文章编号：1005-8451（2018）11-0040-03

基于铁路出行数据的旅客常住地智能识别算法研究

郭根材
（中国铁道科学研究院集团有限公司　电子计算技术研究所，北京　100081）

摘　要：常住地是判断旅客消费能力与收入水平的重要因素，利于根据旅客常住地进行个性化产品推荐。针对铁路客票发售与预订系统的海量出行数据，依据逻辑判断与概率计算设计了铁路旅客常住地智能识别算法；最后利用Scala语言在铁路客运大数据平台上实现算法，并针对最近两年铁路旅客出行数据进行案例验证，结论表明：该算法有效，旅客常住地信息的识别率为67.7%。

关键词：铁路出行数据；大数据技术；常住地；智能识别算法
中图分类号：U293：TP39　**文献标识码：**A

Intelligent recognition algorithm of passengers permanent residence based on railway travel data

GUO Gencai
(Institute of Computing Technologies, China Academy of Railway Sciences Corporation Limited, Beijing　100081, China)

Abstract: Permanent residence is an important factor of passenger's consumption ability and income level, which is conducive to accomplish personalized recommendation. According to the travel data of Railway Ticketing and Reservation System, this article designed an intelligent recognition algorithm to infer the railway passenger's permanent residence through logical judgment and probability statistics, used Scala language to implement algorithm on large data platform of railway passenger transport. Based on the case study of railway passenger travel data in recent two years, the conclusion shows that the algorithm is effective and the recognition rate of passenger permanent residence is 67.7%.

Keywords: railway travel data; big data; permanent residence, intelligent recognition algorithm

人口流动性是经济社会发展的一个重要指标，人口以流动方式追求经济社会目标而形成的较长时间的自由迁徙和异地生活状况。由于升学、工作等原因，我国居民身份证号中包含的居住地信息与居民实际的常住地有较大差异。掌握旅客常住地信息有助于根据居住地人均收入推断旅客的消费水平，为个性化产品推荐提供基础。

目前，获取常住地信息的方法主要有常住人口、户籍登记、人口普查、大数据分析等。文献 [1] 以年度人口变动调查为基础，通过调查指标之间的关系、人口变动自身特征与抽查的情况，分析我国各地区常住人口的推算方法；文献 [2] 以第六次上海市流动人口普查数据为对象，通过分析变量离散趋势、空间分布等探讨了上海流动人口的分布特征；文献 [3] 描述和分析了区域人口迁移流动的实际状况，构建了常住 - 户籍人口缺口指标来观察我国分地区人口迁移流动态势；文献 [4] 结合人口普查数据与 GIS 数据，系统分析了武汉城市圈常住人口空间分布特征；文献 [5] 提出分布式存储与计算，大数据技术成为数据分析重要手段，文献 [6] 基于移动通信运营商的即时通话记录数据所表征的用户行为对人口的流动性进行判断和测度，这些研究为常住地识别提供了较好的基础。

本文参考上述研究结果，分析了利用铁路出行数据推断旅客常住地的主要影响因素，结合大数据技术设计了基于逻辑判断的旅客常住地智能识别算法，并进行了案例验证。

收稿日期：2018-05-09
基金项目：国家重点研发计划项目（2018YFB1201404）；中国铁路总公司科研计划课题（2016X005-D）。
作者简介：郭根材，助理研究员。

1 常住地界定

根据联合国经济和社会事务部统计司在《人口和住房普查原则与建议》中的建议，常住地可按照以下标准界定：(1) 在最近 12 个月的大部分时间一直居住的地方，不包括因度假或工作引起的短暂出行；(2) 至少在最近 12 个月一直居住的地方，不包括因度假或工作引起的短暂出行[1]。

旅客出行一般是从常住地出发经过一个或多个目的地后返回常住地，完成一次出行。对于普通旅客，旅客在目的地的停留时间要远小于在常住地停留的时间。铁路出行数据可以描述旅客乘坐火车的出行轨迹，通过分析旅客的出行记轨迹、在目的地的停留时间，利用逻辑判断、概率计算等方法可以判断旅客每次出行的起点，从而可以利用旅客一年以上的出行数据推断旅客的常住地。

