出行轨迹演绎性时序聚类分割算法

王 瑾[1],张小垒[1],韩 勇[1],张 涛[2],陈 戈[1]
(1. 中国海洋大学 信息科学与工程学院,山东 青岛 266100;
2. 中国科学院 空间应用工程与技术中心 系统工程部,北京 100000)

摘 要:对于大规模出行轨迹数据进行出行方式研究时,除了要对行程中 OD(起点-终点)进行提取,还要对其间交通方式进行判别。然而,一段行程中,往往包含多种交通方式,如何更精细地从中提取多种交通方式,提升最终交通规划效果,是目前出行方式研究的关键问题。在使用移动智能终端采集出行轨迹的基础上,对出行轨迹进行不同交通方式的转换点提取,最终提出了演绎性时序聚类分割算法进行出行轨迹的分段,并对其进行实验验证。结果表明算法对于分割不同交通方式段,达到了很高的精度。

关键词:移动智能终端;出行轨迹;交通方式;演绎性时序聚类分割算法
中图分类号:TP399 **文献标识码**:A **文章编号**:1673-629X(2014)08-0022-04
doi:10.3969/j.issn.1673-629X.2014.08.006

Apriority Sequential Clustering Segmentation Algorithm of Travel Paths

WANG Jin[1], ZHANG Xiao-lei[1], HAN Yong[1], ZHANG Tao[2], CHEN Ge[1]
(1. School of Information Science & Engineering, Ocean University of China, Qingdao 266100, China;
2. System Department, Technology and Engineering Center for Space Utilization of Chinese Academy of Sciences, Beijing 100000, China)

Abstract: When studying transportation mode with large-scale travel trajectory data, need to not only extract the origin and destination model from the trip, but also determine the transportation mode. However, a journey often contains a variety of transportation modes. How to extract the travel patterns and enhance the effect of the final transportation planning is the key problem of transportation research. Based on mobile intelligent terminal to collect travel paths and the point of the travel path of different transportation mode, put forward the apriority sequential clustering segmentation algorithm for segmentation of the travel path. The method is verified by experiment. Results show that the algorithm for separating different transportation period is of very high accuracy.

Key words: mobile intelligent terminal; travel path; transportation mode; apriority sequential clustering segmentation algorithm

0 引 言

一直以来,出行方式研究在交通规划与交通系统管理中都发挥着重要的作用。近年来,随着移动智能终端的普及和 GPS 技术的发展,基于移动智能终端采集数据为出行方式研究提供了新的技术手段[1-2]。

然而,制约利用该技术进行出行方式研究发展的主要问题是一段行程中不同交通方式的分段提取。在移动智能终端采集数据领域,国外的移动定位技术已较成熟,已设计出将定位信息转化为交通信息的系统,得到出行轨迹、交通流等相关数据,但较少的在出行方式划分等方面进行研究[3]。而国内,多利用移动通信基站[4]或者信令数据进行出行信息采集来进行 OD 调查研究,杨飞等曾针对传统出行调查的局限性,指出利用移动智能终端获取 OD 信息的优势[5-6];刘淼等使用移动智能终端获取 OD 信息[7],但两者均未涉及出行轨迹的分割。关于出行轨迹分段大都在提供外设(GPS 便携设备)的基础上进行[8]。

收稿日期:2013-10-15 **修回日期**:2014-01-18 **网络出版时间**:2014-05-21
基金项目:科技部科技型中小企业技术创新基金(12C26214204374)
作者简介:王 瑾(1988-),女,河南漯河人,硕士研究生,研究方向为地理信息系统、智能终端应用、数据挖掘;韩 勇,教授,研究方向为虚拟地理环境及海洋地理信息系统;张 涛,研究员,博士生导师,研究方向为高可靠软件测试验证技术、系统仿真与虚拟现实技术;陈 戈,教授,博士生导师,研究方向为卫星海洋遥感、海洋地理信息系统。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20140524.2149.021.html

