文章编号: 1672-8262(2016) 03-46-04 中图分类号: P208. 1 文献标识码: B

Hadoop 支持下海量出租车轨迹数据预处理技术研究

吕江波[1,2],张永忠[1,2]

(1. 兰州交通大学,甘肃 兰州 730070; 2. 兰州市勘察测绘研究院,甘肃 兰州 730030)

摘 要: 海量出租车轨迹数据预处理是轨迹数据挖掘和应用的前提。出租车轨迹数据是典型的大数据,传统的数据处理技术无法解决大规模出租车轨迹数据误差分析和处理问题,文章在分析轨迹数据误差来源和误差类型的基础上,提出基于 Hadoop 的海量出租车轨迹数据预处理模型,使用 Hive 实现轨迹数据误差统计分析,设计 MapReduce 并行处理程序实现轨迹数据预处理。实验结果表明,该模型可以有效解决大规模出租车轨迹数据预处理问题,处理方式可靠性较高,大大提高了轨迹数据预处理效率,为后期轨迹数据深入挖掘和分析奠定了基础。

关键词: 轨迹数据; Hadoop; 大数据; 数据预处理; 并行计算

1 引 言

随着卫星定位技术、无线通信技术和地理信息技术的迅速发展,定位设备在车载以及移动终端上得到广泛使用。许多城市的出租车都装载了定位设备,这些设备会定时将其位置信息传至服务中心,由此汇聚而成大规模的出租车轨迹数据。这些轨迹数据包含大量信息,已经开始应用于很多重要领域,如城市规划、智能交通、人类行为模式研究以及能源消耗等。近年来,由郑宇主导的“城市计算”,掀起了对 GPS 轨迹数据处理和分析的热潮。他们通过出租车在某区域的连通性评判区域规划的好坏; 利用出租车轨迹数据感知交通流量,为用户提供最快驾车路线和最佳拼车方案; 利用出租车轨迹数据为出租车司机提供最短时间拉到乘客的方案以及为乘客推荐最可能打到车的地点[1]。童晓君利用出租车轨迹数据分析居民出行热点区域和出行行为[2]。张富峥利用出租车在加油站等待时间估计加油站的排队长度,从而估计出此时加油站内车辆数量以及加油量。将全城加油站数据汇总,便可以计算出任意时刻有多少燃油被消耗掉[3]。与此同时,轨迹数据应用也面临着诸多挑战,首先,GPS 定位误差和人为因素导致轨迹数据存在许多不合理数据,这些数据严重影响数据分析结果,因此,数据预处理成为轨迹数据应用首先要解决的问题; 其次,轨迹数据是典型的大数据,以一个城市为单元,一天的出租车轨迹数据量大小从几 GB 到几十 GB 不等,多日的数据更可达 TB、PB 量级,常规的数据处理方式要处理如此大规模的数据几乎是不可能的,即使勉强可以处理,也需要花费很高的时间成本。大数据时代的到来,为海量轨迹数据处理提供了解决方案,Hadoop 作为目前主流的开源大数据分析平台之一,为海量数据分布式并行处理提供强大的平台支撑。Hadoop 可以运行在廉价硬件构建的计算机集群上,能够对大量数据进行可靠的、高效的、可扩展的分布式处理。

针对上述出租车轨迹应用中存在的数据误差和数据量大难处理两大问题,本文以 Hadoop 平台为基础,通过分析轨迹数据误差来源,总结误差类型,研究误差统计分析方法和处理方法,在此基础上提出基于 Hadoop 的轨迹数据预处理模型,实验证明该模型可以有效分析和处理海量轨迹数据误差,解决轨迹数据量大的处理瓶颈,处理方式更加可靠、高效。

