第27卷 第1期 干 旱 区 资 源 与 环 境 Vol. 27 No. 1
2013年1月 Journal of Arid Land Resources and Environment Jan. 2013

文章编号:1003 - 7578(2013)01 - 130 - 07

基于 MapReduce 模型的生物量遥感并行反演方法研究*

付天新[1,2],刘正军[1],闫浩文[2]
(1. 中国测绘科学研究院,北京 100830;2. 兰州交通大学 数理与软件工程学院,兰州 730070)

提 要: MapReduce 模型是一种基于云计算平台下新型的并行编程模型。文中将 MapReduce 并行编程模型应用到遥感影像并行化处理中,以 2005 - 2009 年 5a 生长季期(5 - 10 月) MODIS13Q1 数据产品为数据源,对青海省三江源地区的生物量(草地总生物量和可食草量)进行并行化反演,研究基于该模型的生物量遥感并行反演方法。实验分析结果表明:基于该模型的并行生物量遥感反演结果与经过精度验证的串行反演结果一致,并行化反演结果准确、可信;并行化反演效率较串行化反演效率有大幅提高,并随着计算节点的增加,并行效率不断提高。

关键词: 云计算;MapReduce 模型;生物量;并行计算
中图分类号: TP79;S812 **文献标识码:** A

自 20 世纪 60 年代国际生物圈(International Geosphere - Biosphere Program,IGBP)[1] 实施以来,生态系统的生物量研究一直是生态学研究的重要方向,遥感手段也越来越广泛的应用于生物量监测[2],如冯险峰利用 TM 数据综合地学、生态学信息,建立了中国陆地生物量的遥感动态监测模型[3];徐斌等应用 MODIS 相关产品及地面数据对草地生产力进行了估算研究[4];李昌凌[5]等应用 GMMS NDVI 数据对地表植被生物量进行趋势分析和空间分布特征研究。但随着遥感技术的发展,应用于生物量监测的遥感数据量和数据类型剧增,不同应用需求对遥感数据的处理速度和效率有了更高的要求,这使得现有的遥感数据处理系统面临着严峻挑战。

为了提高海量遥感数据的处理速度,满足生物量监测对遥感数据处理效率的要求,国内外学者开始把并行计算技术引入到遥感数据处理中,并产生了基于多线程、集群、CPU、GPU 等众多并行计算方法[6-9]。MapReduce 模型[10,11] 是一种基于云计算平台下新型的并行模型,由 Google 公司于 2004 年提出并首先应用于大型集群系统中,用于 TB 级数据集的并行运算,并以其文件读写、容错机制、数据组织与管理、数据传输、网络带宽以及简便的编程模式迅速得到人们的青睐。

目前应用 MapReduce 编程模型进行生物量遥感并行反演的研究并不多见,论文旨在分析 MapReduce 并行编程模型的执行机理,参考遥感数据反演的理论与技术方法,提出了一种基于 MapReduce 模型的生物量遥感并行反演方法,并以青海三江源区草地总生物量和可食草量两个生态参数反演为例,评价该模型下并行化反演的性能与反演效率。

1 MapReduce 模型介绍

MapReduce 模型(简称 MR 模型)是一个基于云计算平台下的并行模型,它能够在大型集群系统上以并行的方式处理海量数据,性能稳定,容错性高。它的基本思想是将要执行的问题分解为 Map(映射)和 Reduce(规约)两步操作,Map 是把一组数据一对一映射为另外一组数据,映射规则由一个函数指定,Reduce 对映射后的结果进行规约与合并。MR 模型的核心是 Map 和 Reduce 两个函数,这两个函数由用户负责实现。基于 MR 模型运转在 < key,value > 键值对上,模型把作业的输入看成一组 < key,value > 键值对,同样也产生一组 < key,value > 键值对作为作业的输出。图 2 显示了一个 MR 模型的计算流程,一个 MR

