# 基于 Hadoop 的出租车服务策略①

段宗涛[1,2], 陈欣欣[1], 康 军[1,2], 温兴超[1], 尉青青[1]

[1](长安大学 信息工程学院, 西安 710064)
[2](陕西省道路交通智能检测与装备工程研究中心, 西安 710064)

**摘 要**: 出租车服务策略是出租车司机群体智慧的体现, 隐藏在大规模的出租车 GPS 轨迹数据中. 基于大数据分析工具, 针对出租车 GPS 轨迹数据进行服务策略挖掘, 提取好的服务策略指导司机营运可以提高司机收入和营运效率. 乘客搜索策略是出租车服务策略的主要内容, 在对 GPS 轨迹数据进行清洗之后导入 HDFS, 首先提取司机个人轨迹, 并对其收入进行量化, 然后对乘客搜索策略建模, 根据模型利用 hadoop 平台统计出司机对各种策略的使用情况, 结果表明, 收入较高的司机在乘客搜索策略选择上与收入一般的司机有显著差异.

**关键词**: 智能交通系统; 服务策略挖掘; 出租车 GPS 轨迹; hadoop; mapreduce

## Taxi Service Strategy Based on Hadoop

DUAN Zong-Tao[1,2], CHEN Xin-Xin[1], KANG Jun[1,2], WEN Xing-Chao[1], YU Qing-Qing[1]

[1](The Information Technology School, Chang'an University, Xi'an 710064, China)
[2](Shaanxi Road Traffic Detection and Equipment Engineering Research Center, Xi'an 710064, China)

**Abstract**: Taxi service strategy is a group of taxi drivers' wisdom embodied, hidden in the large-scale taxi GPS data. Mining GPS traces using big data analysis tools to find and understand the service strategies of skilled taxi drivers to guide other drivers can increase their salaries and improve the efficiency of taxi operation. Passenger searching strategy is the main content of taxi service strategies, which loads GPS traces data into HDFS after the data pre-processing; splits the data to get each driver's personal GPS traces; calculates driver's salary; models taxi driver's service strategies; and then studies how the service strategies influence the driver's salary. A case study indicates that, the differences between the taxi drivers who have better salaries and the drivers who have ordinary salaries are significant in terms of passenger searching strategy.

**Key words**: intelligent transportation systems; service strategies mining; taxi GPS traces; hadoop; mapreduce

## 1 前言

移动传感设备(比如智能手机和 GPS 导航仪)的普遍存在, 使得用大量的电子足迹描述人的行为成为可能, 这些电子足迹给我们提供了理解人在各种情景下的行为模式和发现潜在智能的独特视角[1].

在很多城市, 出租车都配备了 GPS 装备, GPS 装备将定时给服务器上传出租车的实时信息, 包括出租车 ID、经纬度、时间戳、瞬时速度、方向角、是否载客等. GPS 轨迹数据中隐含着出租车司机的服务行为, 包括他们从空车状态采取什么策略搜索乘客和在载客状态采取什么策略输送乘客等. 每个司机的服务行为都不同, 取决于司机在特定情况(时间、空间)下个人的服务策略. 比如, 在下客以后, 有的司机会在较近的地方等待新乘客, 而有的司机会到较远的地方搜索新乘客. 出租车司机采取的策略对载客时间和载客距离产生直接的影响, 从而引起收入和燃料消耗以及碳排放量的不同. 好的服务策略不仅会带来较高的运行收入, 还会提高整个出租车服务系统的效率, 更好地满足乘客的出行需求. 因此, 研究出租车司机的服务策略将有益于司机、乘客和交通管理、规划部门[2].

① 基金项目:国家自然科学基金(61303041);交通运输部基础研究项目(2014319812150);陕西省工业攻关项目(2014K05-28, 2015GY002);中央高校创新团队项目(310824153405)

收稿时间:2016-04-25;收到修改稿时间:2016-06-07 [doi: 10.15888/j.cnki.csa.005567]

本文的数据是西安市出租车调度系统采集的 GPS 数据; 由于传统关系型数据库在数据量不断增加时,往往只能向上扩展, 而且代价非常昂贵; 而 Hadoop 可以线性向外扩展. 在 Hadoop 平台上, 首先通过数据预处理、提取白班、夜班司机个人 GPS 轨迹; 然后对单个司机进行性能量化, 挖掘有效的服务策略(主要探讨乘客搜索策略).

