文章编号:1672-5328(2019)03-0011-08　中图分类号:U491.1+2　文献标识码:A　DOI:10.13813/j.cn11-5141/u.2019.0302

基于手机数据的出行链推演算法

吴子啸
(中国城市规划设计研究院，北京 100037)

摘要：利用手机数据推演居民出行特征是近年交通研究中的一个热点。尽管已有众多研究成果发表，但现实应用中仍有诸多问题需要解决和改进。提出一种基于手机数据的出行链推演算法，通过构造时空贪婪同化流程来处理手机数据的空间不确定性，并对传统聚类算法进行改进以提高活动地点识别效率。通过个体实际数据验证了算法的有效性。结果显示，与其他方法相比，提出的算法能够高效地锚固居住地、工作地等停留较长时间的出行端点，从而提高出行链推演的效率和准确性。该算法适用于多天、手机基站定位和三角算法定位的混合位置数据，对现实数据有很好的适应性。

关键词：大数据；手机数据；出行链；贪婪同化；聚类算法

Travel Chain Estimation Based on Cell Phone Data

Wu Zixiao
(China Academy of Urban Planning & Design, Beijing 100037, China)

Abstract: Analyzing the characteristics of residents' travel demand using cell phone data is a popular topic among studies on transportation in recent years. Although there have been a great number of publications in this area, there are still numerous problems needing to be solved and improved in practical application. This paper proposes an algorithm for travel chain estimation based on cell phone data. This algorithm can effectively deal with the spatial uncertainty of cell phone data by developing the greedy assimilation process, and can increase the accuracy of activity location by improving traditional clustering algorithm. The efficiency of the algorithm is demonstrated through the application of actual data. The results show that compared with other methods, the proposed algorithm can efficiently identify travel origins/destinations with longer stay such as the residence and workplace. Accordingly, it improves the efficiency and accuracy of travel chain deduction. The algorithm is applicable to the multi-source location data from multi-day positioning, positioning by cell phone base station, and triangle algorithm positioning, and is well adaptable to the actual data.

Keywords: big data; cell phone data; travel chain; greedy assimilation; clustering algorithms

收稿日期：2018-12-26
基金项目：中国城市规划设计研究院科技创新基金项目“基于多元数据的长三角城市群交通特征研究”(C-201727)
作者简介：吴子啸(1970—)，男，陕西岐山人，博士，教授级高级工程师，主要研究方向：交通调查、交通模型。E-mail: wuzx@caupd.com

1 研究概述

1.1 手机数据类型及特征

与其他类型大数据类似，手机数据也是一种被动产生的痕迹数据，即用户使用手机进行通讯活动(通话、短信等)和运营商提供通讯服务的痕迹数据。由于手机痕迹数据包含位置和时间信息，从而提供了人们出行和活动的线索。与传统交通调查数据相比，手机数据具有大样本、可持续、成本低等优点，为未来交通研究和交通建模提供了全新的数据资源。

相关研究中通常将手机数据分为两类：一类是通话详单数据(Call Detail Record Data, CDR)，另一类是信令数据(Sightings Data)。CDR数据是由用户使用手机进行通讯活动产生的痕迹数据，通常包含通讯活动的用户ID、起始时间、时长信息以及服务基

站信息；信令数据是运营商提供通讯服务的痕迹数据，除了由用户进行通讯活动触发外，还可以由服务区切换等网络活动触发。与CDR数据不同，信令数据中的位置信息通常为几个基站间通过三角算法计算的结果。因此，信令数据的时间密度和位置信息精度均高于CDR数据[1]。

手机数据的时间密度因用户而异，并且非常不规则。通常在某一小段时间内，手机数据密集出现，在较长且不规律的时间后，手机数据再次密集发生。为了表示手机数据的时间密度，通常将一天划分为以0.5 h为单位的时段，在一个时段内有一个或多个手机数据即表示该时段出现手机数据，而以出现手机数据的时段数表示时间密度并据此计算手机数据出现的时间间隔。CDR数据的时间间隔可以达到数小时之久[2]，而信令数据的时间间隔通常也超过1 h[3]。本文采用某城市2015年的手机数据，以手机数据出现时段最多的15 h(大多数出行和活动集中的时段7:00—22:00)为研究周期，统计得到手机数据平均出现时段数量分布(见图1)。手机数据出现的时段数量平均为9.1个，平均时间间隔为84 min，数据特征与已有文献中的描述基本吻合。预计未来随着智能手机用户比例的上升，手机数据的时间密度将会显著提高。