* 收稿日期:2011 - 12 - 13;修回日期:2011 - 12 - 29。
基金项目:国家级基础测绘项目课题"三江源区生态环境遥感动态监测地理信息系统"资助。
作者简介:付天新(1986.6 ~),男,内蒙古人,硕士研究生,主要从事遥感与 GIS 应用研究。Email:fu_tianxin@163.com

工作(job) 通常是先将输入的数据集划分为若干独立的数据块，根据具体的计算要求，在 Map 中以完全并行的方式处理。框架对 Map 任务输出的结果进行分区与排序，并将分区与排序后的结果作为 Reduce 的输入，根据 key 值进行数据块规约和合并，一个 MR 过程结束。

以统计单词出现次数为例，MR 模型按照行首先把图 1 数据集切分为 < Hello World Bye World > 和 < Hello Hadoop Goodbye Hadoop > 两部分，并将切分的数据块映射为 < key，value > 键值对，其中 key 为单词名称，value 为单词出现的个数。然后将切分的行分发给 Map 进行并行化的字数统计，分别得到{ < Hello，1 > ，< Bye，1 > ，< World，2 > } 和{ < Hello，1 > ，< Goodbye，1 > ，< Hadoop，2 > } 两个 Map 输出的中间结果，然后对中间结果进行分区与排序，并将具有相同 key 值的中间结果组成数据集列表，即{ < Hello，1 > ，< Hello，1 > } 、{ < Bye，1 > } 、{ < World，2 > } 、{ < Hadoop，2 > } 、{ < Goodbye，1 > } 五组数据子集，将这五个数据子集分发给 Reduce 进行相同单词出现次数求和，最终输出{ < Hello，2 > 、< Bye，1 > 、< World，2 > 、< World，2 > 、< Hadoop，2 > } ，得到输入数据中各个单词出现的次数，一个 MR 过程结束。

