Sciences, Beijing*, 100101)

[3)](*Institute of Computing Technology, Chinese Academy of Sciences, Beijing*, 100080)

[4)](*LREIS, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing* 100101)

Abstract In this paper, some modern computer software techniques are introduced. However, those techniques can not be put into practice of GIS software design. Especially, spatial data distributed computation, combining query and operation of both spatial and attribute data are focused on. In this paper. The weaknesses of GIS implementation technique, like the spatial data modeling and storing, the completed framework of spatial relationship and declaration, the relationship presentation between the spatial data and the attributes are pointed out. Also, the technical directions to solve those problems are presented. The layer-based spatial data storing and operation model has to be changed. The synchronization mechanism of spatial data operation has to be established. The UNIX semantics of spatial data modification has to be implemented. The GSQL has to be researched so that the ability of spatial data administration can be increased. etc.

Keywords RPC, Synchronization, Spatial data, Spatial relationship, Distributed computing, SQL, Spatial data mining, Data warehouse

0 引 言

现代计算机技术,特别是软件技术的发展十分迅速,为解决各类计算机应用问题提供了有力的手段. 但是,由于空间数据及空间实体相互关系的复杂性以及 GIS 设计思想和体系结构上的局限性,虽然 GIS 软件也在努力地跟上 IT 主流技术的发展,也取得了许多有意义的进展,但相当一部分现代软件技

收稿日期:2001-06-30;**改回日期**:2001-07-16

术至今没有能够在 GIS 软件实现中得到有效的使用,导致许多空间信息的应用问题无法或很难得到满意的解决. OPENGIS 就当前 GIS 软件发展提出了地理信息的抽象规范[1],但没有就其中的软件实现技术展开深入的讨论和研究. 结合 IT 技术的发展和空间数据获取技术的快速发展以及 GIS 日益旺盛的应用需求,认真分析 GIS 软件在设计思想和体系结构方面存在的问题,求得合理的解决途径是一件相当紧迫的事情.

1　GIS 的分布式计算技术

网络技术,特别是 Internet 技术的发展和普及应用,传统的计算机应用解决方案基本上都依赖于网络与分布式计算. 分布式计算环境(DCE)技术已经成熟[2],作为商品软件被广泛地使用. 在此基础上,多数领域应用软件系统已经成功地实现了 C/S、B/S 结构,无论在系统规模、处理能力、维护代价和使用方便等方面都有了十分明显的进步.

一般来说,实现分布式计算,其中关键的技术实现方式之一是远程过程调用(RPC)[3]. RPC 实现模型如图 1.

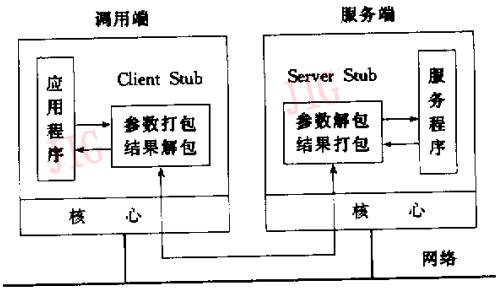

图1　RPC实现模型

对每一个 RPC,在调用端和服务端分别有一个辅助模块(Stub),放在各自的程序库中,在服务端还有一段相应的服务例程. 应用程序执行 RPC,实际上相当于用本地方式调用相应的 Stub,后者将要求的服务名及有关参数打包成消息,通过核心发送至服务端并等待结果回送;消息到达服务端后,相应的 Stub 将消息解包,根据要求调用本地相应的服务例程,服务例程执行结束后将结果返回 Stub,Stub 再将结果打包成消息并通过核心发送到调用端;调用端相应的 Stub 将结果解包并返回给应用程序,从而完成整个过程. 这里,所有的工作都是隐式进行的,从用户角度看,调用远程过程就和调用本地过程一样. 当然,当多个 RPC 同时到达服务端时,服务端需要配备相应的队列管理机制来管理服务请求,有序地进行服务.

对 GIS 软件来说,空间数据的 RPC 不仅是空间数据分布式计算的基础技术,也是不同 GIS 系统之间互操作(除了操作规范以外)的主要实现技术. 同时,空间 RPC 也是分布式协同工作的重要基础. 实现空间 RPC 的难点在于两个方面. 首先,为了直观,往往在调用端需要有图形的显示作参考,例如在图形上进行编辑,用户是依据图形而不是依据空间实体的内部名进行操作,然后由系统转化为 RPC 进行远程修改. 由于空间数据结构和空间实体之间关系的复杂性,同一个空间实体可能是多个其他实体的组成成分. 如图 2 所示,节点 a 既作为一个实体,又是多个线段、弧和多边形的组成成分,a 位置的变化将涉及多个多边形、弧和线段形状的变化.

