文章编号：2095-2163(2019)02-0196-04 中图分类号：TP391 文献标志码：A

Spark 的并行处理技术在岩石薄片图像的研究与应用

王 康
(西安石油大学 计算机学院，西安 710065)

摘 要： 随着岩石图像规模的不断增长，快速、有效地分割处理各类岩石图像的算法得到应用。文章将传统的图像分割处理方法与 Spark 整合起来，提出了基于 Spark 的岩石薄片图像分割处理方法。首先，采用基于二进制的图像预处理转换方法，存储图像到分布式文件系统 HDFS 中；其次，应用传递函数的方法，避免了图像分割处理算法进行 MapReduce 转化，实现了快速的通用图像分割处理，最后，以 DBSCAN 图像分割算法为实例证明了基于 Spark 岩石薄片图像分割处理有较好的适应性和较高的效率，并适应大规模图像的分割处理。

关键词： 岩石图像；分割处理；Spark；Hadoop；大数据；DBSCN 算法

Research and application of Spark´s parallel processing technology in rock sheet images

WANG Kang
(School of Computer Science, Xi´an Shiyou University, Xi´an 710065, China)

【Abstract】 With the growing number of rock images, various quick and efficient image segmentation processing algorithms are applied. This paper integrates the traditional image segmentation processing method with Spark, and proposes a Spark-based rock slice images segmentation processing method. Firstly, the binary image preprocessing conversion method is used to store images into the distributed file system HDFS. Secondly, the transfer function method is applied to avoid the image segmentation processing algorithm to perform Mapreduce conversion, and the fast general image segmentation processing is realized. The simulation using DBSCAN image segmentation algorithm proves that Spark-based rock slice image segmentation processing has better adaptability and higher efficiency, which is suitable for segmentation processing of large-scale images.

【Key words】 rock image; segmentation processing; Spark; Hadoop; big data; DBSCAN algorithm

0 引 言

随着岩石图像规模的不断增长，在岩石图像处理中很容易会遇到数百张序列图像、甚至是上千张高分辨率的岩石图像的情况，尤其是特征点的提取、构建描述符算法比较复杂，导致计算量非常大[1]。在有效利用当下提出的岩片图像处理手段的前提下，结合云计算和大数据处理技术的发展，快速有效的岩片图像处理和应用即已成为亟待解决的问题之一。

大数据时代的到来使得数据平台处理技术能够应用于更多领域，包括各种日志分析、行为分析和流量分析。在大数据应用中，Hadoop 和 Spark[2] 是最活跃的。在 Hadoop 架构下的图像处理方法是基于 MapReduce 的批处理方式实现图像并行处理的。其具有高吞吐量和高延迟，但是缺乏任务和资源分配的公平性，也未能考虑到对多任务与少量任务的区分，影响分割后图像效果[3]。同时处理效率太低，并且要进行各种岩石薄片图像处理算法间的转换，在代码编写上任务较为繁重。与 Hadoop 体系下的 MapReduce 相比，Spark 提供了更好的数据共享抽象，解决了 MapReduce 的高延迟缺陷，并给出了 Scala、Java、Python 三种编程接口。这 3 个 API 可与其它程序有效集成，从而合理分配任务和资源，最终提高图像处理效率。

本文利用分布式文件系统 HDFS 和图像处理界面，通过传递函数与 Spark 平台集成，可以编写各种图像处理算法，实现图像的并行处理[4]。实验证明，该方法可以实现图像的并行处理，适应大规模的图像处理。

1 Spark 平台

Spark 是伯克利大学的 AUMPAB 实验室推出的一个热门实验项目，代码很少，且是一个轻量级框架。Spark 是一个类似于 Hadoop[5] 的开源集群计算环境，但两者之间存在一些差异。这些设计上的差异使 Spark 在某些工作负载上表现更好，换句话说，Spark 已启用了更为出色的架构机制。除了提供交互式查询外，Spark 还优化了迭代工作负载。Spark

作者简介：王 康(1993-)，男，硕士研究生，主要研究方向：图像处理、大数据、人工智能。

收稿日期：2018-12-06

以 Scala 语言为研发基础，使用 Scala 作为其应用程序框架。与 Hadoop 不同，Spark 和 Scala 紧密集成，Scala 可以像本地集合对象一样轻松地运行分布式数据集。

