Computer Science and Application 计算机科学与应用, 2020, 10(8), 1471-1479
Published Online August 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.108154

Hans汉斯

Traffic Spatiotemporal Big Data Analysis and Mining System Based on Mobile Phone Signaling Data

Mengli Lu, Siqiang Wu, Changchao Chen, Shichao Feng, Wei Li*

School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou Jiangxi
Email: *10209167@qq.com

Received: Jul. 26th, 2020; accepted: Aug. 10th, 2020; published: Aug. 17th, 2020

Abstract

This paper builds a pseudo distributed big data environment, and takes the user's mobile phone signaling data as the data set. The data processing mainly includes six steps: data input, data cleaning, data buffering, data processing, data storage and data visualization. Firstly, flume's tail-dir source is used to clean the mobile phone signaling data, and the data processing algorithm is designed and processed by spark programming. Then the processing results are stored in redis, MySQL and HBase databases to analyze the population thermal density, user residence and travel behavior. It creates the springboot background, references the Gode map API, and visualizes the processed data, which is convenient for the traffic management department to control the traffic situation and dispatch the vehicles.

Keywords

Thermal Diagram, OD Diagram, Dwell Diagram, Travel Analysis, Big Data, Mobile Phone Signaling

基于手机信令数据的交通时空大数据分析挖掘系统

卢梦丽，邬思强，陈长超，冯时超，李　伟*

江西理工大学信息工程学院，江西　赣州
Email: *10209167@qq.com

收稿日期：2020年7月26日；录用日期：2020年8月10日；发布日期：2020年8月17日

*通讯作者。

文章引用: 卢梦丽, 邬思强, 陈长超, 冯时超, 李伟. 基于手机信令数据的交通时空大数据分析挖掘系统[J]. 计算机科学与应用, 2020, 10(8): 1471-1479. DOI: 10.12677/csa.2020.108154

摘　要

搭建大数据伪分布式环境，将用户手机信令数据作为数据集，数据的处理主要经过数据输入、数据清洗、数据缓冲、数据处理、数据存储和数据可视化六步。首先采用Flume的Taildir Source清洗手机信令数据，通过设计数据处理算法，利用Spark编程处理，之后将处理结果存入redis、mysql、Hbase数据库中，来分析人口热力密度、用户驻留和出行行为。创建SpringBoot后台，引用高德地图API，将处理所得的数据可视化，方便交通管理部门对交通状况的掌控和交通工具的调度。

关键词

热力图，OD图，驻留图，出行分析，大数据，手机信令

1. 引言

近年来，全世界对大数据应用的关注日益提高，不断在数据中挖掘其内在潜力和价值。大数据在集成和组合数据上具有优势，可帮助形成高效的智能调度能力，优化公共交通信息资源的配置，同时辅助制定出较好的统筹协调方案。

交通信息化的核心技术逐渐转向智能交通。采用信息化的技术方法，让政府以及交通管理部门能及时了解整体的交通状况，根据实际数据做出准确的客流分析并提出切实可效的策略，解决现有的交通问题。

手机信令数据是一种新型的大数据源，与其他类型的数据相比，其具有实时性、完整性、出行时空全覆盖性等其他数据源所不拥有的优势，在各类规划中尤其是交通大数据分析中具有独特的应用优势。另外，作为人们生活中必不可少的交流工具，手机通常一直在工作，故数据记录时间长。而且生活中手机的普及率较高，几乎所有城市居民都可以通过移动手机进行监控，而无需额外的成本[1]。

国外方面，文献[2]提出能够识别用户驻留地点的方法，将测试用户作为实验样本研究用户的出行特征和规律。文献[3]中作者开发一种基于手机数据的智能工具，帮助交通管理机构探索市民的移动规律和优化公共交通。

国内方面，吴乃星[1]等将手机信令数据作为数据集，分析出行需求的空间结构、连续空间分布特征和人口区域运动规律，将分析结果以 OD 图、密度图和流线图的形式可视化。杨飞[4]通过手机定位平面坐标对用户进行追踪，获取居民运动状态，分析活动位置的集中特征，来得到用户的出行 OD 数据。