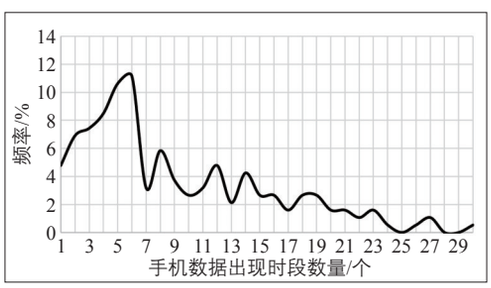

图1 手机数据的时间密度分布

Fig.1 Time-density distribution of cell phone data

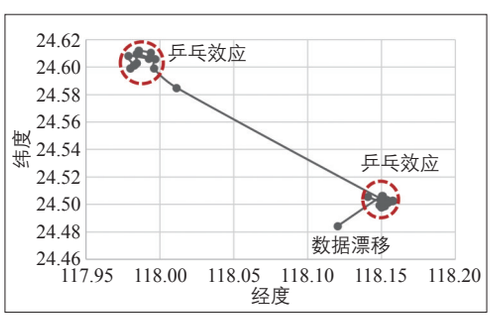

图2 手机数据误差示例

Fig.2 Example of cell phone data error

根据大多数城市的居民出行调查结果，城市居民单次出行平均出行时耗大多集中在20~30 min。结合手机数据的平均时间间隔可以判断，由手机数据推演出行时耗的可靠性不高，并且一些出行活动也不能被手机数据所反映。

1.2 手机数据处理的关键问题

1）乒乓效应和数据漂移。

在通信网络覆盖的区域，通常一个具体地点会被周围几个基站的信号重叠覆盖。该区域内的手机所联系的基站会随各基站信号强度的变化而不断改变，从而产生在两个或多个基站间来回切换的现象，该现象被称为“乒乓效应”。另外，为保证手机用户通讯的效率与质量，服务基站往往有负荷优化的调节机制。当手机邻近的基站负荷较大时，手机信号会被自动切换至较远且负荷较小的基站，从而产生信号的漂移。因此，手机数据对地点的代表具有很大的不确定性。文献[4]对于信令数据的研究认为其在城市区域的位置误差在300 m以内。文献[5]认为CDR数据的位置误差可以从几百米到几千米，主要取决于基站的密度。如图2所示，数据在两个主要活动地点有大量的乒乓效应并伴随数据漂移现象。文献[6]指出乒乓效应和数据漂移的数据约占数据总量的30%。

乒乓效应和数据漂移并不代表用户的真实移动，对这类数据进行处理是利用手机数据进行出行链推演的重要一环。在这方面研究中，已有三种方法被提出和应用。文献[7]提出基于速度的方法：即当基站(或位置点)A夹杂在基站(或位置点)B中并且由A至B的转换速度超过一个设定的极限值，则定义为乒乓效应。在手机数据的时间密度较低并存在空间不确定性的情况下，选择一个速度极限值是该方法的一个挑战。文献[6]和文献[8]则提出基于模式的方法，即根据乒乓效应的特征定义几个基站(或位置点)间切换的模式，当检测到该模式时即视为乒乓效应进行处理。由于实际中乒乓效应的形式非常多样，有时发生在两个基站间、有时发生在多个基站间，并间杂有数据漂移现象，因此，将所有乒乓效应的模式进行列举十分困难；另一方面，实际发生在两点之间的真实高频移动可能会被误以为乒乓效应。将以上两种方法结合而成的混合方法也是比较常见的应

