计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(9):176−180 [doi: 10.15888/j.cnki.csa.005955]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于基站数据挖掘个人驻留规律①

齐　帅, 单桂华, 田　东, 刘　俊

(中国科学院 计算机网络信息中心, 北京 100190)

摘　要: 个人移动通讯设备和位置感知设备的广泛应用, 使得运营商积累了大量的用户位置数据. 目前对位置数据的研究大都关注于活动轨迹的挖掘, 而少量对于个人驻留规律的研究也只停留在识别出驻留点, 却缺乏进一步的挖掘. 本文基于基站采集的位置数据进行研究, 依据基站数据的特点, 提出了一种简单的识别驻留点的方法. 继而提出了两种挖掘驻留规律的方法. 最后使用真实数据对算法效果进行了验证.

关键词: 基站数据; 活动停留; 密度聚类; 最大频繁项集挖掘算法

引用格式: 齐帅,单桂华,田东,刘俊.基于基站数据挖掘个人驻留规律.计算机系统应用,2017,26(9):176–180. http://www.c-s-a.org.cn/1003-3254/5955.html

Mining the Pattern of Personal Stay Based on the Base-Station Data

QI Shuai, SHAN Gui-Hua, TIAN Dong, LIU Jun

(Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: With the widespread use of personal mobile communication devices and location-aware devices, the mobile communication service provider has accumulated a lot of its users' location data. At present, most researches on location data are focused on the mining of active trajectories. A small amount of researches on the pattern of personal stay only determine activity stops, but lack further mining. We conduct researches based on the base station data and propose a simple method to identify the activity stops according to the characteristics of the base station data. Then we propose two methods for mining the pattern of personal stay. Finally, the real data are used to verify the effectiveness of the algorithm.

Key words: base-station data; activity stops; density-based clustering; mining algorithm for maximum frequent itemsets

随着跟踪定位技术的迅速发展, 人们可以通过很多方式获取客体位置的数据, 从而激发了位置数据在诸多领域中的应用. 手机作为采集人们位置数据的天然信号接收装置, 在国内被普遍使用. 一般情况下一部手机只被一个人使用, 手机便成为个人随身携带的定位器. 运营商根据自身的需求, 会采集大量用户的位置信息, 这为挖掘人们的出行规律提供了丰富的数据源. 另一方面, 大量的研究已经证实, 人们的出行是有规律的. 这些研究发现, 尽管个体存在差异, 但他们大多数时间只访问少量的几个地方. 更确切的说, Schlich and Axhausen 的研究揭示 70% 的出行是到 2 到 4 个不同的地方, 90% 的出行是以 8 个不同的地方为目的地; Song 的研究显示, 人们大多数时间停留在少量几个地方, 具体一点说, 75% 的时间用在最频繁访问的 5 个地方. 这些研究为我们挖掘出有意义的结果提供了理论支持.

目前关于驻留规律的研究基本停留在识别出驻留点的阶段. 关于识别出驻留点的研究方法大致分为以下几种: 行进速度、方向变化、信号缺失、轨点密度、K-中值算法、DJ-Cluster 算法、CB-SMoT 算法. 对于在驻留点驻留的时段和时长的研究却非常缺乏. 本文提出了两种方法来挖掘出个人在驻留点驻留的时

① 基金项目: 国家自然科学基金(91530324); 国家高技术研究发展计划(2015AA01A302)
收稿时间: 2016-12-28; 采用时间: 2017-01-18

段和时长, 填补了这方面的空白.

1 基站数据特点

基站数据即通过基站采集的数据, 主要提供了以下三方面的信息: 个人加密后的 ID, 采样时用户的位置(经度, 纬度)和时间. 基站数据有以下两个特点:

(1) 用户在某个基站的信号覆盖范围内活动, 基站会定位到同一个位置点.

(2) 采样时间间隔长且随机.

