

一种基于通用模型的遥感影像并行处理算法 ——以 PCA 融合为例

史园莉, 李海涛, 宋朝达, 韩颜顺

(中国测绘科学研究院, 北京 100830)

摘要: 空间应用需求对遥感数据处理的时间和精度提出新的要求, 为了高速、高效地解决应用需求, 提出一种基于通用模型的粗粒度遥感影像并行处理算法, 采用“分块驱动”和“影像处理链驱动”两种策略进行遥感影像的并行算法设计, 并对其进行 MPI 的实现。利用 PCA 融合算法在集群平台上进行并行性能测试, 分析结果表明该算法在集群系统上获得了良好的视觉效果和近似线性的加速比, 具有较好的扩展性和移植性。

关键词: 通用模型; 分块驱动; 影像处理链; PCA 融合; MPI

doi: 10.3969/j.issn.1000-3177.2011.03.003

中图分类号: TP79 **文献标识码:** A **文章编号:** 1000-3177(2011)115-0014-05

A Remote Sensing Data Parallel Processing Algorithm Based on General Model ——with an Example of PCA Fusion

SHI Yuan li, LI Hai tao, SONG Chao da, HAN Yan shun

(Chinese Academy of Surveying and Mapping, Beijing 100830)

Abstract: The amounts of remote sensing data of global coverage will be grown exponentially. To study and achieve a fast and effective processing of these digital data with high accuracies has become a critical problem in remote sensing. This paper presents one coarse grained method of parallel algorithm of remote sensing data processing based on general model. It inherits the modularization mechanism and uses both data partition and image processing chain to implement parallel processing with MPI library in cluster system. The experiments of PCA fusion algorithm is employed to test the efficiency and effectiveness of the parallel performance, which show that the new algorithm has the fine visual results, approximately linear speedup ratio and has fairly extensibility and transplantation.

Key words: general model; data partition; image processing chain; PCA fusion algorithm; MPI

1 引言

遥感是地球信息科学技术的一个重要组成部分, 从遥感影像中提取的信息对科学研究和社会发展有着重要的作用^[1]。然而, 面对数据量呈指数增长的全球遥感图像^[2], 传统的单处理器串程序的处理能力面临着巨大的挑战, 发展高性能遥感数据集群处理系统是快速高效处理遥感数据的必然趋势。相对于其他高性能计算系统, 集群是一种廉价有效的并行计算系统, 可以满足地球空间科学快速

计算的要求^[4]。本文以集群平台为基础, 提出一种基于通用模型的遥感影像并行处理算法, 采用“分块驱动”和“影像处理链驱动”两种策略进行遥感影像的并行算法设计, 并进行基于 MPI (Message Passing Interface) 的实现。因为遥感影像融合技术中 PCA (Principal Component Analysis) 算法在遥感图像处理领域具有应用广泛、运算量大、计算复杂度高^[5]的特点, 所以用来进行本文并行算法的性能测试与效率分析。

收稿日期: 2010-04-21 修订日期: 2010-05-17

基金项目: 国家科技支撑计划 (2008BA C34B07-04); 地理空间信息工程国家测绘局重点实验室基金 (6101002)。

作者简介: 史园莉 (1984-), 女, 研究方向为遥感数据高性能集群处理。

E mail: shiyuanl001@163.com

2 并行处理算法的通用模型

由于遥感影像数据量大、本身具有规整的几何结构,且大部分处理过程中空间上的彼此交互不多,因此可以将待处理的遥感影像分解为多个能够相对独立操作的影像块,让多个进程同时处理,提高数据处理速度。通常在处理大容量遥感数据时,采用粗粒度的并行处理,尽量减少并行执行进程之间的通信,提高并行执行的加速比,特别是在已有串行程序修改最小的情况下实现并行处理^[6-8]。

