基于出租车GPS移动轨迹数据预测热点区域

作者：张文星

（宁夏大学 数计学院，宁夏 银川）

摘要：随着科技的不断发展，各种移动设备记录着大量的人类移动轨迹数据。出租车是城市交通工具的重要组成部分，几乎每个城市都会大量的出租车，安装了GPS的出租车可以大范围的、24小时进行车辆移动轨迹数据的采集，而对出租车的GPS移动轨迹数据进行分析和研究，不仅仅可以发现人类的生活轨迹，同时通过对人群聚集以及移动规律的发现，可以分析和预测热点区域，为出租车的运营以及城市交通管理提供一定的参考依据。

关键词：出租车GPS；移动轨迹数据；热点区域

Abstract: With the continuous development of technology, a variety of mobile devices recorded a large number of human movement trajectory data. Taxis are an important part of urban transport, almost every city will be a lot of taxis, a taxi can install a wide range of GPS, a 24-hour collection vehicle movement trajectory data, while the taxi GPS movement locus data analysis and research, not only can be found in human life trajectory, while moving through the crowd as well as the law of discovery, you can analyze and predict hot spots, provide a reference for the taxi operators and urban traffic management.

Key words: Taxi GPS ; Taxi trajectory data ; Areas of interest

前言

出租车是城市交通工具中十分重要的一个部分，因为具有方便、快捷等一系列优点，出租车成为人们外出的首选交通工具之一。每年社会都会投入大量出租车进行营运，然而仍然存在乘客打车难，等车难，甚至不知道在哪里可以打到成等问题。

本文采用的方式是通过研究大量出租车在运营过程中车载GPS所记录下来的行车轨迹，找出顾客们的移动规律，发现人群移动规律，进而进行人群聚集热点区域预测工作。

之所以采用出租车GPS移动轨迹数据作为本文的实验数据，是因为手机和出租车是目前产生移动轨迹数据量最大的两种工具。而手机移动轨迹数据因为其保密性，数据获取十分困难，同时数据的使用和研究在许多方面都有一定的限制。比如说，通过基站获取的用户移动轨迹数据会因为基站的分布密度而导致用户定位精准度收到很大影响；而通过手机GPS获取的数据，因为手机用户在一般情况下是不会使用GPS定位系统，同时要想获取大量的手机GPS数据也不是一件容易地事情。因此，相较之下，出租车的GPS移动轨迹数据的获取相对容易一些，精准度也高一些。

近年来，国内外许多学者都对出租车的移动轨迹领域进行了许多研究。比如，微软研究院中国区的郑宇等学者通过对装有GPS传感器的出租车的载客点和顾客下车点的研究而研发出T-Finder系统，为出租车司机以及乘客提供了一种双向服务。而微软研究院国外研究专家则利用出租车的行车轨迹分析并计算当前道路情况。

随着大数据技术的发展，也为出租车的移动轨迹领域的研究更为高效。而大数据的可视化也能够实时的再现出租车的动态分布，也能够帮助人们了解道路情况，觉得打车问题。

1、移动轨迹数据的预处理

热点区域指的是在一定时间内某区域人流量较大，而这些区域通常需要一定的交通运营方案来保证人群的流通，而这些区域正是出租车需求量大的区域。本文进行研究的区域为北京市，将北京市进行区域进行不同的划分，以便后续对热点区域的分析与预测。本文所使用的数据是关于北京市内一万辆出租车的半年的移动轨迹数据，移动轨迹数据包含以下主要内容：出租车的匿名编号、当前时间以及当前时间出租车所处位置的经纬度，如图1所示。

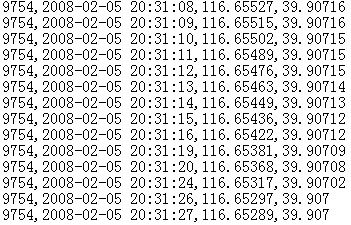


图1 出租车移动轨迹数据

在收集到的大量出租车移动轨迹数据中，我们可以发现，并不是所有的数据都是有效的，因此，在对移动轨迹数据进行研究和分析之前，需要对收集到的数据进行数据预处理工作，而数据预处理部分主要是去除以下几个部分的数据：

（1）超出研究区域范围的数据

由于北京市的的经纬度范围是北纬39.26至41.03,东经115.25至 117.30,因此，在移动轨迹数据中，有超出该范围的数据均认为该数据为无效数据。如经纬度为117.196561,39.112872的数据就必须进行处理，因为该地点为石家庄市，超出北京市范围。

（2）在一定时间内经纬度未发生变化的数据

由于本文所研究的内容是人群聚集热点区域，是与空间以及时间的变化有关系的。在一定时间内未发生经纬度变化的出租车一般情况下是属于停运或者等待状态，此时并没有带着顾客进行转移，也没有产生移动轨迹，因此此类数据也应处理为无效数据，如图2所示。