Hadoop 是一个分布式系统架构, 由 Apache 软件基金会开发, 广泛应用于大数据存储和分析处理场景[3]. 主要包括 HDFS(Hadoop 分布式文件系统)和 MapReduce(分布式计算框架). HDFS 的优点是高容错性, 高扩展性、并且对硬件的要求比较低. 它提供对应用程序数据的高吞吐量访问, 适用于超大数据集的存储. MapReduce 编程框架的优点是, 用户不需要了解分布式系统底层的细节, 就能够开发分布式应用程序[4]. MapReduce 计算过程: 1)读取输入数据, 对数据进行“分片”(分片大小一般为 HDFS 块大小, 默认 128M), 每一个 map 任务处理一个“分片”, 多个 map 同时工作. 2) map: 每一个 map 任务处理一个“分片”根据 map 函数处理数据, 对每条记录以<key,value>形式输出到本地文件. 3) shuffle: 将各个 map 输出数据按 key 分组归结到一起, 发往一个 reducer. 此过程非常耗费资源, 如设置了 combiner, 先在各节点按 key 本地合并, 减少网络 I/O; 默认使用 hash partitioner 均匀分配数据到不同的 reduce 节点. 4) reduce: 对 key 相同的多个 value 进行规约操作. 本文的司机个人轨迹提取、收入量化、服务策略挖掘等都是在 Hadoop 平台上开发完成.

## 2 数据预处理

本文基于西安市出租车调度系统采集的 GPS 数据; 记录的每个字段都为 varchar, 数据格式依次为: “序号” “车辆牌照” “时间” “经度” “纬度” “水平速度” “方向” “状态位”(0 无状态位 1 防劫 2 签到 3 签退 4 空车 5 重车 6 点火 7 熄火). 数据示例如图 1.

```
120,陕AT0408,2011-06-01 00:20:01,108.931912,34.220804,33,174,5
42,陕AU6924,2011-06-01 00:19:59,108.880245,34.245129,34,268,4
43,陕AT0328,2011-06-01 00:19:28,108.939135,34.269748,51,88,4
```

图 1 GPS 轨迹数据示例

原始数据保存在 Oracle 数据库中, 采用 Sqoop(Sqoop 是一个用来将 Hadoop 和关系型数据库中的数据相互转移的工具, 可以将关系型数据库中的数据导进 Hadoop 的 HDFS 中, 也可以将 HDFS 的数据导进到关系型数据库中)将数据库中的数据导进到 Hadoop 的 HDFS 中进行存储, 以便我们的数据处理和分析.

由于城市中建筑物对信号传输过程的影响, 以及 GPS 测量精度等, 造成 GPS 在定位时存在较大误差, 使得原始 GPS 数据中存在一些异常数据, 包括速度远高于正常行驶速度的异常数据等; 以及在没有 GPS 信号的地方, 比如地下车库、隧道等地点, 会产生重复值、字段空缺值问题. 这样的数据无法直接进行数据挖掘, 或挖掘结果差强人意[5].

在数据预处理过程中, 首先对重复数据进行去重; 对包含异常速度: *speed*<0 或 *speed* >120 的值进行剔除. 对于 GPS 轨迹误差: *time1* 时刻: 坐标(*longitude1*, *latitude1*); *time2* 时刻: 坐标(*longitude2*, *latitude2*); 经纬度之差>阈值;则认为是异常值.