2.1 通用模型概述

遥感影像并行处理算法的通用模型是一种粗粒度的并行实现方法,具有两层含义:第一,对于用户要求的任何一种遥感影像处理功能,如影像滤波增强、影像融合、多源遥感影像配准、多景影像镶嵌、专题图制作等,均可以利用空间建模工具通过功能建模的方法实现,并将此过程保存成影像处理链文件供并行执行调用,具有通用性的特点;第二,并行执行时,将遥感影像分成若干影像块发送给多个进程同时执行影像处理链文件,具有快速性的特点。

本文采用主从式(Master Slave)并行编程模式^[9],利用“分块驱动”和“影像处理链驱动”两种策略协同进行遥感影像处理的并行化。主进程负责控制整个并行程序,根据“分块策略”将影像分成若干块并分配给各个从进程,负责收集从进程的处理结果,最后汇总得到最终结果;从进程接收主进程发来的影像块消息,根据“影像处理链”并行进行数据处理,处理结束后向主进程发回各自的处理结果。从进程处于一个循环的过程,直至分配的影像块全部处理完成并发送完毕。主进程分配完影像块后,处于接收状态,直至分配的影像块全部接收并写入输出文件为止。

2.2 影像处理链驱动策略

影像链方法将整个遥感过程看作一个相互联系的系统^[10]。本文将遥感影像处理的过程看作一个影像处理链,作为遥感影像并行处理的基本单位。它描述了生成某一特定影像产品或影像相关产品所需的一系列有序的处理过程,其中的每一个环节均是影像处理链中的一个模块,包括影像源文件、相关函数运算和生成产品的相关参数等。这一过程的具体信息保存为影像处理链说明文件,各个模块之间由影像处理系统自动分配的输入ID和输出ID相互连接,环环相扣,构成影像处理的整个过程。该文件是并行处理引擎生成相应产品的必选参数。根据该文件中关键字的值,并行处理引擎驱动集群系统中

各个从进程相互协作生产所需产品。影像处理链的具体工作过程如图1所示:

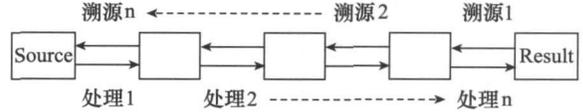


图1 影像处理链及其驱动模式

(1) 创建影像处理链。利用空间建模工具创建需要的影像处理链,并保存成影像处理链说明文件。创建影像处理链并生成影像处理链说明文件所需的时间很少,几乎等于人工操作的时间;

(2) 逆向溯源。从进程从分配的每块空白的输出影像阵列出发,按行列的顺序依次对每个输出像素点位反向查询影响其结果的相关参数,直至所有的参数都已知或查找到原始影像上相关位置的像素值;

(3) 正向处理。从进程按照影像处理链的操作求取相关参数,将原始影像相关位置的像素值按照正向处理顺序逐个进行计算,将得到的结果填回到输出空白影像块点阵中对应的像素点位上去。

2.3 分块驱动策略

分块机制是降低单次数据处理量的常用策略。在海量遥感影像处理中,数据并行是常用的处理模式^[11]。本文采用静态分配的方法将影像利用虚拟网格分成若干块并分配给各个从进程,如图2所示。

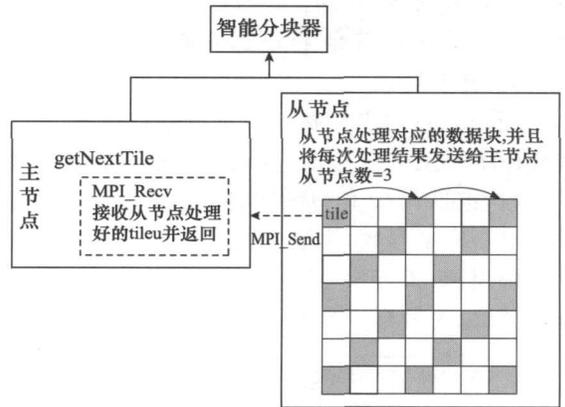


图2 影像并行处理智能分块机制

具体方法如下:

(1) 划分影像块并编号。根据影像及其投影获得影像边界,将影像边界包含的区域根据分块大小按从上到下、从左到右的顺序分成若干大小相同的矩形块;若右边界和下边界的剩余区域不能再分,则保留剩余块的大小。对划分好的各个影像块,按从左到右、从上到下的顺序进行编号,为1、2、3...i...;