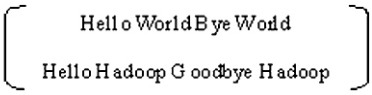


图 1　单词出现次数统计模拟数据集

Fig. 1 Word frequencies statistic model dataset

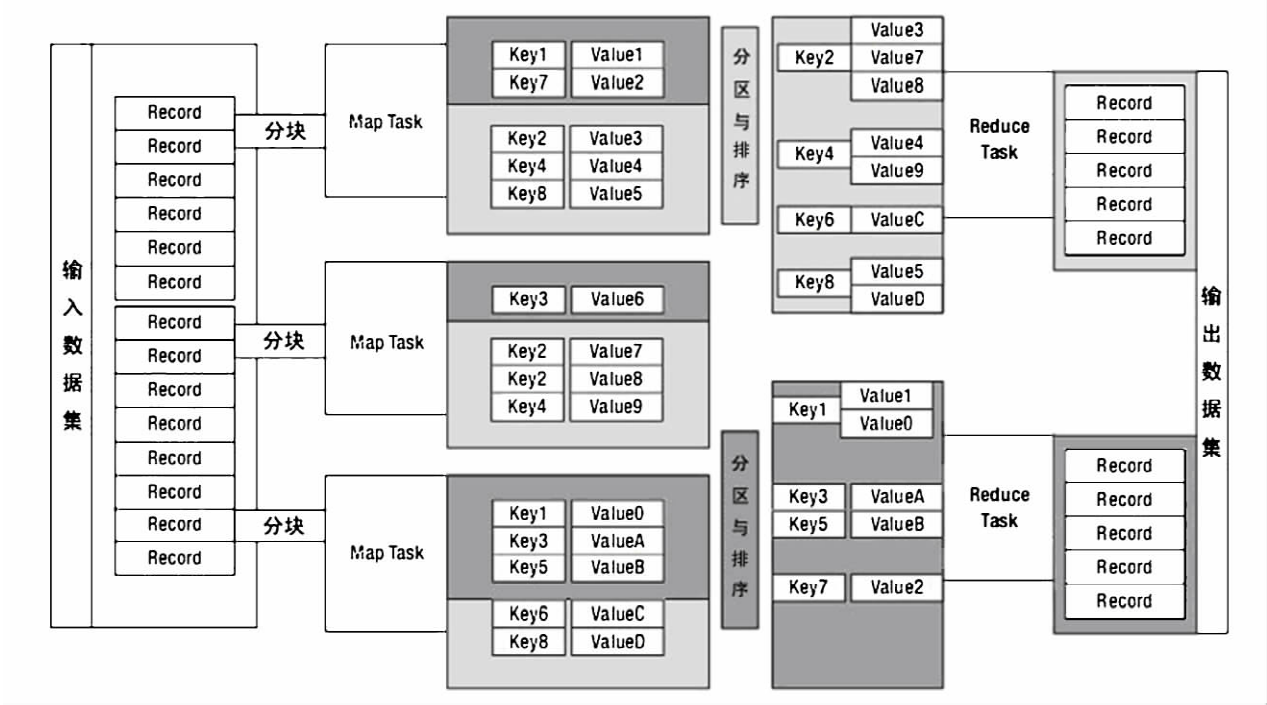


图 2　MR 模型计算流程

Fig. 2 MR modeling process

2　技术流程与方法

2.1　数据块划分

在海量遥感数据并行计算中，遥感数据块的划分是并行化处理的重要部分，数据块划分的方式、数据分块的大小与并行计算效率有着密切联系。目前，遥感图像数据分块的策略主要有：水平条带、竖直条带、矩形块以及不规则划分四种[12,13](图 3) ，由于拟选研究区域的数据为单波段影像，单景影像数据量较小，因此选择水平均匀条带切分每景影像，采用默认的数据分块大小(64M) 作为拟选研究区域影像集的划分方法，研究海量遥感数据反演方法。

遥感数据并行化反演需要数据块的属性信息，以保证分块数据计算、中间反演结果分区与排序、影像块合并、影像头文件生成等各个计算过程正常执行。根据并行化反演过程对影像块的属性信息的需求，对影像块追加数据分块信息、地理参考信息、影像信息，数据块元数据(表 1) 。

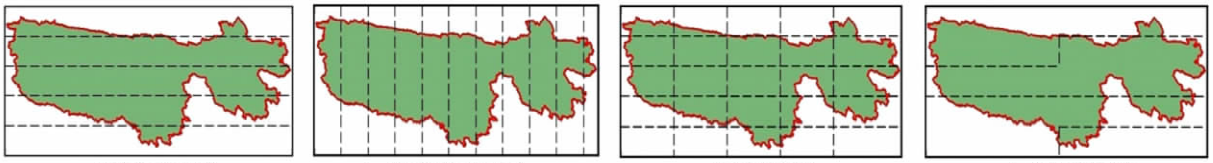


图3 数据分块方式

Fig. 3 Block partitioning method

表1 数据块元数据信息

Tab. 1 Data blocks metadata information

类别	属性与描述					
影像信息	影像类型	年份	月份	反演类型	–	–
数据分块信息	行号	行间距	每景影像数据块个数	影像高度	影像宽度	–
地理参考信息	地图单元中的一个像素在X方向上的X分辨率尺度	平移量	旋转量(角度)	地图单元中的一个像素在Y方向上的Y分辨率尺度的负值	像素(1,1)左上方的X地理坐标	像素(1,1)左上方的Y地理坐标

2.2. 并行化反演

采用HDFS分布式文件系统作为海量遥感数据的文件管理系统,提取单景或多景基于时间序列的遥感影像。进行遥感数据的有效性规则检查,有效性规则定义为数据一致性、数据有效性、数据完备性以及数据格式支持等规则。通过数据有效性规则检查的数据进入MR并行框架,对遥感数据进行分块,并将每个数据块一一映射成<key,Value>键值对。采用基于MODIS数据产品的3类植被指数,以便获得准确的生物量遥感反演模型,指数描述如下:

(1)归一化植被指数(Normolized Differential Vegetation Index,NDVI)。NDVI是最早提出的基于物理知识,将电磁辐射、大气、植被覆盖的相关作用结合在一起的植被指数,也是目前应用最广泛的植被指数,但容易受大气、土壤背景的影响,对研究植被覆盖稀疏的地区效果较差。其表达式为:

$$NDVI = \frac{P_{NIR} - P_R}{P_{NIR} + P_R} \tag{1}$$

式中:P_{NIR}和P_R分别为近红外波段和红光波段的反射率。

(2)修改的土壤调整植被指数(Modified Soil Adjusting Vegetation Index,MSAVI)。MSAVI是QiJ等[14]对土地调整植被指数(Soil Adjusting Vegetation Index,SAVI)改进后提出的植被指数。MSAVI不需要有研究区的先验知识,并且可根据实际情况对土壤调整因子进行调整,对每一点来说,土壤调整因子的取值都是最佳值,尽管MSAVI比NDVI要精确,但依然不能消除土壤背景的影像,并对大气有一定的敏感性。其表达式为:

$$MSAVI = \frac{2^* P_{NIR} + 1 - \sqrt{(2^* P_{NIR} + 1)^2 - 8^* (P_{NIR} - P_R)}}{2} \tag{2}$$

式中:P_{NIR}和P_R分别为近红外波段和红光波段的反射率。

(3)增强型植被(Enhanced Vegetation Index,EVI)3类植被指数。EVI是Huete等[15]对NDVI改进后提出的植被指数,较好的消除土壤背景、大气对其的影响。其表达式为:

$$EVI = 2.5^* \frac{P_{NIR} - P_R}{P_{NIR} + C_1 P_R + C_2 P_B + L} \tag{3}$$

式中:L为土壤调整因子,一般取1;参数C_1和C_2分别为红光和蓝光的大气修正参数,一般分别取6.0和-7.5;P_{NIR}、P_R和P_B分别为近红外波段、红光波段和蓝光波段的反射率。

利用2005-2009年每年8月份的野外采样数据和MODIS数据产品建立草地总生物量、可食草量与NDVI、MSAVI、EVI之间的3种关系模型(线性、乘幂、指数),进行比较,选取相关性较好的关系模型作为拟选区域的生物量反演模型[16],选取模型(表2)。

Map中集成生物量反演模型,并根据数据块的元数据信息,确定Map输出中间结果的key值。框架对集群中的计算机节点分发Map任务,进行拟选区域生物量的并行化反演。Map输出计算后的中间结果,该中间结果的key值由数据块元信息生成,经过中间结果的分区与排序。框架将具有相同key值的中间

计算结果作为 Reduce 的输入，再次根据数据块元信息和 Reduce 输入数据集序列中的 key 值，将同一景影像反演后数据块合并。最终形成反演后的遥感影像，并将反演后的影像存储到 HDFS 文件系统中。

表 2　草地总生物量(TBIO)、可食草量(EBIO)并行化反演模型

Tab. 2 The parallel retrieval model for total biomass of grassland (TBIO) and edible grass (EBIO)

年份	类型		回归模型		R^2		F 值		Sig.	样本数
	TBIO	EBIO	TBIO	EBIO	TBIO	EBIO	TBIO	EBIO		
2005	乘幂	指数	$y=9774.1x^{1.5651}$	$y=210.08e^{3.631x}$	0.7255	0.631	153.264	92.308	0	60
2006	乘幂	指数	$y=9774.1x^{1.5651}$	$y=210.08e^{3.631x}$	0.7255	0.631	153.264	92.308	0	60
2007	指数	乘幂	$y=191.01e^{5.5238x}$	$y=16047x^{2.509}$	0.7555	0.788	142.168	167.433	0	48
2008	乘幂	乘幂	$y=10570x^{1.7396}$	$y=7886.7x^{1.628}$	0.7525	0.638	191.502	110.867	0	65
2009	乘幂	指数	$y=12477x^{1.7222}$	$y=316.22e^{4.433x}$	0.713	0.654	183.8	140.124	0	76