2 基于 Hadoop 的出租车轨迹数据预处理模型研究

2.1 轨迹数据误差来源

由于 GPS 定位本身存在误差,加之出租车在实时动态获取数据,道路交通状态复杂性等原因,在海量的出租车轨迹数据中存在许多不合理数据,虽然大数据分析中有少许错误数据不会对分析结果产生影响,但也要具体问题具体分析,少许的错误数据也会使结果相差很多。例如: 在计算出租车行驶距离时,因为位置偏离使用错误的 GPS 定位坐标计算的距离肯定相差很多,严重影响计算结果。我们将这些导致不合理数据的原因大致分为两类: 一类是与 GPS 设备有关的误差,一类是与人为因素有关的误差。

(1) 与 GPS 设备有关的误差。主要有多路径效应误差、GPS 信号遮挡误差和 GPS 设备故障。多路径效

* 收稿日期: 2016—03—04

作者简介: 吕江波(1989—) ,男,硕士研究生,主要研究方向: GIS 应用与开发。

应误差产生的原因是当出租车行驶到有高大建筑物或水面附近时，建筑物和水面对于电磁波具有强反射作用，产生的反射波进入接收天线时与直接来自卫星的信号产生干涉，从而使观测值偏离真值产生的误差。GPS 设备因建筑遮挡或外界有较强的电磁干扰等因素导致接收装置无法获取卫星信号，随机产生与真值相差较大的位置数据，产生“偏离现象”，这种现象在隧道行驶时特别严重。GPS 设备出现故障后未及时排除，设备采集的位置、时间和出租车状态等信息都会出现错误[4]。

(2) 与人为因素有关的误差，由于司机关闭车载设备，导致 GPS 数据间断传输，这样数据就会不连续，在关闭车载设备的时间段 GPS 数据空白，在连续计算行驶距离或时间时出现错误。司机未规范使用计价器，导致数据中出租车行驶状态与实际不符，分析轨迹数据发现个别出租车全天的车辆行驶状态都是空车或载客，这明显与实际不符。

2.2　轨迹数据误差分类

通过对轨迹样本数据分析，对计算结果产生较大影响的误差类型有以下几类:

(1) 经纬度出界。用经纬度描述轨迹点的位置，由于 GPS 设备误差导致轨迹点严重偏离超出研究区域范围的数据均为不合理数据[2]。

(2) 采集时间错误。主要有时间格式错误和时间无效。

(3) 车辆状态错误。车辆状态 0 表示空驶，1 表示载客。如果出租车全天空驶、全天载客或车辆状态存在非 0 或非 1 的值，则这些都是不合理数据[2]。

(4) 数据丢失。出租车轨迹是由许多在时间上相对连续的轨迹点构成，超出 15 min 不连续的轨迹点数据应该作为两条轨迹的分割点。

(5) 其他轨迹数据错误，主要有瞬时速度和行驶方向数值异常等。

2.3　Hadoop 技术体系

Hadoop 是一个分布式计算框架，它能在大量廉价的硬件设备组成的集群上运行海量数据并进行分布式计算。他处理的海量数据能达到 PB 级别，并且可以让应用程序在上千个节点中进行分布式处理。Hadoop 优点主要有: Hadoop 是低成本的，Hadoop 是开源软件，这样就可以降低成本，此外，不必购买服务器级别的硬件，便可以搭建一个强大的 Hadoop 集群; Hadoop 是可靠的，它假设计算过程和存储会失败，因此它维护多个工作数据副本，对失败的节点重新处理; Hadoop 是高效的，通过并行处理加快处理速度; Hadoop 还是可伸缩的，如果数据量增大或要求提高数据处理效率，Hadoop 集群可以通过提升硬件性能或增加节点数量实现扩展。Hadoop 主要由分布式存储 HDFS 和分布式计算 MapReduce 两部分构成。HDFS 是一个类似于 Google GFS 的开源分布式文件系统，它提供一个可扩展、高可靠、高可用的大规模数据分布式存储管理系统，基于物理上分布在各个数据存储节点的本地 Linux 系统的文件系统，为上层应用程序提供一个逻辑上成为整体的大规模数据存储文件系统。MapReduce 并行计算框架是一个并行化程序执行系统。它提供了一个包含 Map 和 Reduce 两个阶段的并行化处理模型和过程，提供一个并行化编程模型和接口，让程序员可以方便快速地编写大数据并行处理程序。此外，随着 Apache Hadoop 系统开源化发展，Hadoop 平台已经演进为一个包含许多相关子系统的完整的大数据处理系统，这些子系统有: HBase、Hive、Pig、Zookeeper、Avro 等[5~7]。