用[9]，但也无法完全避免各自方法内在的缺陷。

2）活动地点识别。

利用手机数据进行出行链推演的另一个关键点是对于活动地点的识别。一般认为，停留超过一定时间(通常为10 min)的位置点可以视为一个活动地点。由于手机数据的空间不确定性，需要将空间临近的位置点聚合在一点才能形成停留时间的累积。聚类算法通常用来实现这一目标。在众多聚类算法中，无须预设参数的基于距离的聚类算法[10-11]最受青睐。文献[5]提出另一种不需预设参数的聚类方法，即基于模型的聚类方法。但该方法对于时间密度稀疏的手机数据的运行效果并不理想[1]。显而易见，基于距离的聚类算法的结果容易受到数据漂移的影响。例如，在几个时间相继、空间临近的位置点中夹杂一个较远距离的数据漂移，聚类的结果将由一个类(位置点集合)变成三个类，漂移数据成为其中单独的一类。这显然会影响位置点停留时间的累积，进而影响活动地点的识别。

文献[12-15]提出一种改进的基于距离的聚类算法，即随着类中位置点的增加，重新计算类的型心。这种改进方法可以在一定限度上提高算法对乒乓效应和漂移数据的容忍度，但算法有效性仍然会受到数据空间不确定性的影响。

手机数据的预处理与活动地点的识别相互依赖、相互影响。大量乒乓效应和数据漂移会影响活动地点识别算法的效率和准确性。因此，在大多数处理CDR数据的流程中，先进行手机数据的预处理[16]，再执行聚类算法进行活动地点识别。而对于信令数据，文献[1]认为应先执行聚类算法消除空间不确定性，才能有效识别和处理乒乓效应。迄今为止，众多研究在城市层面上基于手机数据进行活动和出行链的推算[1, 16]，但算法在个体数据层面的有效性从未详细探讨。手机数据乒乓效应和漂移的内在规律目前仍不明确，对这些数据预处理方法的有效性也难以评价。与以往研究不同，本文提出一种新的方法——时空贪婪同化法，该方法更多地利用了人们出行和活动的一般规律以及手机数据在某一时段密集出现的特征，最大限度地消除手机数据的空间不确定性。另外，改进的空间聚类算法在进行类型心更新时，考虑了位置点的停留时长，能够更加有效地锚固居住地、工作地等人们长时间停留的活动点，从而更加准确地推算出行链。

2 算法描述

本文提出基于手机数据推演出行链的算法流程，如图3所示。

2.1 数据初始化

数据初始化的步骤主要包括：1)对手机数据进行栅格化处理，以栅格型心经纬度取代落在栅格内各位置点的经纬度。手机数据位置点从本源上均是基于基站定位的，而手机基站经纬度通常有保密性要求，栅格化处理一方面可以实现保密的目的，另一方面，不同数据源(如基于基站定位的数据和基于三角定位的数据)可进行融合处理。2)建立各位置点的开始时间和结束时间项。有些手机信令数据既给出信令开始时间也给出了信令结束时间。在只给出信令开始时间时，可以在开始时间上加一个极小的时间段(如20 s)生成信令结束时间，这样手机时空数据的一条记录可表示为$\{l_i, t_{is}, t_{if}\}, 1 \leqslant i \leqslant n, n=|N|$，$l_i$表示第$i$条数据记录的位置点，$t_{is}$和$t_{if}$分别表示第$i$条数据记录的开始时间和结束时间，数据记录集合为$N$。3)按开始时间对所有数据记录进行排序，合并同一位置点的相邻数据记录。也就是说，将同一位置点的所有相邻数据记录合并为一条新记录，新记录的开始时间设为被合并数据记录中开始时间的最小值，结束时间设为这些数据记录中结束时间的最大值。

2.2 时空贪婪同化

时空贪婪同化的基本思想是依次从当前数据记录集中找出最频繁出现的位置点(即数据记录最多的位置点)，识别短时间内夹

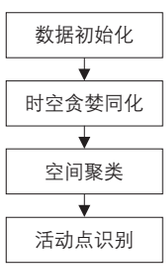

图3 算法流程

Fig.3 Algorithm process

杂在该位置点的两条数据记录间的其他位置点，形成位置点集合进行同化。由于借鉴了贪婪算法的思想，该步骤被称为时空贪婪同化流程。具体算法如下：

1）将初始化的数据记录集 N 设为当前数据记录集，时间段集合 T 设为空集；

2）从当前数据记录集中找出数据记录最多的位置点 m，提取位置点 m 的所有数据记录生成数据记录子集 $N_m=\{m_j, t_{js}, t_{jf}\}$，并生成位置点集合 $C_m=\{m\}$；

