DOI:10.11721/cqnuj20190608

大数据环境下的车牌统计算法的数据分片研究*

汤 星[1], 范永胜[1], 冯 骥[1], 钟 贞[1], 孔亚迪[2]
(1. 重庆师范大学 计算机与信息科学学院, 重庆 401331;
2. 南京理工大学 计算机技术与工程学院, 南京 210094)

摘要:【目的】在大数据环境下,寻找最佳的输入数据分片数量,以便改进统计算法的运行效率。【方法】以车牌统计为例,调整相关多个参数以改变算法中输入数据的分片数量,并分析不同参数下算法运行时间的差异。【结果】最佳的分片数量下的运行效率近似于最差的分片数量下运行效率的 2 倍。【结论】大数据处理中,合理的输入数据的分片数量有助于提高算法的运行效率。同时也分析了分片数量与算法运行时间的函数关系,以期找到最佳的分片数量区间。

关键词:车牌统计算法;分片数量;算法运行时间
中图分类号:TP391.4 **文献标志码:**A **文章编号:**1672-6693(2019)06-0098-06

随着智能交通系统的广泛应用,交通摄像头拍摄到的车辆信息也在迅速增加。基于这些海量的车辆信息,国内的学者们进行了大量研究。其中主要的研究方向包括车牌定位技术和车牌字符识别技术[1-6],也有部分学者针对真假车牌判断技术和车牌提取技术进行研究[7-8]。大数据平台在处理分布式海量数据方面比普通的单机运算具有更明显的优势,因此目前也有一部分学者的研究将大数据平台与车牌识别技术相结合。例如,沈文枫等人[9]将深度学习算法与大数据平台结合,提出了高准确率的车牌汉字识别技术,而曹波等人[10]在车牌识别大数据处理与分析的基础上,提出了挖掘伴随车辆组的方法。以上研究工作虽然有小部分结合大数据平台对交通信息处理相关算法和技术进行了改进,但并未对大数据平台的算法实现过程进行研究。本文对输入数据分片问题的研究,也正是对该领域研究的有效探索和补充。

本文对大数据车辆信息输入数据分片的研究,是以大数据平台的 Hadoop 分布式系统架构为基础,采用 MapReduce 模型的车牌统计算法开展研究[11]。文中的车牌统计算法包含地区车牌统计和套牌检查两部分,前者根据车牌中代表区域名称的第一个汉字将不同地区的车牌进行分类计数,后者检查是否存在两个或两个以上完全相同的车牌号码,若存在则将此车牌号码记录为套牌。在车牌统计和检查过程中,若改变输入数据的分片数量,算法的运行效率会相应的发生改变,分析两者之间的变化规律是本文研究的重点。

1 车牌统计算法

为便于查找和比对实验中的图像,使得在算法的数据处理阶段能根据存储路径或编号快速查到对应的车牌信息,在算法进行分类和检查前,必须对相关的信息进行预处理。预处理的过程如下:

第 1 步,创建一个 txt 文件;

第 2 步,依次为每个通过道路图像采集设备采集到的包含有车牌信息的彩色图像按顺序编号命名,如 1.jpg,2.jpg,3.jpg;

第 3 步,在采集过程中,将每张图片的存储路径按行记录在第 1 步创建的 txt 文件中,格式为:存储路径 车道

* **收稿日期:**2019-06-25 **修回日期:**2019-07-04 **网络出版时间:**2019-11-25 10:35
资助项目:重庆师范大学(人才引进/博士启动)基金项目(No. 17XCB008);教育部人文社会科学研究项目(No. 18XJC880002);重庆市教育委员会科技项目(No. KJQN201800539)
第一作者简介:汤星,女,研究方向为大数据与云计算,E-mail:txmail2016@163.com;**通信作者:**范永胜,男,副教授,博士,E-mail:yongsheng_fan@yeah.net
网络出版地址:http://kns.cnki.net/kcms/detail/50.1165.N.20191125.1034.028.html

