基于大数据技术的气象算法并行化研究

李永生[1],曾 沁[1,2],杨玉红[1],陈 晋[3]
(1. 广东省气象探测数据中心,广东 广州 510080;
2. 广东省气象台,广东 广州 510080;
3. 同济大学 数学系,上海 200082)

摘 要:在气象数值预报解释应用业务中,传统数值算法的应用呈现逐步增加的趋势,但是随着算法输入数据种类和数据量的增加导致算法的完成时间大幅增长,甚至出现了算法完成时间性能瓶颈。为了突破算法时间上的性能瓶颈,基于 OpenCV 算法库,实现了多元逐步回归和卡尔曼滤波算法的执行模块,采用 Map-Reduce 计算框架设计和实现了多站点输入数据分割的并行化执行模块;规范了算法输入和输出的数据格式,设计了并行算法的 Web 服务流程以及实现了基于 Rest Web Service 的算法访问接口。业务应用实验测试表明,并行算法能够很好地满足气象业务实际需求。

关键词:大数据分析技术;并行化;气象数值算法;Web 服务
中图分类号:TP312 **文献标识码**:A **文章编号**:1673-629X(2016)09-0047-03
doi:10.3969/j.issn.1673-629X.2016.09.011

Research on Parallelism of Typical Meteorological Algorithm Based on Big Data Technology

LI Yong-sheng[1], ZENG Qin[1,2], YANG Yu-hong[1], CHEN Jin[3]
(1. Guangdong Meteorological Data Center, Guangzhou 510080, China;
2. Guangzhou Central Observatory, Guangzhou 510080, China;
3. Department of Mathematics, Tongji University, Shanghai 200082, China)

Abstract: The traditional application of numerical algorithms is widely used in the interpretation application of numerical weather prediction. However with the increasing of the type and amount of algorithm input data, the completion time of it presents exponential growth, which faces the bottleneck in the algorithm completion time performance. In order to break it, an execution module of stepwise multiple regression and Kalman filter algorithm is developed based on the OpenCV library (Open Source Computer Vision Library). A distributed data storage model and parallel data access service are designed based on Hadoop framework, and the parallel strategy is designed based on the Map-Reduce framework, then the parallelism execution module is achieved based on the implement of a multi-site input data partition. The algorithm input and output data format are standardized. A Rest Web Service interface of parallel algorithm is designed. Operational trial in multi-user environment shows that this algorithm can meet the actual requirements of meteorological business greatly.

Key words: data analysis technology; parallelism; meteorological numerical algorithm; Web Service

0 引 言

在气象预报领域将多元逐步回归、卡尔曼滤波等数值算法引入到对数值预报解释中已得到广泛的应用,这是对数值预报这一综合性的结果运用动力学、统计学技术再一次进行加工、修正,使预报精度得到进一步提高,以达到有价值的要素预报水平。实践证明:通过数值预报的解释应用,确实可以使要素预报比客观数值预报模式直接输出的预报有明显提高[1]。陆如华等从气象应用角度介绍了卡尔曼滤波的基本原理及其递推算法[2];相关研究也进一步表明,数值预报解释应用算法在数值预报业务中得到了广泛应用,但这些研究对算法性能和算法的并行化设计和实现涉及较少,

收稿日期:2015-06-10 **修回日期**:2015-10-15 **网络出版时间**:2016-08-23
基金项目:中国气象局气象关键技术集成与应用项目(CMAGJ2014M40);广东省气象局重点项目(2012A01);广州市科技计划项目(2012Y2-00031,2013Y2-00053,2013Y2-00074)
作者简介:李永生(1980-),男,硕士,工程师,研究方向为云计算及海量数据存储、气象大数据分析应用技术;曾 沁,高级工程师,研究方向为精细化预报技术与应用、气象大数据分析应用技术。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.tp.20160823.1112.012.html
万方数据

随着应用的逐步深入以及数值预报解释应用算法输入数据的逐步增加,出现了一定的计算瓶颈。因此如何解决数值算法在数值预报解释应用中的计算瓶颈问题是气象预报业务迫切需要解决的。目前大数据处理技术的逐渐成熟为这一问题的解决提供了契机。