3）从 N_m 中第一条记录开始，若当前数据记录的结束时间距下一条记录的开始时间的间隔小于等于 τ，即 $\left(t_{(j+1)s}-t_{jf}\right)\leqslant\tau$，则更新时间段集合 $T=\bigcup_j\left(t_{jf}, t_{(j+1)s}\right)$，执行完毕生成时间段集合 T；

4）对于当前数据记录集 N 中的每一条数据记录 $\left(l_i, t_{is}, t_{if}\right)$，若 $\left(t_{is}, t_{if}\right)\subseteq T$ 并且 l_i 不在位置点集合 C_m 中，则将 l_i 加入 C_m 中，执行完毕生成位置点集合 C_m，并计算 C_m 的型心经纬度，即为其所包含的各位置点经纬度的加权平均值，权重为各位置点在数据记录集 N 中出现的频率或累积停留时间；

5）去除当前数据记录集 N 中位置点属于 C_m 的所有记录，返回步骤2)；循环结束生成位置点集合 $C=\{C_m\}$；

6）将原数据记录集 N 中位置点属于 C_m 的所有记录替代为 C_m 的型心，然后将同一位置点的所有相邻数据记录合并为一条新记录，新记录的开始时间设为这些数据记录中开始时间的最小值，结束时间设为这些数据记录中结束时间的最大值。执行完毕生成数据记录集 N'。

在上述算法中，依次从空间位置点最多的数据记录开始处理，这些位置点通常为手机用户居住和就业所在地以及经常访问地点，这样不仅保证算法有很高的计算效率，也有助于锚固这些经常访问的活动点。步骤3)中时间间隔 τ 可取15 min。根据各城市的居民出行调查结果，一次出行的时耗通常在5 min以上，而在出行目的地活动所花时间一般在5 min以上。从一个地点出行再返回原地的过程涉及一个活动和两次出行，因此最小时间间隔应为15 min以上。另外，由于手机信令数据通常会在某些时段密集发生，尤其是发生乒乓效应和数据漂移时，那么通过上述步骤对夹杂数据进行搜索同化后，手机数据的空间不确定性将在很大限度上被消除。

2.3 空间聚类

空间聚类是对当前数据记录集的空间位置点按聚类算法进行分类，改进的聚类算法如下：

1）将上一阶段结果 N' 设为当前数据记录集，计算各位置点 l 的总停留时间 $T_l=\sum_{i\in I_l}\left(t_{if}-t_{is}\right)$，$I_l$ 为数据记录集中位置点为 l 的序列号集合，按总停留时间由大到小对各位置点排序，形成位置点集合 $L=\{l_1, l_2, \cdots, l_k\}$；

2）在 L 中按顺序取 l_i 进入类 C_h，l_i 为 C_h 型心，依次计算 L 中其他位置点 l_j 至 C_h 型心的距离 $Dis\left(l_j, C_h\right)$，若 $Dis\left(l_j, C_h\right)<\delta$，则将 l_j 加入 C_h，更新 C_h 型心的经纬度为 C_h 中各位置点经纬度的加权平均值，权重为各位置点的总停留时间；否则继续进行，直到 L 中任意一点至 C_h 型心的距离不小于 δ，生成类 C_h；

3）从 L 中删除 C_h 中所含位置点，若 L 不为空，返回步骤2)。否则输出类 $\{C_1, C_2, \cdots, C_f\}$；

4）在 N' 数据记录中，将各位置点替换为其所属的类型心；然后将同一位置点的所有相邻数据记录合并为一条新记录，新记录的开始时间设为这些数据记录中开始时间的最小值，结束时间设为这些数据记录中结束时间的最大值，最后得到结果数据记录集 N''。