早期,大都通过携带 GPS 便携设备开展出行调查,由于设备采集方式的限制,往往通过设定不同的时间阈值[9]、基于方向、借助路网等探索性方法[10]或者根据基于聚类算法识别停留[11]。这些方法仅仅对停留及非停留(行程)部分进行分割,只是对目的地(活动点)进行提取,而不能对活动点之间的交通方式转换点进行识别,故也不能提取行程中复杂的交通方式。

Zheng 等认为换乘和步行密切相关[12-13],针对 GPS 便携设备获取的原始数据进行交通方式转换点的提取,再设定距离阈值对不确定段进行合并,最终进行步行段与非步行段的划分。这种分段方式针对 GPS 原始数据进行分割,不依赖于数据的采集方式以及采集环境,但是庞大的 GPS 原始数据量会给服务器承载提出更高的要求。其次,该方法在交通方式转换点的提取,全部在数据采集后期进行,未在数据采集过程中进行,加大了后期的运算量。此外,方法中设置了速度阈值以及加速度阈值进行步行段以及非步行段的判别,但每个转换点的速度等统计量全是通过公式计算而来,计算产生的误差在一定程度上增加了分段判别误差。

针对以上问题,文中针对移动智能终端采集的数据,对出行轨迹进行分段算法研究,在数据采集的同时进行数据的预处理,初步识别交通方式,避免了后期计算的误差,减轻了后期运算量。此外,对不同交通方式进行分类组合,基于轨迹数据提出演绎性时序聚类分割算法进行出行轨迹分段,理论及探索性方法结合使用,大大提高了对不同交通方式段提取的精度。

1　出行数据采集

针对移动智能终端开发的数据采集器采用等距采集方法,每隔 15 m 采集一个轨迹点,以 10 个数据为一组进行上传。数据上传到服务器端采用 SQL Server 数据库存储,字段说明如下:

(1)设备号:Name。移动智能终端的设备号作为出行用户标识;

(2)位置:Lat,Lon。出行的经纬度信息;

(3)时间:Time。时间精确到秒,记录格式如:例 2013/05/25 18:25:21;

(4)瞬时速度:Speed。通过移动智能终端中的 GPS 采集出行的瞬时速度;

(5)平均速度:AverageSpeed。该数据采集器以 10 个数据为一组进行上传,计算的 10 个数据的速度平均值作为该组定位数据的平均速度;

(6)根据速度识别的出行方式:SMode。根据瞬时速度,参考以往用户出行经验,结合出行方式自身特征,选取合理的经验阈值进行初步分段,初步分为 a,b,c,d 四大类,依次为步行(速度小于 2 m/s)、自行车(速度介于 2 m/s 和 6 m/s 之间)、公交(速度介于 6 m/s 和 15 m/s 之间)、其他机动车(速度大于 15 m/s)等。SMode 值确定如下:

$$S(v)=\begin{cases}a & 0\leqslant v<2\\ b & 2\leqslant v<6\\ c & 6\leqslant v<15\\ d & v\geqslant 15\end{cases}$$

(7)每组数据的出行方式:Mode。该数据采集器以 10 个数据为一组进行上传,后台判别 10 个数据中的平均速度赋给 AverageSpeed。依据(4)中的判别规则判别该组数据的出行方式。

该数据采集方法在采集过程中就进行了阈值判别的预处理过程,避免后期计算统计量产生的误差,此外,分组上传出行数据,减少通信时间,提高获取数据的精度,有利于后期分段算法的实现。

2　出行轨迹分段

表 1 是微软亚洲研究院根据 45 个被访者长达 6 个月的出行记录得出的换乘矩阵[10]。

表 1　被访者出行换乘矩阵　　%

交通方式	步行	小汽车	公共汽车	自行车
步行	/	53.4	32.8	13.8
小汽车	95.4	/	2.8	1.8
公共汽车	95.2	3.2	/	1.6
自行车	98.3	1.7	0	/

