GPS轨迹中活动停留点识别的多层分割算法

张治华，季民河

（华东师范大学地理信息科学教育部重点实验室，上海 200062）

摘要：个人移动通讯和位置感知设备的广泛使用产生了大量可用于信息服务的出行轨迹数据。从轨迹数据挖掘出行信息的关键在于停留识别和语义标注。已有的停留点识别方法在抗噪能力和计算效率上有所不足，识别精度有待提高。重新分析了轨迹的停留和移动两大组成要素，发现其状态存在的基础在于其在时间或空间上的连续性，并基于这一理念提出了一种多层次分割算法实现对轨迹停留点的识别。方法的实证检验使用了GPS模块收集的上海市11位受访者一周的出行活动轨迹及问卷调查表。实验结果表明：多层分割法在精度和计算效率上均显示出较好效果。

关键词：模式识别；GPS语义轨迹；多层分割；活动停留

中图分类号：TP391.4　　**文献标识码**：A　　**文章编号**：1674-2850(2011)04-0673-10

Hierarchical segmentation for identifying activity stops from GPS trajectories

ZHANG Zhihua, JI Minhe

(*Key Lab oratory of Geographic Information Science, China's Ministry of Education, East China Normal University, Shanghai* 200062, *China*)

Abstract: In this paper, currently available methods for identifying activity stops from GPS travel trajectories were examined, and their inability to accommodate data noise and lack of computational efficiency were identified. Based on the observation of multi-level activities imbedded within the trajectory, a hierarchical segmentation method was proposed to cope with the issue of varying spatial scales in which different activities take place. By recognizing the interweaving relationship between stops and moves, the new method starts out from generating basic segments from adjacent GPS points, ten combines adjacent segments into state segments according to their duration similarity, and further brings in some other attributes to determine activity stops and trips at different spatial scales. The algorithm was tested with sample GPS track data collected from 11 survey respondents over a week in Shanghai. Results indicated that the multi-level segmentation method could improve both classification accuracy and computational efficiency over the traditional density-based segmentation methods.

Key words: matrix recognition; semantic GPS trajectories; hierarchical segmentation; activity stops

0 引言

轨迹数据使用近年来受到广泛关注，先进而低廉的定位设备使过去难以想象的客体移动数据获取变得简易可行，从而激发了诸多领域中的应用，如社会学用于观察日常活动对青少年健康的影响[1]和儿童户外活动的环境暴露问题[2]；生态学中观察鸟类迁徙路径和气候等环境因素对迁徙过程的影响[3]；交通管理中实时路况信息的获取、车队管理、出行调查手段革新等[4]。时空连续数据获取手段的个体化、易用性及实时传输能力，从根本上改变了半个世纪以来时间地理学实证研究裹足不前的状况。这

基金项目：国家自然科学基金（40771138）
作者简介：张治华（1983—），男，博士，主要研究方向：地理信息处理、时空数据挖掘
通信联系人：季民河，教授，主要研究方向：空间分析与决策、地理软计算、人地关系模拟. E-mail：mhji@geo.ecnu.edu.cn

已由 HÄGERSTRAND[5]于 1970 创建，研究在物质环境制约下人的时空活动规律的人文地理学分支，因数据规模和计算能力不足长期停留在概念层次。而这些瓶颈的有效解决促进了时间地理学的复兴和实用化[6]。

定位导航设备获取的轨迹数据只具有几何与时空的记录，而基于位置服务或时间地理学研究需要的是轨迹的语义信息。这一信息尺度的转变通常利用轨迹数据挖掘来完成。因此，出行轨迹的行程分段和语义标注成为当前研究的焦点。出行轨迹在概念上可分为两组基本要素：即停留与移动；二者的语义标注及其之间的关系构建形成智能化交通出行信息提取的主要任务[3]。例如 ALVARES 等[7]通过数据查询实例演示了直接利用原始轨迹做数据挖掘存在计算的低效率和语义理解的曲折性，提出在原始轨迹和数据挖掘之间增加语义轨迹层以提升效率。他指出背景地理信息在语义轨迹建立中的作用，并设计了 SMoT 算法，利用轨迹与预定义的活动范围中的空间相交查询建立语义轨迹。ALEGRE[8]在此基础上设计出语义轨迹的查询语言 ST-DMQL，作了形式化和实用性方面的改善和扩展。由此可见，语义轨迹的构建基础是停留与移动，而语义既可产生于轨迹本身的时空特征，也可以结合背景地理信息和社会经济信息综合分析提取。地理信息和人文社会经济属性的收集和数据库建立是一项及其繁琐的工作，需另文专述。

1 停留提取方法回顾

传统的轨迹语义探测一般只考虑停留与移动的时空特征差异，包括：

行进速度。停留时对象静止或者接近静止，速度远比移动时低。因此，一些研究使用持续一定时间的静止点作为识别停留的依据[9~10]。

方向变化。停留时方向变化较之移动时要大。如在一些研究中，180°左右的方向转变是停留的重要标志之一[10~11]。

信号缺失。表征活动场所的停留所固有的特征。比如当受访者活动发生在室内时，携带的 GPS 仪器无法获取卫星信号，导致停留伴随信号缺失。信号丢失和再获取之间的时间差特征也被用来标志和识别停留[12~13]。

轨点密度。在等时记录的情况下，停留时轨迹点的空间分布密度比移动时高。因此，可通过定义密度阈值来区分停留和移动[14]。

前面三类方法均基于单个轨迹点时空特征，阈值的逻辑算法比较简单，抗噪能力差，往往难以达到理想精度，需要其他方法或人工辅助的弥补。密度分割法则通过检验邻近多点的共同特征来划分轨迹点的归属。主要算法包括 K-中值[15~16]，DJ-Cluster[17]，以及 CB-SMoT[18].

