doi:10.3799/dqkx.2010.043

空间数据无缝集成版本机制及其关键技术

胡茂胜[1,2],方 芳[1],黄胜辉[3]

1. 中国地质大学信息工程学院,湖北武汉 430074
2. 地理信息系统软件及其应用教育部工程研究中心,湖北武汉 430074
3. 武汉中地数码科技有限公司,湖北武汉 430074

摘要:为解决分布式异构环境下空间数据无缝集成及其增量更新的数据组织和查询效率问题,利用地理数据库版本机制所具备的多分量存储和存储增量等特性,通过把有缝数据与版本机制中删除分量相对应,把无缝数据与版本机制中添加分量相对应,把无缝一致性维护过程与版本机制中的数据更新过程相对应,提出并建立了空间数据无缝集成版本机制.该机制把无缝集成问题转化成初始无缝化版本的建立问题以及版本数据的一致性维护问题,把无缝查询及动态拼接过程转化成分布式版本查询过程,从而降低了问题的处理复杂度,提高了系统效率;该机制还能为增量更新提供存储支持.

关键词:版本管理;无缝集成;无缝一致性;地理信息系统.

中图分类号:TP311.13 **文章编号**:1000-2383(2010)03-0380-05 **收稿日期**:2010-01-15

Mechanism and Key Technologies of Spatial Data Seamless Integration Version Management

HU Mao-sheng[1,2], FANG Fang[1], HUANG Sheng-hui[3]

1. *Faculty of Information Engineering, China University of Geosciences, Wuhan* 430074, *China*
2. *Engineering Research Center of GIS Software and Applications, Ministry of Education, Wuhan* 430074, *China*
3. *Wuhan Zondycyber Co., Ltd., Wuhan* 430074, *China*

Abstract: In order to solve the data organization problem and improve the query efficiency of seamless integration of spatial data and its incremental update in distributed and heterogeneous environment, some features of geo-database mechanism can be used, such as storage in multi-components and incremental storage. Through corresponding seam data to the deleted component in version mechanism, and corresponding seamless data to the added component in version mechanism, and corresponding seamless consistency maintenance process to the data version update process, it establishes a new mechanism called spatial data seamless integration version management mechanism. By this method, seamless integration issue is transformed into initial seamless version establishment issue and version data seamless consistency maintenance issue, and seamless spatial query and dynamic splicing process is transformed into the distributed version query process, which reduces the computational complexity and enhances system efficiency; with the seamless version mechanism, natural storage support for the incremental update is also provided.

Key words: version management; seamless integration; seamless consistency; geographic information system (GIS).

0 引言

随着空间信息技术的发展,空间数据在人类生活的各个方面都得到了广泛应用.然而分布式异构性也逐渐成为空间数据的基本特征,如何在分布式异构环境下对空间数据及非空间数据进行有效的集成管理,成为一个亟待解决的问题(李德仁等,2005),无缝集成是集成的特殊情况也是集成的高级形式.与此同时,空间数据本身更新的频率越来越频繁,数据量也越来越大,引发了增量信息识别及增量

基金项目:国家重点"863"项目(Nos. 2007AA12Z204,2007AA120503).
作者简介:胡茂胜(1981-),男,博士,主要从事GIS理论及软件开发、空间数据库等研究. E-mail:humosen@gmail.com

更新的问题，而分布式异构环境下对增量信息的存储与管理又不同于单机情况，迫切需要引入新思想和新方法来建立解决方案. 另一方面，地理数据库版本管理机制提供了对增量信息有效的存储管理方式，并且某一版本下的数据可以看成是对版本各分量的集成(陈波等，2006)，因此版本各分量也具有分布式特征. 在这种背景下，扩展地理数据库版本管理机制以适应分布式异构环境，来解决空间数据无缝集成和增量信息存储问题，就成为自然的选择；这种新的版本机制就是分布式异构空间数据无缝集成版本管理机制，简称无缝集成版本机制.

本文在地理数据库版本管理机制及一般的分布式版本机制的基础上，考虑无缝集成的特点，建立无缝集成版本机制，解决在该机制下的无缝查询问题；讨论了实现无缝集成版本机制的关键技术，并用其解决初始无缝版本的建立问题以及版本数据的无缝一致性维护问题，进而讨论了通过版本归档机制提高无缝集成版本机制的查询效率.