基站采样效果如图 1 所示, 图中每个蓝色短线表示一个采样点. 本文为了避免采样点重合, 将短线设置成 360 度随机摆动. 特点(1)意味着采集的位置数据和真实位置存在一定偏移, 且一个采样点标识的是用户在一定范围内的活动, 范围大小由最近基站的信号覆盖范围决定. 也就意味着地图由于各个基站的信号覆盖范围不同, 被划分成大小不规则的块. 特点(2)意味着用户发生位置变化的时间点很不明确. 基站数据采样时间间隔平均在 20 分钟以上, 具体的采样时间间隔因个人会有差异. 比如某人规律性的在早晨八点离开家去公司, 但由于采样的随机性, 七点采样一次, 用户在家, 下次采样间隔两个小时, 九点采样时用户在公司, 我们只能得到用户在七点到九点的时间段内离开家, 而不能得到更准确的离开家的时间点信息.

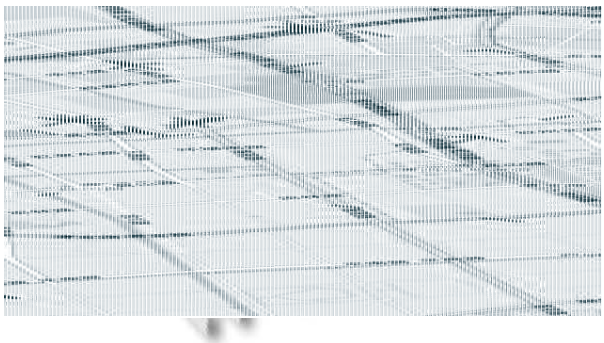


图 1 基站对一个人七个月的采样结果展示

2 识别驻留点

驻留点就是一个人长期频繁驻留的区域, 比如家, 公司, 餐厅等. 目前识别驻留点的方法很多, 但是都不能直接用来解决基站数据的驻留点识别问题. 根据基站数据的特点, 本文提出了一种简单的统计方法来识别驻留点.

因为用户在一定范围内的活动, 基站会定位到同一个位置点上, 采样点有自动聚类的效果. 因此我们可以轻松的得到用户在某个区域停留的时长. 比如某人 7:50 在 A 点, 8:02 在 B 点, 8:10 在 B 点, 8:30 在 C 点, 我们可以简单的理解为 8:02 此人出现在 B 点, 8:30 离开了 B 点. 那么此人在 B 点的停留时长为 28 分钟. 统计一天中停留时长超过半小时的点, 我们把这些点定义为一天的停留点.

假如我们设置一个阈值(比如 0.5), 统计一个人被基站记录以来, A 点作为一天中的驻留点的天数超过了设置的阈值乘以记录天数, 那么我们就可以把 A 点定义为驻留点. 简单来说, 设阈值为 0.5, 若记录时长为六个月, 一个人有 90 天以上在 A 点停留时间超过了半个小时, 我们就把 A 点定义为驻留点. 我们使用 openstreetmap 开源地图来验证找到的驻留点, 发现大都定位在居住区和工作区内. 这说明我们的方法是有效的.

3 挖掘驻留规律

3.1 分割时间段挖掘驻留规律

驻留规律是指个人在驻留点频繁的驻留时段和时长. 因为时间是一个连续的变量, 如果要对时间进行频繁模式挖掘, 首先需要把它转换为离散的变量. 因此本部分的思路为, 先将一天的时间划分为时间段, 然后将一天的位置和时间点归到相应的时间段内. 最后使用频繁项集挖掘算法找到频繁的驻留时段和时长.

(1) 划分时间段

将一天的时间划分为时间段, 以一小时为间隔举例, 那么一天划分为 0~1, 1~2, 2~3…23~24 共 24 个时间段.