2 基于出行数据识别常驻地

2.1 影响因素

利用铁路旅客出行数据推断常住地信息，受出行数据质量影响，主要有：

(1) 出行次数过少。部分旅客在统计周期内的出行次数过少，不能形成有效的出行回路，无法在出行起点与出行终点之间确定常住地，这些旅客的常住地不能通过铁路出行数据进行识别。

(2) 行程不连续。综合交通背景下，旅客可组合多种交通方式完成出行，导致铁路出行数据在整个行程上是不连续的，该类型旅客需要结合其他交通方式的出行数据进行判断。

(3) 多出行起点。铁路出行数据可能构成多个出行回路，旅客出行时可能存在多个不同的出行起点，该情况下可选取比重最大的出行起点作为常住地。

(4) 目的地最大停留时间。根据不同的出行目的，旅客在目的地的停留时间一般会有一个时间上限，当旅客在目的地的停留时间过长时旅客可能存在多个常住地，该情况有效无法识别旅客常住地。

2.2 基本概念

根据铁路出行数据识别旅客常住地的影响因素，通过统计判断、概率计算推断铁路旅客常住地，设计了基于铁路旅客出行数据的常住地智能识别算法。为描述算法，给出了行程、差旅、差旅集合的定义。

(1) 行程是指旅客从一个城市到达另一个城市的出行信息，包括出发城市、到达城市、出发时间和到达时间。

(2) 差旅是指旅客从常住地出发通过乘坐多趟列车到达目的地，最后返回常住地的行程集合，由多个行程构成。

(3) 差旅集合是指旅客在指定时间段内的差旅集合，差旅出发城市是影响常住地判断的重要因素。

2.3 算法流程

单名旅客的常住地智能识别算法流程如下：

(1) 选取某一旅客在指定时间内的行程数据，并按照旅客的出行时间排序，构建行程集合；初始化识别参数，设置行程判断序号 $i=0$；

(2) 设置 $i=i+1$，从旅客的第 i 个行程进行深度搜索；如果 $i<$ 行程集合数量，执行下一步，否则执行 (7)；

(3) 设置深度搜索序号 $j=i$；

(4) 选取行程 j 和行程 $j+1$，判断行程 j 的到达城市与行程 $j+1$ 的出发城市是否相同，如果相同，执行下一步，否则 $i=j$ 执行 (2)；

(5) 判断行程 j 到达与行程 $j+1$ 出发的间隔时间是否小于最大停留时间，如果是，执行下一步，否则设置 $i=j$ 并执行 (2)；

(6) 判断行程 i 的出发城市与行程 $j+1$ 的到达城市是否相同，如果相同，根据 i 至 $j+1$ 的所有行程构成一个差旅，并添加在差旅集合中，设置 $i=j+1$ 并执行 (2)；如果不同，设置 $j=j+1$ 并执行 (4)；

(7) 统计差旅集合的差旅个数，如果差旅集合的差旅数量为 0，旅客常住地为未知，否则执行下一步；

(8) 统计差旅集合的差旅出发城市及次数，选取次数最大的出发城市为常住地。

3 案例

3.1 计算平台

铁路作为大众化交通工具，服务旅客数量庞大，传统的单个服务器程序很难在短时间内推算所有旅

客的常住地。本文利用 Scala 语言在铁路客运大数据平台上实现常住地识别算法。铁路客运大数据平台分为外部系统层、数据层、存储层、分析层、展示访问层和应用层 [7]，可实现铁路旅客群体分析应用 [8]。该平台由 1 个控制节点、2 个管理节点、19 个数据节点组成，SPARK 版本为 1.6。利用铁路旅客出行记录推断旅客常住地，2017 年旅客出行记录条数为 30.46 亿，旅客数量约 4.45 亿人。

3.2 算法实现

算法参数目的地旅客最大停留时长为 30 天，基于 SPARK 的旅客常住地识别核心伪代码，如图 1 所示。

```
hiveContext                              //建立连接
    .sql(“select … from …”)              //获取数据
    .rdd                                 //创建 RDD
    .map(x=>{( id_no,id_name…)})         //对数据进行清洗
    .repartition(“分区数”)               //重新分区
    .groupBy(“id_no”)                    //按身份证号分组
    .mapValues(rdd =>jundge_czd)         //识别旅客的常住地
    .toDF()                              //生成数据集
    .registerTempTable()                 //建立临时表
hiveContext.sql(”insert into …”)         //将数据插入最终结果表
```