1　地理数据库版本管理机制

地理数据库版本机制可以实现多个用户并发地编辑某个地理数据库而不用明确地锁定要素或者复制数据(陈波等，2006；万波等，2006). 地理数据库版本管理基于状态和版本这两个概念. (1)状态是地理数据库变化过程中某一瞬间的标识. 任何改变地理数据库的操作都产生新状态. 地理数据库中的这些状态可以组织成一棵树，在这棵树的线性结构中描述了各个状态的父子关系. (2)版本则是一个命名的状态. 地理数据库中的每一个版本都明确指向一个具体的状态. 这样，在进行地理数据库编辑时，可以为同时进行分工合作的人员定义各自的版本，每个人都在自己的版本空间下工作，不受其他人编辑的干扰，完成编辑后，再进行版本的合并. 图 1 中每个版本指向一个具体的地理数据库状态. 版本机制易于实现对地理信息系统(geographic information system，GIS)空间数据的协同处理(李伟等，2005).

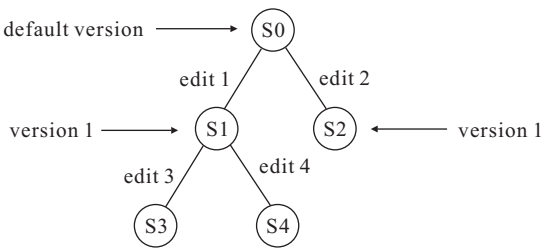

图 1　地理数据库版本演化

Fig. 1　Version states evolvement in geodatabase

分布式版本机制在分布式数据库中多有研究，Andjelic and Worboys(1996)研究了在分布式环境下的 GIS 版本机制.

2　无缝集成版本机制

分布式异构空间数据无缝集成是在数据集成的基础上产生的，甚至一些情况下数据的集成就称为无缝集成(宋关福等，2000)，空间数据集成的方法主要有空间数据交换、数据互操作、数据文件直接访问、GIS 中间件(周顺平等，2008)以及基于空间数据中心的“数据—功能”多层次集成(徐世武等，2006；吴信才，2009)等；单纯从无缝化的角度看，侧重于研究各式各样的无缝化方法、无缝空间数据存储与组织形式(李爱光等，2005)、无缝 GIS、无缝空间数据库等(朱欣焰等，2002)；而在分布式多空间数据库系统中数据集成已经开始考虑数据间的缝隙问题(邬伦和张毅，2002). 本文的无缝集成是指在一般分布式异构环境下的空间数据连续无缝的集成.

通过分析地理数据库的版本机制可以对比建立无缝集成版本机制，考虑到空间数据的分布异构性，版本信息由专门的增量数据服务器维护. 设增量数据库中添加记录表为 A，删除记录表为 D，按照地理数据库的版本机制，一个版本由若干个状态组成；删除操作只需要在 D 表中添加一条对删除要素的状态记录，添加操作也只需要在 A 表中添加一条对添加要素的版本记录，而更新操作需要在 D 表和 A 表针对该更新要素都添加一条相同版本状态的记录. 上述规则说明版本机制是由原始的非版本数据与增量数据库中状态化的数据等多个分量组成，如果把这些分量看成是由根据无缝化过程中要素的处理情况来划分的，就可以建立起无缝集成版本管理机制.

在分布式异构环境下，原始数据由 N 个分布式异构数据节点中的数据构成，相当于版本机制中原始数据也被划分出 N 个分量；但是这 N 个分量之和并不等于初始无缝化版本，因为这 N 个分量内部或 N 个分量之间都可能存在各种各样的缝隙. 要建立起初始无缝化版本还需要有初始的增量部分，如图 2 所示，该增量部分是由删除原始数据中的有缝数据，并增加无缝化处理后的新数据两种操作建立

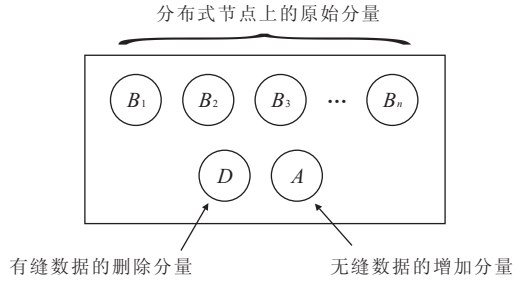

图 2 初始无缝版本的数据分量构成

Fig. 2 Components in initial seamless version

起来的，因而也包括 A 表和 D 表两个部分，这称为无缝化增量数据库. 与版本机制相比，原始数据中存在某种类型缝隙的要素相当于删除记录，需要在 D 表中添加一条对应版本状态的记录；原始数据中有缝隙的要素数据经过无缝化处理形成的新要素数据相当于添加记录，需要在 A 表中添加一条对应版本状态的记录；而对无缝化数据更新相当于版本的更新，需要在 D 表和 A 表中同时添加一条相同版本状态的记录.