3　实验与结果分析

3.1　研究区域与数据源

三江源区位于我国的西部、青藏高原的腹地、青海省南部，北纬 31°39′～36°12′，东经 89°45′～102°23′，是长江、黄河和澜沧江的源头汇水区。行政区域涉及包括玉树、果洛、海南、黄南四个藏族自治州的 16 个县和格尔木市的唐古拉乡，总面积 30.25 万 km^2，约占青海省总面积的 43%。该区以山地地貌为主，地势高耸、地形复杂，平均海拔 4200m。区内气候属青藏高原气候系统，为典型的高原大陆性气候，冷热两季交替，干旱两季分明，年温差小，日温差大，日照时间长。三江源区具有独特而典型的高寒生态系统，为中亚高原高寒环境和世界高寒草原的典型代表。高山草甸和高寒草原是主要植被类型，高山冰缘植被也有较大面积分布。

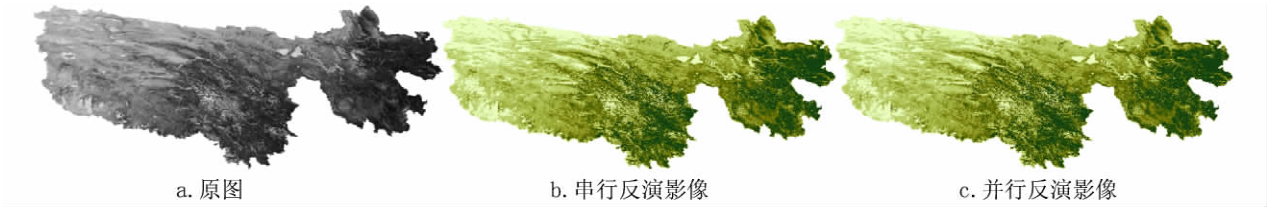
文中所采用的数据源是美国 EOS－MODIS 地球观测系统中陆地专题产品 MODIS13Q1 中 2005－2009 五年 5－10 月获得的植被指数数据产品，该数据源的合成周期为 16 天，分辨率为 250m，共 135 景，总数据量约为 2.5GB。为了能够准确的进行三江源区生物量反演，利用 MRT(MODIS Reprojection Tools)软件将 Sinusoidal 投影转换为地理经纬度坐标，基准为 WGS－84。

3.2　实验环境

应用课题组的软硬件环境，在千兆局域网、Linux Ubuntu10.10 系统、Java 环境、JDK1.60.18 和 Hadoop0.20.2 开源包的支持下搭建 3－4 台 PC 机组成的海量遥感数据反演的集群系统。包括 1 个主节点和 2－3 个计算节点。其中主节点 PC 机配置为：CPU：Intel E6750；内存：4G DDR3；网卡：100Mbps 以太网；硬盘：500G SATA。计算节点配置为：CPU：Intel E6750；内存：2G DDR3；网卡：100Mbps 以太网；硬盘：250G SATA。

3.3　结果分析

应用 MR 并行框架和上述的反演方法，得到三江源区 2005－2009 年各年中 5 月－10 月份生物量反演影像。图 4(a)为预处理后 2009 年 8 月份影像，图 4(b)为串行程序反演后三江源区草地总生物量反演影像，图 4(c)为应用 MR 模型并行化反演三江源区草地总生物量反演影像。由图可见，并行化反演后的影像与串行反演的影像一致；同时，经比较两个反演结果的直方图一致，并行反演结果可信。



a. 原图　　b. 串行反演影像　　c. 并行反演影像

图 4　并行与串行反演结果对比

Fig. 4 Comparison of parallel and serial retrieval result

为了验证基于 MR 模型生物量并行化反演的效率，将串行反演和并行化反演进行对比分析(表 3 和表 4)，表 3 和表 4 分别表示草地总生物量和可食草量在串行化反演、两个计算节点并行化和 3 个计算节点并行化反演时间对照。通过表 3 和表 4 可以看出，由于基于 MR 模型的并行计算数据分块是在主节点中进行，数据分块完成后，通过主节点向各个计算节点分发任务，在各个计算节点上进行并行化计算。因此，数