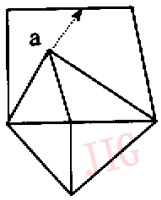

图 2　空间实体及关系

首先,参数的传递成了一个复杂的问题,除了实体名的传递外,还可能需要拓扑关系的传递. 解决这个问题就要求空间实体不仅在图廓范围内有唯一标识,还要求在全库范围内有唯一标识(实体可能移出图廓),拓扑关系可以放在服务端维护,利用数据库的触发机制作相应的修改. 显然,空间数据在文件系统中存储的方式将难以实现上述要求. 而如何在空间数据库中存放并高效地查找和重建拓扑关系的问题必须解决;高效的、动态的索引机制必须建立.

其次,为了保证多用户同步关系,RPC 应该作为一个"事务"来处理,这里就涉及空间实体的封锁机制. 显然,加锁不能只涉及实体本身,而且应该遍及与它相关的空间实体. 考虑到 RPC 可能嵌套,这样,多个 RPC 就可能造成空间实体循环上锁而出现死锁的现象,问题将变得更加复杂. 另一方面,空间数据的"两阶段递交"机制必不可少,必须引起高度的重视.

除了上述问题,RPC 的执行还可能引起空间实体的增减和相应属性数据的变更(如引起相交、切割等),所以必须进行适当的处理以保证数据的完整性

和一致性. 当然,这些问题过于复杂,不可能完全自动实现,但必须引起高度重视,制定合理的规则,同时建立适当的机制,提供相应的手段.

地球上 80%以上的信息都与其空间位置有关,GIS 软件应该成为在 OS、DBMS 之上的重要应用集成平台,在一套空间数据上集成多个应用系统的情况需求强烈. GIS 的分布式计算能力、互操作能力方面的弱势已经成为应用的瓶颈,为了有效地扩充 GIS 的计算能力,上述问题是难以回避的. 所有这一切都要求 GIS 软件的空间数据组织与存储结构发生大的变革,以适应分布式计算的需要.

第四代 GIS 软件首先应该是一个分布式空间数据计算的平台. 它必须针对空间数据的特点,解决多用户环境下的同步操作机制,实现空间数据修改的 UNIX 语义,即"任一用户对空间数据的修改都应为其他用户所可见";它必须提供空间数据处理事务的定义、组织和实现机制,从而允许用户在其支持下组织协同工作的环境;它必须提供一系列远程空间数据计算的手段,包括空间数据中间件和一致的空间实体定义、命名和操作等. 只有这样,才能与主流应用软件的体系结构和技术实现接轨,才能成为名副其实的应用集成平台.

2 结构化的空间数据查询语言

实现分布式计算的另一种方式是使用跨机的查询语言. 传统的 DBMS 对所管理的结构化数据提供了结构化的查询语言 SQL,用户可以方便地在全库范围内进行查询,OBDC 的出现还允许实现跨计算机的、不同数据库的互操作,从而实现了更高级的、面向问题的计算方式. 使用分布式 SQL 已经成为分布式计算的一种重要的底层实现技术. 但是,传统的 GIS 软件对空间数据的处理是面向过程的. 即使使用商用的 DBMS 来存放和管理空间数据与属性数据,甚至将他们存放在同一个数据库中时,通常也是把空间数据与属性数据作为相对分离的存储实体来对待;虽然也定义了具有一定空间数据处理能力的查询语言,但数据操纵能力较弱,应用的组织也很不方便. 因此,设计一种统一处理空间数据与属性数据的结构化查询语言十分必要. 空间结构化查询语言[4],有时也称 G/SQL 或 S/SQL. 第三代 GIS 已经实现了利用商用的 DBMS 管理空间数据与属性数据,这就需要用一种统一的工具来操纵这两类数据,并在此基础上实现分布式计算. 但是,这种一体化的查询语言目前还处于比较初等的阶段,对复杂的空间数据计算表达能力还不够强,还不能满足实际应用的需要.