Spark 提出了一种新的弹性分布式数据集（Resilient Distributed Datasets，RDD）。RDD 是一个并行、容错的数据结构，方便用户将数据集显式地存储于内存中。RDD 与 Hadoop 不一样的地方在于这些数据集是缓存在内存中，因而尤其利于数据的迭代计算。RDD 操作类型可以分为转换操作（Transformation）和控制操作（Actions）。其中，转换操作是惰性求值，即通过在 RDD 之间构造相互依赖的非循环图（DAG），最后传递动作。该操作会触发任务以返回结果。Spark 中的任务运行过程如图 1 所示。

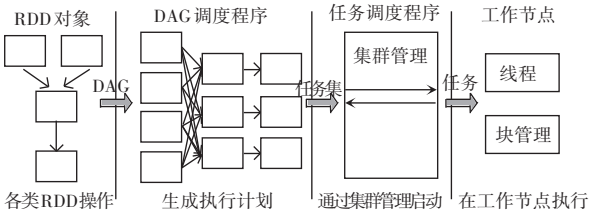

图 1　Spark 任务运行流程

Fig. 1　Spark task running process

2　基于 Spark 的处理架构与流程

基于 Spark 的岩石薄片图像分割处理可分为 3 个部分：HDFS、Spark 集群和图像处理接口。整体架构如图 2 所示。对每一部分的设计功能可阐释如下。

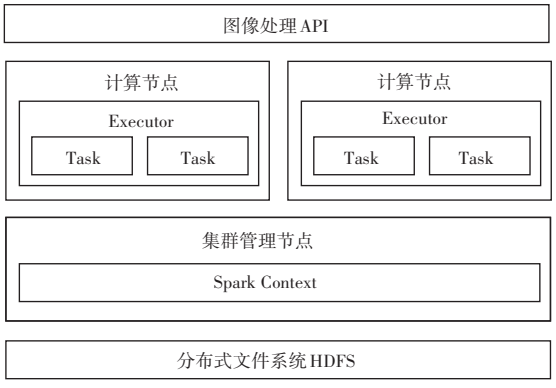

图 2　整体架构

Fig. 2　Overall architecture

（1）分布式文件系统 HDFS。负责预处理图像和各种输出结果的存储，并支持根据图像大小增加存储节点，以确保读取速度和存储规模。

（2）Spark 集群。图像数据读取、并行化处理和集群上的作业调度和资源分配。

（3）图像处理接口。用于与 Spark 主驱动器集成的图像预处理、转换操作和图像处理算法。

2.1　图像的预处理与转换

在现有的大规模岩石薄片图像中，由于不同的岩石图像存储在不同的服务器文件系统中，因此存在大量不同类型的图像数据。同时，Spark 不支持直接读取 JPEG、JPG 等格式的图像数据源。为了使每个 worker 节点能够访问不同格式的图像数据并通过 Spark 成功读取，本文运用程序代码将每个图像转换为二进制文本文件，并将其写入 HDFS。

对于远程图像数据源，Spark Streaming[6] 可用于通过读取网络流来处理和转换远程图像。Spark Streaming 预处理依赖于 Spark 集群，未经预处理的图像数据可通过流传输直接传输到 Spark 集群。不同区域的岩石图像数据不能得到有效的处理和集成，对网络带宽的要求也越来越高，不可避免地增加了图像处理的难度。

通过将图像转换为二进制文本文件、再进行存储，而且每个 worker 节点都可以方便地访问图像，解决了图像源数据同构的问题。在确定编码模式的情况下，通过读取二进制流可以将二进制值文本文件恢复为图像，并将图像存储在许多图像数据库中。

2.2　岩石图像并行处理的实现

Spark 具有用于文本文件的统一数据源文本接口，并将整个文本文件读取为一组行，这些行定义了基本 RDD，然后执行一系列 RDD 操作。

将预处理后的岩石薄片图像读取到 Spark 平台中。读入后，每个二进制文件对应于基本 RDD，即每个图像对应于一个 RDD，并且图像处理操作可以被视为简单的 RDD 转换操作。