据分块速度并没有随计算节点的增加而发生明显变化。

表3 草地总生物量(TBIO)并行化反演时间(S)对照表

Tab. 3 Time table for total grassland biomass of (TBIO) parallel inversion

年份	单线程串行反演时间	2 NODES 并行时间				3 NODES 并行时间			
		数据分块时间	模型反演时间	反演所占时间百分比(%)	并行反演加速比	数据分块时间	模型反演时间	反演所占时间百分比(%)	并行反演加速比
2005	322	26	204	87.25	1.58	25	160	84.38	2.01
2006	315	23	190	87.89	1.66	26	169	84.62	1.86
2007	306	28	251	88.84	1.22	27	205	86.83	1.49
2008	318	25	201	87.56	1.58	26	158	83.54	2.01
2009	316	27	205	86.83	1.54	25	157	84.08	2.01

表4 可食草量(EBIO)并行化反演时间(S)对照表

Tab. 4 Time table for edible grassland biomass of (EBIO) parallel inversion

年份	单线程串行反演时间	2 NODES 并行时间				3 NODES 并行时间			
		数据分块时间	模型反演时间	反演所占时间百分比(%)	并行反演加速比	数据分块时间	模型反演时间	反演所占时间百分比(%)	并行反演加速比
2005	315	26	210	87.62	1.50	25	162	84.57	1.94
2006	308	23	205	88.78	1.50	26	160	83.75	1.93
2007	317	28	217	87.10	1.46	27	162	83.33	1.96
2008	315	25	204	87.75	1.54	26	156	83.33	2.02
2009	306	27	208	87.02	1.47	25	156	83.97	1.96

综合草地总生物量和可食草量并行化反演,在两个计算节点并行化反演中,并行化反演所占MR框架运行时间百分比在(86.83%,88.84%)之间,并行化反演所占MR框架运行时间百分比均值为87.66%,并行化反演加速比在(1.22,1.66)之间,并行化反演加速比均值为1.51。在3个计算节点并行化反演中,并行化反演所占MR框架运行时间百分比均在(83.54%,86.83%),并行化反演所占MR框架运行时间百分比均值为84.24%,并行化反演加速比在(1.49,2.02)之间,并行化反演加速比均值为1.92。随着计算节点的增加,并行化反演所占MR框架运行时间百分比均值由87.66%变化为84.24%,并行化反演加速比均值由1.51变化为1.92,说明在数据分块不变的情况下,并行化反演的时间减少,并行效率提高,并行化反演较串行反演在时间效率上有大幅提高。草地总生物量、可食草量并行反演节点数与平均加速比的对应关系(图5)。

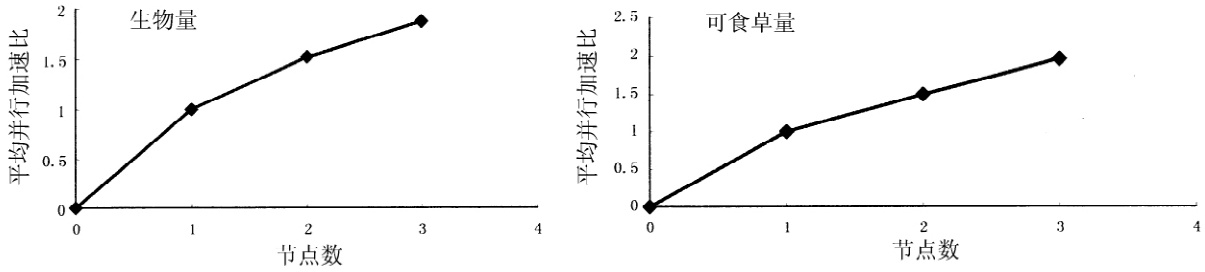


图5 草地总生物量、可食草量并行反演节点数与平均加速比对应关系

Fig. 5 The relationship between calculate nodes and parallel retrieval speed in total biomass and edible biomass