根据表中数据可得,交通方式换乘与步行密切相关[12-14]。在非步行的交通方式之间换乘概率极小,即使换乘,中间也会伴随短暂的步行。

综上,针对一段出行中的交通方式转换点(为了叙述方便,以下简称“转换点”)进行提取,基本等同于行程中的所有步行段起点和终点的提取,换言之,对于不同交通方式的分段,其实是对行程中步行段与非步行段进行分割提取。

演绎性时序聚类分割算法如下:

演绎性时序聚类分割算法是通过对数据特征的观察以及对数据规律的总结,针对出行轨迹数据进行的探索性分割方法。不同于 k 均值聚类、层序聚类分析算法,该方法依赖于数据采集方法,顺应了轨迹数据的时序特征,运用严格的逻辑推理—演绎,进行算法过程的推断。算法假定的前提是数据无失真,主要思路如下:

1)出行方式二值化。

二值化是图像处理的基本操作。随着彩色图像二值化算法的应用探索,扩展了传统意义的黑白图像二

值化的概念,使二值化的应用更加广泛[15]。在本研究中,为了后期轨迹数据分段的方便,也采用二值化方式对出行方式段进行处理。

结合手机定位数据分组采集的特点,根据采集数据过程中每组数据判断出的出行方式(Mode)对出行数据进行二值化(1 为步行段,0 为非步行段),根据出行数据采集中 Mode 的识别规则,将 Mode 为“a”的值设为 1,将 Mode 为“b”,“c”,“d”的值设为 0,初步将其分为步行与非步行两大类集合。

2)转换点预提取。

在出行方式二值化生成的集合中提取临时转折点位置 P(Mode 前后不一致的点),进行模糊转换点的预提取。之所以称之为模糊转换点,是由于数据采集方式为分组采集,在交通方式转换时,转换点可能为正在采集的一组数据(10 个轨迹数据点)中的任一点,并不一定是分界点,把这类转换点标记为“S”。与模糊转换点相对应的是行程转换点,当用户在一个地方停留超出了一定的时间阈值(文中研究选定的时间阈值是两个小时),则视为一段出行结束。行程转换点也作为转换点存放,用于后期活动点的识别,这类转折点标记为“T”。考虑到算法的实现效率,模糊转换点列表中只存放交通方式转换的首个分界点。

3)出行轨迹分割。

根据先验知识提取转换点相关特征:

(1)标志为“T”的点,说明其前后间隔超出了设定的时间阈值,视为转换点;

(2)转换点处,SMode 和 Mode 的值须一致;

(3)转换点前后是以该点为分界的不同交通方式聚类集合。

预提取生成的模糊转换点并不全是严格意义上的转换点,根据轨迹数据的时序特征辅助模糊转换点 P_n 的前后两个点 P_n-1,P_n+1 进行真实转换点的演绎推理,用肯定前件论证法论证:

前继式:

m:NUM(P_n-1. SMode = P_n. Mode, P_n. SMode = P_n. Mode, P_n+1. SMode = P_n. Mode) >= 2

n:真实转换点自 P_n(包括 P_n)向后遍历寻找

($m\rightarrow n$);$m \vdash n$

描述:将 P_n-1,P_n,P_n+1 三个点识别的出行方式(SMode)与 P_n 以及 P_n+1 所在组的出行方式(Mode)作对比。若三个点的出行方式大都与 P_n 组内出行方式相同,由于 P_n 是生成的模糊转换点,P_n 与 P_n+1 出行方式(Mode)不同,即:P_n. Mode $\neq P_n+1$. Mode,根据数据采集方法的特点“采集 10 个轨迹点中出行方式类别个数较多的值赋给 Mode”,且出行轨迹数据具有连续性及时序性,得出 P_n 点之后仍然有出行方式与其相同的轨迹点,故真实转换点应从 P_n(包括 P_n)向后遍历寻找。

