

引用格式:李剑斌,陈忠辉,苏志鹏. 基于区域密度划分的车辆轨迹数据分析方法研究 [J]. 微电子学与计算机, 2019, 36(2): 53-56. [LI J B, CHEN Z H, SU Z P. Taxi trajectory data analysis method based on regional density division [J]. Microelectronics & Computer, 2019, 36(2): 53-56.]

基于区域密度划分的车辆轨迹数据分析方法研究

李剑斌, 陈忠辉, 苏志鹏

(福州大学 物理与信息工程学院, 福建 福州 350100)

摘 要: 车联网是目前世界交通领域的一个热点. 车联网可以通过实时处理和挖掘这些数据中的新信息而变得更加智能化. 随着车联网的发展, 从设备监控中采集到的海量数据需要更高的数据存储和查询性能. 本文将 Apache Spark 作为基于 HDFS 高可用集群的统一集群计算平台, 提出并设计了一种基于 Spark 的车辆轨迹数据分析方法, 通过区域划分对海量数据进行了挖掘, 解决道路堵塞问题, 对辅助城市路径规划有一定的研究意义.

关键词: 车联网; 数据存储; 集群; 区域划分; 路径规划

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2019)02-0053-04

Taxi Trajectory Data Analysis Method Based on Regional Density Division

LI Jian-bin, CHEN Zhong-hui, SU Zhi-peng

(College of Physics and Information Engineering, Fuzhou University, Fuzhou 350100, China)

Abstract: Vehicle networking is a hot spot in the world transportation industry. Vehicle networking can become more intelligent by processing and mining new information in these data in real time. With the development of Vehicle networking, the massive data collected from device monitoring requires higher data storage and query performance. This paper uses Apache Spark as a unified cluster computing platform based on HDFS high-availability clusters, proposes and designs a Spark-based vehicle trajectory data analysis method, and uses area division to mine large amounts of data to solve road congestion problems and to assist urban roads. Planning has certain research significance.

Key words: vehicle networking; data storage; cluster; regional division; path planning

1 引言

目前基于轨迹数据的研究多利用大数据来处理城市交通面临的问题, 最早提出城市计算的概念是微软亚洲研究院, 城市规模的扩张以及人口数量的剧增等带来的一系列问题, 如环境污染、道路拥挤等, 在 Hadoop 分布式集群上部署 Spark 计算框架作为数据处理层, 加快车联网平台计算速度, 支持更多计算模式, 用于更多的应用场景^[1,2]. 通过大数据计算能够对诸多问题提供更好的解决方案, 基于城市轨迹数据和基于位置服务的研究工作显得至关重要,

如从海量轨迹数据中挖掘出最为密集的地点从而进行更合理的路径规划, 利用出租车轨迹数据预测并分析交通路段拥挤程度, 为辅助城市建设等提供参考, 也为出租车司机提供更加便利的驾驶路线.

2 车联网数据(分布式)存储处理平台

2.1 平台总体架构

本文针对车联网数据在实际应用中的特点与需求, 同时结合了海量数据处理技术, 构建了以分布式存储和处理为主的车联网数据分布式平台, 其总体架构如图 1 所示. 分为数据采集部分、存储分析部分

以及应用服务部分.

本文架构主要依据大数据解决方案而提出,利用 Hadoop 搭建车联网数据^[3]基础平台,在平台上采用分布式集群部署方式以及资源管理系统 Yarn,由 HDFS 和 HBase 建立数据存储层,分别对离线的数据和实时在线数据进行存储,配置 Zoo-keeper 对 Hadoop 进行协调服务. 为了避免发生离线数据块太大以至于不能完全读入内存的情况,Spark 仍需与 Hadoop 和 MapReduce 一同工作,因此,在平台上使用并行化计算框架 MapReduce 和基于内存的并行化计算框架 Spark 作为数据处理层,实现完整的、稳定的分布式平台构建,对海量车联网数据进行高效的分析.