K-中值算法。首先，规定簇内最少的点数 m 和聚类半径 d，从轨迹中的第一点开始，计算连续 m 个点中任意两点间的最大距离，如果该距离小于 d，则建立一个簇，判断该簇的中位点和簇外下一个点之间的距离，如果小于 $d/2$，则将该点加入簇中，否则结束该簇，直到所有的点都被遍历到，最终建立的各个簇标记为停留。该算法的不足在于采用了簇内最远两点的距离界定一个簇，易受噪声影响，而且不能适应形状不规则的簇。

DJ-Cluster 算法。是经典密度聚类算法 DBSCAN[19]的改进。该算法首先给定点数阈值 m 和聚类半径 d，计算每个点 d 邻域内的点数，如果点数小于 m，则标记当前点为噪声点，否则建立一个新簇；如果该密度簇和已有的密度簇相交（至少有一个点重合）则合并相交的簇。该算法假定的前提是 GPS 轨迹等时记录，没有考虑数据缺失的情况，在实际情况中，GPS 轨迹数据缺失现象相关普遍，使用该方法会产生较大误差。

CB-SMoT 算法。将 DBSCAN 中邻域的设定方法从点计数改为时间阈值，解决了之前等时距假设存在的问题，可以处理带有缺失的数据。但该方法在处理邻域时采用相邻点之间的距离累加，GPS 轨迹中的漂移现象对其影响较大，尤其是停留时产生的较远的漂移对该方法的精度会产生致命影响。基

于DBSCAN思路的DJ-Cluster算法的抗噪能力比K-中值算法有所增强，但因采用邻域的距离累加，不可避免长距离漂移的干扰，将单个停留割裂为多个，识别精度仍然有限。

从计算效率上看，无论哪种密度算法均趋于偏低。由于引入邻近多点，对每点的计算需遍历其他点，计算复杂度为$O(n)$，即使采用空间索引机制仍有$O(n\log n)$，内存开销也较大。

根据状态的连续性特征，提出了一种自下而上逐层合并的轨迹分割思路（bottom-up trajectory segmentation，BUTS），旨在提高识别精度和改善计算效率。传统的轨迹分割把研究焦点放在停留与移动的静态特征差异上。而实际上，轨迹中的移动作为停留的相对过程，也可为停留标志提供有用信息。停留和移动在本质上同属移动状态，其持续一定时间或跨越一定距离的特点是一种状态区别于另一种状态的基本表征。因此对状态连续和变化的区别，也可以作为停留和移动划分的基本依据。而另一方面，由于人类活动具有多尺度特征，对不同层次的活动和应用，界定停留和行程的准则可以有所不同。例如在大尺度上的停留，本身也许包含有小尺度的移动和停留（如图1所示）。多尺度分割算法的目的就在于通过时间或距离阈值设定，实现对不同尺度活动的探测与合并。

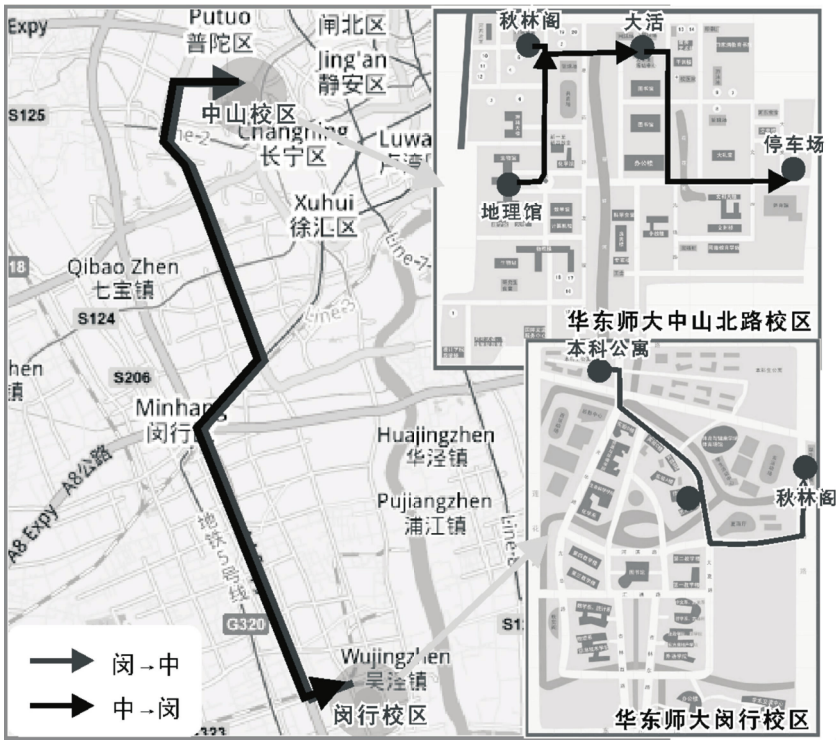

图1 人的活动地点的多尺度特征示意图：校区之间的出行与校区内部的出行比较。

Fig. 1 Example of human activities at multiple scales: inter-campus versus intra-campus travels.