## 3 个人轨迹提取

首先, 对出租车 GPS 轨迹数据进行切分, 提取单个司机个人轨迹数据. 一辆出租车一般由白班、夜班两个司机共同运营; 两个司机的行为模式不同, 所以提取单个司机的 GPS 轨迹以便分析单个司机的服务策略[6]. 在 MapReduce 编程框架下实现, 主要分为 map 过程(对整个文件进行分片, 然后对每个分片执行 map 函数中的操作)和 reduce 过程(接收 map 产生的数据块, 按 reduce 函数中的操作对数据进行规约), 以车牌后 3 位为 partition 依据重写 partitioner, 尽量使数据均匀的分配到集群各个节点进行计算, 重写多文件输出格式 MultipleOutputFormat, 使每个 key 输出一个文件. 算法伪代码:

算法 1

输入: 清洗后的 GPS 轨迹数据
输出: 个人(白班司机)的 GPS 轨迹数据

```
1.  GetGpsMap:
2.  If time stamps between 05:00 and 07:00 {
3.  If(state==7){早换班}
4.  得到早换班时间后的轨迹
5.  }
6.  If time stamps between 17:00 and 18:00 {
7.  If(state==7){ //晚换班}
8.  得到晚换班时间前的轨迹
9.  }
```

10. If time stamps between 07:00 and 17:00{得到时间段内的轨迹}
11. GetGpsReduce:
12. 按车牌、日期及白班为文件名输出白班轨迹

## 4 收入量化

利用 Hadoop 对出租车司机收入进行量化; 计算出租车司机每趟乘客输送过程中累积的距离, 然后计算司机的大致收入. 即通过司机的 GPS 轨迹数据中相邻时间经纬度变化, 计算此间隔的距离, 累积一天(白班或夜班)得到该司机的行驶距离, 从而得到司机大致收入. 数据为前文所述的 GPS 轨迹数据, 提取出白班司机轨迹, 量化白班司机个人的收入. 需要注意 linux 系统默认的 ulimit 参数中 open files、stack size、max user processes, 以及 limits.conf 中的 nproc 限制等需要根据数据适量调大以满足大规模数据读写和产生大量文件需求[7]. 算法 2 为链式作业, 第一个 job 完成后, 结果作为第二个 job 的输入; 分布式计算过程: 首先用 TextInputFormat 从 HDFS 读取数据, 然后数据分片, 对各分片进行 map, 然后把 key 相同的 map 结果拉取到同一节点, 进行 reduce.

算法 2

输入: 输入按时间排序的单个司机 GPS 轨迹数据
输出: 累积计算产生的收入

1. SalaryMapper1:
2. 切分数据
3. while(state==5){
4. 累积计算处于载客状态的每两个相邻时间间隔 GPS 点的距离， GetDistace(olon, olat, lon, lat)
5. }
6. SalaryReducer1:
7. 累积计算每一趟载客的距离，并计算收入
8. SalaryMapper2:
9. 切分数据
10. SalaryReducer2:
11. 累积计算白班司机一天白班收入

## 5 乘客搜索策略挖掘

### 5.1 乘客搜索策略模型

分别从收入较高和一般的司机群体的 GPS 轨迹数据里挖掘乘客搜索. 出租车实际服务策略建模, 主要探讨即乘客搜索策略. 乘客搜索策略, 考虑司机在搜索新乘客时影响出租车司机决策的实际因素. 出租车司机可能会倾向于在一个的区域等待(酒店和火车站)或者在一个较小的范围内搜索或者去一个较远的熟悉的区域搜索乘客. 基于此把乘客搜索策略分为: 本地等待、本地搜索、远距离搜索三种策略. 定义 *tr* 为: 出租车司机在当前乘客下车之后, 搜索下一个乘客的累积轨迹长度(当前乘客下车后到下一个乘客上车, 每两个相邻点距离之和). 如果 *tr*<=1500m 且等待时间 *w*<=5min 即认为本地搜索; *tr*<=1500m 且等待时间 *w*>5min 即认为本地等待; *tr*>1500m 认为远程搜索. 一个有经验的出租车司机可能会综合考虑交通状况和时空特性等因素, 快速搜索到乘客[8].

### 5.2 乘客搜索策略挖掘

算法 3 通过两个 job 来完成乘客搜索策略的挖掘. 在 GetStrategyMap1 中, 计算乘客下车后出租车司机搜索乘客时离下车点的距离, 然后在 GetStrategyReduce1 中确认搜索策略类型, 并在 GetStrategyMap2 中使用 GetStrategyReduce1 的输出结果进行计算, 改变输出格式, 最后在 GetStrategyReduce2 中使用多文件输出格式输出. 分布式计算过程同算法 2.