上面的空间聚类方法实质上是一种渐进聚类算法[13]的改进算法。渐进聚类算法会因位置点的不同排列次序而产生不同的聚类结果，文献[1]选择K-means聚类方法来克服这一缺点。但K-means聚类方法的预设参数需要经过试算才能确定，会极大地影响计算效率。本文提出的方法先对位置点按停留时间进行排序，聚类始终从长时间活动的位置点发起，保证了聚类结果的唯一性和合理性。另外，类型心更新的机制中引入了位置点停留时间的权重，更加锚固了长时间停留点的位置。大多数的出行距离为500 m以上，步骤2)中 δ 可取200~500 m。

表1 手机数据原始记录

Tab.1 Original raw cell phone data

序号	位置点编号	经度	纬度	开始时间	合并标识	序号	位置点编号	经度	纬度	开始时间	合并标识
1	3	118.097 59	24.525 48	0:00:00	C_1	40	15	118.118 90	24.519 95	13:30:57	C_2
2	2	118.095 85	24.524 31	0:59:39		41	12	118.117 00	24.521 60	13:31:31	
3	3	118.097 59	24.525 48	6:35:54		42	15	118.118 90	24.519 95	13:31:58	
4	1	118.095 82	24.524 30	6:41:36		43	12	118.117 00	24.521 60	13:32:35	
5	2	118.095 85	24.524 31	6:42:24		44	15	118.118 90	24.519 95	13:32:54	
6	1	118.095 82	24.524 30	6:44:13		45	14	118.118 06	24.525 22	13:46:17	
7	2	118.095 85	24.524 31	6:46:40		46	15	118.118 90	24.519 95	13:47:08	
8	6	118.099 00	24.523 10	6:54:53		47	16	118.122 33	24.525 55	14:00:40	
9	2	118.095 85	24.524 31	6:56:17		48	15	118.118 90	24.519 95	14:01:24	
10	11	118.108 82	24.528 56	7:38:21		49	14	118.118 06	24.525 22	14:14:05	
11	15	118.118 90	24.519 95	7:55:26	C_2	50	15	118.118 90	24.519 95	14:15:38	
12	12	118.117 00	24.521 60	7:59:59		51	16	118.122 33	24.525 55	14:15:56	
13	15	118.118 90	24.519 95	8:00:17		52	15	118.118 90	24.519 95	14:17:57	
14	14	118.118 06	24.525 22	8:18:00		53	16	118.122 33	24.525 55	14:18:18	
15	14	118.118 06	24.525 22	94:6:49		54	14	118.118 06	24.525 22	14:19:45	
16	16	118.122 33	24.525 55	9:50:07		55	12	118.117 00	24.521 60	14:20:52	
17	14	118.118 06	24.525 22	9:50:32		56	15	118.118 90	24.519 95	14:35:08	
18	14	118.118 06	24.525 22	9:58:44		57	12	118.117 00	24.521 60	15:24:32	
19	12	118.117 00	24.521 60	1:004:40		58	15	118.118 90	24.519 95	15:26:04	
20	15	118.118 90	24.519 95	10:21:18		59	13	118.118 04	24.525 20	17:03:02	
21	14	118.118 06	24.525 22	10:42:58		60	10	118.108 33	24.525 56	17:12:05	
22	15	118.118 90	24.519 95	10:43:58		61	9	118.102 00	24.524 70	17:13:39	
23	12	118.117 00	24.521 60	10:55:23		62	2	118.095 85	24.524 31	17:14:41	C_1
24	15	118.118 90	24.519 95	10:57:24		63	1	118.095 82	24.524 30	17:36:23	
25	14	118.118 06	24.525 22	11:01:58		64	2	118.095 85	24.524 31	17:38:01	
26	15	118.118 90	24.519 95	11:03:03		65	1	118.095 82	24.524 30	17:44:25	
27	12	118.117 00	24.521 60	11:09:14		66	2	118.095 85	24.524 31	17:45:52	
28	15	118.118 90	24.519 95	11:25:39		67	1	118.095 82	24.524 30	17:46:09	
29	12	118.117 00	24.521 60	11:44:27		68	8	118.100 50	24.525 51	17:47:36	
30	15	118.118 90	24.519 95	12:15:45		69	2	118.095 85	24.524 31	17:54:21	
31	14	118.118 06	24.525 22	12:41:45		70	6	118.099 00	24.523 10	17:54:58	
32	15	118.118 90	24.519 95	12:44:15		71	2	118.095 85	24.524 31	17:56:06	
33	14	118.118 06	24.525 22	12:44:54		72	1	118.095 82	24.524 30	18:14:39	
34	15	118.118 90	24.519 95	12:48:30		73	2	118.095 85	24.524 31	18:15:51	
35	14	118.118 06	24.525 22	13:06:26		74	3	118.097 59	24.525 48	18:57:54	
36	15	118.118 90	24.519 95	13:06:58		75	1	118.095 82	24.524 30	19:03:06	
37	14	118.118 06	24.525 22	13:19:36		76	2	118.095 85	24.524 31	19:05:16	
38	15	118.118 90	24.519 95	13:25:14		77	1	118.095 82	24.524 30	19:06:31	
39	16	118.122 33	24.525 55	13:26:13		78	7	118.099 10	24.526 57	19:10:53	