如图 3 所示，在无缝集成版本机制下，无缝集成问题转化为无缝初始版本的建立问题以及数据更新过程中的无缝一致性维护问题，这两个问题本文稍后加以讨论.

基于集合论和数据库关系理论，对某一个无缝版本的查询过程相当于对各数据分量实行子查询，并合并生成无缝查询结果集的过程，而且能够通过数据库的基本查询操作得以实现；无缝查询结果集是实现逻辑无缝的基础，也是实现无缝空间分析处理的基础.

定义两个集合的差运算为 $DIFF_{R,S(\mathrm{KEY})}=(R-S)_{\mathrm{KEY}}$，其中 KEY 为集合 R 与集合 S 进行比较、进而求差的关键字；另一方面定义某个无缝版本由 N 个分布式节点分量以及增量数据库中 m 个版本状态分量(State1，State2，…，State$m$)构成. 则无缝查询结果集相当于最终有效要素集($Valid_{\mathrm{W}}$)，满足

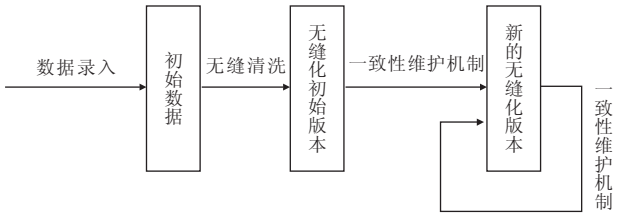

图 3 无缝版本状态演化与维护

Fig. 3 Seamless version states evolvement and maintenance

如下关系式：

$$Valid_{\mathrm{W}}=Valid_{\mathrm{B}}\cup\Delta Valid_{(\mathrm{State1},\mathrm{State2},\cdots,\mathrm{State}m)}=(B_1+B_2+\cdots+B_N-D_{(\mathrm{STATE\ IN(State1,\ State2,\ \cdots,\ State}m))})\cup DIFF_{\mathrm{A,D(FID,STATE)}}\ ,\quad(1)$$

(1)式中，$Valid_{\mathrm{B}}$ 为原始数据在本版本下的有效分量；$\Delta Valid_{(\mathrm{State1},\mathrm{State2},\cdots,\mathrm{State}m)}$ 为增量数据库中在本版本下的有效分量；$B_1,B_2,\cdots,B_N$ 为原始数据在 N 个分布式节点上的完整分量；$D_{(\mathrm{STATE\ IN(State1,State2,\cdots,State}m))}$ 表示在 D 表中属于当前版本的记录，也即原始数据在本版本下的无效分量；$DIFF_{\mathrm{A,D(FID,STATE)}}$ 表示 A 表与 D 表在当前版本中的要素号和要素状态号为关键字的集合差，也即当前版本在 A 表中存在但在 D 表中不存在的记录，也即增量数据库中的有效分量.

基于版本机制实现对分布式异构空间数据的无缝集成，使各分布节点仍然能够保留存储自治性，保持系统中原有空间数据在存储格式及存储方式上的不变性；同时对有缝数据的剔除及新的无缝数据的建立也没有直接提交到各分布式节点，而是以增量的形式集中存放在独立的无缝增量数据库，保证各节点数据在逻辑上的一致性. 对给定某无缝版本的空间数据，包括增量数据库中的分量在内的各个无缝版本分量都是完整的要素数据，不存在割裂现象；因此可以通过集合的和、差等运算实现无缝查询结果集的生成，计算复杂度低，避免了计算量大的动态拼接处理，但是数据需要首先经过无缝化预处理过程来建立初始的无缝化版本，并由一致性维护机制保证后续数据版本始终处于无缝状态.

3 无缝集成版本机制的关键技术

3.1 初始无缝化版本的建立

根据不同的标准可以定义不同的无缝化方式，从图幅的拼接到图层间的合并，从属性字段的映射到不同数据质量数据的融合，包含十分广泛的内容. 因此初始无缝化版本的建立是一个开放的面向问题与需求的数据加工过程，但是也遵循一般的处理步骤：(1)数据资源的描述以及元数据和定位信息的录入；(2)定义无缝化标准、选择无缝处理例程，必要情况下需要通过编码对无缝处理例程进行扩展，实现新的无缝化语义；(3)建立无缝化规则，并生成处理流程；(4)根据无缝化标准判断存在缝隙的空间要素或非空间数据记录，在无缝集成增量数据库中予以

删除;(5)对所有判断存在缝隙的空间要素或非空间数据记录进行无缝化处理,生成若干无缝的完整要素或记录,以新增记录的形式在增量数据库中进行记录;(6)定义初始版本的各分量,完成初始无缝化版本的建立过程.