(2) 获得所要处理的影像块。每个从进程都有相同的影像块划分方法,根据分块策略获得所要处

理的影像块对应的编号。假设总共有 P 个从进程, 第 p ($1 \leq p \leq P$) 个从进程所要处理的影像区域对应的影像块的编号是:

$$i = p + n \times P$$

式中 $n = 0, 1, 2, \dots$;

(3) 分块处理。从进程按影像块的编号从小到大依次处理各个影像块, 并将处理结果发送给主进程, 直至将分配的影像块全部处理完成为止。

3 PCA 融合并行实现

遥感影像融合是指将多源遥感图像按照一定的算法, 在规定的地理坐标系, 生成新的图像的过程。全色图像一般具有较高空间分辨率, 多光谱图像信息较丰富, 为提高多光谱图像的空间分辨率, 可以将全色图像融合进多光谱图像。通过图像融合既可以提高多光谱图像空间分辨率, 又保留其多光谱特性^[12]。

基于主成分分析 (PCA) 的数据融合首先对多光谱图像进行主成分变换, 变换后的第一主分量含有变换前各波段图像的信息, 而各波段中其余对应的部分, 被分配到变换后的其他波段; 其次将高分辨率图像和第一主分量进行直方图匹配, 使高分辨率图像与第一主分量图像有相似的均值和方差; 最后用直方图匹配后的高分辨率图像代替主分量中的第一主分量和其余主分量一起进行主分量逆变换, 得到融合影像^[12]。

当并行处理引擎执行影像处理链说明文件时, 首先由下到上进行参数追溯, 查找处理过程中需要的参数, 包括两种类型的参数: 影像参数和常量参数。对于常量参数, 需要预先求出放入内存供后续处理调用; 对于影像参数, 用于分块并行处理。图 3 描述了 PCA 融合中参数的传递过程。

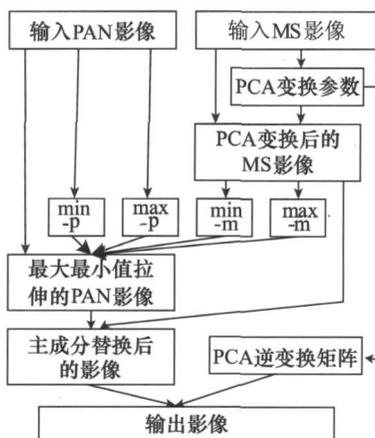


图3 PCA 融合参数的传递过程

(1) 求解常量参数。根据“分块驱动”策略将输入影像进行分块, 分发给从进程求解影像处理链说明文件所需要的各个常量参数。由于采用分块驱动的方式, 每个从进程处理不同的影像块, 对于需要求取全局结果, 如 PCA 正变换参数、PCA 逆变换参数、最大最小值拉伸参数(全色影像的最大最小值、多光谱影像 PCA 正变换后第一主分量的最大最小值)的运算来说, 构建一个进程组包括所有从进程的通信器 (communicator)。利用 MPI 的归约操作汇总位于相同或不同处理机上的多个进程的局部结果, 得到所需的新结果, 存储在所有进程的变量中。

(2) 求解影像块参数。根据输入影像及其投影参数确定输出影像的边界, 并根据“分块驱动”策略对输出影像进行分块。每个从进程确定所要处理的影像块的编号。不同从进程分别根据对应的输出影像块的范围, 在影像处理链说明文件中逆向查找所需的各个参数, 直至查找到所需的输入影像块; 各个输入影像块按照影像处理链说明文件正向处理顺序分别进行处理, 得到输出影像块的结果, 并将其发送给主进程, 如图 4 所示。在此过程中, 不同从进程调用内存中相同的常量参数, 处理不同的影像块, 进程之间不相互干预, 完全处于独立的状态; 一个影像块处理结束后将结果发送给主进程进行存储。然后进行下一个影像块的计算, 直至分发的输出影像块全部处理结束为止。