号，如 txt 文件的第一行/user/mapreduce/platerecog/images/1.jpg(存储路径) 2(车道号)。

1.1 算法运行流程

本文中采用的车牌统计算法包含两组 MapReduce 任务，分别完成对地区车牌统计算法和套牌检查算法的实现。具体步骤如下：

第 1 步，算法调用实验平台提供的车牌识别程序识别图片中的车牌号码；

第 2 步，将识别出的车牌号码与包含了存储路径的 txt 文件传入地区车牌识别 MapReduce 任务中，完成不同地区的车牌分类数量统计，并将结果输出至对应文件中；

第 3 步，修改一些变量，将车牌号码与 txt 文件传入至套牌检查的 MapReduce 任务中，检查是否存在套牌并输出套牌号码和套牌对应的图片名称，输出结果保存在对应的文件中。

1.2 地区车牌统计原理

算法实现的基本原理是：利用大数据实验平台提供的车牌识别程序来识别图片中的车牌号码，由 Map 端将车牌号码的第一个汉字(省市自治区简称)作为 key 值，value 值设为 1，如(“苏”，1)，解析成 key/value 对；然后将相应的 key/value 对传入 Reduce 端，由 Reduce 端将具有相同 key 值的 value 集中的数字累加，获得该地区的车牌数总量，最终将结果写入到输出文件中进行统计。算法原理实现的数据流图如图 1 所示。

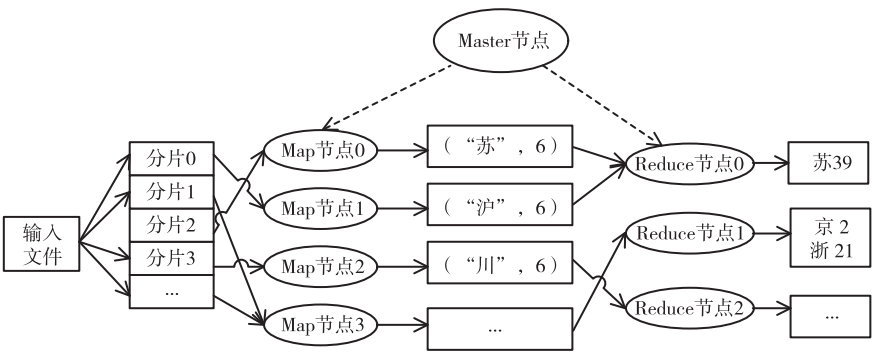

图 1　地区车牌统计算法数据流

Fig. 1　Regional license plate statistical algorithm data flow diagram

1.3 套牌检查原理

套牌检查的基本数据流处理与地区车牌统计算法基本相同，也是首先调用大数据实验平台提供的车牌识别程序识别图片中的车牌号码，由 Map 端将车牌号作为 key 值，该车牌图像的编号作为 value 值，如从某个车牌图像中识别出的车牌号码为“浙 A12345”，该车牌图像对应的图像编号为“78.jpg”，则解析出的 key/value 为(“浙 A12345”，78)；将 Map 端解析出的 key/value 对传入 Reduce 端，由 Reduce 端比较传入的 key 值，若出现两个或两个以上相同的 key 值(即车牌号)，则视为套牌，记录该车牌号码及对应的图像编号，并将结果写入输出文件中进行统计分析。

1.4 算法执行过程中数据的分片与处理

大数据中的 MapReduce 技术是一种基于数据密集型的并行计算模型，它由一个主节点 Master 及若干个从节点 Worker 组成。本文的车牌统计算法也采用了 MapReduce 技术，因此在算法执行的初始阶段，MapReduce 模型将输入数据按指定大小分成若干个数据片，然后将数据片存储于各 Worker 节点，并由 Master 节点按相应的规则为每个数据片分配 Map 任务和 Reduce 任务；任务执行完毕后，结果保存于内存中。