(2) 时间点转化为时间段

将时间点和位置用相应的时间段表示, 将有驻留点存在时间段标记为此驻留点, 没有记录的时间段标记为 None, 有记录但不在驻留点的时间段标记为 M. 比如 A 点和 B 点为驻留点, 以一小时为间隔划分时间, 某人一天的记录如下所示:

00:01 A, 01:10 A, 06:05 A, 08:10 C, 09:20 B, 11:15 B, 13:12 D, 15: 16 B, 17:10 B, 19:02 E, 20:10 A

那么这一天的记录转化为:

A, A, None, None, None, None, A, None, M, B, None, B, None, M, None, B, None, B, None, M, A, None, None, None.

通过转化, 可以直观的得到一天中一个人在相应的时段所在的位置.

(3) 填补无记录时间段

为减少随机记录造成的影响, 我们将前后有相同驻留点标记的, 而中间没有记录的时间段, 标记为此驻留点. 如上例所示, 1~2 时间段为 A 点, 6~7 时间段也为 A 点, 中间的时间段都为 None, 则中间的时间段全部标记为 A, 上面的例子转化为:

A, A, A, A, A, A, A, None, M, B, B, B, None, M, None, B, B, B, None, M, A, None, None, None.

因为基站数据的特点是, 对于对象发生移动时, 基站更可能去采样. 而对于长时间的静止, 基站则不会去记录. 通过这种方法, 来填补记录的一部分空白.

(4) 标记时间属性

为降低算法的复杂度, 为每个时间段标记上时间属性, 把求频繁序列问题转化为求频繁组合问题. 上面的例子转化为:

0 A, 1 A, 2 A, 3 A, 4 A, 5 A, 6 A, 7 None, 8 M, 9 B, 10 B, 11 B, 12 None, 13 M, 14 None, 15 B, 16 B, 17 B, 18 None, 19 M, 20 A, 21 None, 22 None, 23 None.

(5) 挖掘频繁时段和时长

将每天的记录按照上述步骤转化为步骤四的样式, 之后对长期的记录采用最大频繁项集挖掘算法就可以得到在驻留点频繁的驻留时段和时长, 效果如下所示:

0 点~7 点 A, 9 点~17 点 B, 20 点~24 点 A

通过挖掘的结果, 我们可以推测 A 点为此人的居住地, B 点为此人的工作地.

分割时间段挖掘驻留规律的方法优点是简单, 计算量小. 缺点也很明显, 只能挖掘出粗糙的结果. 比如某人每天在 8:00~8:10 离开家, 我们只能得到其在 8 点到 9 点发生了位置变化. 而且此方法切割了时间的连续性. 比如某人每天到家的时间在 18:50~19:20 之间, 离开家的时间在 8:40~9:10 之间, 造成同样的规律可能被分割成不同的记录, 继而造成需要降低频繁项的阈值来挖掘最大频繁项集.

3.2 密度聚类挖掘驻留规律

在上一部分中, 我们为了将连续型的时间变量转化为离散型, 采取了分割时间段的方法, 但是效果并不好. 为了克服这个方法的缺点, 我们提出了密度聚类挖掘驻留规律的方法. 本方法的思路是: 首先将离开或到达驻留点的时间点进行密度聚类. 然后将一天的记录转化为用类名来表示. 最后使用最大频繁项集挖掘算法找到在驻留点频繁的驻留时段和时长.

(1) DBSCAN 密度聚类

我们对离开或到达一个驻留点的时间点进行聚类, 在聚类方法的选择中, 我们发现 DBSCAN 密度聚类更适合解决我们的问题. DBSCAN 算法的目的在于过滤低密度的区域, 发现密度高的区域. 跟传统的基于层次聚类和划分聚类的凸形聚类簇不同, 该算法可以发现任意形状的聚类簇. 与传统的算法相比, 它有如下优势能更好的解决我们的问题.

一: 聚类簇的形状没有偏倚;

二: 与 K-MEANS 比较, 不需要输入要划分的聚类个数.