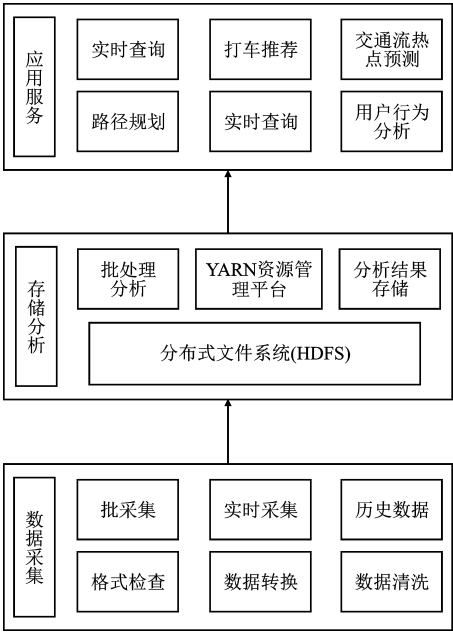


图 1 车联网数据存储处理平台架构

2.2 车联网数据分布式存储和处理

Spark 采用 Master/Slave 的架构,Spark 架构中有两种类型的节点,一种是 Master 节点,另一种是 Worker 工作节点. Master 相当于整个集群中的控制器,管理整个集群使得其能正常工作;Worker 听从主节点 Master 的命令进行计算工作并对自己进行汇报,Worker 中包含执行器 Executor 用于执行任;Client 用于给用户的客户端提交应用;Driver 中包含了定义了集群上的分布数据集应用的 main 函数用于进行了相关操作,主要用于派发 Task 任务,Spark 处理数据架构如图 2 所示.

本文的架构中,分别部署了一个 Master 节点和多个 Worker 节点构成集群进行分布式计算,同时利用 Spark 计算框架可以缩短访问延迟,具有可行

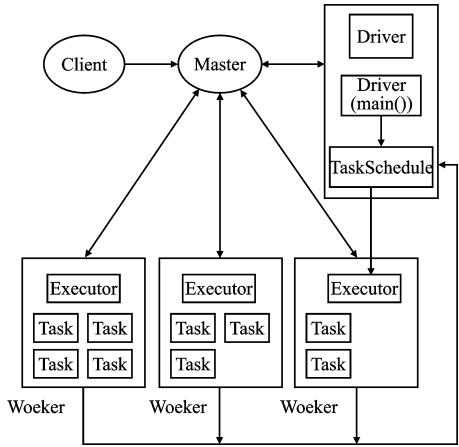


图 2 Spark 框架架构图

性的优点,对成都市出租车数据进行分析,实现交通流热点预测^[4]、用户行为分析等应用.

本实验数据集采用真实的成都市出租车 GPS 数据,数据类型复杂,数据量庞大,每一条数据包括:司机 ID、订单 ID、上车时间和下车时间、上车经度和纬度、下车经度和纬度. 如表 2—1.

表 2—1 GPS 数据示例

数据字段	数据示例
司机 ID	3f4. . . 2-uF2n
订单 ID	75d. . . 2dg4v
上车时间	1478500331
下车时间	1478502302
上车经度	104. 062460
上车纬度	30. 664400
下车经度	104. 094930
下车纬度	30. 619990

在实际的交通数据采集过程中由于受到天气、通信设备等客观因素影响,车辆网数据存在异常数据,因此在对数据进行处理分析前,需要对异常轨迹数据进行判定,包括对数据进行去重,异常数据剔除等,本文针对车联网数据挖掘的需要,利用 Spark 大数据计算框架,实现了基于区域划分的方法,对采集到的车联网数据进行了交通流热点区域分析,分析结果见实验部分.

3 基于区域划分的聚合方法

3.1 混合高斯密度算法

混合高斯模型^[5]是利用多个高斯分布函数的线性组合,拟合出任意类型的分布,通常用于解决同一集合下的数据包含多个不同的分布情况,本文采用

混合高斯密度算法来对海量数据进行初步的分类.

混合高斯模型公式为:

$$p(x)=\sum_{k=1}^K\pi_kN(x|\mu_k,\sum k)$$

(1)

式中, $N(x|\mu_k,\sum k)$ 称为混合模型中的第 k 个分量; π_k 表指的是概率,即是一个加权系数. 本文研究采用 k 值为 20,利用混合高斯密度算法对某市出租车 GPS 数据进行处理,结果如图 3 所示.