1 算法并行化设计与实现

算法平台实现了多元逐步回归和卡尔曼滤波算法,基于跨平台的通用算法库 OpenCV(Open Source Computer Vision Library)实现了多元逐步回归和卡尔曼滤波两种数值算法的执行模块。文中重点讨论基于大数据分析技术的算法并行化设计实现和算法应用设计,对算法本身的原理和算法在数值预报解释应用中的参数选择等不做详细的描述和说明[1]。

1.1 算法并行化策略

针对数值算法在数值预报解释应用中的并行化可以从基于算法本身的并行化处理方案和基于数据分割的并行化处理方案两方面着手。算法本身的并行化需要强大的硬件资源支持,需要一定规模的资金投入,在应用方面受到一定的限制[3],因此设计主要采取基于数据分割的并行化处理方案,这也是 MapReduce 编程模型的典型应用。

首先,数值预报解释应用中的数据大都涉及多年资料的统计分析和回归算法应用,而算法应用过程的统计中间量,可以根据时间长度和计算节点数目进行适当分割,从而分解到各个计算节点(这是 Map 函数过程),并行计算后,对各个时间段的统计量[4]在 Reduce 函数过程中进行汇总分析,从而得到全时段的计算统计量,实现算法运算的并行化。

其次,格点数值预报解释应用的过程中,首先还是乡镇站点的单站点解释应用,而广东省共有 1 500 多个行政乡镇,因此对应了 1 500 个方程建模,而建模过程对每个站点而言,回归算法的应用是独立的,因此,可以随机分配给若干的计算节点进行处理[5]。

基于数据分割的并行化处理需要重点解决的关键问题为:

(1)合理确定多元回归和卡尔曼滤波等算法的 MapReduce 中的 Map 和 Reduce 策略。

(2)Map 阶段的数据本地化的考虑和数据流转策略。

(3)Reduce 阶段算法结果数据拷贝和数据结果的获取。

1.2 算法的 MapReduce 实现

基于数据分割的并行化计算框架如图 1 所示。

系统将输入数据和输出结果统一成 JSON 格式,所有的输入站点可以存储在一个文件,也可以分多个文件存储,在 Map 算法中会将每一个站点 ID 作为 map key 处理。由于此过程无需 reduce 处理,所以 map 之后数据会直接生成<key,value>数据保存在给定的输出目录中,一般输出文件数目与 map 个数一致。在 map 函数中,算法会根据输入参数类型确定是运行 Kalman 滤波器算法还是多元逐步回归算法。

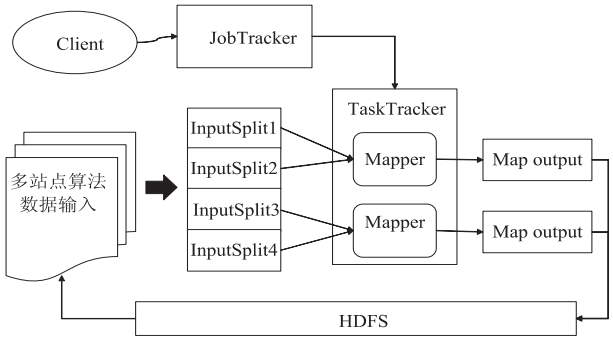

图 1 算法并行化 MapReduce 执行框架

2 算法 Web 服务设计与实现

平台主要通过 Web Service 模块向用户提供分布式计算服务。考虑到应用实际,系统将数据上传,任务执行与结果获取分开,通过多站点的观测数据、数值预报产品要素数据以及任务类型和算法名称作为输入启动数据上传模块,数据上传后存储在系统集成层中的 HDFS 中并返回一个任务 ID,通过任务 ID 启动具体算法的任务执行模块,算法成功执行后返回一个结果 ID,业务用户通过算法结果 ID 启动任务结果获取模块获取具体算法的执行结果。具体流程如图 2 所示。