2 研究方法

2.1 相关概念定义

定义1 轨迹（Trajectory）：按时间顺序有向相连的时空点集，表示为

$$T_p = \{\boldsymbol{p}_i \mid \boldsymbol{p}_i = (x_i, y_i, t_i), i \in I\}, \tag{1}$$

其中，x，y为位置；t为时间；I为组成轨迹的点个数。

定义2 停留（Stop）：为轨迹T_p的子轨迹（Sub-trajectory），表示为

$$S = \{\boldsymbol{p}_k \mid (x_k, y_k) \cap C_s, t_m - t_1 > t_{\text{thresh}}, i \in I\}, \tag{2}$$

其中，C_s 是由组成 S 的轨迹点共同定义的空间范围；t_m 为停留轨迹末点时间；t_1 为首点时间；t_{thresh} 为预先设定的时间阈值。定义表明 S 中的任一点必须落在 C_s 之内，且 S 的首末点时间差必须大于 t_{thresh}. 此外，轨迹的起点和终点也定义为停留。

定义 3 移动（Move）：相邻 2 个停留之间的子轨迹，其轨迹点位置自外于 C_s. 表示为

$$M = \{\boldsymbol{p}_k \mid \boldsymbol{p}_k \notin S, \max(t(S_{j-1})) < t_k < \min(t(S_j)), j \in J, k \in I\}, \tag{3}$$

其中，S_j 和 S_{j-1} 为任意 2 个相邻的停留；$t(S)$ 为停留内轨迹点的时间集，J 为子轨迹的总点数。

定义 4 数据缺失（Missing Link）：在 GPS 轨迹等时记录的情况下，如果任一对相邻点之间的时间差大于给定阈值 T，则认定为一个数据缺失。表示为

$$\text{ML} = \{\boldsymbol{p}_k, \boldsymbol{p}_{k-1} \mid t_k - t_{k-1} > T\}. \tag{4}$$

定义 5 数据漂移（Drift）：由于数据接收质量等问题造成 GPS 轨迹点偏离超过一定距离阈值 D，即构成数据漂移。表示为

$$\text{DF} = \{\boldsymbol{p}_k \mid \text{distance}(\boldsymbol{p}_k, \boldsymbol{p}_{k,\text{true}}) > D\}. \tag{5}$$

漂移属轨迹中的噪音，对停留点识别产生很大干扰。漂移常与数据缺失相混杂，无法在预处理中通过简单的速度阈值去除，只能在停留识别的同时予以去除。漂移使得一次停留被割裂成多次缺失和多次假的移动，并且对人工判读停留建立训练样本造成较大干扰。

定义 6 出行或行程（Trip）：在从 GPS 轨迹中提取出行信息时，可以用轨迹中的停留和移动对出行加以定义。如果一段移动同时满足如下条件，则可以视为一次出行。

1）移动的距离超过给定阈值 d；

2）移动前后相连的停留具有特定的活动目的。

在居民出行调查中，距离阈值 d 一般取 400 m 或 500 m，文中考虑到方法的扩展性，比如研究人的行为特征，则采用更为精细的值，为 200 m.

定义 7 行程端点（Trip End）：行程两端的停留。因此，停留包括行程端点和无目的停留。前者是受访者发生活动的时间和地点，具有目的性；后者是无活动目的的暂停，比如交通堵塞、站台候车、熟人路上偶遇驻足交谈等，时间一般较短，两者之间多可以通过停留时间进行区分。

定义 8 活动场所（Activity Location）：指行程端点所在的语义位置，比如“家”。活动场所是一个空间概念，而行程端点是一个时空概念。因此，在一个活动场所可以有多个行程端点，比如在受访者连续 3 天的出行轨迹中，“家”这个活动场所到访了 3 次，就有 3 个活动端点。

2.2 GPS 轨迹的多尺度分割算法

GPS 轨迹中因为噪声的存在，表现为移动中有静止，静止中有移动。很多情况下移动和静止被对方所割裂，无法表现出应有的特征。因此，文章的多尺度分割思路为：先将整体的轨迹分割为基本单元的轨迹段，通过速度阈值设置对轨迹段进行动静状态标记；然后，采用较小的时间阈值对相邻同态的轨迹段做连续性合并以形成较高级的状态段；最后，根据状态段的其他属性特征进行语义合并，得到最终的停留识别结果，如图 2 所示。

