第 33 卷 第 2 期 计 算 机 工 程 2007 年 1 月
Vol.33 No.2 Computer Engineering January 2007
·博士论文· 文章编号：1000—3428(2007)02—0034—03 文献标识码：A 中图分类号：TP39

GIS 模型网格计算原理及算法

扈海波[1]，刘伟东[1]，李 京[2]，朱文泉[2]
(1. 中国气象局北京城市气象研究所，北京 100089；2. 北京师范大学资源学院，北京 100875)

摘 要：GIS 模型计算逐步向数据处理海量化及过程复杂化方向发展。网格计算为解决 GIS 应用这一难题提供了契机。GIS 模型计算从运算模式上可分解为并行、串行和并行串行模式，该文在这些模式的基础上提出了实现 GIS 模型网格计算的原理性方法：分块加工、分步加工和立体加工方法，并以 GIS 和 RS 中的常用模型(直方图和矢量地图插值)作为算法实例，论述其并行算法实现，同时给出矢量地图数据的分块规则。该并行算法实现可作为实现其它同类算法的基础和扩展。

关键词：GIS 模型；网格计算；并行算法

Principle and Algorithm to Fulfill Grid Computing for GIS Models

HU Haibo[1], LIU Weidong[1], LI Jing[2], ZHU Wenquan[2]
(1. Institute of Urban Meteorology, China Meteorological Administration, Beijing 100089;
2. Institute of Resources Sciences, Beijing Normal University, Beijing 100875)

【Abstract】 The problem of the model computing for GIS application is that the processed-data is being more and more gigantic, and its procedure is much more complex than ever. The grid computing is the key to solve it. The computing mode of GIS model could be listed as the parallel, the serial and the parallel-serial. Based on these modes, the paper brings forward three prototypes to fulfill grid computing, such as the pieced-processing, the stepped-processing and the cubic-processing. Consequently, it also gives the principles of dividing vector data, and the parallel algorithm of histogram, convolution, and curving-fitting, which could be the basic and extension of the other more parallel algorithm for GIS application.

【Key words】 GIS model; Grid computing; Parallel algorithm

1 概述

GIS模型应用需要网格计算的原因是：GIS模型处理数据量逐步海量化，有些模型的处理量已经达到TB级别；模型计算过程的复杂程度也在不断加深，需要多道处理加工才能完成[1]；网格为GIS模型应用提供超强的计算能力。

网格计算的目的和要求是实现网格节点机之间的并行计算。网格计算不完全等同于并行计算。Ian Foster认为网格计算是在虚拟协同的环境中执行，能够高效、充分地获取网格资源，如数据网格、数据元网格等[2]。

并行算法在数值计算领域的研究较丰富，而在空间领域中则亟待深入。遥感栅格数据采用矩阵排列方式，比较适合分块方式的并行处理，其并行算法可借用数值计算的并行原理和方法，相对易实现，如蒋艳凰等采用局部数据分块处理的办法，实现了遥感数据空间域的并行几何校正等[3]，这类算法隶属局域计算群范畴，其方法和原理有待引入网格中。Wang Saowen利用Globus Toolkits做了空间数据插值的网格计算，发现数据的切分和整合对并行计算的效果影响较大，其算法实例用“散列点”来完成空间插值，没有针对“线”和“面”空间实体的并行操作，也没有提出具体数据切分规则[4]。Hoel以R树结构作为并行计算时提取待处理空间数据的组织排列方式，但这样的操作结果将产生大小不等的数据分块，造成并行机之间的计算快慢不一，产生计算的相互等待和延迟，从而影响并行计算效果[5]。本文针对空间数据，尤其是“矢量地图数据”结构及组织特点，提出GIS模型网格计算的下述实现要点：

(1)处理数据分布存储，这是数据网格的基本要求和实现；

(2)在数据超大的情况下，整体数据按均衡方式分块处理，每块数据可交给对应的网格节点机处理；

(3)整个计算流程分割成多道可独立运行的数据加工过程，每道工序可交给相应的节点机处理；

(4)融合分块或者分步处理结果。

实现原则：对单一复杂问题的求解切分为对多个问题的求解，然后，将各个子问题的求解分配到网格上的每个节点机上，实现网格计算。

总之，GIS 模型网格计算是把数据分块，加工步骤分步，把模型单一求解分散到各个节点上处理，最后融合计算结果。

2 网格计算方法

定义 1(模型分块加工法) 模型的某一计算步骤对相应的一对空间或非空间数据(或单个数据)进行操作，并产生一组计算结果，称这种加工方式为分块加工法。

定义 2(模型分步加工法) 模型的每个操作符仅对相应的一对空间或非空间操作数进行操作，或对一个空间或非空间操作数进行操作，并产生一个计算结果，称这种加工方式为

基金项目：国家“863”计划基金资助项目(2002AA130020，2002AA134090)；国家科技攻关计划基金资助项目(2003BA808A16)；环境大气综合探测资料四维数据库技术研究项目(H020620250230)
作者简介：扈海波(1970－)，男，博士，主研方向：GIS算法，GIS模型库；刘伟东，博士、副研究员；李 京，硕士、教授；朱文泉，博士、讲师
收稿日期：2006-01-23 **E-mail：**lijing@bnu.edu.cn