G/SQL 的关键技术之一是空间算子集合的设计[5],希望空间算子集合尽可能完备,具有强大的描述能力. 通常,空间算子又包括空间关系谓词和空间函数两个大类. 其中,空间关系最值得注意的是拓扑关系和方向关系. 拓扑关系描述空间实体之间的邻接、包含、叠置等在平移、缩放、旋转变化下保持不变的关系,方向关系则涉及空间实体之间的序. 目前,空间谓词主要集中在拓扑关系的表达,基于九交叉模型,可以得到多达 512 种拓扑关系,目前通常从中选取 8 种(相离、外接、相等、交叠、包含、包含于、内接、内接于)关系建立空间谓词;对于方向关系,尚未建立适当的谓词. 空间函数则作用于空间实体,更确切地说是几何体上,通过量测、变换等手段得到数值或新的几何体. 目前实现的空间函数有:

一元空间函数

作用于点的空间函数:取点的 X 坐标与 Y 坐标.

作用于曲线的空间函数:计算曲线的长度;计算线的斜率;求曲线的第一个点;求曲线的最后一个点.

作用于多边形的空间函数:计算多边形面积;计算多边形周长;计算多边形最小外包矩形;计算多边形质心计算多边形内部某点;取多边形的所有线段.

作用于点、曲线、多边形的空间函数:计算缓冲区.

二元空间函数

计算两个几何体之间的最短距离;计算两个面状几何体(交)叠置的结果;计算两个面状几何体(查)叠置的结果;计算面状体被线段切割后的左部;计算面状体被线段切割后的右部;计算线、面相交后线段的面内部分.

显然,这些空间函数功能的覆盖范围有限,特别是集合运算的能力更加有限,不能满足应用要求.

G/SQL 涉及的关键技术还包括扩充的、支持空间的"关系"及其运算和操作. 在传统的 GIS 中,空间实体与关系是一对一的,即

空间关系 $O=\langle$关系 R,空间数据结构 $S\rangle$

特别是在以 GIS 为基础的多种应用集成环境中,为了有效地处理空间数据库的连接操作,在查询过程中应该允许建立临时的扩展的空间型"关系",即允许一个空间数据结构对应多个属性表. 即

扩展的空间型关系 $X_O=\langle$关系数据 R,$\{$空间数据结构 $S\}$,连接表 $L\rangle$.

不仅如此,还要解决“嵌套表”的定义和操作来满足多种类型的集成需要.

由此,数据库的基本操作应该加以扩充,不仅需要扩充操作的类别,还需要扩充操作的语义. 空间数据库的基本操作应该包括关系型选择、空间型选择、关系型连接和空间型连接. 关系型选择作用于基于关系成份 R 上,得到新的关系 R',进而得到连接表 L',用 L' 过滤 S,得到 S 的子集 S'. 空间型选择作用于空间成分 S 上,得到新的 S', 进而得到新的连接表 L', 用 L' 过滤 R,得到 R 的子集 R'. 关系型连接作用于空间型关系 O_1 和 O_2 上,得到新的关系 R_3,进而得到新的连接表 L_3,用 L_3 过滤 S_1 和 S_2,得到 S_1 和 S_2 的子集 S'_1 和 S'_2. 空间型连接作用于空间型关系 O_1 和 O_2 上,可能得到 S_1 和 S_2 的子集 S'_1 和 S'_2,进而生成新的连接表 L_3,用 L_3 过滤 R_1 和 R_2 的全连接,得到新的关系 R_3;也可能得到包括两个源空间数据结构的某些组合信息的新的空间数据结构 S_3,同时生成新的连接表 L_3,用 L_3 过滤 R_1 和 R_2 的全连接,得到新的关系 R_3.

目前 G/SOL 存在的更为关键的问题是以空间数据的查询与属性数据查询的不对称性,空间对象之间的关系与空间数据的操作只能局限在当前图廓之内. 这无疑将极大地限制查询的范围,成为查询的瓶颈,这就需要对空间数据模型和存储组织进行彻底的改革,这无疑是第四代 GIS 软件应该集中解决的问题.

不仅如此,一体化的空间查询语言还应该支持用自定义的数据类型及其操作;支持可加载的查询优化和执行模块;支持可变的数据组织和索引结构. 所有这些都是第四代 GIS 软件需要研究的问题.

3　以空间数据为基础的数据挖掘技术

数据挖掘,就是从大型数据库的数据中提取人们感兴趣的知识[6],也就是说,数据挖掘的过程实际上就是在一些事实或观察、测量数据中寻找模式的决策过程. 数据挖掘一般分为 3 个阶段,即数据准备、挖掘操作、结果表达与解释[7]. 数据准备包括数据集成、数据选择和数据预处理等工作,其目的是:归并和清洗数据以解决语义的模糊性;辨别需要进行分析的数据集合以缩小处理范围;回避或缩小由于挖掘工具的局限性对挖掘质量的影响等. 数据挖掘包括产生假设、选择工具、挖掘操作和结果证实等步骤. 数据挖掘有多种分类方法,包括总结规则挖掘、特征规则挖掘、关联规则挖掘、分类规则挖掘、聚类规则挖掘、趋势分析、偏差分析、模式分析等,可以在原始层次上,也可以在高层次上或多层次进行数据挖掘. 常用的挖掘技术包括人工神经网络、决策树、遗传算法、最近邻技术、规则归纳、可视化分析等. 结果表达与解释是对提取的信息进行分析,分离有价值的内容并用各种方式交给决策者或决策工具[8]. 在传统的计算机应用领域里,数据挖掘的技术已经日趋成熟,并得到了广泛的应用,这有助于人们透过现象认识本质,从浩瀚的数据海洋中发现趋势和规律,从而进行科学的决策.