1）基本轨迹段生成

在出行调查中，观察对象的运动状态（移动和静止）体现在相邻点连接形成的基本轨迹段上，而不在离散的轨迹点本身。因此使用轨迹段为分析的基本单元。

对于某移动对象轨迹，$T_p = \{\boldsymbol{p}_i\}$，将相邻的轨迹点链接成为轨迹段，则有

$$T_s = \{\boldsymbol{s}_i \mid \boldsymbol{s}_i = (\boldsymbol{p}_i, \boldsymbol{p}_{i+1}), i \in [1, n-1]\}, \tag{6}$$

其中，n 为轨迹点个数。每一个轨迹段如图 1 所示的两点之间的连线。

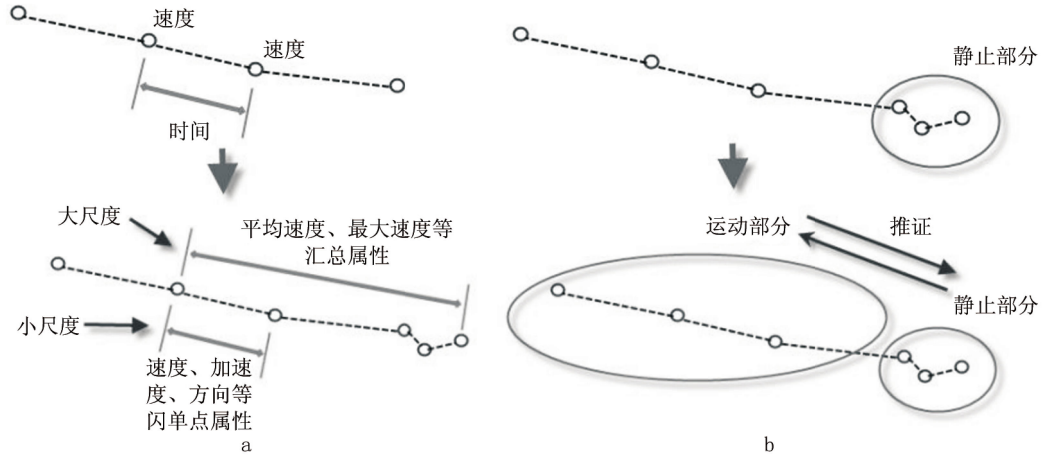

图 2 GPS轨迹的多尺度分割概念图

Fig. 2 Conceptual diagram of multi-scale GPS trajectory segmentation

a—从点基到段基的停留点探测以及不同尺度轨迹段综合；b—静止段与移动段的相互推证关系

a-Conversion of point-based to segment-based stop detection and segment assembly at different scales;
b-Supplementary reasoning of dwell segments and move segments for stop dentification.

每个原始轨迹点表示的是即时位置。转化为段之后，就可以给每个段赋予时长（duration）、距离（length）、平均速度（velocity）、方向（direction）等各项属性，使之有更加明确的意义，可在后续处理中做进一步的合并。

2）确定轨迹段状态

计算每条轨迹段 s_i 的平均速度 $s_i \cdot \bar{v}$

$$\boldsymbol{s}_i \cdot \bar{v} = \frac{\boldsymbol{s}_i \cdot \text{length}}{\boldsymbol{s}_i \cdot \text{duration}}, \tag{7}$$

其中，$\boldsymbol{s}_i \cdot \text{length} = \sqrt{(\boldsymbol{p}_i \cdot x - \boldsymbol{p}_{i+1} \cdot x)^2 + (\boldsymbol{p}_i \cdot y - \boldsymbol{p}_{i+1} \cdot y)^2}$；

$$\boldsymbol{s}_i \cdot \text{duration} = \boldsymbol{p}_{i+1} \cdot \text{time} - \boldsymbol{p}_i \cdot \text{time}.$$

平均速度低于速度阈值 v_{thresh} 的轨迹段 s_i 分为静止段，否则分为移动段：

$$\boldsymbol{s}_i \cdot \text{type} = \begin{cases} 0, & \boldsymbol{s}_i \cdot \bar{v} \leqslant v_{\text{thresh}} \\ 1, & \boldsymbol{s}_i \cdot \bar{v} > v_{\text{thresh}} \end{cases}, \tag{8}$$

其中，v_{thresh} 取步行速度的下限，0 为静止，1 为移动。

3）基于连续性的合并

轨迹分割后形成的每个轨迹段 s_i 都被赋予移动状态（即移动或静止）。将相邻同向的同状态轨迹段合并在一起便构成了状态段：

$$\boldsymbol{s}_I = \{\boldsymbol{s}_k \mid \forall \boldsymbol{s}_k \in T_s, \boldsymbol{s}_k \cdot \text{type} = c\}, \tag{9}$$

其中，c 取值 0 或 1，当 $c=0$ 时，状态段 S_I 为静止段；当 $c=1$ 时，状态段 S_I 为移动段。如图 3 所示“状态段”中的虚线段和实线段。状态段的相应属性通过合并计算赋值。