2.4　数据预处理模型

轨迹数据误差不可避免，而且在海量轨迹数据中误差数据的总量不容小觑，严重影响计算结果。为了剔除海量轨迹数据中不合理数据，提高处理效率，保证分析结果的正确性，本文提出了基于 Hadoop 的出租车轨迹数据预处理模型，具体模型如图 1 所示:

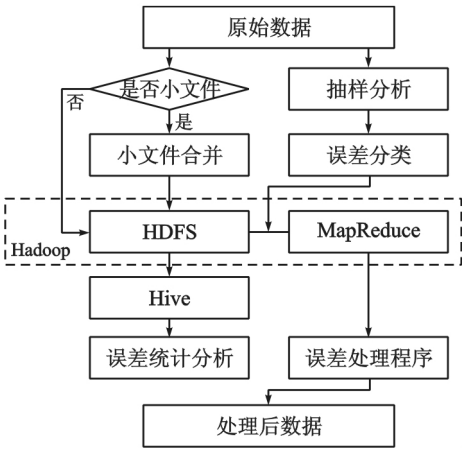

图 1　基于 Hadoop 的出租车轨迹数据预处理模型

首先对原始数据进行抽样分析，找出数据存在的问题，结合误差来源，进行误差分类，误差分类的结果是误差统计分析和误差处理的直接依据。为了处理海量轨迹数据，该模型运行在 Hadoop 集群上，其中 HDFS 负责数据分布式存储，MapReduce 负责数据并行处理。轨迹数据源文件通常是由许多小于 64 M 的小文件组成，为了避免大量小文件引起的 Hadoop 运行效率低问

题,编写小文件合并程序,合并后的轨迹文件直接存储在 HDFS 上。然后,在 Hadoop 集群上部署 Hive 组件,Hive 组件管理 Hadoop 中存储的数据,并提供类似 SQL 的查询语言,快速实现数据抽取、转换和加载,实质是将用户定制的类似 SQL 查询语言转换为 MapReduce 程序[8]。根据误差分类结果,使用 Hive 工具对轨迹数据误差进行统计分析。最后,针对各类误差类型编写 MapReduce 数据预处理程序,完成数据清洗。MapReduce 程序主要有 Map 函数和 Reduce 函数组成,Map 负责把任务分解成多个任务,Reduce 负责把分解后多任务处理的结果汇总起来,一些简单的数据预处理可以交给 Map,例如:数值超界、数值异常、格式校验等。复杂一些的数据预处理需要 Map 和 Reduce 相互配合。

3 应用实例

为了验证本文所提出的基于 Hadoop 的出租车轨迹数据预处理模型的可行性,以深圳市出租车轨迹数据预处理为应用案例进行测试。

3.1 数据概况

本文采用深圳市 13 799 辆出租车 2011 年 4 月 18 日 ~2011 年 4 月 26 日共 9 天的轨迹数据。轨迹数据文件均以车牌号命名,数据文件采用 csv 格式存储,共 13 799 个文件,约 2 亿条记录,数据量大小约为 11 G。数据文件记录了车牌号、采集时间、经度、纬度、行驶速度、行驶方向和车辆状态。表 1 为轨迹数据文件结构,表 2 为轨迹样例数据。