3.2　无缝一致性维护

分布式异构环境下经过集成及无缝化的空间数据面临着一致性维护问题,主要包括在编辑和更新中拓扑关系的一致性以及空间位置的一致性. 一般来说,在分布式异构环境下各个数据节点既要参与全局视图的生成,又有一定的自治性;从数据全局视图对版本数据进行编辑更新不会破坏数据的全局无缝一致性,而这种数据编辑更新操作对单个数据节点来说也是隐藏的,改变只发生在无缝增量数据库中,因此局部单数据节点的一致性也没有遭到破坏,但看到的数据仍然是有缝的;从局部数据节点对版本数据原始分量编辑更新会破坏数据的全局一致性. 简单的做法是对全局无缝化过程中涉及的原始数据进行只读锁定,只允许从全局视图进行编辑更新.

3.3　无缝集成归档机制

无缝集成空间数据的一致性维护建立在版本的机制之上,因此版本机制效率高低影响到系统的整体效率. 时空地理数据库系统一般采用基态修正法来避免存储研究区域中每个状态的全部信息,只存储某个时刻的数据状态(称为基态),以及相对于基态的变化量,这样做可使时态数据量大大减少. 基态修正法一般把历史上某个历史事件后的状态作为"基态",把用户最关注的"现在"状态,即系统最后一次更新的数据状态,作为"现状". 无缝集成版本机制在原始数据节点上存储数据的初始化版本,在集中式的增量信息数据库中存储增量信息,也可以看成是一种关于时态数据的基态修正模型,此时的时态信息是不断增长的版本状态号.

随着对数据的增加、更改和删除操作的累积,增量信息将变得越来越庞大,从原始数据节点上获取到的初始版本需要经过较大的调整才能演变为"现状"数据,或者初始版本的数据包含越来越多的对"现状"查询无用的数据,但是这些数据的存在会影响查询和数据传输效率. 因此需要提供归档机制,在增量数据库变大时,对初始数据与增量数据进行合并,压缩无效状态,形成新的初始版本.

具体的算法为:(1)对状态分支表和状态表中的记录进行排序和比较,确定状态树的分支节点,同时对末端分支进行遍历,判断其是否属于"悬挂"分支,如果是则"剪除";从而状态树被分成没有分支的多段,可以分别进行压缩操作(经过"剪除""悬挂"分支操作后,剩余的分支节点都要予以保留);(2)对这些段进行遍历操作,判断各状态是否有版本指向;如果有版本指向,进一步分成更多段;(3)对各段进行压缩操作分为两种情况:首状态为 base 状态和非 base 状态;如果为 base 状态,将所有状态压入 B 表,否则将分支上的所有状态压入末端状态.

4　无缝集成版本机制的实现

无缝集成版本机制主要是实现初始无缝化版本建立过程、无缝查询过程以及在编辑更新过程中的无缝一致性维护过程. 初始无缝化版本建立需要实现专门的预处理工具,按照一定的无缝化标准在增量数据库中逻辑上剔除原始数据中有缝要素或数据记录;由于无缝化处理例程本质上说是开放的,在实现上可以建立工具框架,以统一的方式实现数据资源录入、无缝化流程定义、剔除操作以及无缝新记录的添加操作,而对开放过程中变化的步骤可以通过插件式的方式提供给用户扩展机制,例如无缝化规则的制定、无缝化处理例程等. 无缝查询过程有明显的层次性,在从提交查询到查询获取数据的过程中,需要完成查询任务的分解、数据的粗定位、分布式任务的分发等处理过程,而在从查询数据的获取到查询结果集的生成过程中,又需要处理异构数据的互操作、参照系与投影变换、精过滤、版本分量的合成等处理过程,因此无缝查询过程可以建立起多层次查询中间件来实现. 编辑更新过程中的无缝一致性维护在实现上需要对增加、删除、更新等函数进行机制性调整,以相互配合,能够建立新的版本状态在增量数据库中来描述这些操作所涉及到的记录.

5　总结

利用分布式版本机制实现无缝化分布式异构空间数据的无缝化存储与管理,可以最大限度地保持数据本来的存储与管理形式,并通过一次性的无缝化预处理过程避免了每次查询过程中的动态无缝化过程;无缝查询过程转化为版本各分量的合并过程,计算复杂度低,保证了系统的响应效率;同时该机制还提供了对增量信息的存储与管理能力.