连续性体现在静止段和移动段都应该持续一定的时长，低于阈值的应视作为另一状态的噪声，转化为另一状态。

$$S_I \cdot \text{type} = \begin{cases} 0, & \text{when } S_I \cdot \text{type} = 1 \cap S_I \cdot \text{dur} < \text{dur}_{\text{thresh2}} \\ 1, & \text{when } S_I \cdot \text{type} = 0 \cap S_I \cdot \text{dur} < \text{dur}_{\text{thresh1}} \end{cases} \tag{10}$$

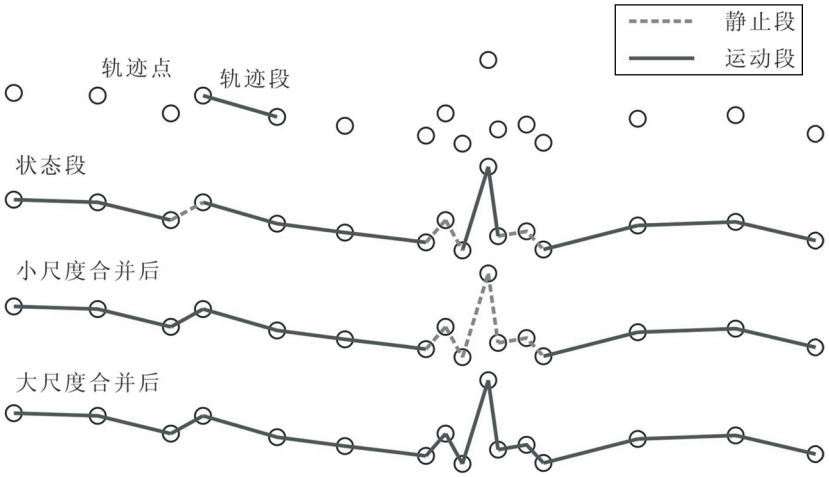

图 3　轨迹逐级合并示意图

Fig. 3　Multi-scale segmentation of a GPS trajectory

4）基于属性特征的停留段和移动段提取

经过连续性合并得到的状态段已具有了一定语义。比如某段 S_I 的状态是静止，持续时间 30 min，则该段实际上代表一个停留（即活动地点）。状态段的其他属性如：平均速度、方向、直线距离、路径距离，绕路指数等各种统计特征，均可作为停留/移动区分的进一步依据。另一方面，由于人类活动具有多尺度特征，对不同层次的活动和应用，界定停留和移动的准则可以有所不同。多尺度分割算法的优势体现在这一阶段：通过不同时间或距离阈值的设定，实现对不同尺度活动的探测与轨迹段合并。

2.3　效率和精度评价

从计算效率上看，该算法只需对数据做若干单循环遍历，因此计算复杂度只有 $O(n)$. 而对停留点精度的评价比较复杂。定义 2 给出的停留概念是一组连续有向、在空间上相对聚集的轨迹点，因具有一定空间覆盖范围，不适于停留点的精度评价。因此这里取停留子轨迹的中位点为停留点代表，记为 p_{derived}，计算其相对真实停留点的位置误差。由于停留点是具有持续时段的空间点，因此评价时不仅要求位置相近，而且要求时段相近，开始时刻和结束时刻都要在一定阈值之内，即同时满足如下条件：

$$\text{distance}(\boldsymbol{p}_{\text{real}}, \boldsymbol{p}_{\text{derived}}) < \Delta d_{\text{thresh}}, \tag{11}$$

$$|\boldsymbol{p}_{\text{real}} \cdot \text{start\_time} - \boldsymbol{p}_{\text{derived}} \cdot \text{start\_time}| < \Delta t_{\text{thresh}}, \tag{12}$$

$$|\boldsymbol{p}_{\text{real}} \cdot \text{end\_time} - \boldsymbol{p}_{\text{derived}} \cdot \text{end\_time}| < \Delta t_{\text{thresh}}. \tag{13}$$

停留精度的评价指标采用了 ZHOU 等[20]的查准率和查全率。设识别出的停留为 D（Derived stops），真实停留为 R（Real stops），识别停留中包含真实停留的个数为 RD（$R \cap D$），则有

查准率——所有识别出来的停留（D）中真实停留（RD）所占的比例为

$$P1 = \frac{RD}{D} \times 100\%, \tag{14}$$

查全率——是真实停留（R）中被识别出来的比例为

$$P2 = \frac{RD}{R} \times 100\%. \tag{15}$$

3 实验与结果

3.1 数据来源

实证分析使用了该研究在上海市小范围居民出行GPS调查实验中收集的轨迹数据。实验招募了11位受访者携带GPS数据收集模块记录三天出行，记录密度为每秒一个轨迹点，样本共含352 837个轨迹点。调查获得的GPS轨迹经过简化处理被叠加在Google地图上，供受访者通过提示回忆出行过程在轨迹上互动式标注停留。标注结果共得到160次真实停留。图4为研究区域背景和出行轨迹分布情况。表1为经归一化整理后的受访者标注的真实停留表一部。

图4 上海市区小范围GPS出行调查实验数据分布
Fig. 4 Distribution of GPS trajectories collected from a small-scale personal travel survey in Shanghai

表1 受访者自我报告的真实停留存表格式
Tab. 1 Formatted records of real stops self-reported by the GPS survey respondents