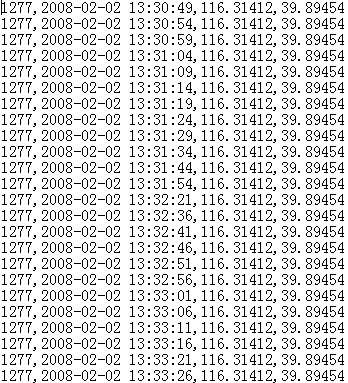


图2 一定时间内经纬度未发生变化的数据

在相关数据进行预处理后，若数据合理就可以将数据存入数据库中进行下一步操作，如图3所示。

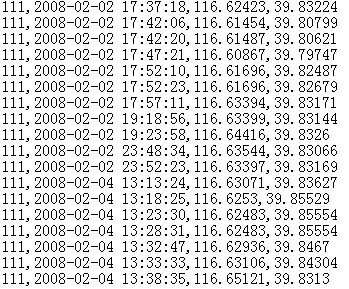


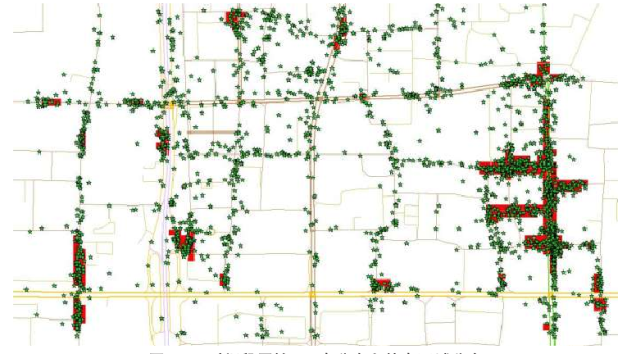
图3 经过处理后的可以使用的数据

2、研究区域的划分

将北京市进行划分，并将划分好的区域进行功能归类，比如说政府行政区、学校区、商业区、火车站、飞机场、地铁站等等不同的功能区域，由于人们出行的习惯有所不同，不同的功能区人群流动的时间范围也会有所不同。比如说学校区域一般会有人群流动的时间点不会晚于晚上11点，而机场或者火车站则到凌晨都有很大可能有大量人群流动。因此在研究不同功能区域的热点区域时，可以选择研究不同的时间段，这样有利于提高研究的效率，也有利于对热点区域预测命中率的提高。

例如，北京市的王府井就是一个著名的购物圣地，且占地面积比较大，每天都有大量的人在此地进行购物休闲，因此，我们可以知道人流量巨大的王府井是十分具有研究价值的购物区域

因为出租车的行驶轨迹遍布了整个北京城，为了更好更方便地进行研究，我们采用经纬度交叉定位的方式定点出租车的在当前时刻的停车的位置，于是我们可以得到该功能区域内所有出租车所处的位置，且都会以点的形式呈现。将该功能区域划分成若干个面积为R\*R的区域，由于环境的原因，比如地形的因素、道路宽窄的因素等都会对该地区出租车停留数量的有所影响，因此，对R的取值应有所不同。在这里我们对R的取值为R={20,25,30,35,40,45,50}，其效果如图4所示。在每个小区域内，若出租车停车，则该区域的测试值加1，若测试值大于认为设定的阈值时，该区域标记为红色，同时也表明该区域为人群聚集热点区域。同时也将热点区域的分析问题转换成为了每个区域的测试值的聚类分析问题。



图四 功能区域划分图

图中的绿色点点即为出租车的在这一时刻停车的位置，而整个区域图则为北京市的平面底层图，红色区域则为人群聚集热点区域。

3、热点区域的发现

热点区域的定义是在一段时间内人流量大，对出租车需求量大的区域。在对出租车移动轨迹数据进行预处理后，便可以将出租车移动轨迹数据进行分析。因为出租车的每个停靠点都有极大可能是顾客上下车的地点，对于停靠点的判断方法是：若在一段时间内出租车的经纬度未发生改变的地点，且时间不宜过长，也不应该过短，因为长时间未发生移动可以认为此事出租车属于停运状态，若时间过短可以判断出租车处于等待红绿灯或者堵车状态。当在同一时间段有大量出租车在同一个区域停靠，且停靠数量大于我们认为设定的阈值的时候，我们可以判断该区域为热点区域，同时在地图上将该区域标记起来。

3.1、算法分析

4、热点区域的预测

5、总结与不足

出租车是公共交通工具中的一个部分，而不合理的出租车分布会导致城市拥堵、出租车载客量低下、乘客难打车等种种问题出现。因此研究分析人群流动量大的区域有利于出租车的合理分布，并可以解决一部分相关问题。

研究存在一些不足，首先因为研究数据中不包含出租车的载客情况，导致无法判断出租车停靠时是载客状态还是送客状态，因此会导致测试中命中率的降低。其次是研究数据的时间跨度较小，因此不能很好地训练预测模型，这也是会导致热点区域预测命中率降低的一个原因。还有一个导致命中率降低的原因是数据不是最新数据，2008年与现在在交通状况与道路状况都有着不小的差异，因此会导致2008年的数据并不一定完全适合现在来使用。

参考文献

[1] 童晓君，向南平，朱定居.基于出租车GPS数据的城市居民出行行为分析[J].电脑与电信，2012(1):56.

[2] 常凯，王爱平.基于功能区划分的出租车热点区域发现方法研究[J].电脑知识与技术，2013(9):25.

[3] Yuan J,Y Zheng ,et al. Discovering regions of different functions in a city using human mobility and POIs[C].Proceedings of the 18th ACM SIGKDD international conference on Knowladge discover and data mining ,ACM. 2012.

[4] 杜慧勇.出租车智能调度系统的设计及应用研究[D].西安：长安大学硕士学位论文.2006,05