我们首先找到所有的离开或到达驻留点的时间点, 其中离开驻留点的时间点由发生位置变化后的那个时间点来确定, 比如 A 点为驻留点, 若 8:02 在 A 点, 8:15 在 B 点, 则离开 A 点的时间点为 8:15. 对于到达驻留点的时间点也由发生位置变化后的那个时间点来确定. 比如 A 点为驻留点, 20:02 在 B 点, 20:16 在 A 点, 则到达 A 点的时间点为 20:16. 对离开或到达驻留点的时间点进行 DBSCAN 密度聚类后, 用类中的最小值和最大值组成的区间来表示这个类. 比如, 对某人离开 A 点的时间点进行 DBSCAN 密度聚类得到的效果为:

类 1: 7: 30, 7:35, 7:40, 7:32, 7:45, 7:39, 8:00, 8:02

类 2: 8: 30, 8:32, 8:35, 8:40, 8:29, 8:42, 8: 45

那么类 1 表示为[7:30, 8:02], 类 2 表示为[8:29, 8:45].

挖掘结果说明, 此人频繁的在[7:30, 8:02]和[8:29, 8:45]这两个时间区间内离开 A 点.

(2) 时间点转化为类

我们把一天的记录转化为用聚类后的类名表示. 比如某人有 A, B 两个驻留点, 对离开 A 点的时间点聚类后分成两个类: 类 1[7:30, 8:02], 类 2[8:29, 8:45]. 对到达 A 点时间点聚类后分成两类: 类 3[19:45, 20:10], 类 4[20:30, 20:59].离开或到达 B 点的时间点聚类都为一个类, 分别为类 5[17:02:17:30], 类 6[9:10, 9:45]. 若此人一天的记录为: 7:35 离开 A 点, 9:10 到达 B 点, 17:10 离开 B 点, 20:35 到达 A 点. 那么此人一天的记录应转化为: 类 1, 类 6, 类 5, 类 4.

(3) 挖掘频繁的驻留时段和时长

将每天的记录用类名表示后, 对长期的记录采用最大频繁项集挖掘算法, 得到的效果如下所示:

[7:35, 8:10]离开 A 点, [9:12, 9:25]到达 B 点, [17:40, 18:03]离开 B 点, [20:02, 20:34]到达 A 点.

由此我们可以知道此人在 A 点频繁的停留时间段大约在晚上八点到第二天八点之间, 在 B 点频繁的停留时间段为上午九点到下午六点. 并且可以推测 A 点为此人的居住地, B 点为此人的工作地.

4 实验分析

目前我们的数据为运营商基站数据, 因为涉及到保密问题, 不可能通过运营商获得特定某个人的数据. 为了验证我们提出方法的效果, 我们通过 GPS 数据来模拟基站数据. 根据基站数据的特点一, 用户在某个基站的信号覆盖范围内活动, 基站会定位到同一个位置点, 我们将地图分块来模拟基站将地图分块的效果. 我们以纬度跨度 0.003 为高, 经度跨度 0.0025 为宽将地图分块, 落在某个块的采样点, 将这个采样点定位到这个块的中心. 通过这种方式, GPS 数据满足了基站数据的第一个特点. 根据基站数据的特点二, 采样时间间隔长且随机, 我们将一天 24 小时分割成半小时为一个单位, 其中每个单位内随机选取一个采样点. 这样我们一天之内可以采样 48 个点, 且采样点的时间间隔随机且保持在小于一个小时的范围内. 通过这种方式, GPS 数据满足了基站数据的第二个特点.

我们征集了十个志愿者, 在他们的手机上下载 GOOGLE 开发的“我的足迹”APP 来记载他们每天的轨迹. 记录时间为 2016 年 11 月 01 号到 2016 年 12 月 01 号一个月的时间. 我们从中选取周一到周五的数据, 经过上面介绍的两项处理之后, 将 GPS 数据转变为基站数据. 经过我们提出的方法处理得到的结果与志愿者后期自己填写的规律性表格来对比, 从而来评估我们方法的有效性.