本文以手机信令数据作为数据集，搭建大数据伪分布式环境，将数据清洗后，根据设计的算法处理得到用户分布密度数据、出行轨迹数据、出行方式数据，之后开发数据可视化后台系统将数据直观地呈现给系统的用户。

2. 系统整体架构

以用户手机信令数据作为待处理数据，数据的处理主要分为数据输入、数据清洗、数据缓冲、数据处理、数据存储和数据可视化六步。首先在 Linux 系统下配置数据处理每个步骤所需的环境，然后设计数据处理的数据流向，具体的设计如图 1 数据处理过程设计图所示。

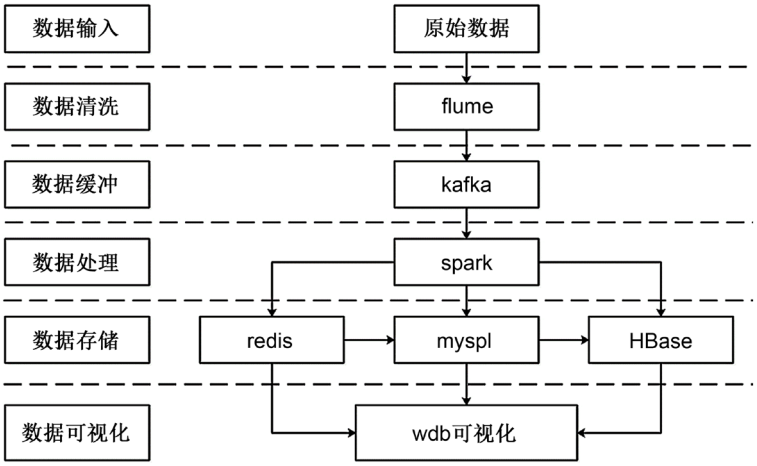

Figure 1. Data processing process design diagram
图 1. 数据处理过程设计图

数据清洗：将原始数据输入到 flume 进行数据清洗，得到所需的数据并存入 kafka 中准备处理。

数据处理：spark 从 kafka 中读取数据并对其进行处理。

数据存储：将处理后的不同类型的数据分别输出到实时性要求不同的 redis、mysql 和 HBase 中存储。

数据可视化：创建 springboot 架构的 web 后台，分别从 redis、mysql 和 HBase 中读取需要的数据传到前端页面进行展示。

3. 系统算法设计

3.1. 数据说明

本文采用手机信令数据。数据内容包含原始数据.csv、基站经纬度数据.csv 和出行方式静态数据.csv 三份数据。数据的详细内容如表 1 所示。

Table 1. Data details table
表 1. 数据详细信息表

原始数据.csv	信息记录开始时间(timestamp)，用户识别码(imsi)，基站位置区码(lac_id)，扇区编号(cell_id)，信息记录结束时间(timestamp1)
基站经纬度.csv	经度(longitude)，纬度(latitude)，基站信息(laci)
出行方式静态数据.csv	经度(longitude)，纬度(latitude)，交通方式(mode)，站名(mode_name)，线路(mode_num)

3.2. 数据清洗

数据清洗使用 Flume 的 Taildir Source 从原始数据.csv 文件读取数据并使用自定义的 flume Interceptor 对其进行清洗，数据清洗操作分为如下三类：

1) 筛选出需要的数据项并去除数据项缺失的数据；
2) 数据不合法与日期不是 2018 年 10 月 03 日的数据；
3) 将原始数据中的时间戳转换为日期格式。

设计了三个步骤的判断：

Step1：判断数据项是否完整，用于筛除数据项有缺失的数据；

Step2：判断数据项是否合法，刷去数据项格式不正确的不合法数据；

Step3：最后判断是否为指定日期的数据，得到目标数据存入 Kafka。

3.3. 数据处理算法设计

数据处理判断对象均为坐标，故以 k，v 键值对的形式存储[基站位置，基站 id]，对从 kafka 获取的数据中的基站位置进行匹配并替换为基站坐标。数据处理的数据流均为从 kafka 中获取，在 Spark 中处理，之后实时数据存入 Redis，离线部分数据存入 Mysql。具体的处理步骤和算法设计如下所示：