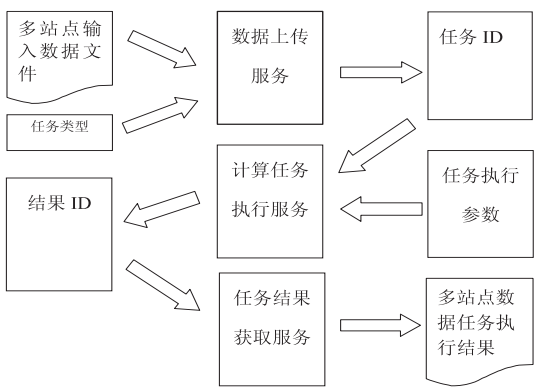

图 2 分布式计算 Web 服务流程

2.1 输入数据上传服务设计

此模块设计的目的是为了将多站点的输入数据文件上传至 HDFS,同时也考虑到输入文件规模大时的情况。主要特点如下:

(1)系统采取 HTTP PUT 模式上传数据文件,此模式在客户端更容易操作,可以上传已存在文件,而无须创建新的文件流。

(2)系统设计了两个输入参数,算法服务类型(regression,表示多元逐步回归服务;Kalman,表示 Kalman

万方数据

滤波器算法)与任务 Id(即 jobId,与返回参数相同)。设计任务 Id 的目的是可以多次上传同一次任务所需的输入数据,以达到上传大规模输入数据的目的。

2.2　计算任务执行服务设计

在这一模块中,主要通过上传文件后获得的任务 Id 以及算法输入参数来执行给定类型的分布式算法任务。设计要点如下[6]:

(1)将常用的各个站点统一的参数通过 URL 输入,达到无需修改输入数据文件的目的,如在多元逐步回归算法中,将 F 检验参数通过 URL 输入。

(2)Web 服务通过执行脚本来运行算法任务。在脚本里通过 ssh 到 Hadoop 平台上某一个节点来执行算法任务。将脚本放在 tomcat 目录下。

其中,MapReduce 任务 jar 文件放在指定目录,输入参数必须包括算法类型、输入数据路径以及输出结果路径。

(3)在任务执行完成之后系统没有马上返回结果,而是将结果放在 HDFS 上,通过下载服务下载数据。同样系统也可以对此任务多次运行,也会得到不同的结果 Id。

(4)结果 Id 里包含了算法类型信息,输出文件路径信息。

3　平台业务应用测试

业务应用测试包括多元逐步回归算法和 Kalman 滤波算法的功能测试和算法性能测试两个部分。其中,功能测试主要测试算法通过 Web 提供分布式计算功能,包括算法数据接入模块、算法执行模块以及算法结果获取模块;算法性能测试主要测试算法的分布式执行时间性能[7-9]。

3.1　业务应用测试方案设计

测试选取欧洲中心 0.25 度分辨率的数值预报产品(以下简称 EC_FINE)中的 2 米温度、海平面气压、相对湿度和 850 百帕温度共 4 个预报要素,实际观测要素数据是地面温度,测试的时间范围是 2013 年 5 月至 9 月。其中,EC_FINE 要素预报 0 ~ 24 小时内是 3 小时间隔的,因此实际观测数据也选取 24 小时内隔 3 小时的观测数据,这样 EC_FINE 的要素预报值和观测数据值二者之间能够实现一一对应[10-11]。

系统首先将测试数据转换成多元逐步回归算法输入数据,然后进行算法效果测试与单点时间性能测试[12],之后将数据以及输出量转换成 Kalman 滤波算法输入量,然后进行单点测试[13],最后将多站点数据输入进行集群性能测试。

3.2　业务应用分析

基于 Hadoop 技术在 IBM 3650 服务器上构建大数据的算法应用平台,系统实现了能自动双向失效切换的 HDFS NameNode HA 机制,实现了平台的可靠性、动态可扩展以及安全一体化等功能。选取 MapReduce 2.0(yarn)实现 MapReduce 计算引擎,并在此基础上部署了针对气象业务数据应用的数据接入模块、算法执行模块和结果获取模块。同时对算法性能进行了测试,测试时输入数据分别是 1 000,3 000,5 000,8 000,10 000个站点的数据,分别对单节点计算的算法性能和多计算节点并行化执行时的算法性能进行测试。图 3 的测试结果表明,算法在分布式并行化的执行环境下性能较优,特别当测试数据站点越多时算法时间性能改善效果越明显。