轨迹数据文件结构 表 1

字段名称	字段说明
NAME	车牌号
TIME	采集时间点(格式:YYYY/MM/DD hh:mm:ss)
JD	经度
WD	纬度
STATUS	车辆状态(0 表示空载,1 表示载客,非 0 或 1 表示异常值)
V	车速(单位:km/h)
ANGLE	行车方向(0 = 东;1 = 东南;2 = 南;3 = 西南;4 = 西;5 = 西北;6 = 北;7 = 东北)

轨迹样例数据 表 2

NAME	TIME	JD	WD	STATUS	V	ANGLE
粤 B00G25	2011/04/18 00:02:20	114.116 798	22.546 200	1	0	0
粤 B00G25	2011/04/18 00:12:28	114.118 401	22.595 966	1	0	2
粤 B00G25	2011/04/18 00:17:20	114.119 720	22.594 049	0	43	3
粤 B00G25	2011/04/18 00:22:20	114.111 153	22.586 884	0	2	3
粤 B00G25	2011/04/18 00:27:20	114.100 266	22.571 383	1	44	6
粤 B00G25	2011/04/18 00:32:22	114.098 648	22.580 183	0	22	1
粤 B00G25	2011/04/18 00:37:22	114.079 086	22.558 434	0	61	5
粤 B00G25	2011/04/18 00:42:22	114.067 734	22.553 917	0	76	3
粤 B00G25	2011/04/18 00:47:22	114.081 619	22.543 118	0	1	2
粤 B00G25	2011/04/18 00:52:22	114.088 501	22.543 182	0	0	0

3.2 环境搭建

本研究使用 VMware 在一台高性能的服务器上搭建 7 台虚拟机集群,其中 1 台为主节点,其余 6 台为数据节点。主节点配置 8 核中央处理器,8 G 内存,数据节点配置 4 核中央处理器,4 G 内存,操作系统均为 64 位 CentOS7,并行计算环境基于 Hadoop2.6,在 Hadoop 上部署 hive1.2.1 组件[8]。

3.3 技术路线

对实验数据进行抽样查看,发现存在经纬度超界、采集时间错误、车辆状态错误和数据丢失问题,按照上述出租车轨迹数据预处理模型对实验数据进行预处理,具体内容如下:

(1) 由于实验数据是由 13 799 个文件组成,单个文件大小 1 MB 左右,为了避免大量小文件引起的 Hadoop 运行效率低问题,编写程序实现小文件合并。具体思路是:从本地文件夹中读取文件,为了保证每行数据的完整性按照逐行读取方式读取数据,循环累计到单个文件达到阈值直接将文件保存到 HDFS,新建另一个文件开始输出,直到所有文件读取结束[9,10]。

(2) 按照轨迹数据误差类型,分别构造经纬度超界、采集时间错误、车辆状态错误 Hive 查询规则,并在 Hadoop 集群上运行,统计各类误差类型总数,抽取错误数据样例。

(3) 针对各类误差数据,编写 MapReduce 并行处理程序剔除这些不合理数据。轨迹大数据分析时经常需要进行路径分析,数据丢失问题会导致路径起始点错误,需要单独编写程序进行处理,本次实验处理的方法是将超出 15 min 不连续的轨迹点作为两条轨迹的

分割点。

3.4　实验结果

根据上述技术路线，对深圳市 13 799 辆出租车 9 天的轨迹数据进行预处理，分析得出，错误数据约占 6.68%，其中车辆状态错误约占 6.21%，经纬度出界约占 0.4%，采集时间错误约占 0.08%。编写数据预处理程序共剔除 1.14 千万条错误数据，耗时约 10 min。实验证明基于 Hadoop 的出租车轨迹数据预处理模型可以有效处理大规模轨迹数据中的常见的错误数据，运行可靠性较高，大大提高了轨迹数据预处理效率。实验数据误差统计分析结果如表 3 所示。