在实际的实验过程中，通过修改 FileInputFormat.setMinInputSplitSize(job,size)和 FileInputFormat.setMaxInputSplitSize(job,size)中参数 size 的数值可分别设置 minSize 和 maxSize 的数值大小，使数据的指定分片大小在 blockSize 上下自由调整，从而改变输入数据的分片数量。

在 MapReduce 模型中改变输入数据的分片数量是否会对算法产生影响，是本文探索的一个重要问题，先分析两种极端情况。

第1种极端情况,是将整个输入数据只分为1个数据片。Master节点会在若干个Worker节点中选择1个来处理该数据片,导致其他没有被分配任务的节点处于饥饿状态,最后大数据分布式处理效果没有真正地得以实现,算法的运行效率低。

第2种极端情况,是输入数据的分片数量过多。如按照车牌数量进行分片,一个车牌就对应一个数据片,则会在启动和关闭Map任务上花费过多的时间,从而降低算法执行效率。

为此寻找一个合适的分片数量,将会使得各个Worker均有任务可执行,不会再出现某些Worker节点负载过重的情况;除此之外,也能使得Map任务和Reduce任务上花费的时间趋于合理化。

以下实验将通过修改输入数据的指定分片大小来调整输入数据的分片数量,并记录不同分片数量情况下算法的运行时间,进而寻找最为合理的分片数量,从而分析分片数量与运行时间之间存在的函数关系。

2 实验与分析

2.1 实验环境与数据集

实验使用南京云创大数据科技股份有限公司提供的一套大数据实验平台。软件平台使用的Hadoop集群由1个Master节点和4个Slave节点构成,运行在Linux操作系统中。Hadoop设置的总空间大小(Configured capacity)为119.94 GB,非Hadoop文件系统所使用空间(Non DFS used)为16.28 GB。实验数据集的对象为云创公司提供的数据环境,该数据环境中的数据信息均来自实际案例且已经过数据预处理,其中包含有较清晰的车牌信息图像以及记录图像存储路径的txt文件1份。

2.2 实验结果及分析

2.2.1 单次实验及结果 分别改变输入数据的分片数量(1~31),对应记录下车牌统计算法中地区车牌统计算法和套牌检查算法的运行时间,采用Origin工具对两种算法的实验结果绘制散点图,并进行曲线拟合。两种算法的输入数据分片数量与算法运行时间的拟合结果分别如图2和图3所示。

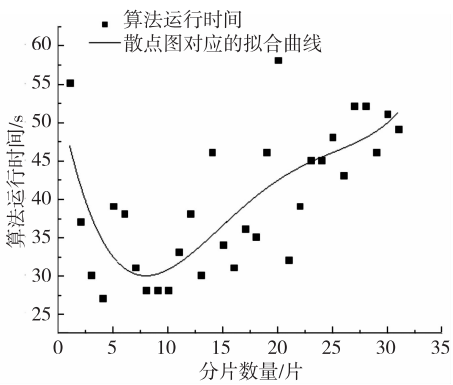

图2 地区车牌统计算法的散点与拟合曲线

Fig. 2 Scatter and graph of regional license plate statistical algorithm

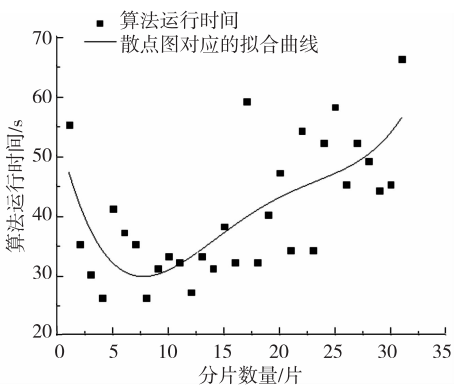

图3 套牌检查算法的散点与拟合曲线

Fig. 3 Scatter and graph of the vehicle deck inspection algorithm

从图2和图3可以看出对应的点非常分散,难以看出其中的规律。这是由于分片处理时,数据的分片具有偶然性所致。在数据处理中,通常进行多次实验来减少统计涨落的影响。

2.2.2 多次实验的结果 考虑到单次试验所得出的结果存在偶然性,影响拟合的准确性。因此,增加实验次数至5次,并对这5次的试验结果求数学期望(或均值),即:

$$E(x)=\sum x_i p_i \ , \tag{1}$$

其中,i表示实验的次数,x表示算法的运行时间,p表示x对应取值的概率。

将此期望值作为最终数据进行拟合。地区车牌统计算法的部分(取分片数分别为5,10,15,20)运行时间结果及计算完成后的运行时间期望值形成表1,以表1第1行为例,E(x)=1/5×39+1/5×38+1/5×36+1/5×

32＋1/5×27＝34。

套牌检查算法实验结果处理过程与地区车牌统计算法相同。

采用多次实验求得统计平均之后的输入数据分片数量与算法运行时间的拟合结果分别如图 4 和图 5 所示。

表 1　部分分片数量实验结果

Tab. 1　Partial slice number experiment result

输入数据分片数	第 1 次运行时间/s	第 2 次运行时间/s	第 3 次运行时间/s	第 4 次运行时间/s	第 5 次运行时间/s	运行时间期望值/s
5	39	38	36	32	27	34
10	28	30	38	30	26	30
15	34	39	38	51	34	39
20	58	31	34	53	50	45

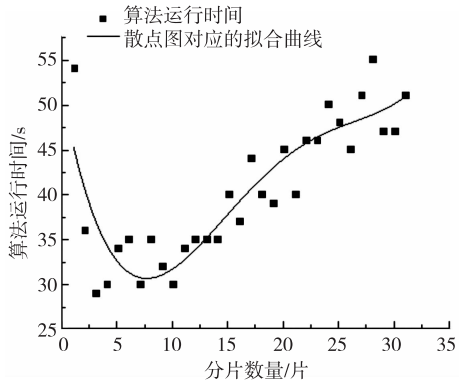

图 4　改进后地区车牌算法散点与拟合曲线

Fig. 4　Improved regional license plate algorithm scatter and graph

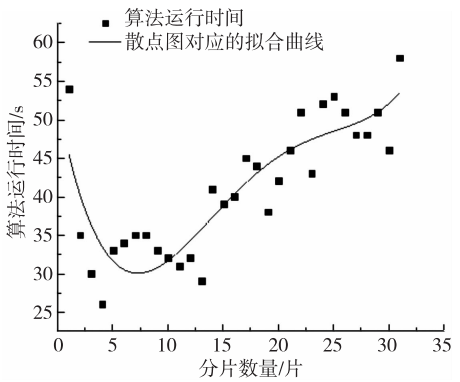

图 5　改进后套牌检查算法散点与拟合曲线

Fig. 5　Improved deck check algorithm scatter and graph

从图 4 和图 5 可以看出，经过多次实验得出的数据绘制的图形，散点更加集中，拟合曲线也能显示出需要的结果。

2.2.3 实验结果分析　地区车牌统计算法和套牌检查算法实验统计数据改进前拟合出的表达式分别为：

$$y=(3.252\ 44\mathrm{e}-4)x^4-0.025\ 92x^3+0.721\ 21x^2-7.212\ 13x+53.394\ 95, \tag{3}$$

$$y=(4.108\mathrm{e}-4)x^4-0.030\ 79x^3+0.814\ 17x^2-7.816\ 78x+54.185\ 59。 \tag{4}$$

地区车牌统计算法和套牌检查算法实验统计数据改进后拟合出的表达式分别为：

$$y=(2.850\ 74\mathrm{e}-4)x^4-0.023\ 4x^3+0.660\ 92x^2-6.498\ 63x+51.132\ 1, \tag{5}$$

$$y=(3.719\ 32\mathrm{e}-4)x^4-0.029\ 04x^3+0.779\ 51x^2-7.290\ 99x+51.948\ 25。 \tag{6}$$