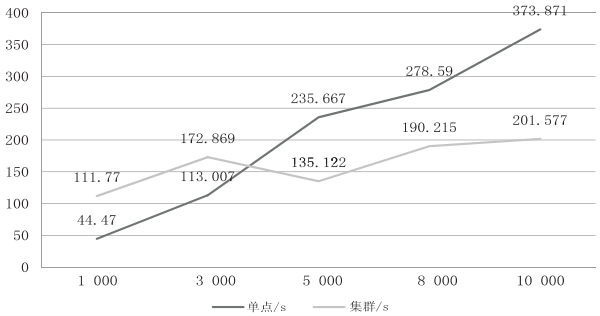

图 3　Kalman 滤波算法单点与集群时间性能对比图

4　结束语

文中基于大数据分析技术实现了典型气象算法应用平台,实现了非结构化大数据的分布式存储和处理,借助 OpenCV 算法库,实现了多元逐步回归和卡尔曼滤波算法的执行模块,采用 MapReduce 计算框架设计和实现了多站点输入数据分割的并行化执行模块。考虑业务应用的规范,选取 JSON 数据格式规范了算法输入和输出数据,设计了并行算法的 Web 服务流程以及实现了基于 Rest Web Service 的算法访问接口,并进行了算法功能测试和算法性能测试。测试结果表明,算法功能满足数值预报解释应用的业务实际需求,同时随着算法输入数据中站点量的线性增加,算法时间性能曲线突破了传统计算方式的近似线性增长,而是平缓增长并趋于稳定。业务实验结果表明,利用大数据相关技术对典型气象算法进行并行化设计和实现能够取得很好的业务效果,是气象大数据技术应用的一个重要发展方向。

参考文献:

[1]　李玲娟,张　敏. 云计算环境下关联规则挖掘算法的研究[J]. 计算机技术与发展,2011,21(2):43-46.

[2]　郭　苑,张顺颐,孙雁飞. 物联网关键技术及有待解决的问题研究[J]. 计算机技术与发展,2010,20(11):180-183.

(下转第 55 页)

万方数据

定位标准差用式(15)来衡量:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i - \bar{r})^2} \tag{15}$$

其中,$\bar{r}$ 为实验中待定位点的定位误差平均值;N 为实验中待定位点的数量;σ 为定位标准差。

最终获得的定位结果:采用欧氏距离法获得的定位误差平均值为 1.792 7 m,标准差为 0.873 1 m;使用文中方法获得的定位误差平均值为 1.414 6 m,标准差为 0.814 8 m。

通过图 5 可以看出,欧氏距离法对环境噪声比较敏感,在受环境噪声影响较大的待定位点定位误差较大,利用文中提出的定位算法可以比欧氏距离更好地对抗环境噪声,以整体提高定位精度。

3　结束语

文中提出一种基于卡方距离和灵敏度法结合的算法(CSKNN)来实现人或物的 WLAN 定位。CSKNN 算法是一种基于距离计算方式改进的 KNN 算法,其将距离的计算方法从欧氏距离变为了卡方距离。算法能够根据不同的室内环境特点在算法中给各 AP 赋予不同的权值,有效降低了复杂室内环境对定位精度的负面影响,整个定位系统不用添加任何额外的硬件即可实现。实验结果表明,该算法的定位精度明显优于原始的欧氏距离法。

参考文献:

[1] Khan A U, Al-Akaidi M. A distributive algorithm for WLAN localization[C]//Proc of 6th international conference on e-merging technologies. [s. l.]:IEEE,2010:388-393.

[2] Gu Y, Lo A, Niemegeers I. A survey of indoor positioning systems for wireless personal networks[J]. IEEE Communications Surveys & Tutorials,2009,11(1):13-32.