行程序号	交通方式	到达时间	到达地点	出行目的
1	步行	13：08	华轻购物商城	购物
2	步行	13：31	石头记	购物
3	步行	13：42	凌云路50路车站	换乘
4	公交	14：10	天钥桥路徐家汇站	购物
5	步行	15：47	市四中学前车站	换乘
6	公交	16：44	虹梅南路729车站	上学

注：日期：2009-08-21 始发时间：13：03 始发地点：天等路430弄

3.2 行程识别结果

行程识别分为数据预处理和轨迹段分类两阶段进行。轨迹数据预处理步骤包括：通过数据库查询剔除定位不准确［即卫星颗数少于3颗，位置精度因子（position dilution of precision，PDOP）＞5）］的轨迹点，把轨迹点逐点记录格式转化为邻点相连的轨迹段格式，并计算每段的时长、距离和平均速度（如图5所示）。

PID	时刻	经度/°	维度/°
1	2009-3-22 15:52:47	121.401 435	31.226 678
2	2009-3-22 15:52:48	121.401 433	31.226 683
3	2009-3-22 15:52:49	121.401 430	31.226 687
…	……	……	……
n	2009-3-22 18:16:54	121.399 669	31.227 851

初始ID	结束ID	时长	距离	均速
1	2	1	0.19	0.23
2	3	1	0.00	0.00
3	4	4	0.29	0.07
…	…	…	…	…
n-1	*n*	1	1.15	0.35

图5 GPS数据预处理后获得的轨迹点至轨迹段转换和存表示例
Fig. 5 Point-to-segment conversion: sample trajectory segments after GPS data preprocessing

在轨迹段分类阶段，先通过速度阈值设置对所有轨迹段做状态标记，即静止或移动。速度阈值取步行速度的下限。人正常步行的速度在3～6 km/h之间，下限约为0.8 m/s. 实际划分中考虑到GPS定位随机性，通常取更低的速度下限（如DU等[10]使用0.51 m/s）。在初步试验的基础上取与其相近的值（0.6 m/s），随后使用时长标准对相邻轨迹段做连续性合并，时长阈值通过优化选择，移动段的连续下限取10 s，静止段的连续下限取30 s. 图6的左表展示了连续性合并的结果，包括了重新计算的相关统计特征，并增加了状态段的绕路指数（路径长度和行程端点间的直线距离之比），作为下一阶段分类的基础。

在停留点识别阶段，首先通过真实活动和GPS轨迹对比观察，得出如下停留判别规则：

初始 PID	结束 PID	移动状态	轨段时长	直线距离	绕师指数
1	5 842	静	4 902	5.6	8.6
5 842	5 861	动	26	11.7	1.4
5 861	5 862	静	595	12.6	1
…	…	…	…	…	…
42 796	42 822	动	27	11.6	1.1

初始 PID	结束 PID	移动状态	轨段时长	直线距离	中位经度/°	中位经度/°
1	15 374	静	79 021	5.6	121.401 431	31.226 646
5 842	15 743	动	408	11.7	121.401 395	31.228 429
5 861	15 825	静	200	12.6	121.401 584	31.230 343
…	…	…	…	…	…	…
42 796	42 822	动	272	11.6	121.399 333	31.228 391

注：时长＝s，距离＝m，速度＝m/s，经纬度＝段中位点坐标

图 6　经过合并形成的状态段存表示例（左），以及最终判别获得的轨迹停留标记和指针（右）

Fig. 6　Sample state segments after initial low-level combination (left) and stops and their pointers to the raw GPS trajectory data after the duration-based assembling process (right)

1）真实出行中连续移动的直线距离不小于 200 m；

2）真实出行中单个静止的持续时间不小于 120 s；

3）真实出行中单个行程的绕路指数不大于 5.

停留判别的最后结果同样记录成表（如图 6 右所示），其中未满足规则的状态段判定为相反状态。最终从轨迹中识别出 211 个停留，而 160 个真实停留点中未被识别 11 个。根据式（14）和式（15）计算得出查准率和查全率分别为 71%和 93%. 与 ZHOU 等[17]的 K-中值和 DJ-Cluster 算法的结果相对比（前者的查准率和查全率分别为 24%和 28%，后者为 71%和 83%），研究方法精度有较大改善，但查准水平仍不能令人满意。究其原因，可能是过高估计了受访者自报活动时间的精度。通过对 t_{thresh} 的优化可以确定对本案例的合适阈值。图 7（3）表明，查准率在 t_{thresh} 达在 120 s 之后呈上升趋势，直至 180 s之后基本持平。在 180 s 阈值点识别出 190 个停留点，未识别出的真实停留点有 16 个，查准率和查全率分别为 76%和 90%，较之 120 s 的结果更为合宜。另外通过个案分析，发现导致查准误差的原因主要有如下类型：

1）停留时间低于设定阈值。如在小尺度活动情况下，受访者步行至单位大门传达室取信，仅停留了 67 s，低于 180 s 的阈值，故未能标记为停留。这是活动本身的性质所决定，需要引入其他信息来处理。