分步加工法。

定义 3(模型立体加工法) 模型的操作序列有T个，$T\geq 2$，$\theta_1, \theta_2, \theta_3,\ldots,\theta_i$，$I$=1,2,…,$T$。

每个操作均对应一对数组(或单个数组)执行操作，并分别产生一个(或一组)计算结果，则称这种加工方式为立体加工法。

地学模型计算中大多数数据是可以进行分块加工处理的，但有些计算不能简单地分块处理，需要作进一步的算法调整，比如矩阵乘法运算等。而且也有一些问题属于串行计算，原则上也不能作分步计算处理。

定义 4(模型串行计算) 一个模型的计算方式是按单一步骤(单向数据流的加工方式进行的，相邻的操作$step_i$,$step_{i+1}$(i=1,2,…,N)之间的关系是后一操作($step_{i+1}$)依赖于前一操作($step_i$)的计算结果，则这种类型的计算问题称为串行计算问题或相关模型计算问题。

定义 5(模型并行计算) 一个模型的计算的求解可以组织成多数据流(分块数据)及多计算步骤的加工方式进行，则这类模型问题的求解称为模型的并行计算。

定义 6(模型串行混合计算) 还有一类模型运算是属于上述两种情况的混合型。它既有串行计算部分，又有并行计算部分。这类问题称为串并行混合计算问题。

串并行计算问题是空间模型数值计算中的主要类型。对于这类问题的最终划分及取舍，主要考虑总的计算量中是并行计算的成分大，还是串行计算的成分大。解决此类问题的最终目的是充分挖掘它的并行计算能力，提高这类模型运算的并发性能。

3 网格环境下各类空间模型算法分析

3.1 遥感数据直方图并行算法(Parallel Histogram)

直方图是遥感数据图像增强的一种重要方法。直方图运算包括 2 大部分：直方图统计和直方图变换(线性、非线性变换)，直方图并行算法主要采用分块加工法，即把待处理的遥感图像数据进行分块处理。直方图不适合作分步加工，直方图计算是典型的串行计算问题。

并行算法步骤为：

(1)数据分块，即

$$\boldsymbol{A}=\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{11} & \ldots & a_{1n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{bmatrix}$$

划分为

$$\boldsymbol{A}=\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \ldots & \boldsymbol{A}_{1m} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \ldots & \boldsymbol{A}_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ \boldsymbol{A}_{m1} & \boldsymbol{A}_{m2} & \ldots & \boldsymbol{A}_{mm} \end{bmatrix}$$

其中

$$\boldsymbol{A}_{11}=\begin{bmatrix} a_{11} & \ldots & a_{1k} \\ \ldots & \ldots & \ldots \\ a_{k1} & \ldots & a_{kk} \end{bmatrix} \qquad \boldsymbol{A}_{12}=\begin{bmatrix} a_{1k+1} & \ldots & a_{12k} \\ \ldots & \ldots & \ldots \\ a_{k1} & \ldots. & a_{k2k} \end{bmatrix}$$

$$\boldsymbol{A}_{1m}=\begin{bmatrix} a_{1n-k+1} & \ldots & a_{1n} \\ \ldots & \ldots & \ldots \\ a_{k1} & \ldots & a_{kk} \end{bmatrix} \quad \boldsymbol{A}_{mm}=\begin{bmatrix} a_{n-kn-k} & \ldots & a_{1n} \\ \ldots & \ldots & \ldots \\ a_{n1} & \ldots & a_{nn} \end{bmatrix}$$

其中 $k=n\div m$，k 取整。

(2)在N个节点机上，分别对分块数据$A_1,\ldots,A_n$作直方图统计，即

$$A_1 \rightarrow Histogram \rightarrow \{C_1^{\,1}, C_2^{\,1}, ..., C_L^{\,1}\}$$
$$A_2 \rightarrow Histogram \rightarrow \{C_1^{\,2}, C_2^{\,2}, ..., C_L^{\,2}\}$$
$$A_3 \rightarrow Histogram \rightarrow \{C_1^{\,3}, C_2^{\,3}, ..., C_L^{\,3}\}$$
$$\ldots$$
$$A_n \rightarrow Histogram \rightarrow \{C_1^{\,n}, C_2^{\,n}, ..., C_L^{\,n}\}$$