(3) 写入输出文件。主进程不断接收从进程发送的结果影像块, 并将其写入输出文件, 直至将分发的影像块全部回收完毕并成功写入文件为止。

(4) 结束。

4 性能测试与分析

根据上述算法分析和设计, 在中国测绘科学研究院高性能集群处理系统平台上进行实现, 利用 TM 影像进行了测试与性能分析。测试平台的配置为: Sun Fire X4600 高性能集群处理系统, 它由 4 个双核处理器组成, 每个处理器主频 2.8 GHz/1MB Cache; 系统包括 8 块内存, 每块容量为 2GB。X4600 同时兼做管理节点和计算节点。遥感影像存储在机柜式磁盘阵列存储器 HSD AMS 1000 上。X4600 和 HSD AMS 1000 通过总带宽上限为 128GB/s 的 FC 光纤存储子网连接。软件环境为 Redhat Linux Enterprise Server 5.0 操作系统, 基于 LAM/MPI 消息传递并行库。

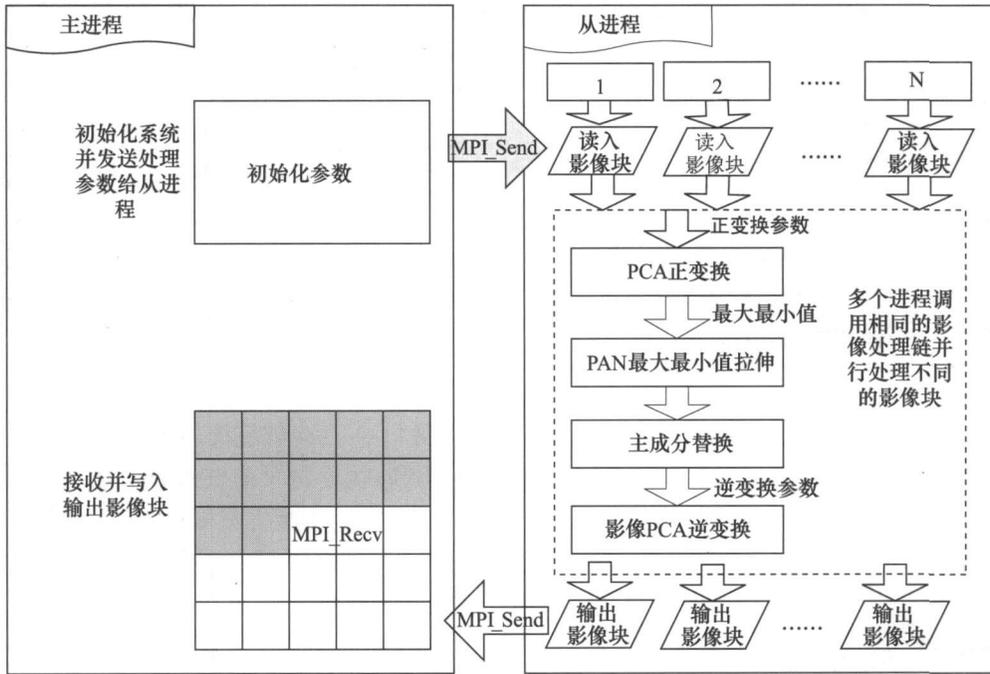


图4 PCA融合并行执行过程

4.1 实验结果

图5给出了在集群平台上进行PCA融合实验的处理结果,其中图5(a)为输入的pan影像(9968 × 9133),图5(b)为输入的MS影像(9968 × 9133),图5(c)为PCA融合影像结果。同时,利用较小影像(3905 × 3109)进行了相同的测试。表1列出了两种数据在集群系统上分别利用1~8个进程进行PCA融合所耗时间,图6描述了执行时间变化折线图;表2分别列出了两种数据PCA融合在不同进程的加速比,图7描述了并行执行的加速比变化折线图。其中影像分块大小为64 × 64。与传统串行算法相比,将9968 × 9133影像在普通PC机上利用传统遥感影像处理软件系统ERDAS IMAGE进行相同影像的实验所耗时间为900s,3905 × 3109影像所耗时间为110s。