Spark 中 RDD 的转换操作可以视为已建立方法下的集合或类型转换操作。本文提出转移所需的图像处理算法函数来构造 RDD 转换操作，实现各个区域的岩石图像操作。岩石薄片图像与 Spark 平台并行算法的结合过程如图 3 所示。

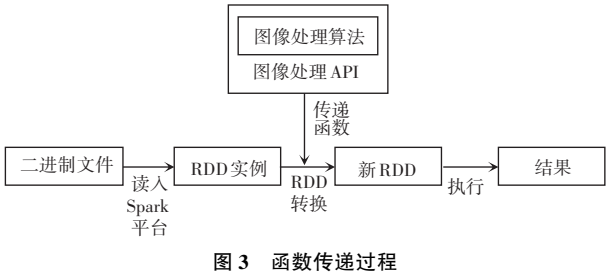

图 3　函数传递过程

Fig. 3　Function transfer flow

通过功能转换和集成，图像处理界面将图像处

理算法传输到 Spark Big Data 平台。在 RDD 转换完成之后,触发操作执行实例,Spark 通过预处理输入图像。此数字决定了任务的数量,并且取决于集群的大小,任务负载均衡并分发到每个工作节点。处理完工作节点后,结果将写入分布式文件系统(HDFS)或分布式数据库(HBase[7])。这样,确保了图像与任务之间的一一对应关系,同时,当执行图像并行处理时,需要对图像进行序列化,并且能够快速有效地读取每个图像。

3 基于 Spark 的并行处理技术在岩石薄片图像的应用

将本文提出的基于 Spark 的并行图像处理技术应用在岩石图像分析领域,通过粒度分析和矿物识别实现岩石薄片图像分割。本文将基于密度的算法添加到图像处理界面,即 DBSCN 图像分割[8]。图像处理方法的主要思想是使用粒子中心来表示以简化模型,然后使用 DBSCAN 算法聚类目标中心以标记不同的粒子目标。最后,结合改进的数学形态学方法,近似了粒子边界。实验结果表明,该方法对粒子分割和边界提取具有良好的效果,为粒子边界表征提供了有效途径。

3.1 实验环境及数据来源

基于 Spark 的岩石薄片图像并行处理方法在搭建的 Spark 集群中实现。实验集群由 1 个主节点和 5 个从属节点构成。其中,主节点在 Spark 中担任 Master,从节点担任 Worker。操作系统均为 Centos7.5,硬件配置均为 Intel(R) Core(TM) i7-8700 CPU @ 3.20 GHz 3.19 GHz 内存。所有节点均已成功配置 Spark2.3.0 并行处理框架、Java 以及各类 Java 下的图像处理包,如 JMagick,Sanselan 等。实验中使用 Spark 的 Java 接口,通过 Java 语言撰写 Spark 主驱动程序及岩石薄片图像实现方法,任务部署模式采用 Spark on Yarn。

本文中所有的图像数据来自鄂尔多斯盆地砂岩薄片。薄片图像均用高分辨率显微镜拍摄,每个图像都是以 JPEG 格式存储,并且图像大小是 10 M 内。在本文中,从每个区域选择相似大小的1 000张图像用于处理。