从图6中可以直观的看出,2007年的草地总生物量(TBIO)反演分别在两个计算节点和3个计算节点并行化反演时间对照曲线中出现波峰。对照表3发现,2007年草地总生物量在两个计算节点和3个计算节点的并行化反演中,并行加速比仅为1.22和1.49,反演时间较其他四年中的均值有较大变化,这是因为在2007年草地总生物量并行化反演的过程中,集群中的某个计算节点执行Map或Reduce任务失败,MR并行模型通过自身的容错机制,调度集群中的其他计算节点进行并行计算。从而导致并行化反演时间长,并行化反演效率低的问题。

4 结束语

文中应用MR并行编程模型,结合生物量遥感反演理论与技术方法,提出了一种云计算平台下的遥感影像生物量并行化反演方法。通过增加计算节点个数研究并行化反演的效率,并以青海省三江源区生物

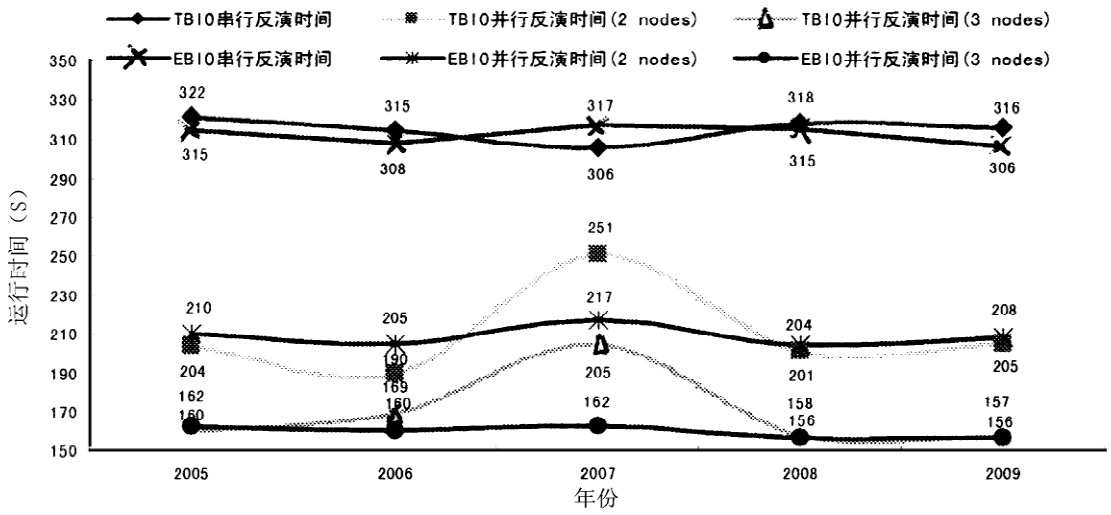


图 6　草地总生物量(TBIO) 与可食草量(EBIO) 串行、并行反演时间对照

Fig. 6 Serial and parallel inverse time comparison between total biomass of grassland and edible grass

量(草地总生物量和可食草量) 为例，进行实验分析。分析结果表明: 1) 并行化反演结果与经过精度验证的串行反演结果一致，并行反演结果可信; 2) 并行化反演较串行反演在时间效率上有明显提高，并且当计算节点小于最大并行效率下所对应的节点数量的情况下，反演效率与集群环境中计算节点的个数成反比，随着计算节点的增加，并行化反演时间减少，反演效率提高。基于 MR 模型的并行化反演表现良好的并行加速比和并行化反演效率，具有可行性，适合应用海量遥感数据快速、高效的并行化反演与处理。

但是，随着计算节点的不断增加，集群中的系统资源和主控节点的调度与通信压力不断增大，当计算节点超过最大并行效率下所对应的节点数量时，并行加速比将不再增大。因此，论文的下一步工作是: 进一步增加集群环境中计算节点数量，研究并行化反演效率随计算节点数量增加的变化趋势，组建满足生物量遥感并行化反演要求的最优集群环境。

参考文献

[1] Steffen W L，Walkr B H，Ingran J I，et al. Global Change and Terrestrial Ecosystems: The Operation Plan [M] . Stockholm: IGBP Secretariat，1992.