4.1 识别驻留点

我们首先选取其中一个志愿者的数据来具体分析方法的效果. 通过简单的统计方法来识别驻留点, 得到两个驻留点, 其中一个落在青年公寓所在的方格, 一个落在腾达大厦所在的方格. 经过与此志愿者填写的表格对比, 发现得到的结果是正确的. 在分别对这十个志愿者的数据做处理后, 得到的正确率为 90%, 即十个志愿者的结果中有九个是正确的. 其中有一个错误是因为此志愿者在这段时间在外地出差. 错误是由于采样时间太短造成的.

4.2 分割时间段挖掘驻留规律

为验证分割时间段挖掘驻留规律的方法, 我们还是首先选取其中一个志愿者的数据来具体分析效果. 以一小时为间隔对一天的时间分段. 以总记录个数*0.2 为最大频繁项集挖掘算法的阈值, 得到最大频繁项集有四十多条记录. 数据量比较大, 这是由于时间间隔设置带来的问题. 我们限定挖掘出的记录长度大于等于 16, 则结果如下所示:

[‘0 116.34296, 39.98840’, ‘1 116.34296, 39.98840’, ‘2 116.34296, 39.98840’, ‘3 116.34296, 39.98840’, ‘4 116.34296, 39.98840’, ‘5 116.34296, 39.98840’, ‘6 116.34296, 39.98840’, ‘7 116.34296, 39.98840’, ‘9 116.33312, 39.94396’, ‘10 116.33312, 39.94396’, ‘11 116.33312, 39.94396’, ‘14 116.33312, 39.94396’, ‘15 116.33312, 39.94396’, ‘16 116.33312, 39.94396’, ‘17 116.33312, 39.94396’, ‘19 116.34296, 39.98840’, ‘20 116.34296, 39.98840’, ‘21 116.34296, 39.98840’, ‘22 116.34296, 39.98840’, ‘23 116.34296, 39.98840’]

语言表述为: 晚上七点到早晨七点在家, 九点到达工作地点, 到上午十一点, 午休, 下午两点到五点在工作地点, 晚上七点到家.

通过这个实验, 我们可以看到这个方法带来的问题, 挖掘出的频繁项过多, 并且此人大约在七点多到八点多到家, 通过这个方法把一个规律割裂成两种不同的规律. 由于这个方法存在太多的缺陷, 我们不再对这个方法的有效性做评估.

4.3 密度聚类挖掘驻留规律

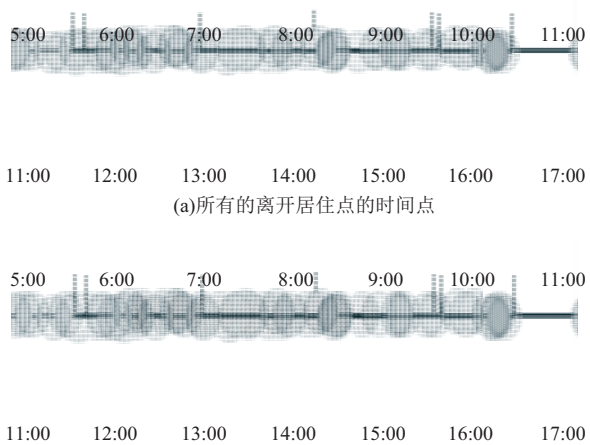
我们选取一个志愿者的数据, 以离开居住地点为例, 首先找到所有离开居住地点的时间点, 可视化在时间轴上如图 2(a)所示, 使用 DBSCAN 密度聚类, 我们设置半径为 5 分钟, 最小点数为 10. 对所有离开居住地点的时间点进行密度聚类, 聚类效果如图 2(b)所示.