图1　旅客常住地识别算法伪代码

在铁路客运大数据平台提交作业，平台将计算作业划分为 3 个任务、5 个阶段的运算过程，SPARK 作业流程图，如图 2 所示。SPARK 运算通过一系列弹性分布式数据集（RDD，Resilient Distributed Datasets）的转换，实现分布式读取数据、数据重新分区、常住地识别、结果转换等计算流程，最终将计算结果写入铁路客运大数据平台的分布式数据仓库 HIVE 中。图 2 中黄色部分为常住地核心算法，实现分布式推算旅客常住地。

在运算过程中通过对数据的重新分区与分组降低了作业的内存使用规模与执行单元数量。通过 SPARK 运算，推算出我国铁路近两年服务旅客的常住地信息，识别率为 67.7%。

4 结束语

本文设计了基于铁路出行数据推算旅客常住地的识别算法，该算法可以推算出铁路旅客的常住地

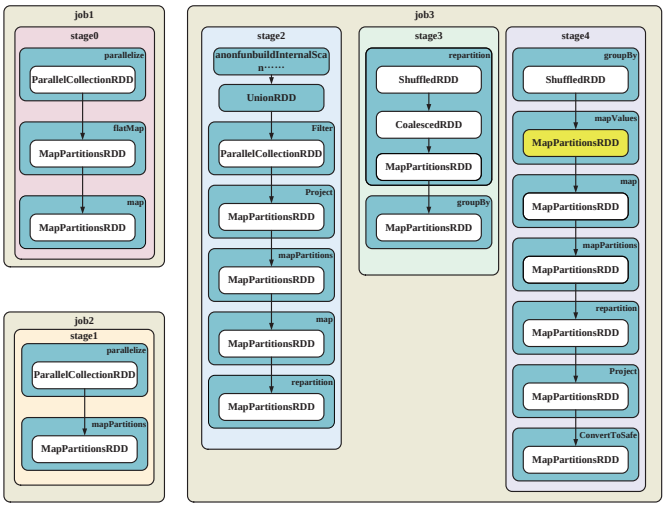

图2　推算旅客常住地SPARK作业流程图

信息，识别率为 67.7%，为常住地的获取提供了一种新思路。受旅客出行次数、行程是否连续等因素影响，算法的识别率可结合其他交通方式的出行数据进一步提高，并利用计算结果进行常住人口分析与预测 [9-10]。

参考文献：

[1] 武　洁，李桂芝．我国各地区常住人口总量推算方法探讨 [J]. 统计研究，2011，28（2）：76-80.

[2] 杨　杰．上海流动人口特征及其空间差异性分析 [D]. 上海：华东师范大学，2013.

[3] 余欣甜．我国分区域人口迁移流动的动态和特点 [D]. 上海：复旦大学，2014.

[4] 许　亮，邓文胜．基于 GIS 的武汉城市圈常住人口空间分布研究 [J]. 资源开发与市场，2007，23（2）：112-115.

[5] Jeffrey Dean, Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters[J]. COMMUNICATIONS OF THE ACM, 2008, 51(1): 101-113.

[6] 靳鑫元．基于移动通信大数据的人口流动性测度研究 [D]. 太原：山西财经大学，2017.

[7] 单杏花，王富章，朱建生，等．铁路客运大数据平台架构及技术应用研究 [J]. 铁路计算机应用，2016，25（9）：14-16.

[8] 吕晓艳，刘彦麟，单杏花．基于大数据平台的铁路旅客群体分析应用研究 [J]. 铁路计算机应用，2016，25（9）：28-30.

[9] 逄锦波，武　博．基于 Logistic 模型的青岛常住人口预测 [J]．山东科技大学学报（自然科学版），2008，27（3）：102-108.

[10] 李　翔，陈振杰，吴洁璇，等．基于夜间灯光数据和空间回归模型的城市常住人口格网化方法研究 [J]. 地球信息科学学报，2017，19（10）：1298-1305.

责任编辑　徐侃春