反之,如果三个点大都与 P_n+1 组内出行方式等同,则转换点在 P_n 之前(包括 P_n),自 P_n 向前遍历寻找。

根据上述推理规则,遍历存放 P_n 的转折点列表,以预提取生成的临时转换点为起点,设定聚类半径为 10(一组数据为 10 个),寻找这个半径内所有的轨迹点,提取转换点前后的两个点辅助进行分类判断,直至出现最优转换点,把转换点 P_1,P_1+1,P_2,P_2+1 等进行保存,为以后轨迹分段做准备。转换点在出行轨迹上的分布示意图如图 1 所示。

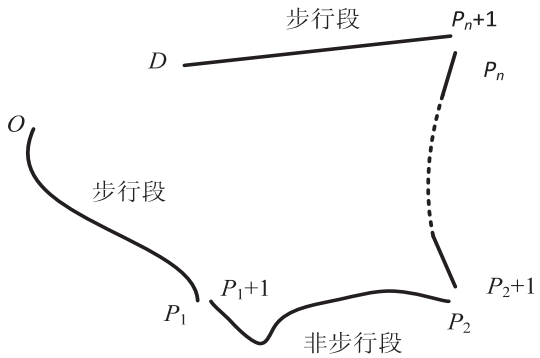

图 1 转换点在出行轨迹上的分布示意图

4)轨迹段后处理。

演绎性时序聚类分割算法是在理想状况(交通正常)下对样本进行归纳推理,未考虑交通严重堵塞等特殊情况,所以需要对分割的轨迹段进行后处理。

后处理过程有两种特殊情况需要考虑,交通换乘或者交通堵塞,这都会导致非步行段被识别为步行段。对交通换乘情况来说,设定最低行走距离阈值 210 m 进行进一步判断。每隔 15 m 采集轨迹点的采集方法避免了后期根据经纬度计算距离的误差,只需判断该步行段的采集的点个数是否小于 15 个,决定是否将此步行段合并到前后的非步行段中。

在交通堵塞期间步行段前后公交车的速度都小于正常速度,且交通堵塞的行程在一定的距离范围,故文中研究需判断步行段前后组判别方式(Mode)均为 b,设定堵塞路段距离阈值(文中设为 360 m),即步行段的采集点个数小于 25 个,判断是否将此步行段合并。

经过上述步骤,最后生成的转换点列表如下:P_1,P_1+1,P_2,P_2+1,…,P_n,P_n+1,见表 2。附加行程的起点 O 和终点 D 的 id 值,可以将其转换成段的格式,并添加其他属性数据进行存储,以便于后期进行分段判别出更精细的出行方式(包括出租车、公交车等)。

这些转换点附加中间轨迹点最终以出行轨迹形式呈现到地图上,用不同的颜色显示不同方式段(步行段及非步行段),直观清晰表达人的出行信息,以便进

一步精确判断更细致的交通方式。

表 2　出行方式段示意表

Start id	End id	Startpoint	Endpoint	…
O	P_1	$\text{lat}_O, \text{lon}_O$	$\text{lat}_{P_1}, \text{lon}_{P_1}$	…
P_1+1	P_2	$\text{lat}_{P_1+1}, \text{lon}_{P_1+1}$	$\text{lat}_{P_2}, \text{lon}_{P_2}$	…
…	…	…	…	…
P_n+1	D	$\text{lat}_{P_n+1}, \text{lon}_{P_n+1}$	$\text{Lat}_D, \text{lon}_D$	…

3　实　验

此次实验招募了 15 个志愿者,在手机上安装数据采集器,采集了连续三周的出行数据。将采集到的手机定位数据传送到服务器中的数据库。每个志愿者在进行数据采集的同时,需要记录以下信息,包括出发点及目的地,何时出发,何时换乘,以及其间的交通方式等(直接填写步行段与非步行段即可)。用文中提出的分段算法对采集的轨迹数据进行分段,将分段结果和之前记录的信息进行匹配验证,算法的划分效果如表 3 所示。