可以得到, 离开家的时间分为两类, 分别为[6:29, 7:08]和[7:35, 7:48]. 以同样的方法对其他驻留点的离开到达时间进行密度聚类, 将每天的记录替换为类号, 使用最大频繁项集挖掘算法, 最后得到的结果为:

6:29-7:08 离开居住地点 7:40-8:12 到达工作地点 11:18-11:34 离开工作地点 14:02-14:34 回到工作地点 18:20-19:10 离开工作地点 19:15-19:46 回到居住地点

而此志愿者填写的规律性表格为:

6:30-7:00 离开居住地点 7:45-8:00 到达工作地点 11:10-11:30 离开工作地点 14:20-14:40 回到工作地点 18:40-19:00 离开工作地点 19:30-19:50 回到居住地点.



(a)所有的离开居住点的时间点

(b)对离开居住点的时间点进行密度聚类结果

图 2 志愿者数据的聚类分析结果

我们以挖掘的结果与表格的结果重合的时间段长度除以挖掘结果的时间段长度来评估算法的准确率. 那么此志愿者离开居住地点时间段的准确率为重合时间段长度 30 min 除以挖掘结果的时间段长度 39 min, 即为 77%. 其他时间段的准确率分别为 47%, 80%, 43%, 40%, 52%. 最后我们取他们的平均数作为最终的准确率, 其结果为 57%. 其他九个志愿者的准确率分别为 34%, 60%, 45%, 43%, 33%, 55%, 38%, 41%, 52%.

此方法消除了分割时间段方法的缺点, 非常详尽的挖掘出对象的驻留规律.

5 结语

本文首先分析了基站数据的特点, 根据基站数据的特点, 提出了一种简单的统计方法来识别驻留点. 然后提出了时间段分割挖掘驻留规律的方法, 但是这个方法出现了挖掘的频繁项太多, 割裂时间连续性的缺点. 为了消除这些缺点, 本文又提出了密度聚类挖掘驻留规律的方法. 最后通过实验验证, 发现密度聚类的方法能有效详细的挖掘出个人的驻留规律.

参考文献

1 Cao HP, Mamoulis N, Cheung DW. Discovery of periodic patterns in spatiotemporal sequences. IEEE Trans. on Knowledge and Data Engineering, 2007, 19(4): 453–467. [doi: 10.1109/TKDE.2007.1002]

2 Elgethun K, Fenske RA, Yost MG, *et al*. Time-location analysis for exposure assessment studies of children using a novel global positioning system instrument. Environmental Health Perspectives, 2003, 111(1): 115–122.

3 Spaccapietra S, Parent C, Damiani ML, *et al*. A conceptual view on trajectories. Data & Knowledge Engineering, 2008, 65(1): 126–146.

4 Stopher PR. Collecting and processing data from mobile technologies. Proc. of the 8th International Conference on Survey Methods in Transport. Annecy, France. 2008.

5 Hägerstrand T. What about people in regional science? Papers of the Regional Science Association, 1970, 24(1): 6–21. [doi: 10.1007/BF01936872]

6 Goulias K, Janelle D. GPS tracking and time-geography: Applications for activity modeling and microsimulation. Final Report of an FHWA-sponsored Peer Exchange and CSISS Specialist Meeting. Santa Barbara, CA, USA. 2005.

7 Schuessler N, axhausen KW. Processing raw data from global positioning systems without additional information. Transportation Research Record: Journal of the Transportation Research Board, 2009, (2105): 28–36. [doi: 10.3141/2105-04]

8 Stopher PR, Jiang Q, FitzGerald C. Processing GPS data from travel surveys. Proc. of the 2nd International Colloquium on the Behavioural Foundations of Integrated Land-Use and Transportation Models: Frameworks, Models and Applications. Toronto, Canada. 2005.

9 Schuessler N, Axhausen KW. Processing raw data from global positioning systems without additional information. Transportation Research Record: Journal of the Transportation Research Board, 2009, (2105): 28–36. [doi: 10.3141/2105-04]

10 张治华. 基于 GPS 轨迹的出行信息提取研究[博士学位论文]. 上海: 华东师范大学, 2010.

11 张用川. 基于手机定位数据的用户出行规律分析[硕士学位论文]. 昆明: 昆明理工大学, 2013.