ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.11, No.36, December 2015
E-mail: xsjl@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

基于大数据的智能交通分析系统的设计与实现

苏刚,王坚,凌卫青

(同济大学CIMS研究中心,上海 201804)

摘要:随着社会经济的进步和交通运输业的快速发展,面对快速增长的城市道路交通数据,论文提出了基于Hadoop的智能交通分析系统设计方案,论文采用HBase分布式数据库存储城市道路静态RDF数据,采用Hive数据仓库存储城市道路交通数据,采用MapReduce编程模型对海量、异构城市道路交通数据进行分析,最后通过原型系统对整体方案进行验证。

关键词:交通大数据;Hadoop框架;智能交通

中图分类号:TP311 文献标识码:A 文章编号:1009-3044(2015)36-0044-03

DOI:10.14004/j.cnki.ckt.2015.3472

Design and Implementation of Intelligent Transportation Analysis System Based on Big Data

SU Gang,WANG Jian, LING Wei-qing

(CIMS Research Center,Tongji University,Shanghai 201804, China)

Abstract: With the progress of social economy and the rapid development of transportation industry, in the face of the rapid growth of urban road traffic data, the paper puts forward the design scheme of intelligent traffic analysis platform based on Hadoop, the paper adopts HBase distributed database storage static RDF data on urban road traffic data, the paper adopts the Hive data warehouse storage urban road traffic data, using MapReduce programming model to analyze massive, heterogeneous urban road traffic data, finally, using the prototype system to the overall plan for validation.

Key words: traffic big data; the Hadoop framework; intelligent transportation

随着经济发展和人们生活水平的提高,城市汽车数量呈快速增长趋势,交通事故、交通拥堵等都对城市道路交通造成很大负担。在大数据时代,城市道路交通数据也呈指数级增长,面对其多源、异构、数量巨大等特点,如何快速高效进行数据分析,并将分析结果充分利用,从而提高城市交通运行效率是城市交通亟待解决的一大难题[1]。本文在分析交通大数据特点和当前大数据技术基础上,基于Hadoop相关组件和计算模型进行整体架构设计和应用响应流程设计,最终实现了对海量多源异构交通大数据的快速高效处理。

1 系统设计

1.1 交通大数据特点及其面临的问题

交通数据主要包括交通领域的道路信息、车辆信息等,通常将这些数据分为静态数据和动态数据。静态数据包括道路信息数据、道路设施数据、停车场数据等;动态数据包括如线圈设备、视频设备等采集到的交通信息数据。通过总结,可得出交通大数据主要具有以下特征:多源性、多维性、海量性、动态性和异构性[2]。对于静态数据来说,其主要是结构化数据,相比动态数据,其数量较小,且相对固定,但因数据采集机构和设备的差异,造成其具有较大的语义异构性,因此本系统引入本体模型对静态数据进行处理,解决语义异构问题;对于动态数据来说,则因为交通运输的快速发展以及大数据时代对数据需求的不断提高,使其具有更强的多源异构性,并且动态数据体量巨大,如何存储动态数据以及满足快速查询和计算的需求,是目前交通大数据面临的主要问题[3]。对静态数据本体,需要通过RDF三元组来表示,为了满足存储稳定、快速查询和方便扩展等特点,引入Hadoop框架的HBase组件进行存储;对动态数据,由于其体量巨大,且具有多源异构的特点,同时结合语义网技术,引入Hadoop框架的Hive组件进行存储,以满足计算需求。在交通大数据计算方面,除了传统的数据挖掘方法实现外,还要考虑数据挖掘时交通分析相关算法的并行实现,MapReduce计算模型用于大规模数据集的计算,可以方便、高效的将程序运行在分布式系统上[4]。

1.2 系统架构设计

本系统的设计与实现,是基于Hadoop分布式系统架构的,Hadoop框架的核心部分包括HDFS和MapReduce,HDFS为底层数据存储提供支持,MapReduce为海量数据提供计算支持。Hadoop的主要优势体现在其可扩展性、高效性和可靠性等方面。根据交通大数据的特点,基于Hadoop框架设计智能交通分析平台总体架构,如图1所示。该总体架构组要分为数据采集层、数据存储层、数据分析层和应用层。

收稿日期:2015-12-05

作者简介:苏刚(1991—),男,吉林白山人,同济大学硕士研究生,研究方向为大数据分析;王坚,男,同济大学CIMS研究中心主任,教授,博士生导师;凌卫青,男,博士,同济大学CIMS研究中心副研究员。