续表

序号	位置点编号	经度	纬度	开始时间	合并标识	序号	位置点编号	经度	纬度	开始时间	合并标识
79	5	118.097 76	24.527 39	19:18:03		101	2	118.095 85	24.524 31	21:21:50	
80	5	118.097 76	24.527 39	19:20:43		102	1	118.095 82	24.524 30	21:23:55	
81	4	118.097 63	24.525 54	19:32:28		103	2	118.095 85	24.524 31	21:25:08	
82	2	118.095 85	24.524 31	19:33:40		104	1	118.095 82	24.524 30	21:28:36	
83	3	118.097 59	24.525 48	20:09:04		105	2	118.095 85	24.524 31	21:29:53	
84	1	118.095 82	24.524 30	20:38:27		106	1	118.095 82	24.524 30	21:49:19	
85	2	118.095 85	24.524 31	20:40:04		107	2	118.095 85	24.524 31	21:50:33	
86	1	118.095 82	24.524 30	20:40:14		108	1	118.095 82	24.524 30	21:54:20	
87	2	118.095 85	24.524 31	20:44:03		109	2	118.095 85	24.524 31	21:55:47	
88	1	118.095 82	24.524 30	20:45:40		110	1	118.095 82	24.524 30	22:01:24	
89	2	118.095 85	24.524 31	20:46:55		111	2	118.095 85	24.524 31	22:03:10	C_1
90	1	118.095 82	24.524 30	20:49:05	C_1	112	1	118.095 82	24.524 30	22:11:17	
91	3	118.097 59	24.525 48	20:51:02		113	2	118.095 85	24.524 31	22:12:32	
92	1	118.095 82	24.524 30	20:59:11		114	6	118.099 00	24.523 10	22:23:26	
93	2	118.095 85	24.524 31	21:02:26		115	2	118.095 85	24.524 31	22:25:01	
94	1	118.095 82	24.524 30	21:07:01		116	1	118.095 82	24.524 30	22:28:36	
95	2	118.095 85	24.524 31	21:08:24		117	2	118.095 85	24.524 31	22:30:06	
96	1	118.095 82	24.524 30	21:14:27		118	3	118.097 59	24.525 48	22:31:10	
97	2	118.095 85	24.524 31	21:15:34		119	6	118.099 00	24.523 10	22:36:57	
98	4	118.097 63	24.525 54	21:15:59		120	2	118.095 85	24.524 31	22:38:52	
99	2	118.095 85	24.524 31	21:16:57		121	6	118.099 00	24.523 10	23:20:55	
100	1	118.095 82	24.524 30	21:20:10							

2.4 活动点识别

活动点识别指辨认用户活动地点，一般将单次停留时间超过一定值(如5 min)的位置点作为活动点。在时空贪婪同化和空间聚类步骤之后，存在活动的位置点能够最大限度地累积到停留时间，便于活动点的识别。

3 算例

表1为随机抽取的匿名用户一天的手机数据，共121条，表中序号按数据开始时间排序，位置点共计16个(见表2)，位置点编号按各位置点经度由小到大排序。可以看出，出现次数最多的位置点编号为2，表1中蓝色框表示相邻两条位置点为2的数据记录满足开始时间小于15 min的条件，红色框表示相邻两条位置点为15的数据记录满足开始时间小于15 min的条件。