算法 3

输入: 输入按时间排序的单个司机 GPS 轨迹数据
输出: 搜索策略类别

1. GetStrategyMap1:
2. 切分数据
3. if ((oldstate==5)&& (state==4) {
4. 上一个乘客下车，开始搜索乘客
5. While(state==4){
6. 计算每个点距离开始搜索的点的距离
7. Distance=GetDistace(olon, olat, lon, lat)
8. }
9. GetStrategyReduce1:
10. If(Distance> 10000){远距离搜索}
11. If(Distance <= 1000){本地等待}
12. Else{本地搜索}
13. }
14. GetStrategyMap2:
15. 切分数据，保证输出格式
16. GetStrategyReduce2:
17. 按车牌多文件输出

最后, 研究相关服务策略对司机收入的影响情况,

分析收入和出租车服务策略的相关关系. 发现能够提高司机收入和出租车系统效率的服务策略, 更好地满足乘客的出现需求.

## 6 实验与结论

本文基于大规模的 GPS 历史数据(中国某城市约 10000 辆出租车 2011 年 6 月的数据), 发现高效的出租车服务策略. 首先, 我们分离出白班、夜班单个出租车司机的 GPS 轨迹. 第二, 量化出租车司机个人的收入. 第三, 我们探讨出租车服务策略, 主要乘客搜索策略: 本地等待、本地搜索、远程搜索. 最后, 评估服务策略和收入的关联关系, 对比分析得出哪些策略有效. 挖掘出高效出租车司机的服务策略, 使用这些策略来提高司机服务水平. 实验环境: 使用 Hadoop 集群, 其中一台为 NameNode, 一台为 Secondary NameNode; 其余为计算节点. GPS 数据集一天约 3 千万条记录, 包含约 1 万辆出租车.

我们从收入较高和收入一般的白班司机中, 各取样 500 名, 进行数据分析. 对比收入较高与收入一般的司机在搜索策略上的相同与不同之处, 分析收入与策略的关系.

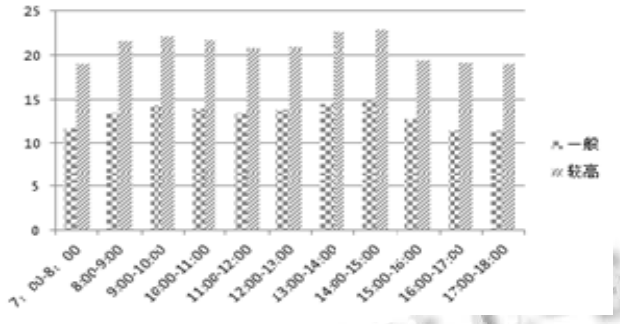


图 2 白班司机各小时收入(表现一般与较好的司机)

图 2 是表现一般与表现较好的白班司机各小时内收入表现情况, 可以看出表现好的司机, 各个小时的收入都显著的高于表现一般的司机.

图 3 是表现一般的白班司机各小时内服务策略使用的次数情况; 可以看出在各个时间段内本地等待策略次数多于本地搜索次数, 多于远程搜索次数. 在 7:00-8:00、13:00-14: 00、15:00-16:00、16:00-17:00 时段, 本地搜索次数与本地等待次数相当, 其余时段相差较大.

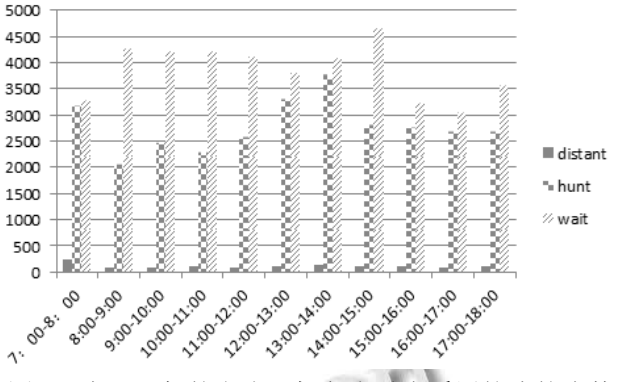


图 3 表现一般的白班司机各小时内采用策略的次数

图 4 是表现一般与表现较好的白班司机各小时内收入表现情况, 可以发现与图 4 类似的情况. 不同之处在于, 远程搜索次数明显多于表示一般的司机.