直方图的灰度范围是 $\{1, 2, 3, ..., L\}$。

(3)对分块统计结果 $\{C_1^{\,1}, C_2^{\,1}, ..., C_L^{\,1}\}$, $\{C_1^{\,2}, C_2^{\,2}, ..., C_L^{\,2}\}$, $\{C_1^{\,n}, C_2^{\,n}, ..., C_L^{\,n}\}$ 作最后的直方图汇总，即直方图统计集合为 $\{C_1, C_2, ..., C_L\}$，其中

$$C_1=\sum_{n=1}^{N} C_1N\,,\ C_2=\sum_{n=1}^{N} C_2N\,,\ldots,\ C_L=\sum_{n=1}^{N} C_LN$$

对于直方图变换算法，原则上也是采用数据分块处理，其步骤(1)与“直方图统计”一致，但在步骤(2)中的算子不一，不是做统计而是作“线性”和“非线性变换”，即

$$A_1 \rightarrow LinearTransformation \mid noneLinearTransformation \rightarrow A_1'$$
$$A_2 \rightarrow LinearTransformation \mid noneLinearTransformation \rightarrow A_2'$$
$$\ldots$$
$$A_n \rightarrow LinearTransformation \mid noneLinearTransformation \rightarrow A_n'$$

最后整个变换结果矩阵 $\boldsymbol{A}'$为 $\{\boldsymbol{A}_1', \boldsymbol{A}_2', ..., \boldsymbol{A}_n'\}$。

直方图的其它应用模型，如直方图均衡化、直方图规定化，均可采用这种分块方式实现模型的并行计算，从行为规范和体系特征上符合模型网格计算标准。

直方图变换如采用串行计算方式，需要作N^2次“线性”或“非线性”变换，而数据分为M块，在M个节点机作直方图并行变换，平均做N^2/M次变换运算。如不考虑节点机的计算资源或者能力的差异，那么参与网格计算的“节点机”越多，计算的执行时间就越少。

3.2 矢量地图数据内插并行算法(Parallel Curving-Fitting)

定义 7(空间数据内插) 根据一组已知的离散或分区数据，按照某种数学关系推求出其它未知点或未知区域的数学过程。

GIS 应用在很多情况下，必须进行空间数据内插，例如数字高程模型的建立、区域边界的分析、空间趋势预测等。空间内插的特点是数据处理量大，算法复杂度较高，为了实现空间数据内插的网格计算，必须得到高效的内插并行算法。原则上内插的并行算法主要采用分块数据加工方法。

矢量图数据分块必需遵循如下规则：

规则 1 分块图幅划分完毕，各个图幅范围内的矢量图实体(点、线、面等)必须是个完整体，即分割时不能采用分割线(分区范围)硬性切分地图实体。规则的目的是为了避免图幅划分操作过程中裁剪(cut)地图实体。同时，图幅拼接时，也不用缝合(seam)地图实体，以减少无谓的运算。

规则 1 的执行过程是：如图 1 所示，图中的矢量图层数据将被分割为 2 个区(一区和二区)，按规则 1，所有与分割边线相交的实体，如 PolygonA、PolygonB 等，都是不能被裁剪

的。解决办法是按一定的判断规则把这些与分区界相交的地图实体划分给对应分区。

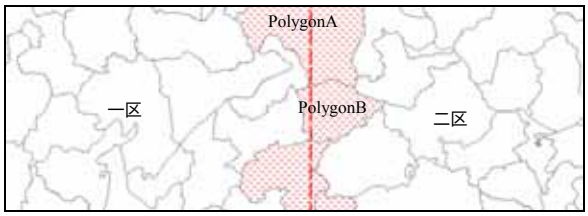

图 1 矢量图切分示意图

判断规则 1(最大面积法) 即多边形(objPoly)与哪个分区的相交面积最大，就把该地图实体(objPoly)划给哪个分区。

图 2 的多边形在“一区”的面积最大，依据判断规则 1，把该多边形实体“仲裁”给“一区”。这种判断规则适合对“面(Polygon)实体”的处理。

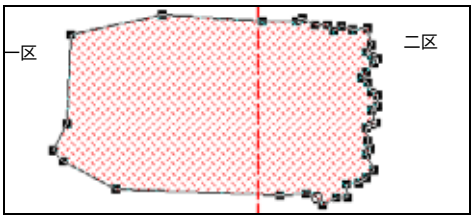

图 2 矢量实体切分示意图

判断规则 2(结点最多法) 对于“线实体(Polyline)”一类的地图实体，是无法计算分割面积的，这时可以根据线的结点(Nodes)数在哪个分区分布较多来判断它隶属哪个分区。同样，图 2 中的多边形落在“二区”中的结点数较多，依据判断规则 2，该多边形实体属于“二区”。