2）实际停留段中的轨迹点漂移过大。较大的信号漂移因时间和距离均超过停留阈值，导致停留被误判为移动（行程）。这一问题与预处理中对数据漂移和缺失未做处理或处理失当有关。GPS 数据漂移处理至今仍是业内技术难题，需要进一步研究。

在 180 s 阈值识别出的虚假停留共计 46 个。造成虚假停留的原因除了之前所述的漂移问题外，大部分是由于交通控制（红绿灯、交通堵塞等）和换乘。其中，交通控制引起的有 8 例，换乘问题引起的有 19 例。其他几例是由缺失处理不当、小行程未识别而导致停留不能完全匹配、起点修正等原因造成的。

3.3　阈值优化

文章算法需要设定数个阈值，包括 1）区分静止段与移动段的速度阈值；2）初步合并时移动段的持续时长；3）初步合并时静止段的持续时长；4）属性合并时静止段的时长。研究分别对 4 项阈值的敏感度进行了测试，通过固定其他 3 项阈值而测试剩余项，即可获得查准率和查全率对受调阈值的敏感情况。敏感度测试结果如图 7 所示。阈值 a 和 c 调整时查准率和查全率同向变化，可以取得最优阈值，分别为 0. 6 s 和 27～35 s；阈值 b 几乎不敏感；而阈值 d 的查准率和查全率相对反向变化，沿时长增加而趋于一致，故可根据精度要求在二者之间权衡。

4　结论

多层次轨迹分割算法可以较好的处理带有数据缺失和信号漂移的 GPS 轨迹，不依赖于事先定义的活动场所，可根据轨迹自身的时空特征产生具有一定语义信息的停留和移动（行程），准确率较之现有

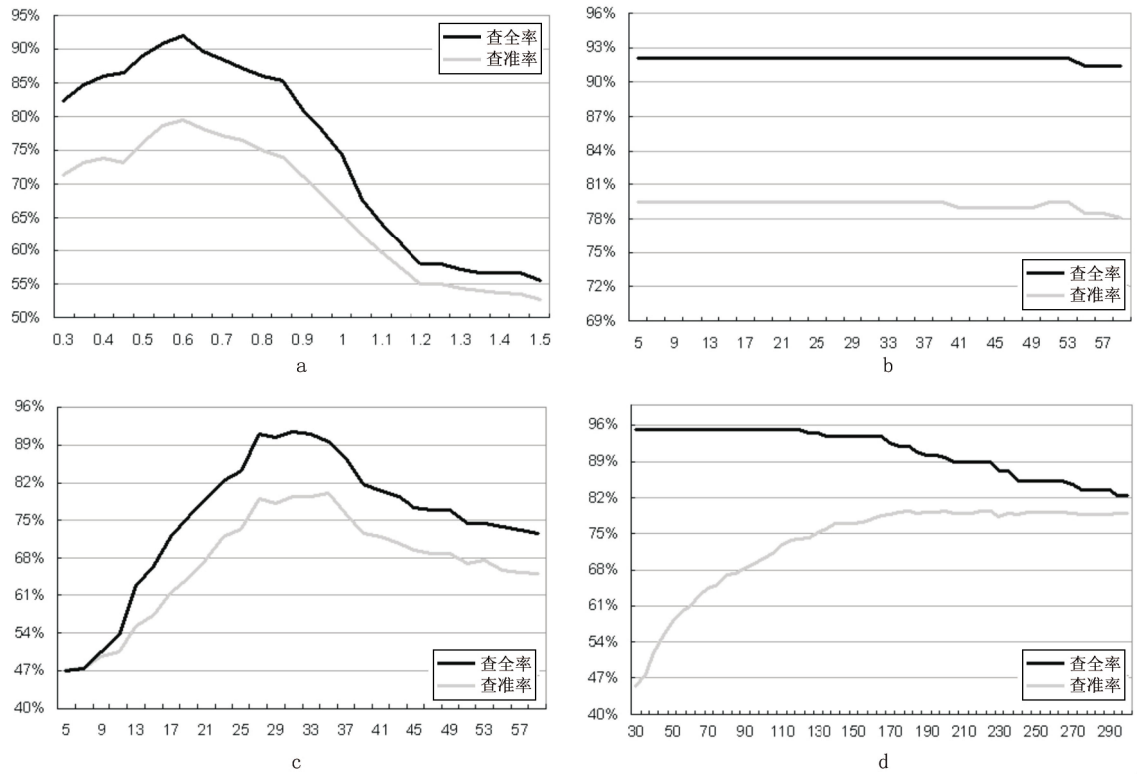

图 7 阈值敏感性测试结果

Fig. 7 The results of threshold sensitivity tests

a—区分静止段与移动段的速度阈值；b—初步合并时移动段的持续时长；
c—初步合并时静止段的持续时长；d—属性合并时静止段的时长

a-Distinguish static section and locomotive section threshold velocity;
b-Duration of locomotive section on preliminany merger;
c-Duration of static section on preliminary merger;
d-Time of static section on attribute merger