1) 热力图数据处理

以基站为单位，统计某一时间点每个基站内包含的人数[5]，同时结合人口的移动情况。具体的算法设计如图 2 所示。经过处理后，得到[坐标，权值]形式的数据。

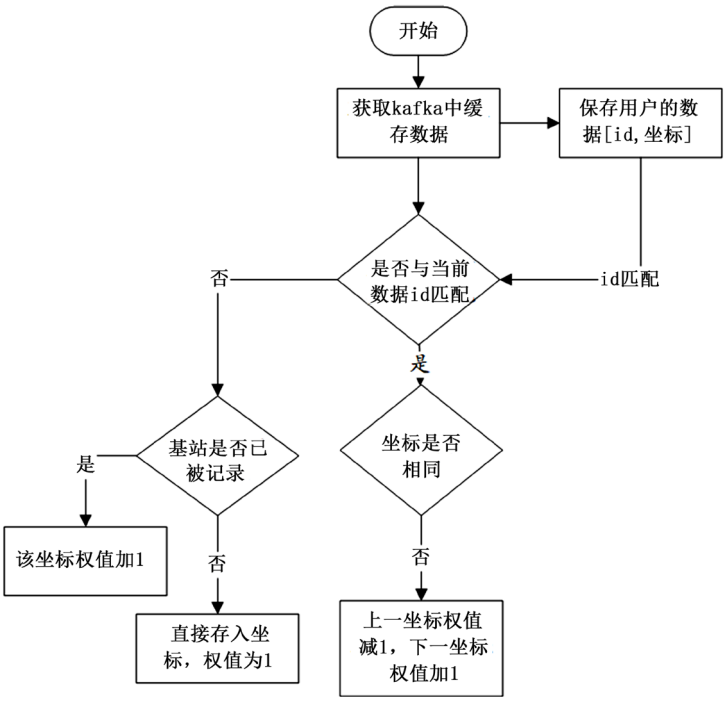

Figure 2. Heat map data processing flowchart
图 2. 热力图数据处理流程图

如图所示，将数据从 Kafka 中取出后保存用户的数据，然后判断 id 是否匹配、坐标是否匹配和基站是否已被记录。根据不同的判断结果对于权值和坐标进行不同的操作。

2) OD 图数据处理

以基站对为单位，统计出某一时段从一个基站到另一个基站的人数[6]，可实时刷新也可查看历史记录。通过对从 Kafka 中取出的数据判断其 id、坐标和时间，来确定是否对地区迁徙权值进行+1 操作。设计具体步骤如下：

Step1：获取 kafka 中缓存数据，存入 Map[id,坐标，时间]格式数据；

Step2：匹配到与当前 id 相同的 Map；

Step3：获取该 Map 的坐标和时间，判断用户上个坐标和当前坐标是否相同；

Step4：判断上一个坐标—>当前坐标的迁徙是否被记录。被记录过，该地区迁徙权值+1，否则存入该地区迁徙，权值为 1。

3) 驻留图数据处理

根据同一用户的坐标变化，判定驻留情况，在一段时间内，坐标无变化则判定为驻留。以基站为单位，统计每一个基站的驻留次数，驻留人数和驻留时间。

算法核心思想：判断同一用户的坐标是否相同，如果相同则为驻留，将该驻留点的权值加 1；否则判定为非驻留点，如果上条数据被记录过驻留且权值不为 0，则该坐标的权值减 1。

4) 出行方式分析数据处理

出行分析是将数据的分析标签化，分析每个用户的出行方式。该分析总体分为轨迹处理和出行方式标签化两个部分，对数据经过判断去除不符合的数据后，计算每次出行的特征值[7]，最终实现出行方式的分析，处理流程图如图 3 所示。