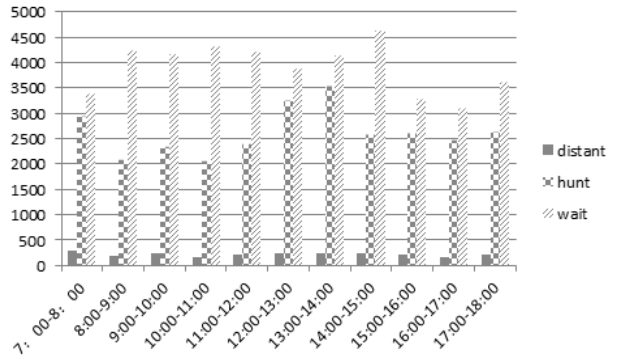


图 4 表现较好的白班司机各小时内采用策略的次数

对图 3、图 4 的数据用 t 检验进一步分析表现一般与表现较好的白班司机采用策略的次数显著性差异, 建立假设 H0: u1=u2 无差异, H1: u1<>u2 有差异, 取显著性水平为 0.05. 根据统计量(如公式 1 所示).

$$t=\frac{\overline{X_1}-\overline{X_2}}{\sqrt{\frac{(n_1-1)S_1^2+(n_1-1)S_1^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \quad (1)$$

计算 P 值如下: Distant: 5.39232E-0.6;Hunt: 0.450986861;Wait: 0.83893276. 对于 Distant 策略, P< 0.05 拒绝原假设, 即表现一般与表现较好的白班司机采用 distant 策略次数有显著差异; 对于 hunt 和 wait 策略接受原假设, 即认为表现一般与表现较好的白班司机采用 hunt 和 wait 策略次数无显著差异.

## 7 总结

出租车服务策略, 作为大量出租车司机的群体智

慧, 隐藏在出租车 GPS 轨迹中. 分析 GPS 轨迹, 发现熟练司机的决策行为, 理解熟练司机的服务策略将给司机、乘客、和城市规划者带来益处. 比如, 好的服务策略可以减少空载率、减少乘客等车时间、减少碳排放等. 传统的数据库难以支撑这样大量数据的分析, 而分布式大数据处理系统 Hadoop 可以存储和处理这些数据. 而且可以根据业务需求不断线性扩展集群存储和计算规模.

分析结果表明, 收入较高的出租车司机和收入一般的出租车司机相比, 采取远距离搜索策略的次数有显著性差异. 收入较高的出租车司机比收入一般的出租车司机在远距离搜索策略上更有经验.

## 参考文献

1 Zhang D, Sun L, Li B, et al. Understanding taxi service strategies from taxi GPS traces. IEEE Trans. on Intelligent Transportation Systems, 2014, 16(1): 123–135.

2 Chen G, Jin X, Yang J. Study on spatial and temporal mobility pattern of urban taxi services. 2010 International Conference on Intelligent Systems and Knowledge Engineering (ISKE). IEEE. 2010. 422–425.

3 Apache 软件基金会.hadoop 官方文档 http://hadoop.apache.org/. [2016-02-13].

4 White T. Hadoop 权威指南.北京:清华大学出版社.

5 杨扬,姚恩建,潘龙,等.基于 GPS 数据的出租车路径选择行为研究.交通运输系统工程与信息,2015,15(1):81–86.

6 何雯,李德毅,安利峰,等.基于 GPS 轨迹的规律路径挖掘算法.吉林大学学报(工学版),2014,44(6):1764–1770.

7 孙翎等.通过 ulimit 改善系统性能. http://www.ibm.com/developerworks/cn/linux/l-cn-ulimit/.

8 李小龙.基于大规模出租车轨迹的乘客移动行为的预测及其应用[学位论文].杭州:浙江大学,2012.