的其他方法有所改善。而且可以通过阈值的取值不同，解析出不同尺度的活动。文章仅表述了研究的初步成果，目前还存在如下问题：

1）对数据缺失和信号漂移仅能达到一定限度。过大的信号漂移仍难以处理。

2）对阈值有较强的依赖性。该方法的合并过程需要定义时间和距离阈值，合并后识别停留点时也需要定义时间阈值，阈值的设定目前还需要人为经验的参与。

针对当前存在的问题，可以考虑从如下方面进行改进：

1）交互式的语义信息建立。停留点识别是一个逐步深入的过程，因此在算法设计上应该具备逐步深化的能力。比如在初步停留之后，较长时间和较为确切的停留点可以自动设定较大的阈值。

2）减少阈值设定的武断性。通过阈值之间的联动关系，减少需要设置的阈值数据；根据数据的内生关系，自动化或半自动化的设定阈值。

此外，研究受访人在自我报告中没有区分各自活动的空间尺度，也对分层分割的结果验证造成一定困难。在某个阈值尺度上显著的活动，在另一尺度上可能受到抑制，反之亦然。这一问题在图 7d 中非常明显。设计不同层次活动的问卷调查比较复杂且投入更大，但也是值得深入探索的问题之一。

[参考文献] (References)

[1] WIEHE S E, HOCH S C, LIU G C, et al. Adolescent travel patterns: pilot data indicating distance from home

varies by time of day and day of week[J]. Journal of Adolescent Health, 2008, 42: 418-420.

[2] ELGETHUN K, FENSKE R A, YOST M G, et al. Time-location analysis for exposure assessment studies of children using a novel global positioning system instrument[J]. Environmental Health Perspectives, 2003, 111: 115-122.

[3] SPACCAPIETRA S, PARENT C, DAMIANI M L, et al. A conceptual view on trajectories[J]. Data & Knowledge Engineering, 2008, 65: 126-146.

[4] STOPHER P R. Collecting and processing data from mobile technologies[A]. International Conference on Survey Methods in Transport[C]. Annecy, France: 2008. 361-391.

[5] HÄGERSTRAND T. What about people in regional science?[J]. Regional Science, 1970, 24: 6-21.

[6] GOULIAS K, JANELLE D. GPS tracking and time-geography: applications for activity modeling and microsimulation[A]. Final Report of FHWA-sponsored Peer Exchange and CSISS Specialist Meeting[C]. SantaBarbara, USA: 2005.

[7] ALVARES L O, BOGORNY V, KUIJPERS B, et al. A model for enriching trajectories with semantic geographical information[A]. Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems[C]. Seattle, USA: ACM, 2007.

[8] ALEGRE B. ST-DMQL: a semantic trajectory data mining query language[J]. International Journal of Geographical Information Science, 2009, 23: 1245-1276.

[9] SCHUESSLER N, AXHAUSEN K W. Processing raw data from global positioning systems without additional information[J]. Journal of the Transportation Research Board, 2009, 2105: 28-36.

[10] DU J, 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, 2007, 41: 220-232.

[11] STOPHER P R, JIANG Q, FITZGERALD C. Processing GPS data from travel surveys[A]. 2nd International Colloquium on the Behavioral Foundations of Integrated Land-use and Transportation Models: Frameworks, Models and Applications[C]. Toronto, Canada: 2005.

[12] WOLF J, GUENSLER R, BACHMAN W. Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data[J]. Journal of the Transportation Research Board, 2001, 1768: 125-134.

[13] MARMASSE N, SCHMANDT C. Location-aware information delivery with commotion[A]. Proceedings of the 2nd international symposium on Handheld and Ubiquitous Computing[C]. Bristol, UK: 2000. 361-370.

[14] SCHUESSLER N, AXHAUSEN K W. Processing raw data from global positioning systems without additional information[J]. Journal of the Transportation Research Board, 2009, 2105: 28-36.

[15] HARIHARAN R, TOYAMA K. Project lachesis: parsing and modeling location histories[J]. Geographic Information Science, 2004 LNCS 3234: 106-124.

[16] LIU J H, WOLFSON O, YIN H B. Extracting semantic location from outdoor positioning systems[A]. 7th International Conference on Mobile Data Management[C]. Nara: MDM, 2006. 73.

[17] ZHOU C, FRANKOWSKI D, LUDFORD P, et al. Discovering personal gazetteers: an interactive clustering approach[A]. Proceedings of the 12th annual ACM international workshop on Geographic information systems[C]. Washington DC, USA: 2004, 266-273.

[18] TIETBOHL A, BOGORNY V, KUIJPERS B, et al. A clustering-based approach for discovering interesting places in trajectories[A]. Proceedings of the ACM Symposium on Applied Computing, Advances in Spatial and Image-Based Information Systems Track[C]. Fortaleza, Brazil: 2008. 863-868.

[19] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[A]. Proc. 2nd int. Conf. on Knowledge Discovery and Data Mining[C]. Portland, Oregon: 1996. 226-231.

[20] ZHOU C, LUDFORD P, FRANKOWSKI D, et al. An experiment in discovering personally meaningful places from location data[A]. Proc. Conference on Human Factors in Computing Systems[C]. New York, USA: 2005. 2029-2032.