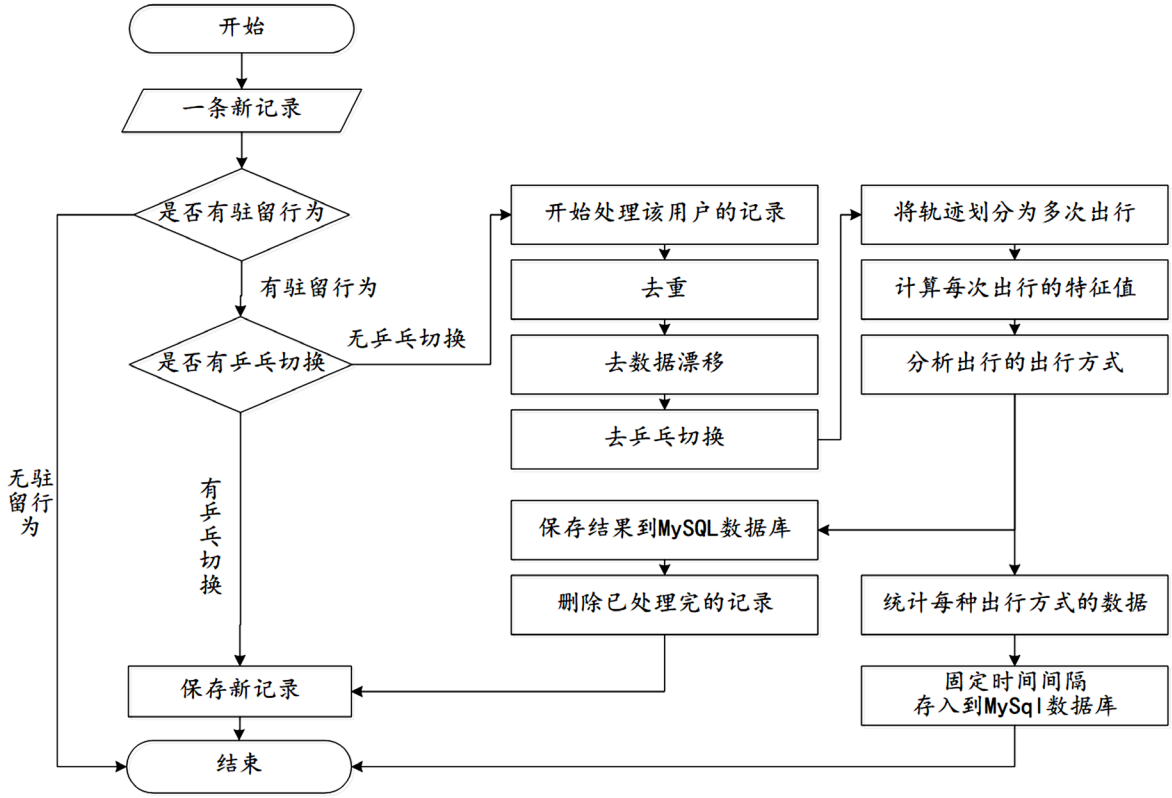

Figure 3. Flow chart of travel analysis processing
图 3. 出行分析处理流程图

a) 轨迹处理

轨迹处理是先对数据清洗后的数据进行去重，去乒乓切换，去数据漂移[8]等处理并将用户一条轨迹通过关键点拆分为多条出行路径，其中关键点包括驻留点与换乘点。

I) 去除重复的数据

用户在同一个基站范围内停留时，可能会产生多条坐标相同而时间不同的数据。对于此类数据选取最开始和最后的一条数据进行保留，中间数据可以舍弃。去除的数据示例如图 4 所示。

II) 去数据漂移

对于记录 A、B、C，A 到 B 的距离 << A 到 B 加 B 到 C 的距离，并且 A 到 B 与 B 到 C 的时间较短，可判定为发生了数据漂移，B 点为数据漂移点，将其去掉，只保留记录 A，C。如示例图 5 所示。

去除

time	imsi	location
2018-10-03 01:02:13	460000095007329090	坐标1
2018-10-03 01:12:13	460000095007329090	坐标1
...	...	...
2018-10-03 06:02:33	460000095007329090	坐标1
2018-10-03 06:13:15	460000095007329090	坐标1
2018-10-03 06:15:50	460000095007329090	坐标2