表1 影像PCA融合执行时间(单位:秒)

| 进程数 | 影像大小 | |
|-----|-------------|-------------|
| | 9968 × 9133 | 3905 × 3109 |
| 1 | 652 | 105 |
| 2 | 637 | 102 |
| 3 | 326 | 52 |
| 4 | 219 | 36 |
| 5 | 164 | 28 |
| 6 | 134 | 22 |
| 7 | 111 | 19 |
| 8 | 99 | 18 |

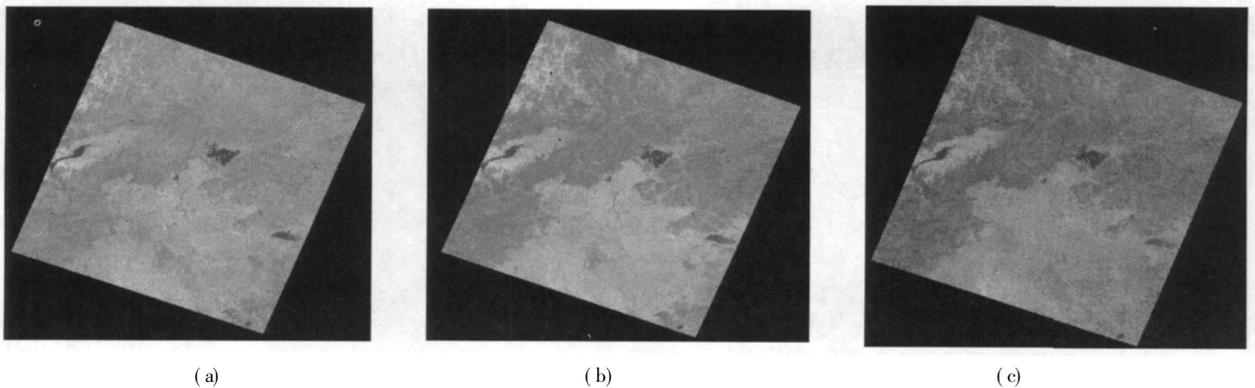


图5 PCA融合并行处理结果

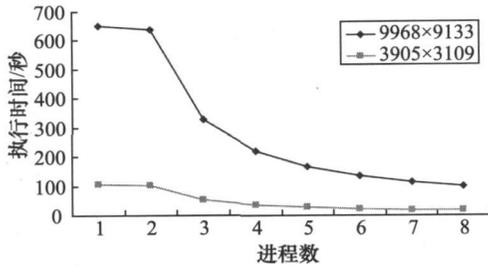


图6 不同大小影像 PCA 融合执行时间变化

表2 影像 PCA 融合加速比

| 进程数 \ 影像大小 | 9968× 9133 | 3905 × 3109 |
|------------|------------|-------------|
| 1 | 1.00 | 1 |
| 2 | 1.02 | 1.03 |
| 3 | 2.00 | 2.02 |
| 4 | 3.00 | 2.92 |
| 5 | 4.00 | 3.75 |
| 6 | 4.87 | 4.78 |
| 7 | 5.87 | 5.53 |
| 8 | 6.59 | 5.83 |

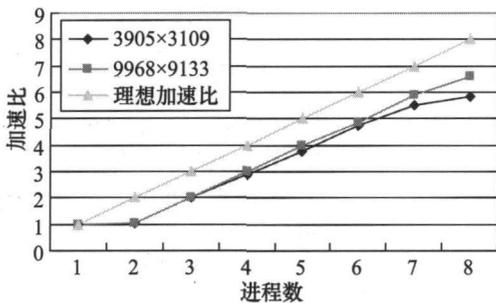


图7 影像 PCA 融合加速比图

4.2 性能分析

由于本文并行算法采用主从模式的程序结构。当进程数等于1时,执行串行计算,此时由一个进程负责整个的数据处理全过程。当进程数大于1时,执行并行计算,主进程负责整个并行程序的控制,分配数据和计算任务给从进程,收集从进程计算的结果写入输出文件;而从进程负责完成分配给它的数