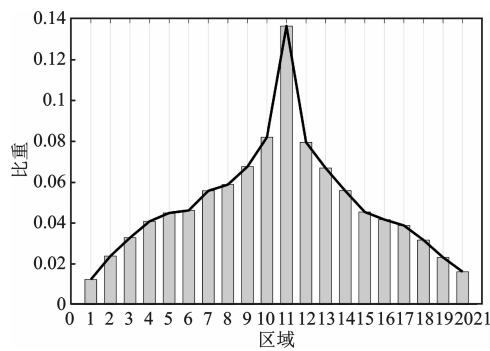


图 3 高斯密度分布

从图 3 中可以看出,将同一数据集分为 20 个部分,各个部分所占的比重不同,反应了各个集群的密集程度各不相同,实验将处理过的数据投射在地图.

从地图中可以发现,车流量密集区域所在经纬度基本符合某市道路建设以及热点区域的分布规律,进而可以初步判断出某市密集位置所在地,为接下来进行区域划分提供帮助.

3.2 基于区域密度的分割方法

城市路网数据^[6]是描述道路交通的一些地图数据,通常用矢量图的形式来表示,矢量模型通过笛卡尔坐标系来确定点、线以及多边形的位置,从而来表示不同的地理数据对象. 本研究采用的是利用矢量模型将各个路段存储的路网数据,是一系列的形状点的坐标构成的多段环形折线,将热点地点相互连接构成了整个路网,不同地点用不同的路网数据表示. 上述路网数据采用 Json(JavaScript Object Notation)格式^[7]进行存储,它基于 ECMAScript (欧洲计算机协会制定的 js 规范)的一个子集,采用完全独立于编程语言的文本格式来存储和表示数据,是一种轻量级的数据交换格式. 可用于描述空间几何对象,如点、线、面等等.

从图中可以看出通过高斯密度算法发现的几个密集目的地,作为路网数据,通过得到的密度集中的地点规划的 Json 环形折线区域,分析出热门区域.

4 实验结果及分析

本实验在搭建的 Spark 大数据处理平台下进行,平台有多台 PC 机(1GZ CPU 1GB 内存)组成,利用多节点来节省计算时间. 通过区域划分将海量车辆轨迹数据进行处理. 基于密度的区域划分领域,将密集区域中心定义为 POI,即兴趣点,指地图上的热门地点,可能是旅游景点、交通设施、大型商场等,构造感兴趣区域,指地图上由兴趣点连接形成的一个区域范围,是一个具体的区域,比如移动对象停留时间长的一个范围,像著名风景区等. 通过对热点区域的划分计算,可以分析出某市出租车的空间分布特征. 整个过程如下所示:

- (1)首先,实验利用高斯混合密度算法计算出车辆轨迹数据的密集点,并将其投射在地图上;
- (2)接着,将密集点作为车辆聚集目的地的兴趣点进行区域的划分;
- (3)然后,构造矢量模型^[8]将划分的路网数据存储在 Json 格式中;
- (4)最后通过 Spark 平台处理海量车辆轨迹数据,利用数据处理得到的实验结果分析出租车空间分布特征;

实验流程图如图 4 所示.

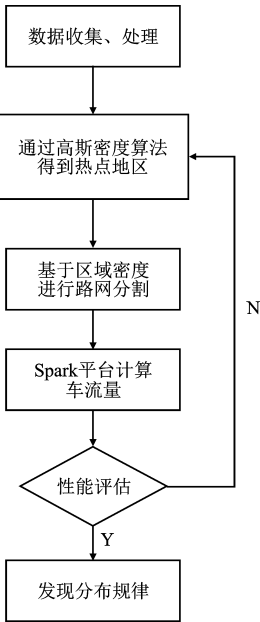


图 4 实验过程流程图

在每天有大量的车辆轨迹点^[9]的情况下,由于计算开销和内存不足带来的 I/O 开销导致了严重的计算结果延滞问题. 为解决该问题,实验采用 Spark 并行式计算框架,来完成高效的数据挖掘工

作. 本实验考虑数据分布问题利用一种新的空间划分^[10]方法来对数据点进行分割计算,统计出各个热门区域的车流量分布情况,如图 5 所示.