Figure 4. Example diagram of removing duplicate data
图 4. 去除重复的数据示例图

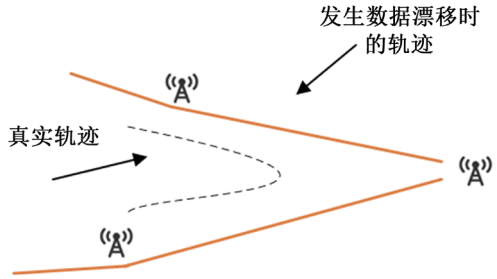

Figure 5. De-drift example diagram
图 5. 去漂移示例图

III) 去乒乓切换

乒乓切换分为高速乒乓切换和低速乒乓切换。高速乒乓切换一般在用户基站覆盖范围重合区移动时发生。低速乒乓切换一般在用户的活动区域同时在多个基站的覆盖范围内时发生。对于乒乓切换保留第一条与最后一条数据。如图 6 所示。

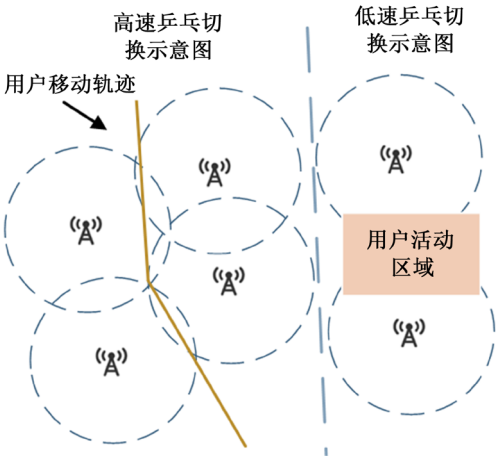

Figure 6. Example diagram of switching to ping-pong
图 6. 去乒乓切换示例图

b) 出行方式标签化

通过计算每一条出行记录特征值，再结合静态交通数据和交通方式的划分对每种交通方式的可能性进行分析，得出每种交通方式的权值，选取权值最大的方式为此处出行的出行方式[9]。实现流程如图 7 所示。

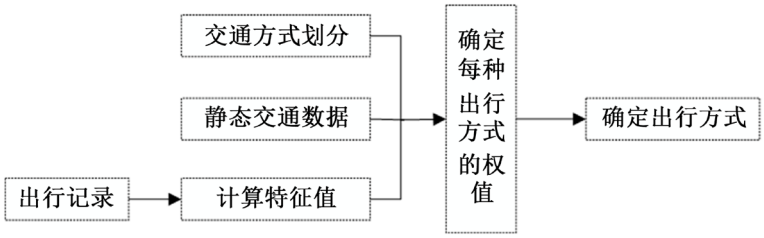

Figure 7. Flow chart for implementing travel labeling
图 7. 出行标签化实现流程图

出行特征值：确定出行记录的 95%位速度、中位速度、平均速度、低速度率作为出行的特征值。

交通方式的划分：根据每种交通工具的速度，将交通方式按照速度划分为步行、自行车、摩托车、公交、汽车、地铁、火车。

静态交通数据：使用爬取到的地铁数据与公交数据增加模型的准确度。

出行方式：根据出行特征值，结合交通方式的划分和静态交通数据对每种交通方式的权值进行计算，取权值最大的为本次出行的出行方式。设每种交通方式的权值为 w_i，则静态交通数据的权值为

$$w = \max\{w_1, w_2, w_3, \cdots, w_i\}$$

4. 系统实现

根据设计的数据处理流程处理数据并存入 mysql 和 redis 中，搭建 SpringBoot 后台框架，引入高德地图 API，实现的效果如图 8~12 所示。