空间数据挖掘是从空间数据库中抽取隐含的知识、空间关系以及其他非显式地包含在空间数据库中但以别的模式存在的信息,供用户使用. 一般来说,空间数据挖掘可以分为 3 个层次:传统式,以属性数据为操作对象进行挖掘,将挖掘结果通过图形和其他可视化的方式显示出来;空间式,以地理空间数据为操作对象进行挖掘,将挖掘结果通过图形、表格或其他方式表示;混合式,以地理空间数据和属性数据的结合体为操作对象进行挖掘,将结果用各种方式加以显示. 由于在空间数据组织模型方面的缺陷,目前的 GIS 软件和空间数据库还不能有效地支持这 3 个层次的数据挖掘.

通常,空间数据库可以表示为:

空间数据库 = ⟨{属性表 A},{地理空间数据结构 D},{连接关系 R}⟩

这里,属性表是对空间实体特性的非空间特性描述,表示成通常的关系数据表形式,其连接对象是空间实体,在数据库中表现为地理空间数据结构;地理空间数据结构包括空间实体的组成部分(也是空间实体)、组成关系和位置数据,按照目前空间数据的组织模型,地理空间数据结构的层次为:图集→图层→空间实体;连接关系则表示空间实体与属性表的连接关系,目前连接关系 R 仅仅通过标识码一种简单的方式表示.

传统式空间数据挖掘的挖掘操作及技术与通常的关系型数据的挖掘并无二致,即在 $\{A\}$ 上进行挖掘操作,挖掘结果将得到新的属性数据表 A'. 进而提取出有用的信息,由此形成规律或知识. 传统式空间数据挖掘的困难之处是如何将挖掘结果通过空间图形表示出来. 结果表达过程中需要对 A' 重新组织、生成新的连接关系 $\{R'\}$、最后映射到 (D) 上,得到其子集,并用各种方式表达. 可以表示成:

$op\_mr(\{\{A\},\{D\},\{R\}\}) = \langle op\_r(\{A\}),\{D'\},\{R'\}\rangle$

由于属性数据的挖掘是整个数据库范围的,新的属性表及其连接关系同样是全库范围的.由于空间数据结构的层次关系,空间实体之间的相互关系以及空间计算只有同处于同一图层才有意义.全库范围的数据挖掘将形成全库范围的新的连接关系 R',空间数据这种连接关系如何映射到空间数据库,进而形成新的显示关系将成为难题.显然,不改变以图层为基础的空间数据组织模型,完整地、有效地实现传统式的空间数据挖掘及其结果显示几乎是不可能的.

空间式数据挖掘在全库范围内对地理空间数据结构(包括其派生信息)进行操作,得到新的地理空间数据结构 D' 及其关系,进而形成新的连接关系 R',映射到属性表,生成新的属性 A'.即:

$op\_ms(\{\{A\},\{D\},\{R\}\}) = \langle\{A'\},op\_s(\{D\}),\{R'\}\rangle$

首先,这类挖掘将涉及跨图层的空间计算和操作,以图层为基础的空间数据组织将使得全库范围内的空间型挖掘成为不可能;其次,地理空间数据不同于传统的关系数据,它们之间不是独立的,相互之间有着许多隐含的依赖关系和彼此影响,传统的数据挖掘技术将不再能够满足这类数据挖掘的要求,必须加以扩展,除了空间推理、空间几何学等已有技术外,需要发展新的挖掘方法.例如,目前对空间对象的分类方法已经进行了大量的研究,出现了诸如空间关联算法、空间聚类算法、泛化的空间描述、空间聚类特征描述等研究[9~11],但是,由于空间数据的复杂性和信息的不完整性,尚未出现完全有效的分类模型及算法.需要说明的是,相当部分的空间式挖掘可以通过数据的预处理转化为传统式的挖掘.但是,后者不能替代前者解决所有的空间数据挖掘问题.