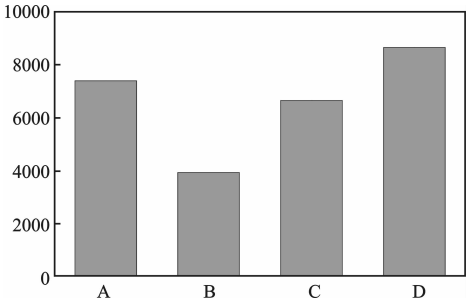


图 5 各个区域车流量分布图

从图 5 中可以看出,D 区以及 A 区的车流量相对较大,这可能是因为这几个区域附近存在一些热门地点或者主要交通路线,C 区这个区域在 10 号地铁线附近,因此可能会有较大的车流量;而 D 区的车流量大小可能跟附近有多所大学有关.

实验进一步将某市的部分区域的车流量通过时间来显示,如图 6 所示.

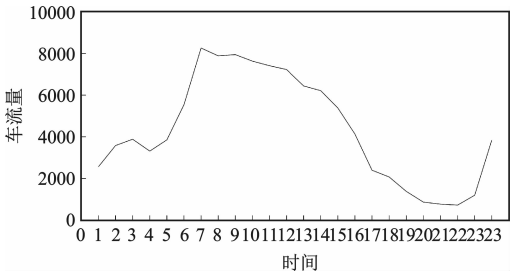


图 6 车流量时间分布图

从图 6 中可以观察到,出租车出行高峰期在早上 7 点到 11 点之间,结合区域流量的分布情况,对照成都市的路况,可以根据实际实验结果,在热点区域内设置一些出租车停靠站、出租车休息站以及打车区域等等,方便市民搭乘出租车,出租车司机换班、休息.

5 结束语

本文针对车联网系统目前所面临的问题和对于数据处理实时性、高效性和可靠性的要求,构建了完整的、易扩展的车联网数据分布式平台并进行了实验.在大数据处理方面,运用 Spark 大数据计算框架,兼容性较强,速率快,实验采用真实的成都市车

辆轨迹数据,利用 Spark 大数据处理平台,提出基于密度的区域划分方法来处理分析.最终,实验结果表明,研究结果能够反映成都市真实的车辆行为与时空分布特征.

参考文献:

[1] 陈滨. 基于 GPS 车辆轨迹数据的地图匹配算法研究[D]. 福州: 福建师范大学, 2015.

[2] 姜晓辉, 郭久武, 王博, 等. 分布式集群处理环境平台软件技术研究[J]. 吉林大学学报(信息科学版), 2015, 33(3):331-337.

[3] 孙川, 吴超仲, 褚端峰, 等. 基于车联网数据挖掘的营运车辆驾驶速度行为聚类研究[J]. 交通运输系统工程与信息, 2015, 15(6):82-87.

[4] 田保慧, 郭彬. 基于时空特征分析的短时交通流预测模型[J]. 重庆交通大学学报(自然科学版), 2016, 35(3):105-109.

[5] 郭伟, 高媛媛, 刘鑫焱. 改进的基于混合高斯模型的运动目标检测算法[J]. 计算机工程与应用, 2016, 52(13):195-200.

[6] Liu Y, Wang F, Xiao Y, et al. Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai[J]. Landscape & Urban Planning, 2012, 106(1):73-87.

[7] 龚建华. JSON 格式数据在 Web 开发中的应用[J]. 办公自动化, 2013(20):46-48.

[8] 郑坤, 付艳丽. 基于 HBase 和 GeoTools 的矢量空间数据存储模型研究[J]. 计算机应用与软件, 2015, 32(3):23-26.

[9] 夏楠, 姜波, 毕冬鸣. 利用路线轨迹点信息进行路线规划的方法及装置:, 中国, CN 101799301 A [P]. 2010.

[10] Anant R, Sunita J, Jalal A S, et al. A Density Based Algorithm for Discovering Density Varied Clusters in Large Spatial Databases[J]. International Journal of Computer Applications, 2011, 3(6):1-4.

作者简介:

李剑斌 男, (1993-), 硕士研究生. 研究方向为数据存储及处理. E-mail: 448163249@qq. com.

陈忠辉 男, (1961-), 教授, 硕士生导师. 研究方向为信号处理与通讯工程.

苏志鹏 男, (1993-), 硕士研究生. 研究方向为计算机图形学.