根据时空贪婪同化流程建立位置点集合 $C_1=\{1,2,3,4,6\}$。同理，对应于位置点15(见表1中红色框)，可以得到 $C_2=\{12,14,15,16\}$。表1中合并标识表示这些数据记录的位置点属于同一个位置点集合，将被集合型心所取代。时空贪婪同化流程的结果如表3所示。为直观对比，表3中序号与表1保持一致，合并的数据记录保留了起始数据记录序号。

为说明算法的有效性和计算效率，同时采用K-means聚类方法对以上数据进行处理，结果如图4所示。K-means聚类方法(δ取500 m)的结果生成3个类，按文献[1]提出的处理框架，类 C_2 与 C_3 将在接下来的步骤中被检测为乒乓效应而进行合并，从而得到与时空贪婪同化流程完全类似的结果。但K-means聚类方法需要计算两点之间的空间距离，本质上是一种 $O(|N|^2)$ 算法($|N|$ 为初

始数据记录条数)，而时空贪婪同化流程为 $O(|N|)$ 算法，计算量远远小于前者。

基于 K-means 聚类方法[1]得到的类 $C_1=\{1,2,3,4,5,6,7,8\}$ 会因 δ 取值而变化，从用户多日的数据分析来看，时空贪婪同化流程得到的 $C_1=\{1,2,3,4,6\}$ 能更客观地反映用户居住地的位置。位置点 1，3，4，6 夹杂在相邻两条位置点为 2 的记录中的最小时间间隔分别为 2 min 52 s，8 min 46 s，1 min 23 s 和 1 min 45 s。也就是说，只要时间间隔 τ 取 10 min 以上时，C_1 即可包含位置点 1，2，3，4，6。另一方面，位置点 5，7 和 8 夹杂在相邻两条位置点为 2 的记录中的最小时间间隔分别为 27 min 9 s，27 min 9 s 和 1 min 20 s，当 τ 取 17~27 min 时，位置点 8 将进入 C_1；当 τ 取 28 min 以上时，位置点 5 和 7 将进入 C_1。由于 C_1 型心计算会以各位置点停留时间为权重，位置点 8，5 和 7 进入 C_1 对于型心位置的影响很小。由此可以看出，只要 τ 在一定范围内，时空贪婪同化流程的结果相当稳定。

从表 3 可以看出，经过时空贪婪同化流程，121 条原始数据记录被精简至 10 条，计算量远远小于各种空间聚类算法，且原始数据记录中各种形式的乒乓效应完全被消除，位置点 C_1 和 C_2 的停留时间得到最大限度的累积。不难推测，C_1 为居住地，而 C_2 为工作地，位置点 11 为上班途中留下的痕迹，而位置点 13，10 和 9 为下班途中留下的痕迹。位置点 7 和 5 可能表征一次基于家的活动，但由于出行距离较短，在接下来的空间聚类流程中可能会被抹去(取决于聚类所采用的参数)。

虽然从这个具体例子来看，在时空贪婪同化流程之后进行空间聚类并非十分必要，但由于数据记录已大幅精简，改进后的空间聚类的计算量非常小。况且手机数据产生漂移和乒乓效应的情形多种多样，保留空间聚类流程可以弥补时空贪婪同化流程可能力不

表 2 各位置点经纬度及出现次数

Tab.2 The latitude and longitude of different location point and its occurrences

位置点编号	经度	纬度	出现次数
1	118.095 82	24.524 30	23
2	118.095 85	24.524 31	29
3	118.097 59	24.525 48	6
4	118.097 63	24.525 54	2
5	118.097 76	24.527 39	2
6	118.099 00	24.523 10	5
7	118.099 10	24.526 57	1
8	118.100 50	24.525 51	1
9	118.102 00	24.524 70	1
10	118.108 33	24.525 56	1
11	118.108 82	24.528 56	1
12	118.117 00	24.821 60	9
13	118.118 04	24.525 20	1
14	118.118 06	24.525 22	13
15	118.118 90	24.519 95	21
16	118.122 33	24.525 55	5