混合型数据挖掘对属性数据和地理空间数据联合进行操作,得到新的地理空间数据 D' 和新的属性数据表 A',进而形成新的连接关系 R',即:

$op\_mm(\{\{A\},\{D\},\{R\}\}) = \langle op\_m(\{A'\},\{D\}),\{R'\}\rangle$

这类挖掘将更为复杂.目前,空间实体与属性数据之间的联系仅仅依赖于标识码,这种一维的连接方式无疑将丢失大量的连接信息,不能有效地表示多维和隐含的内在连接关系,这无疑会极大地增加数据挖掘的计算复杂性,极大地加大数据准备阶段的工作量和人工干预程度;其次,目前的空间数据组织尚不能有效地表达时序数据,只能静态地表达空间实体与属性的联系,对动态的地理空间数据、属性数据及其相互之间复杂的联系既缺乏有效的表达手段,也缺乏高效的存储手段.实际上,空间数据挖掘应用在很大程度上涉及到空间实体及其属性的时序关系,这不能不严重地妨碍空间数据挖掘的应用范围.此外,基于图层的计算模式、不同尺度空间数据之间的完全割裂都对空间数据挖掘设置了重重障碍.第四代 GIS 软件应该在这些理论和技术方面有所突破,处理真正实现面向客观存在的空间实体及其相互关系.

除此以外,目前的 GIS 软件实现还有许多方面不能满足实际应用的需要,例如多维数据的表示和操作问题、矢栅数据的一体化处理问题、与主流软件的集成问题,等等.应该说,空间信息领域在技术上还有许多落后于 IT 主流技术技术发展的地方.现代 IT 技术为人们提供了强有力的计算手段.但是,由于空间数据及其关系的复杂性,在 GIS 软件中还不能有效地应用其中的许多行之有效的技术,这是造成 GIS 还不能"融入"IT 主流和与其他计算机应用软件不能有效集成的一个重要原因.日益发展的空间信息应用对 GIS 软件提出了更高的要求,开展第四代 GIS 软件理论和技术的研究,争取在其中一个或几个方面取得突破,将是一件十分必要、十分有意义的事情.

参 考 文 献

1 OGC. The OpenGIS abstract specification[R]. 1998:121~135.

2 Berson A. Client/Server Architecture[M]. McGraw-Hill, Inc. 1992:347~367.

3 Tanenbaum A. S. Distributed operating system. Prentice Hall, Inc. 1995:68~75.

4 Egenhofor M J. Spatial SQL: a query and presentation language [J]. IEEE Trans. on Knowledge and Data Engineering, 1994: 2~5.

5 方裕,楚放,陈斌. 空间结构化查询语言[J]. 中国图象图形学报, 1999,4(11):902~905.

6 Usama F. Gregory P. Advances in knowledge discovery and data mining[M]. California AAAI/MIT Press, 1996:57~83.

7 IBM White Paper. Data mining: extending the information warehouse framework[R]. 1996:11~15.

8 石云. 数据空间化的决策支持研究——模型、方法及应用[博士学位论文][D]. 北京:中国科学院软件研究所,2000.

9 Koperski K. Didcovery of spatial association rules in geographic information databases [A]. In: Proceedings of thr 6th International Symposium on Spatial Databases[C], Germany, Berlin, 1995:47～66.

10 Xu X. A distribution-based clustering algorithm for mining in large spatial databases[A]. Proceedings of ICDE[C], USA, Florida, 1998:476～483.

11 Lu W. Discovery of general knowledge in large spatial databases [A]. In: Proceedings of Far East Workshop on Geographic Information Systems[C]. Singapore, 1993:275～289.

方　裕　1946 年生,1968 年毕业于北京大学,现为北京大学计算机系教授. 长期从事计算机软件工程、地理信息系统、计算机集成制造系统等方面的研究工作.

田国良　1939 年生,吉林大学物理系毕业,现任中国科学院遥感应用研究所研究员. 主要研究方向为遥感物理及其应用. 发表论文 90 余篇.

史忠植　博士生导师,1964 年毕业于中国科技大学计算机专业,1968 年毕业于中国科学院研究生院,现任中国科学院计算技术研究所研究员. 长期从事知识工程、分布式人工智能、机器学习、神经计算、认知科学等方面的研究. 发表论文 250 多篇,著作 8 本.

周成虎　1964 年生,博士,研究员,博士生导师,国际欧亚科学院通讯院士,中国科学院资源与环境信息系统国家重点实验室主任. 主要研究兴趣包括空间数据库挖掘与知识发现、地理空间单元相互作用、遥感影象的地学理解与分析等.