3.2 实验设计与分析

实验拟通过利用 DBSCAN 分割方法的 Java 研发算法来设计展开,并根据实验的目的选取若干岩石薄片图像作为样本进行实验。实验运行结果及分析详见如下。

3.2.1 与传统方法的性能对比

比较传统的物理机的方法,本文提出的并行处理方法速度较快。在 5 个节点的情况下,基于不同的岩石图像数,实验运行后绘制得到的时间结果曲线如图 4 所示。

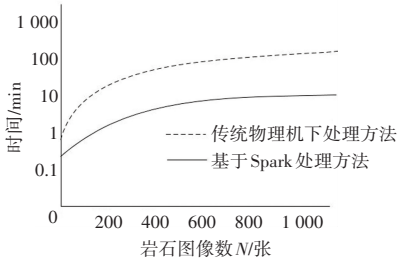

图 4 传统处理方法与并行处理方法测试结果

Fig. 4 Test results between traditional processing and parallel processing

图 4 显示伴随图像的增加,基于 Spark 的并行处理方法的优点变得越来越明显,并且可以实现图片的快速处理。

3.2.2 多节点下的分割效率

本文在拍摄的各地区岩石薄片图像中选取 1 000张、2 000 张、5 000 张图像的情况作为数据源进行实验,通过改变集群工作节点数目,将 1 000 张、2 000 张和5 000张图像选择作为每个区域的岩石图像中的数据源,研究得到的基于 Spark 的并行处理方法实验结果如图 5 所示。

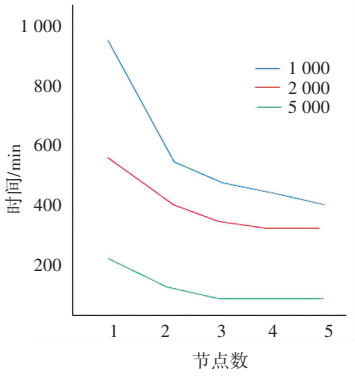

图 5 不同节点数目的处理时间

Fig. 5 Multi-node processing time

由图 5 可以看到,在单个工作节点对于5 000张的图像共需要 16 h 的处理时间,而在 5 个节点情况下却只需 2 h 的处理时间,这就极大地减少了处理相同图像的时间。而当节点的数目增加时,图像的处理速度也在逐渐提升。

4 结束语

基于 Spark 的岩石薄片图像处理架构有效地解决

了岩石薄片图像处理算法在不同区域对不同种类岩石图像的快速并行处理，集群节点数量越多，就越能提升并行处理图像的速度。本文提出了一种基于 Spark 的并行图像处理体系结构，通过预处理后的分布式图像进行存储和转换，可以将岩石薄片图像读取到 Spark 平台中，并应用传递函数方法避免 MapReduce 转换。实验表明，在该平台下可以增加任何图像处理算法，实现并行处理，适应于大规模的图像处理。

参考文献

[1] IWASE Y, IMAMURA H. Medical image processing apparatus, medical image processing method, and program:US ,8560341B2 [P]. 2013-10-15.

[2] ZAHARIA M, CHOWDHURY M, FRANKLIN M J, et al. Spark:Cluster computing with working sets[J]. HotCloud, 2010, 10:10.

[3] MADUSKAR P, HOGEWEG L, PHILIPSEN R H H, et al. Automated localization of costophrenic recesses and costophrenic angle measurement on frontal chest radiographs [J]. Proceedings of SPIE-The International Society for Optical Engineering, 2013, 8670(4):867038.

[4] Message Passing Interface Forum. MPI: A message - passing interface standard [EB/OL].[2009-09-04]. http://ishare.iask.sina.com.cn/f/6669516.html.

[5] WHITE T. Hadoop: The definitive guide—MapReduce for the cloud[M]. USA:O'Reilly Media, Inc., 2009.

[6] ZAHARIA M, DAS T, LI H, et al. Discretized streams: An efficient and fault-tolerant model for stream processing on large clusters [C]// Usenix Conference on Hot Topics in Cloud Computing. Boston, MA:Usenix Association, 2012.

[7] VORA M N. Hadoop-HBase for large-scale data[C] //2011 International Conference on Computer Science and Network Technology (ICCSNT). Harbin, China:IEEE, 2011:601-605.

[8] ACHANTA R, SHAJI A, SMITH K, et al. SLIC superpixels compared to state-of-the-art superpixel methods [J]. IEEE Transactions on Pattern Anlysis and Machine Intelligence, 2012, 34 (11):2274-2282.

[9] 张良将, 宦飞, 王杨德. Hadoop 云平台下的并行化图像处理实现[J]. 信息安全与通信保密, 2012(10): 59-62.

[10] 吴拥, 苏桂芬, 滕奇志, 等. 岩石薄片正交偏光图像的颗粒分割方法[J]. 科学技术与工程, 2013, 13 (31) : 9201-9206.

[11] MACQUEEN J. Some methods for classification and analysis of multivariate observations[C]// Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, CA, USA:University of California Press,1967, 1(14) : 281-297.