References

Andjelic, T., Worboys, M., 1996. Version management for GIS in a distributed environment. In: Parker, D., ed., innovations in GIS, 3. ed. T. J. Press, Great Britain, 65—74.

Chen, B., Zhou, S. P., Wan, B., et al., 2006. Long transactions in GIS. *Earth Science—Journal of China University of Geosciences*, 31(5): 605—608 (in Chinese with English abstract).

Li, A. G., Wang, H., Guo, J., 2005. The spatial data organization of seamless GIS. *Engineering of Surveying and Mapping*, 14(1): 30—32 (in Chinese with English abstract).

Li, D. R., Yi, H. R., Jiang, Z. J., 2005. Introduction and analysis of grid technology. *Geomatics and Information Science of Wuhan University*, 30(9): 757—761 (in Chinese with English abstract).

Li, W., Liu, R. Y., Liu, N., 2005. Research on task-based model and multiversion cooperative GIS spatial data processing. *Journal of Zhejiang University (Sciences Edition)*, 32(4): 475—480 (in Chinese with English abstract).

Song, G. F., Zhong, E. S., Liu, J. Y., et al., 2000. A study on seamless integration of multi-sources spatial-data (SIMS). *Progress in Geography*, 19(2): 110—115 (in Chinese with English abstract).

Wan, B., Zhou, S. P., Chen, B., et al., 2006. Design and realization of MapGIS 7.0 management based on DBMS. *Earth Science—Journal of China University of Geosciences*, 31(5): 600—604 (in Chinese with English abstract).

Wu, L., Zhang, Y., 2002. The integrated framework on distributed multi-spatial database system. *Geography and Territorial Research*, 18(1): 6—10 (in Chinese with English abstract).

Wu, X. C., 2009. Datacenter integration development technology: the next generation GIS architecture and development model. *Earth Science—Journal of China University of Geosciences*, 34(3): 540—546 (in Chinese with English abstract).

Xu, S. W., Xie, Z., Huang, Z. C., 2006. Research and design of isomerism distributed multilevel spatial data center. *Earth Science—Journal of China University of Geosciences*, 31(5): 624—630 (in Chinese with English abstract).

Zhou, S. P., Wei, L. P., Wan, B., et al., 2008. A study of integration of multi-source heterogenous spatial data. *Bulletin of Surveying and Mapping*, 5: 25—27, 39 (in Chinese with English abstract).

Zhu, X. Y., Zhang, J. C., Li, D. R., et al., 2002. Concepts, implementation and problems of the seamless spatial database. *Geomatics and Information Science of Wuhan University*, 27(4): 382—386 (in Chinese with English abstract).

附中文参考文献

陈波，周顺平，万波，等，2006. GIS **中长事务模型. 地球科学——中国地质大学学报**，31(5)：605—608.

李爱光，王卉，郭健，2005. **无缝** GIS **的空间数据组织研究. 测绘工程**，14(1)：30—32.

李德仁，易华蓉，江志军，2005. **论网格技术及其与空间信息技术的集成. 武汉大学学报(信息科学版)**，30(9)：757—761.

李伟，刘仁义，刘南，2005. **基于任务划分和多版本技术的** GIS **空间数据协同处理研究. 浙江大学学报(理学版)**，32(4)：475—480.

宋关福，钟耳顺，刘纪远，等，2000. **多源空间数据无缝集成研究. 地理科学进展**，19(2)：110—115.

万波，周顺平，陈波，等，2006. **基于** DBMS **的** MapGIS 7.0 **版本管理的设计与实现. 地球科学——中国地质大学学报**，31(5)：600—604.

邬伦，张毅，2002. **分布式多空间数据库系统的集成技术. 地理学与国土研究**，18(1)：6—10.

吴信才，2009. **数据中心集成开发技术：新一代** GIS **架构技术与开发模式. 地球科学——中国地质大学学报**，34(3)：540—546.

徐世武，谢忠，黄志超，2006. **分布式异构多级空间数据中心的研究与设计. 地球科学——中国地质大学学报**，31(5)：624—630.

周顺平，魏利萍，万波，等，2008. **多源异构空间数据集成的研究. 测绘通报**，5：25—27，39.

朱欣焰，张建超，李德仁，等，2002. **无缝空间数据库的概念、实现与问题研究. 武汉大学学报(信息科学版)**，27(4)：382—386.