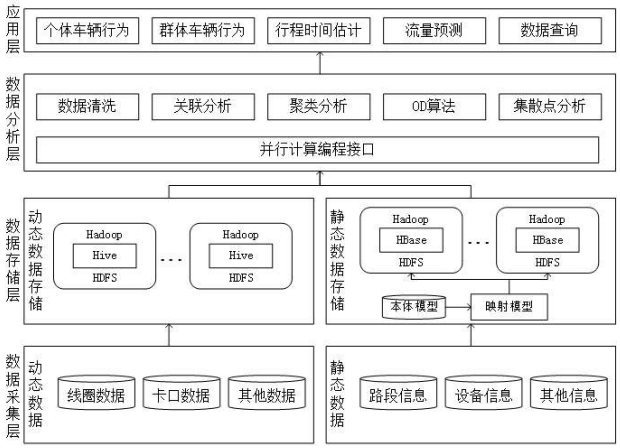

图1 智能交通分析平台总体架构

1)数据采集层

数据采集层分为动态数据采集和静态数据采集,其中动态数据包括线圈采集的流量数据、车型数据和车速数据等;卡口数据包括车牌数据、采集时间等数据。静态数据包括路段信息,设备信息等不会因采集时间不同而改变的数据。动态数据经过整合存储到分布式数据库中,静态数据根据后期本体模型映射成RDF数据进行存储。

2)数据存储层

数据存储层将采集到的动态数据和静态数据存储到Hadoop计算机集群下,计算机集群采用主/从架构,主节点为管理节点,存在一个,负责记录数据存储位置等信息;从节点为数据节点,存在多个,是数据真正的物理存储位置。在本框架下,动态数据整合为一张数据大表,存储于hive数据仓库下;静态数据根据本体模型进行映射,成为RDF数据,然后写入HBase分布式数据库中[5]。

3)数据分析层

数据分析层根据不同的交通大数据分析需求,采用MapReduce编程模型进行数据处理和计算,在本系统中主要引入数据清洗、关联分析、聚类分析、OD算法等模块。由于外界因素的干扰和数据采集设备本身具有的缺陷,首先需要对数据进行清洗,去除不合理和异常的数据,如车牌数据中全为"0"的数据。

关联分析和聚类分析都是数据挖掘中的重要分析方法,并且可以充分利用到交通拥堵等分析中[6]。如利用关联分析,针对拥堵瓶颈点,选取车辆型号、车辆属地、天气情况等信息进行关联挖掘,从而确定影响交通拥堵的主要因素。聚类分析则可以通过多次迭代运算,寻找到合适的中心值,从而将道路交通流量密度进行等级划分,有利于为出行者提供参考。

OD计算则是统计一定区域内每一辆车出行的起始点和终止点,记录每一辆车经过的线圈信息,以及根据起止点计算车辆位移。在MapReduce计算模型下,在Map阶段根据车牌号码进行合并,将结果传递给Reduce阶段;在Reduce阶段,针对每一个车牌的数据的时间字段进行计算,将满足阈值条件的每组数据作为OD计算结果输出。OD计算结果可被二次利用,作为车辆出行行为统计和画像等应用的基础。

4)应用层

利用数据分析层计算出的结果,以接口的形式为交通行业各用户提供服务,根据各用户不同的需求提供不同的计算结果,同时可以在数据分析层增加新的计算模块来满足新的需求。如OD计算结果可以用于进行个体出行行为和群体出行行为的分析,聚类分析的结果可用于行程时间估计。

1.3 应用响应流程设计

根据系统的总体架构对应用响应流程进行设计,如图2所示。应用相应流程主要包括用户请求、请求验证、应用匹配、数据计算模块。

首先,用户根据个人需求进行应用请求,系统自动匹配其应用所需要的算法模块和数据,如果出现不存在此应用、数据不全或者输入条件不足等情况,则将错误信息返回给用户,如果输入条件满足计算需求,则通过请求验证,进入应用匹配阶段。

应用匹配包括数据匹配和算法匹配两方面。根据用户的请求,通过本体推理的方式查找出该应用所需要的数据类型,如针对OD算法,用户需要输入道路范围,时间和车牌集合,然后根据这些范围和OD算法本身所需要的数据,推理出所要查找的线圈代码、车牌号码、线圈坐标和采集时间等信息,然后进入语义网搜索阶段。语义网搜索部分根据数据类型,利用SPARQL查询HBase数据库查询出所需数据所在的实际地址[7],将地址结果传送给大数据搜索模块,通过对Hive中存储的大表数据进行提取,从而获得最终要计算的数据。另一方面,通过将最小力度的算法进行封装和重组的方式获得复杂应用的算法模块。由于已经将各个应用对应的算法进行封装,并采用接口的形式提供服务,算法匹配部分只需在算法库中进行查找,获得此次请求所需要的算法,这里以单车OD算法为例。最后将查找到的数据和算法相结合得出计算结果返回给用户。