[3] Zhou M, Zhang Q, Tian Z, et al. IMLours: indoor mapping and localization using time - stamped WLAN received signal strength[C]//Proc of wireless communications and networking conference. [s. l.]:IEEE,2015:1817-1822.

[4] Ding G, Zhang J, Zhang L, et al. Overview of received signal strength based fingerprinting localization in indoor wireless LAN environments[C]//Proc of 5th international symposium on microwave, antenna, propagation and EMC technologies for wireless communications. [s. l.]:IEEE,2013:160-164.

[5] 孙善武,王　楠,陈　坚. 一种改进的基于信号强度的 WLAN 定位方法[J]. 计算机科学,2014,41(6):99-103.

[6] 徐玉滨,邓志安,马　琳. 基于核直接判别分析和支持向量回归的 WLAN 室内定位算法[J]. 电子与信息学报,2011,33(4):896-901.

[7] 刘洺辛,孙建利. 基于能效的 WLAN 室内定位系统模型设计与实现[J]. 仪器仪表学报,2014,35(5):1169-1178.

[8] 程金晶,魏东岩,唐阳阳. WLAN 指纹定位中 AP 选择策略研究[J]. 计算机技术与发展,2015,25(3):1-5.

[9] Sohn I. Localization performance analysis of KNN in IEEE 802.11 TGn channel[C]//Proc of ICT convergence. [s. l.]:IEEE,2011:219-220.

[10] 牛建伟,刘　洋,卢邦辉,等. 一种基于 Wi-Fi 信号指纹的楼宇内定位算法[J]. 计算机研究与发展,2013,50(3):568-577.

[11] Bosch A. Image classification for a large number of object categories[D]. Girona: University of Girona,2007.

[12] 贾世杰,孔祥维. 一种新的直方图核函数及在图像分类中的应用[J]. 电子与信息学报,2011,33(7):1738-1742.

[13] 郑　倩,卢振泰,陈　超,等. 基于邻域信息和高斯加权卡方距离的脊椎 MR 图像分割[J]. 中国生物医学工程学报,2011,30(3):357-362.

[14] Pele O, Werman M. The quadratic - chi histogram distance family[C]//Proc of ECCV 2010. Berlin: Springer,2010:749-762.

(上接第 49 页)

[3] Kschischang F R, Frey B J, Loeliger Hans-Andrea. Factor graphs and the sum-product algorithm[J]. IEEE Transactions on Inform Theory,2001,47(2):498-519.

[4] 陈　康,郑纬民. 云计算:系统实例与研究现状[J]. 软件学报,2009,20(5):1337-1348.

[5] 余楚礼,肖迎元,尹　波. 一种基于 Hadoop 的并行关联规则算法[J]. 天津理工大学学报,2011,27(1):25-28.

[6] 王　彬,肖文名,李永生,等. 华南区域中心计算资源管理系统的建立与应用[J]. 气象,2011,37(6):764-770.

[7] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science,2007,315:972-976.

[8] 王　彬,常　飚,朱　江,等. 气象计算网格平台资源监视模块的设计与实现[J]. 应用气象学报,2009,20(5):642-648.

[9] Zhen Bin, Wu Xihong, Liu Zhimin, et al. An enhanced relative spectral processing of speech[J]. Chinese Journal of Acoustics,2002,21(1):86-96.

[10] 幸莉仙,黄慧连. MapReduce 框架下的朴素贝叶斯算法并行化研究[J]. 计算机系统应用,2013,22(2):108-111.

[11] Halkidi M, Vazirgiannis M, Batistakis Y. Quality scheme assessment in the clustering process[C]//Proc of 4th European conf principles and practice of knowledge discovery in databases. [s. l.]:[s. n.],2000:165-276.

[12] 应　毅,任　凯,曹　阳. 基于改进的 MapReduce 模型的 Web 挖掘[J]. 科学技术与工程,2013,13(5):1205-1209.

[13] 王　萍,刘　颖,王汉芝,等. 基于格点场数据的沙尘暴双预报模型[J]. 天津大学学报,2006,39(3):329-333.

万方数据