(3)～(6)式中，x 表示输入数据分片数量，y 表示算法运行时间。

残差是指观测值与对应估计值的差，此处指数据点和在拟合曲线上相应位置的差。残差平方和则指所有残差的平方之和，是判断拟合程度一个重要标准，数值越小则表示拟合效果越好。y_i 表示实际测量出的数据，$\hat{y}$ 表示拟合出的数值，残差平方和的计算公式为：$SSE=\sum_i (y_i-\hat{y})^2$。

校正决定系数是为减少实际结果与拟合数据之间的误差而设定的系数，大小在 0 到 1 之间，数值大则拟合结果优。n 为样本数量，p 为特征数量，校正决定系数的计算公式为：$R^2_{\text{adjusted}}=1-\dfrac{(1-R^2)(n-1)}{n-p-1}$，这里 R^2 表示决

定系数,且 $R^2=1-\frac{\sum(y_i-\hat{y})^2}{\sum(y_i-\overline{y})^2}$,$\overline{y}$为 y_i 的平均值。

未改进实验统计数据时地区车牌统计算法与套牌检查算法拟合结果的残差平方和分别为:40.539 22 和64.950 09,校正决定系数分别为:0.501 32 和0.444 8。改进后地区车牌统计算法和套牌检查算法的残差平方和分别为:14.142 75 和20.576 47,校正决定系数分别为:0.759 44 和0.725 83。综上所述,改进实验统计数据后的拟合结果优于改进前的拟合结果,更符合实际测量结果。

地区车牌统计算法和套牌检查算法的运行时间与输入数据的分片数量存在函数关系,运行时间均随分片数量的增加呈现先迅速减少后增加的情况。地区车牌统计算法的运行时间在相邻的两个分片数量之间差距较小,但当分片数量为1和2时,算法的运行时间差距明显。套牌检查算法也存在类似的情况,除此之外,套牌检查算法的实验统计数据多呈块状分布,而地区车牌统计算法的实验统计数据则围绕着拟合出的曲线均匀分布。由实验拟合曲线可知,在输入数据的分片数量取值位于[5,10]区间时,地区车牌统计算法与套牌检查算法的运行时间最短,实际情况与拟合结果基本符合。

3 结束语

本文重点研究了输入数据的分片数量与车牌统计算法的运行时间之间存在的函数关系,并找到了最优的数据分片方案提高算法的运行效率。本次研究为今后的类似研究、应用提供了一个普遍适用思路,在研究以Map Reduce并行计算模型为基础实现的算法时,可参考本文中有关输入数据分片数量与算法运行时间的关系研究,如果实验平台完全相同可以采用已拟合出的公式直接进行时间的计算。但本次实验的拟合结果采用了特定的实验平台(云创),当输入数据、算法运行或实验环境等发生改变时会对表达式的参数造成怎样的影响,是下一步研究的方向。

参考文献:

[1] 胡琛,秦实宏.基于色彩纹理的车牌定位系统设计[J].武汉工程大学学报,2017,39(3):273-280.
HU C,QIN S H.Design of license plate location system based on color texture[J].Journal of Wuhan University of Technology,2017,39(3):273-280.

[2] 李林青,彭进业,冯晓毅.基于边缘分析和颜色统计的车牌精确定位新方法[J].计算机应用研究,2012,29(1):336-339.
LI L Q,PENG J Y,FENG X Y.A new method for precise positioning of license plates based on edge analysis and color statistics[J].Journal of Computer Applications,2012,29(1):336-339.

[3] 郑贵林,吴黄子桑.基于MSER与边缘投影的车牌定位算法[J].计算机工程与设计,2019,40(1):241-244.
ZHENG G L,WU H Z S.Vehicle license plate location algorithm based on MSER and edge projection[J].Computer Engineering and Design,2019,40(1):241-244.