实验数据误差统计分析结果　　表 3

错误类型名称	错误数据示例	错误记录数	错误占比
经纬度出界	粤 BWG943,0.0,0.0 粤 BWG943,0.0,0.0 粤 B57F45,113.760765,22.797701 粤 B57F45,113.764252,22.797783 粤 B57F45,113.769852,22.798983	677 917	0.40%
车辆状态错误	车辆状态有:96,97,32,33,40,-47,-48 等 257 个异常值。	10 647 091	6.21%
采集时间错误	粤 BWG943,1970/01/01 08:00:00 粤 B809Q3,1986/07/04 00:49:43 粤 B809Q3,1987/09/24 10:14:49 粤 B809Q3,2002/09/13 13:21:31 粤 B809Q3,2007/01/21 06:50:45	128747	0.08%
汇总		11 453 755	6.68%

4　结　语

大数据时代的到来给海量出租车轨迹数据分析和应用提供了可能，而 Hadoop 作为目前重要的并行计算平台，为大数据的存储、管理和处理提供了技术支撑。本文认真分析了轨迹数据误差来源，对轨迹数据误差类型进行分类，在此基础上，结合 Hadoop 并行计算平台，提出基于 Hadoop 的出租车轨迹数据预处理模型。最后，本文对所提出的模型进行了验证。实验表明，该模型可以有效解决海量轨迹数据预处理问题，为后期轨迹数据深入挖掘和分析奠定了基础。

参考文献

[1] 王诏远，李天瑞，程尧等. 基于经验分布的打车概率和等待时间预测[J]. 计算机工程与应用，2015(24):254~259.

[2] 童晓君. 基于出租车 GPS 数据的居民出行行为分析[D]. 长沙:中南大学，2012.

[3] Zhang，Fuzheng，et al. "Sensing the pulse of urban refueling behavior."Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing ACM，2013:13~22.

[4] 温雅静. 基于热点载客区域的出租车应急调度方案研究[D]. 北京:北京交通大学，2014.

[5] 黄宜华，苗凯翔. 深入理解大数据、大数据处理与编程实践[M]. 北京:机械工业出版社，2014:31~36.

[6] 万川梅，谢正兰. Hadoop 应用开发实战详解[M]. 北京:中国铁道出版社，2013:11~21.

[7] Chuck Lam. Hadoop in Action[M]. 北京:人民邮电出版社，2011:2~5.

[8] 张岩，郭松，赵国海. 基于 Hadoop 的云计算试验平台搭建研究[J]. 沈阳师范大学学报·自然科学版，2013(1):85~89.

[9] 陈光景. Hadoop 小文件处理技术的研究和实现[D]. 南京:南京邮电大学，2013.

[10] 张丹. HDFS 中文件存储优化的相关技术研究[D]. 南京:南京师范大学，2013.

Based on the Hadoop Massive Taxi Trajectory Data Preprocessing Technology Research

Lv Jiangbo[1,2]，Zhang Yongzhong[1,2]

(1. Lanzhou Jiaotong University，Lanzhou 730070，China;
2. Lanzhou Surveying and Mapping Research Institute，Lanzhou 730030，China)

Abstract: Massive taxi trajectory data preprocessing is the precondition of trajectory data mining and the application. Taxi trajectory data is a typical big data，the traditional data processing technology can not solve the problem of large scale taxi track data error analysis and preprocessing，on the basis of analyzing the trajectory data error source and error type，study of mass trajectory error statistical analysis method and data processing method，the taxi trajectory data preprocessing model based on Hadoop is put forward，using the hive for the realization of the trajectory error statistics，design MapReduce parallel processing procedures for the realization of trajectory data preprocessing. Experimental results show that，the model can effectively solve the problem of large scale taxi trajectory data preprocessing，high reliability，greatly improve the efficiency of the trajectory data preprocessing，late for trajectory data digging and analysis laid a foundation.

Key words: trajectory data; hadoop; big data; data preprocessing; parallel computing