[2] Friedl M A，Machaelsen J，Davis F W，et al. Estimating grassland biomass and leaf area index using ground and satellite data [J] . International Journal of Remote Sensing，1994，15(7) : 1401 – 1420.

[3] 冯险峰. GIS 支持下的中国陆地生物量遥感动态监测研究 [D] . 西安: 陕西师范大学，2000.

[4] 徐斌，杨秀春，陶伟国，等. 中国草地产草量遥感监测 [J] . 生态学报，2007，27(2) : 405 – 413.

[5] 李昌凌，李文军. 基于 NDVI 的锡盟苏尼特旗地表植被生物量的趋势分析和空间格局 [J] . 干旱区资源与环境，2010，24(3) : 147 – 152.

[6] 陈国良，孙广中，徐云，等. 并行计算的一体化研究现状与发展趋势 [J] . 科学通报，2009，54(8) : 1043 – 1049.

[7] 杨靖宇，张永生，张宏兰，等. 基于可编程图形硬件的遥感影像并行处理研究 [J] . 测绘工程，2008，17(3) : 21 – 27.

[8] 方金云，何建邦，池天河，等. 地理数字图像集群并行处理实验 [J] . 计算科学，2001，28(5) : 99 – 101.

[9] 刘信安，李佳. 基于 PC 集群系统的 MPICH 大规模并行计算实现与应用研究 [J] . 计算机与应用化学，2003，20(5) : 577 – 582.

[10] Dean J，Ghemawat S. Mapreduece: Simple data processing on large clusters [J] . Communications of the ACM，2005，51(1) : 107 – 113.

[11] Lammel R. Google's Mapreduce Programming Model – Revisited [M] . Redmom，USA: Data Programmability Team Microsoft Crop，2007.

[12] 黄国满，郭建峰. 分布式并行遥感图像处理中数据划分 [J] . 遥感信息，2001(2) : 9 – 12.

[13] 沈占锋，骆剑承，陈秋晓，等. 高分辨率遥感影像并行处理数据分配策略研究 [J] . 哈尔滨工业大学学报，2006，38(11) : 1968 – 1971.

[14] QI J，Chenbouni A，Huete A R，et al. A modified soil adjusted vegetation index [J] . Remote Sensing of Enviorment，1994，48: 119 – 126.

[15] Alfredo H，Liu H Q. A feedback based modification of the NDVI to minimize canopy background and atmospheric noise [J] . IEEE Transactions on Geosciences and Remote Sensing，1995，33: 457 – 465.

[16] 于秀娟，燕琴，刘正军，等. 基于时间序列的三江源区生物量遥感检测模型研究 [J] . 安徽农业科学，2010，38(29) : 16530 – 16533.

Remote sensing retrieval method for biomass based on MapReduce parallel model

FU Tianxin[1,2], LIU Zhengjun[1], YAN Haowen[2]
(1. Chinese Academy of Surveying & Mapping, Beijing 100830, P. R. China;
2. School of Mathematics, Physic and Software Engineering, LanZhou Jiaotong University, Lanzhou 730070, P. R. China)

Abstract: MapReduce is a new parallel programming model based on cloud computing platform. The MapReduce parallel programming model was applied to remote sensing image parallel processing, and Three – River Source Region biomass (total biomass of grass and grazing capacity) in Qinghai Province was taken as an example to study the remote sensing retrieval method for biomass in a paralleling way by using growing season period MODIS13Q1 data products in 2005 – 2009 as the data source. The experimental analysis shows that: parallel inversion results based on the MapReduce model are consistent with serial inversion results which accuracy is validated and parallel inversion results are accurate and credible. The parallel retrieval efficiency has been greatly improved than the serial inversion efficiency; with the computing nodes increase, the parallel efficiency continues to increase.

Key words: cloud computing; MapReduce model; biomass; parallel computing