Figure 8. System heat map interface
图 8. 系统热力图界面

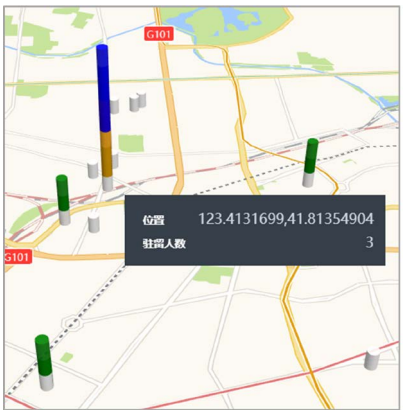

Figure 9. System Resident Map Interface
图 9. 系统驻留图界面

Figure 10. System OD diagram interface
图 10. 系统 OD 图界面

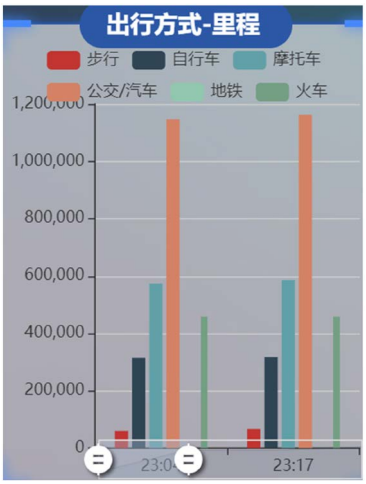

Figure 11. Travel mode-visualmileage
图 11. 出行方式—里程可视化

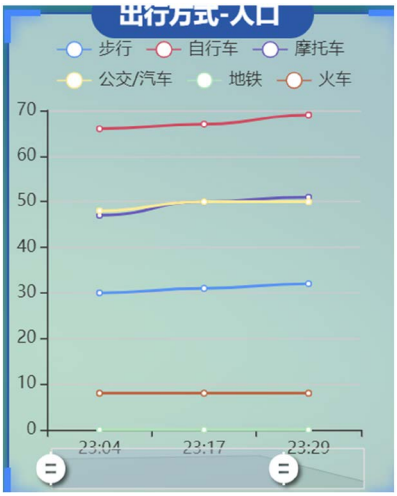

Figure 12. Travel mode-population visualization
图 12. 出行方式—人口可视化

5. 结语

对已给出的手机信令数据，经过数据清洗，数据处理，数据存储，数据可视化等操作最终得到有价值的数据，并将数据存储。编写实现 Web 网站端，提供实时和离线两种方式，以各种图表和地图的形式将数据呈现给用户。该系统挖掘出了数据中潜藏的价值，能够很好地对城市的智能化交通带来一定的贡献。

参考文献

[1] 吴乃星. 基于手机数据的人口出行行为分析研究[J]. 现代计算机, 2018(27): 10-15.

[2] Yoo, B.S., Kang, S.P., Chon, K., *et al.* (2005) Origin-Destination Estimation Using Cellular Phone BS Information. *Journal of the Eastern Asia Society for Transportation Studies*, **6**, 2574-2588.

[3] Widhalm, P., Yang, Y., Ulm, M., *et al.* (2015) Discovering Urban Activity Patterns in Cell Phone Data. *Transportation*, **42**, 597-623. https://doi.org/10.1007/s11116-015-9598-x

[4] 杨飞. 基于手机定位的交通 OD 数据获取技术[J]. 系统工程, 2007, 25(1): 42-48.

[5] 史宜, 杨俊宴. 基于手机信令数据的城市人群时空行为密度算法研究[J]. 中国园林, 2019, 35(5): 102-106.

[6] 孙卓, 刘即明, 阎妮. 基于手机信令大数据的城市居民出行 OD 预测[J]. 数学的实践与认识, 2019(11): 68-76.

[7] 周剑明. 基于手机信令数据的居民出行特征研究[J]. 计算机工程应用技术, 2019, 23(3): 14-16.

[8] 唐杰. 基于手机信令的出行方式识别方法研究[D]: [硕士学位论文]. 重庆: 重庆邮电大学, 2019: 1-67.

[9] 丘建栋, 林青雅, 李强. 基于手机信令数据的居住和出行特征分析[J]. 数据挖掘, 2018, 8(4): 162-173.