下面以矢量地图数据的整体拟合内插，分析其并行算法。

定义 8(整体拟合) 指内插模型是基于研究区域内的所有采样点的特征观测值建立的。

基于整体的拟合算法有：

(1)趋势面法。它是通过一个二元函数来逼近采样数据的整体变化。即

$$F(X,Y)=a_0+a_1X+a_2Y+a_3X^2+a_4XY+a_5Y^2$$

为求出系数$a_0, a_1, a_2, a_3, a_4, a_5, a$，最少需要 6 个均衡分布在图幅中的采样数据，生成一个六元一次方程组，解出方程即得到拟合系数。

(2)最小二乘法。它采用 N 个采样点，用一个二次多项式曲面来拟合，其矩阵表达为

$$F(X)=AX+S+R$$

其中，$F(X)$为曲面高程值列向量，A 为二次曲面的系数矩阵，X 为二次曲面参数列向量，S 为系统误差列向量，R 为偶然误差列向量。在计算出 A、S、R 等特征向量值后，可以分别采用预测法、滤波法、配置法等算法算出插值数据。

整体拟合的并行算法步骤为：

(1)取采样数据N，根据不同的内插方法，算出拟合系数，如趋势面法的$a_1,\ldots a_n$等；

(2)曲面数据分块。假设参与网格计算的节点机数量为 $n\times m$，曲面划分为 $n\times m$ 块，其分块定义为，设在 X-Y 平面上的插值区域 R：$a\le x\le b, c\le y\le d$ 上给定一个矩形块分割：

$$\Delta x: a=x_0<x_1<..<x_{i-1}<x_i..<x_n=b, |\overline{x_{i-1}x_i}|=\frac{1}{n}|\overline{ab}|$$

$$\Delta y: c=y_0<y_1<....<y_n=d, |\overline{y_{i-1}y_i}=\frac{1}{n}|\overline{cd}|$$

如图 3 所示，R_{ij}即是所分的内插块，该区域的内插数据将由对应的一个网格节点机完成；

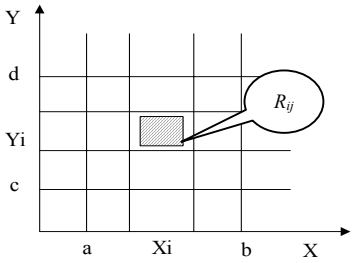

图 3 内插块示意图

(3)把(1)算出的内插系数散播到每个节点机上，节点机利用接收到的内插系数，算出对应内插区域R_{ij}上的内插点 数据；

(4)整合每个分块的内插数据，结束运算。

4 算法实例验证

验证数据：北京市气象局自动站 2005 年 8 月 19 日 15 时采集的雨量数据。

验证过程：

(1)在北京市整体范围内，利用自动站数据，“插值拟合”数据采集时刻的降水量等值图 T(图 4)；

(2)以 116°E 度为边界线，把北京市整体图幅划分为东西两部分；

(3)分别在两台节点机上做东西两部分范围内的“插值拟合”运算，算出各自的降水量等值图 W 及 E(图 4)。

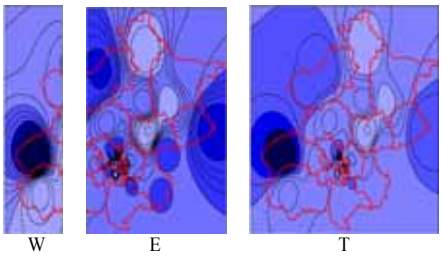

图 5 降雨量插值结果等值线

验证结果：如表 1 所示。

表 1 (W、E、T)结果图所用插值方法及运行结果参数表

执行节点机地址(Site)	图名	方法	插值网格行	插值网格列	执行时间(s)
172.168.8.3	W	距离倒数插值	1 000	500	9
172.168.8.18	E	距离倒数插值	1 000	500	9
172.168.8.200	T	距离倒数插值	1 000	1 000	19

验证分析：

对于数据在节点机之间的传输、分割、整合等操作所需要耗费的执行时间，数据的传输过程属于数据网格管理范畴，其耗费时间基本上可为一常数值$T_{传输}=\theta_1 f_1(n)$，数据分割及整合的耗时为：$T_{分割}=\theta_2 f_2(n)$，$T_{整合}=\theta_3 f_3(n)$，其中$\theta_1, \theta_2, \theta_3$为常数。并行计算时间($T_{并}$)与参与网格计算的节点机数目($N$)呈反比例关系，$T_{并}=T_{串}/N+T_{分割}+T_{整合}+T_{传输}+T_{其它}$。当$T_{串}\frac{N-1}{N}=T_{分割}+T_{整合}+T_{传输}+T_{其它}$，即数据传输、分割、整合等取值太大时，并行与串行计算执行时一样。所以，提高并行计算效果的关键是降低这部分执行时。本文的数据切分规则

(下转第 49 页)