[12] 兰叶芳,邓秀芹,程党性,等. 鄂尔多斯盆地三叠系延长组次生孔隙形成机制[J]. 地质科技情报, 2014, 33 (6) : 128-136.

[13] ZHANG Jie. PyGel: Distributed graph computing engine research & implementation based on DPark[D]. Guangzhou: South China University of Technology, 2013.

[14] OLSON L, SAMSON C, MCKINNON S D. 3- D laser imaging of drill core for fracture detection and rock quality designation[J]. International Journal of Rock Mechanics and Mining Sciences, 2015, 73(1) : 156-164.

[15] GAO G, YAO W, XIA K, et al. Investigation of the rate dependence of fracture propagation in rocks using digital image correlation (DIC) method[J]. Engineering Fracture Mechanics, 2015, 138(4) : 146-155.

[16] SAFARI A, MORADI M, HASSANI A, et al. Numerical simulation and X-ray imaging validation of wormhole propagation during acid core-flood experiments in a carbonate gas reservoir[J]. Journal of Natural Gas Science and Engineering, 2016,30(3):539 - 547.

(上接第 195 页)

[2] 李宇昊. 无人机在林业调查中的应用实验[J]. 林业资源管理, 2007(4):69-73.

[3] 李克,杨姗姗,王正阳,等. 无人机低空摄影在面积量算上的可行性分析[J]. 价值工程,2018(6):164-166.

[4] ZARCO-TEJADA P J, DIAZ-VARELA R, ANGILERI V, et al. Tree height quantification using very high resolution imagery acquired from an unmanned aerial vehicle (UAV) and automatic 3D photo-reconstruction methods [J]. European Journal of Agronomy, 2014, 55(2):89-99.

[5] PARIS C, BRUZZONE L. A three-dimensional model-based approach to the estimation of the tree top height by fusing low-density LiDAR data and very high resolution optical images[J]. IEEE Transactions on Geoscience & Remote Sensing, 2015, 53 (1):467-480.

[6] JING Linhai, HU Baoxin, NOLAND T, et al. An individual tree crown delineation method based on multi-scale segmentation of imagery[J]. Isprs Journal of Photogrammetry & Remote Sensing, 2012, 70(3):88-98.

[7] MATHEWS A J. Object-based spatiotemporal analysis of vine canopy vigor using an inexpensive unmanned aerial vehicle remote sensing system[J]. Journal of Applied Remote Sensing, 2014, 8 (1):1-17.

[8] 何海清,黄声享. 一种稳健估计的无人机机载相机标定法[J]. 测绘通报,2013(2):99-102.

[9] 陈启晟,朱进,丁亚洲,等. 一种便携式的无人机航测非量测相机标定方法[J]. 科学技术与工程,2015,15(25):1-6.

[10] 刘阳,童恒庆,张齐,等. 基于 MATLAB 的数码相机定位研究及实现[J]. 武汉理工大学学报,2009,31(13):130-132.

[11] ZHANG Zhengyou. Flexible camera calibration by viewing a plane from unknown orientations[C]// The Proceedings of the Seventh IEEE International Conference on Computer Vision. Kerkyra, Greece:IEEE, 1999:666-673.

[12] 刘艳,李腾飞. 对张正友相机标定法的改进研究[J]. 光学技术, 2014,40(6):565-570.

[13] 刘杨豪,谢林柏. 基于共面点的改进摄像机标定方法研究[J]. 计算机工程,2016,42(8):289-293.

[14] 邹建成,田楠楠. 简易高精度的平面五点摄像机标定方法[J]. 光学精密工程,2017,25(3):786-791.

[15] 李二森,张保明,周晓明,等. 自适应 Canny 边缘检测算法研究[J]. 测绘科学,2008,33(6):119-120,65.

[16] 王静,彭望碌,郭平. 多源遥感图像区域提取分析研究[J]. 计算机科学,2004,31(b07):66-67.

[17] ANGULO J, SERRA J. Modelling and segmentation of colour images in polar representations[J]. Image & Vision Computing, 2007, 25(4):475-495.