表 3 时空贪婪同化流程处理结果

Tab.3 Results of space-time greedy assimilation process

序号	位置点编号	经度	纬度	开始时间	结束时间	停留时长	直线距离/m
1	C_1	118.096 25	24.524 32	0:00:00	6:56:17	6:56:17	
10	11	118.108 82	24.528 56	7:38:21	7:38:41	0:00:20	1 356
11	C_2	118.118 67	24.522 27	7:55:26	15:26:04	7:30:38	1 218
59	13	118.118 04	24.525 20	17:03:02	17:03:22	0:00:20	332
60	10	118.108 33	24.525 56	17:12:05	17:12:25	0:00:20	983
61	9	118.102 00	24.524 70	17:13:39	17:13:59	0:00:20	648
62	C_1	118.096 25	24.524 32	17:14:41	19:06:31	1:51:50	583
78	7	118.099 10	24.526 57	19:10:53	19:11:13	0:00:20	381
79	5	118.097 76	24.527 39	19:18:03	19:20:43	0:02:40	163
81	C_1	118.096 25	24.524 32	19:32:28	23:20:55	3:48:27	374

a 时空贪婪同化流程　　b K-means聚类方法

图4 算法结果比较

Fig.4 Results of different algorithms

能及的地方。从表3可以看出，位置点13在空间聚类流程中进入 C_2，这虽然无助于提升对于停留点位置的判断，但是能够提升对于离开工作地时间判断的精度。

4 结语

与各种基于空间紧密性的聚类算法不同，本文提出一种基于时间紧密性的新算法——时空贪婪同化流程，用于从手机数据推演出行链。在随机抽取的个体手机数据测试中，新算法表现出很高的效率，对于居住地、工作地等停留较长时间的活动点的锚固作用尤其明显。由于篇幅限制算例是针对一天的手机数据，显然算法无须调整即可应用于多天手机数据。未来的研究将集中在整个城市层面居民出行链的推演以及推演结果与交通调查数据的比对方面。

参考文献：

References:

[1] Wang F L, Chen C. On Data Processing Required to Derive Mobility Patterns from Passively- Generated Mobile Phone Data[J]. Transportation Research Part C: Emerging Technologies, 2018, 87: 58−74.

[2] González M C, Hidalgo C A, Barabási A L. Understanding Individual Human Mobility Patterns[J]. Nature, 2008, 453: 779−782.

[3] Calabrese F, Colonna M, Lovisolo P, et al. Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome[J]. IEEE Transactions on Intelligent Transportation, 2011, 12(1): 141−151.

[4] Jiang S, Ferreira J, Gonzalez M C. Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore[J]. IEEE Transactions on Big Data, 2017, 3(2): 208−219.

[5] Chen C, Bian L, Ma J. From Traces to Trajectories: How Well Can We Guess Activity Locations from Mobile Phone Traces?[J]. Transportation Research Part C: Emerging Technologies, 2014, 46: 326−337.

[6] Lee J K, Hou J C. Modeling Steady-State and Transient Behaviors of User Mobility: Formulation, Analysis, and Application[C]//Conti M, Palazzo S, Sivakumar R. Proceedings of the 7th Acm International Symposium on Mobile Ad Hoc Networking & Computing. Florence: ACM, 2006: 85−96.

[7] Vandenbroucke D, Bucher B, Crompvoets J. Geographic Information Science at the Heart of Europe[C]. Leuven: Springer International Publishing, 2013.

[8] Bayir M A, Demirbas M, Eagle N. Mobility Profiler: A Framework for Discovering Mobility Profiles of Cell Phone Users[J]. Pervasive & Mobile Computing, 2010, 6(4): 435−454.

[9] Wang M. Understanding Activity Location Choice with Mobile Phone Data[D]. Washington DC: University of Washington, 2014.

[10] Ye Y, Zheng Y, Chen Y, et al. Mining Individual Life Pattern Based on Location History [C]//On S T, Hu H B, Li Y, Xu J L. Proceedings of 2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware. Taipei: IEEE Computer Society, 2009: 1−10.

(下转第83页)