[4] 高聪,王福龙.基于模板匹配和局部HOG特征的车牌识别算法[J].计算机系统应用,2017,26(1):122-128.
GAO C,WANG F L.License plate recognition algorithm based on template matching and local HOG features[J].Journal of Computer Systems,2017,26(1):122-128.

[5] 盛兆亮,高军伟.基于区域统计和BP神经网络的车牌识别[J].电子测量技术,2019,42(8):78-82.
SHENG Z L,GAO J W.License plate recognition based on regional statistics and BP neural network[J].Electronic Measurement Technology,2019,42(8):78-82.

[6] 王艳,谢广苏,沈晓宇.一种基于MSER和SWT的新型车牌检测识别方法研究[J].计量学报,2019,40(1):82-90.
WANG Y,XIE G S,SHEN X Y.A new method of vehicle license plate detection and recognition based on MSER and SWT[J].Acta Metrologica Sinica,2019,40(1):82-90.

[7] 杨英仓.基于字符包络和笔画宽度的伪车牌判断方法[J].计算机应用与软件,2017,34(3):222-226.
YANG Y C.Pseudo-license plate judgment method based on character envelope and stroke width[J].Computer Applications and Software,2017,34(3):222-226.

[8] 费继友,谢金路,李花,等.基于字符特征约束的自适应车牌校正提取[J].仪器仪表学报,2016,37(3):632-639.
FEI J Y,XIE J L,LI H,et al.Adaptive license plate correction based on character feature constraints[J].Chinese Journal of Scientific Instrument,2016,37(3):632-639.

[9] 沈文枫,张建蕾,周丁倩,等.大数据时代的车牌汉字识别[J].上海大学学报(自然科学版),2016,22(1):88-96.

SHEN W F, ZHANG J L, ZHOU D Q, et al. License plate Chinese character recognition in the age of big data[J]. Journal of Shanghai University(Natural Science), 2016, 22(1): 88-96.

[10] 曹波,韩燕波,王桂玲. 基于车牌识别大数据的伴随车辆组发现方法[J]. 计算机应用,2015,35(11):3203-3207.
CAO B, HAN Y B, WANG G L. A companion vehicle discovery method based on license plate recognition big data[J]. Journal of Computer Applications, 2015, 35(11): 3203-3207.

[11] 徐焕良,翟璐,薛卫,等. Hadoop 平台中 MapReduce 调度算法研究[J]. 计算机应用与软件,2015,32(5):1-6.
XU H L, ZHAI L, XUE W, et al. Research on MapReduce scheduling algorithm in Hadoop platform[J]. Computer Applications and Software, 2015, 32(5): 1-6.

Data Fragmentation of License Plate Statistics Algorithm in Big Data Environment

TANG Xing[1], FAN Yongsheng[1], FENG Ji[1], ZHONG Zhen[1], KONG Yadi[2]

(1. College of Computer and Information Science, Chongqing Normal University, Chongqing 401331;
2. College of Computer Technology and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: [**Purposes**]In the big data environment, finding the optimal number of input data fragments is an effective means to improve the efficiency of statistical algorithms. [**Methods**]Taking the license plate statistics as an example, several related parameters (minsize, maxsize) were adjusted to change the number of input data fragments in the algorithm, and the different running time of the algorithm was analyzed under different parameters. [**Findings**]The running efficiency at the optimal number of fragments was found 1 times faster than that the worst number of fragments. [**Conclusions**]In the big data processing, the running efficiency of the algorithm is improved by the reasonable number of input data fragments. The functional relationship between the number of fragments and the running time of the algorithm was also analyzed to find the optimal quantity interval of fragments.

Keywords: license plate statistics algorithm; number of fragments; running time of algorithm

(责任编辑 许 甲)