据的处理和计算任务。所以利用两个进程执行并行计算时,主进程负责任务的分发,只有一个从进程进行数据的处理和计算,虽然写入速度比串行执行快,但进程之间的通信增加了执行时间,所以两个进程执行时间和一个进程的执行时间几乎相同。

定性方面,由图5可见,利用本文的算法对遥感影像进行PCA融合处理取得了良好的视觉效果,可以看出,在集群平台上进行PCA融合算法并行化的有效性。

定量方面,较串行算法,利用并行算法在集群上实现PCA融合较传统的串行算法,在性能上有明显的提高,执行时间减少的幅度随着数据规模的扩大而增大。从加速比图可以看出,该并行算法的加速比与进程数基本成线性关系,因为影响并行算法性能的主要因素是进程间的同步和进程间的通信,所以随着处理进程数的增加,并行算法中由于通信时间占计算时间的比例上升,导致处理进程空闲和通信的增多,因而加速比略有减小。随着数据量的增大,加速比曲线越来越接近理想直线,这说明随着问题规模的增大,计算时间的减小幅度会大大高于通信世界的增加幅度,加速比曲线会趋近于直线,表明该算法适合数据量大的计算,具有良好的可扩展性。

5 结束语

本文结合传统遥感处理系统的空间建模技术以及用于并行化的消息传递库,提出了一种基于通用模型的遥感影像并行处理算法,采用“分块驱动”和“影像处理链驱动”两种策略,实现了集群系统上遥感影像的并行处理。该算法保留了传统遥感处理系统的简单可操作性,提高了该并行系统的扩展性。测试表明,该算法在集群系统上效率很高,具有近似线性加速比,具有良好的扩展性,特别适于处理海量遥感影像,完全满足灾害响应等应急情况实时或近实时需要。今后将着重检验影像分块大小、I/O速度以及任务动态分配对并行效率的影响,从而进一步提高数据处理速度。

参考文献

- [1] Ying Luo, Yong Xue. Remote sensing information processing grid node with loose computing parallel structure[J]. LNCS, 2006(3991): 876- 879.
- [2] 周海芳,等. 遥感图像自动配准的串行与并行策略研究[J]. 国防科技大学学报, 2004(2): 56- 61.
- [3] Rhonda D. Phillips. Hybrid image classification and parameter selection using a shared memory parallel algorithm[J]. Computers & Geosciences, 2007(33): 875- 897.
- [4] Plaza, A. J., Chang, C.-I. High Performance Computing in Remote Sensing[M]. Chapman & Hall/CRC, London, 2008: 9~ 10.
- [5] 胡冰. 遥感图像融合并行算法的研究与实现[D]. 长沙: 国防科学技术大学, 2006. (下转第120页)

通过卫星或网络直接传输到数据处理与发射中心。

(2) 水污染遥感监测

系统利用高光谱遥感数据和陆地卫星数据,与来自地面的数据相结合,可以监测淮河流域大型企业的排污状况及江河水污染状态,从整体上反演水体的情况。

(3) 水污染模拟预报预警

运用建立的模型可以模拟预测联防调度下、突发污染事故下和水闸联合调度下 3 个不同方案的水质水量联合调度效果,对各种污染源排放条件变化的水质响应进行预报,给出污染团到达的时间和最高浓度值,对污染团的运行过程提前发出预警。

(4) 水污染可视化显示与辅助决策

系统将区域 DEM、卫星影像、三维信息及数值模拟成果多要素紧密融合,以表格、图形、报表、文字、多媒体等方式表示出来,使相关部门或决策者能够更加直观清晰地了解事故发生地的各类信息;采用静态、动态结合的数值模拟成果展示技术,模拟出污染发生、发展趋势的三维真实场景,通过人为干预以减少污染发生的可能性,并及时制定出污染应急预案与防治对策。