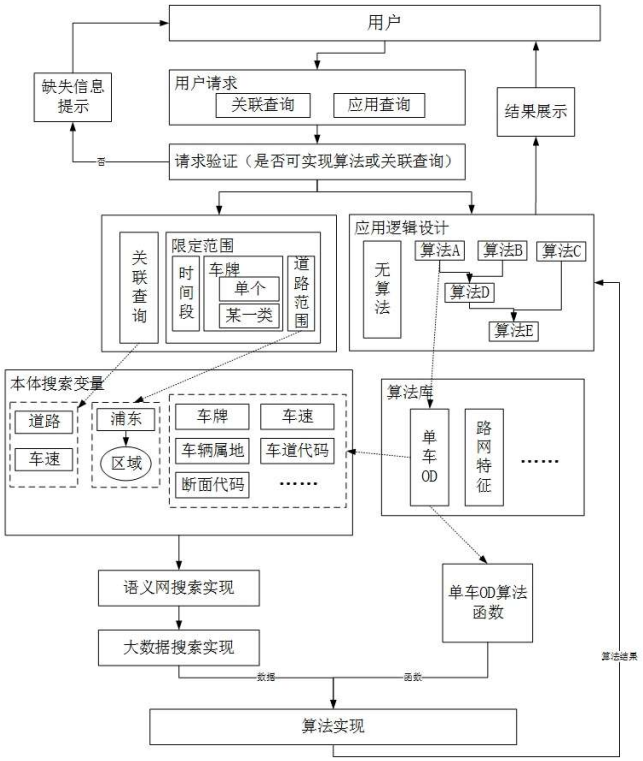

图2 应用响应流程

2 系统实现

根据本文论述,构建了相应的智能交通大数据实验平台,实验室集群采用一主两从,分析平台采用Hadoop-2.5.2,HBase-0.98.6和Apache-hive-0.13.1-bin。数据来自卡口数据,

总计2GB，14000000万条。相应的算法采用Java语言实现。以OD算法为例，实现了多源异构交通大数据的计算。

OD算法实现：

OD算法针对卡口数据，道路静态数据进行统计和计算，获得每辆车的每一条OD记录，一条OD记录包括O点断面代码、D点断面代码、号牌号码、经过时间和位移距离等信息。其中位移距离根据OD两点的坐标，利用公式 $L=\sqrt{(x_1-x_2)^2-(y_1-y_2)^2}$ 进行计算，OD两点坐标则利用SPARQL在HBase中查询得到。最后对计算过程进行封装并以接口的形式提供服务。计算结果如图3所示。

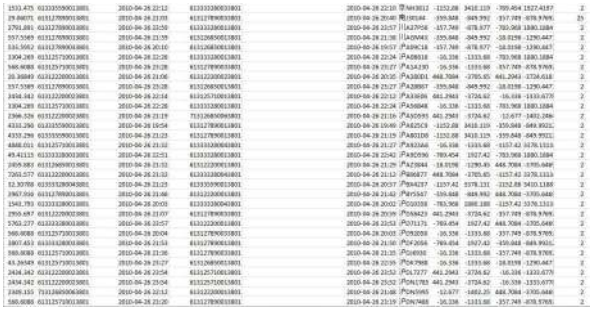

图3 OD计算结果

3结论

本文结合城市交通现状和大数据发展状况，提出了智能交通分析平台构建方法，完成了系统总体架构设计，应用响应流程设计以及原型系统的实现，并初步实现了交通大数据的快速计算。Hadoop自身具有的可扩展性、高效性和可靠性也为快速准确的进行交通大数据分析提供保障。基于此平台得到的分析结果也可以被再次利用，为交通研究、交通管理和交通规划等提供支持，对缓解城市交通拥堵等问题具有一定的帮助。

参考文献：

[1] 陆化普, 李瑞敏. 城市智能交通系统的发展现状与趋势[J]. 工程研究：跨学科视野中的工程, 2014(1):6-19.

[2] 周为钢, 杨良怀, 潘建,等. 论智能交通大数据处理平台之构建[C]// 第八届中国智能交通年会论文集. 2013.

[3] 边伟. 智能交通大数据综合平台的应用浅析[J]. 中国安防, 2015(15):68-71.

[4] 查礼. 基于Hadoop的大数据计算技术[J]. 科研信息化技术与应用, 2012(6).

[5] 朱敏, 程佳, 柏文阳. 一种基于HBase的RDF数据存储模型[J]. 计算机研究与发展, 2013, 50(z1):23-31.

[6] 郝晓飞, 谭跃生, 王静宇. Hadoop平台上Apriori算法并行化研究与实现[J]. 计算机与现代化, 2013(3):1-4.

[7] 谢桂芳. SPARQL-一种新型的RDF查询语言[J]. 湘南学院学报, 2009, 30(2):80-84.