表 3　交通方式段划分结果

	步行段/%	非步行段/%
步行段	98.6	1.4
非步行段	3.6	96.4

在进行交通方式分段划分时,该算法的总体精度平均达到了 96% 以上。

4　结束语

文中针对移动智能终端采集数据,提出了演绎性时序聚类分割算法对出行轨迹进行分段,并面向实际的手机用户出行开展模拟实验,验证基于该算法进行出行轨迹分段的可行性。算法在取得很高精度的同时,也存在更大的改进发展空间。例如,算法是在假设数据无失真的情况下进行的,现实情况下,基于移动智能终端采集数据时会有很多失真值,数据的去噪问题是未来出行方式判别时需要解决的问题。

此外,未来工作中还会注重在轨迹分段结果的基础上,进行更细致的交通方式(步行、自行车、公交车、自驾车)划分判别,提取用户出行信息,进而服务于城市交通规划。

参考文献:

[1] 冯　冲. 基于移动定位数据的用户出行模式识别[D]. 昆明:昆明理工大学,2011.

[2] 黄美灵,陆百川. 基于手机定位的交通 OD 数据获取技术[J]. 重庆交通大学学报(自然科学版),2010,29(1):162-166.

[3] 张　博. 基于手机网络定位的 OD 调查的出行方式划分研究[D]. 北京:北京交通大学,2010.

[4] 魏玉萍,韩　印. 基于手机定位的交通 OD 获取技术[J]. 交通与运输,2011,27(B12):33-36.

[5] 杨　飞,裘炜毅. 基于手机定位的实时交通数据采集技术[J]. 城市交通,2005,3(4):63-68.

[6] 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.

[7] 刘　淼,张小宁,张红军. 基于手机信息的居民出行调查[J]. 城市道桥与防洪,2007(3):18-21.

[8] 张治华. 基于 GPS 轨迹的出行信息提取研究[D]. 上海:华东师范大学,2010.

[9] Du Jianhe, Aultman-Hall L. Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: automatic trip end identification issues[J]. Transportation Research Part A: Policy and Practice, 2007, 41(3): 220-232.

[10] Stopher P R, Jiang Q, Fitzgerald C. Processing GPS data from travel surveys[C]//Proc of 2nd international colloqium on the behavioural foundations of integrated land-use and transportation models: frameworks, models and applications. Toronto: [s. n.], 2005.

[11] Zhou Changqing, Frankouski D, Ludford P, et al. Discovering personal gazetteers: an interactive clustering approach[C]// Proceedings of the 12th annual ACM international workshop on geographic information systems. New York, NY, USA: ACM, 2004: 266-273.

[12] Zheng Yu, Chen Yukun, Li Quannan, et al. Understanding transportation modes based on GPS data for web applications[J]. ACM Transactions on the Web, 2010, 4(1): 1-36.

[13] Zheng Yu, Liu Like, Wang Longhao, et al. Learning transportation mode from raw GPS data for geographic applications on the web[C]//Proceedings of the 17th international conference on World Wide Web. New York, NY, USA: ACM, 2008: 247-256.

[14] Tsui S Y A. An enhanced system for link and mode identifications for GPS-based personal travel surveys[D]. Toronto: University of Toronto, 2005.

[15] 刘玉红,王志芳,杨佳仪,等. 彩色图像二值化算法及应用[J]. 中国医学物理学杂志,2013,30(1):3873-3876.

出行轨迹演绎性时序聚类分割算法

作者： 王瑾， 张小垒， 韩勇， 张涛， 陈戈， WANG Jin， ZHANG Xiao-lei， HAN Yong， ZHANG Tao， CHEN Ge

作者单位： 王瑾,张小垒,韩勇,陈戈,WANG Jin,ZHANG Xiao-lei,HAN Yong,CHEN Ge(中国海洋大学 信息科学与工程学院,山东 青岛,266100)， 张涛,ZHANG Tao(中国科学院 空间应用工程与技术中心 系统工程部,北京,100000)

刊名： 计算机技术与发展 ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2014(8)