本文研建的流域水污染模拟预报和应急处理系统已在淮河流域得到了实际应用。淮河流域 2004 年颍河污染物的下泄过程模拟如图 10 所示。

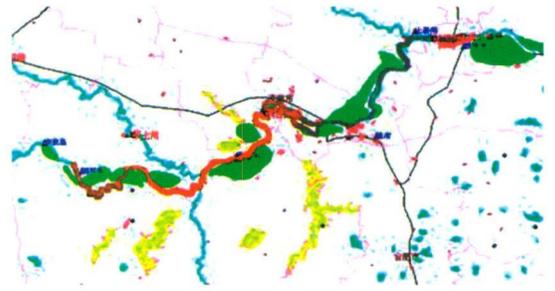


图 10 颍河污染物的下泄过程

5 结束语

流域水污染模拟预报和应急处理涉及到水文、地理、生态、环境等多个方面,本文在集成应用 GPS、GIS、RS、三维仿真、水动力学等方面的多种技术条件下,建立了淮河流域的水污染模拟仿真模型,实现了自动化测报、评估、预警预报、应急处理等一体化的综合应用。本文所建立的水污染模拟预测及应急系统通过在淮河流域的实际应用,有效提高了应对各种突发性水污染问题的应急决策能力。

参考文献

- [1] 姜卫星,张海平,陈玲,林卫青,黄浦江突发性水污染事故预报预警系统[C]//中国环境水力学:第七届全国环境水力学学术研讨会论文集.北京:中国水利水电出版社,2006:131-136.
- [2] 施益强.海上溢油事故应急反应系统的框架研究[J].灾害学,2002(4):87-91.
- [3] 熊德琪.大连海域溢油应急预报信息系统及其应用[J].交通环保,2002(6):5-7.
- [4] 孙俊.基于 OILMAP 的中国舟山港溢油管理信息系统[J].计算机仿真,2002(7):76-78.
- [5] 北京灵图软件公司.VRMp3.0 企业版用户手册[Z].2004.
- [6] 陈蓓青,等.三峡水库突发性水污染事件应急系统的开发[J].人民长江,2006,37(5):89-91.
- [7] 冉圣宏,陈吉宁.区域水环境污染预警系统的建立[J].上海环境科学,2002,21(9):541-544.
- [8] 陈惠君,唐允吉.广西桂江水质预警预报信息系统的研究[J].陕西水力发电,1997,13(2):50-52.
- [9] 崔伟中,刘晨.松花江和沱江等重大水污染事件的反思[J].水资源保护,2006,22(1):1-4.
- [10] 闫志刚,盛业华,左金鑫.3S 技术及其在环境信息管理系统中的应用[J].测绘通报,2001(增刊):17-20.
- [11] 司鹤,毕海普,李晓红.基于虚拟技术的三峡库区应急救援系统[J].中国安全科学学报,2007,17(4):21-24.
- [12] 李茂堂,姜永生,等著.基于 3S 的江河水污染监测与应用[M].现代教育出版社,2009.

(上接第 18 页)

- [6] Serot, J., Ginjac, D., Derutin, J. P.. SKiPPER a skeleton based programming environment for image processing applications[C]//Fifth International Conference on Parallel Computing Technologies. 1999.
- [7] Valencia, D., Plaza, A. Martinez P., Plaza, J.. Parallel processing of high dimensional remote sensing images using cluster computer architectures[J]. International Journal of Computers and Their Applications, 2007, 14(1):23-34.
- [8] Plaza, A. Valencia, D., Plaza, J., Martinez, P.. Commodity cluster based parallel processing of hyperspectral imagery[J]. Journal of Parallel and Distributed Computing, 2006, 66(3):345-348.
- [9] 陈国良,等.并行算法实践[M].北京:高等教育出版社,2004:61.
- [10] Schott, John R. Remote Sensing: The Image Chain Approach[M]. Oxford University Press, Oxford, 2007:17-19.
- [11] Li, G. Q., Liu, D. SH.. Key technologies research on building a cluster based parallel computing system for remote sensing[J]. LNCS, 2005(3516), 484-491.
- [12] 孙家桢.遥感原理与应用[M].武汉:武汉大学出版社,2003:163-164.