基于过滤-精炼策略的用户特定时间段移动轨迹特征提取①

杨东山, 张晓滨

(西安工程大学 计算机科学学院, 西安 710048)

摘 要: 发现移动用户在特定时间段的轨迹特征是实现用户个性化推荐服务的关键之一. 采用过滤--精炼策略, 研究了如何从单用户的大量轨迹数据中发现其在较长时间内的特定时间段的兴趣点. 在过滤阶段, 将用户连续若干天中同一特定时间段内的轨迹数据进行基于密度的聚类, 从而得到用户在这些天中每天的该特定时间段的停留点. 在精炼阶段, 对所有的停留点再一次聚类, 进而得到用户在这些天中该特定时间段的兴趣点. 最后, 通过实验验证了该方法的有效性.

关键词: 轨迹; 聚类; 停留点; 兴趣点

Feature Extraction for Users' Trajectories in a Period Based on Filter-Refinement Strategy

YANG Dong-Shan, ZHANG Xiao-Bin

(School of Computer Science, Xi'an Polytechnic University, Xi'an 710048, China)

Abstract: Finding features of users' trajectories in a period of time is one of the key point to realize user's personalized recommendation service. In this paper, how to find the interests in a period from the large amount of user's trajectories is presented with a filter-refinement strategy. In the filter step, the user's trajectories in the same period for several certain days are clustered based on density to obtain the user's stops; in the refinement step, the stops are clustered to obtain the user's interests. Finally, experiments show the effectiveness of this work.

Key words: trajectories; clustering; stop and move; interest

移动设备的广泛使用和GPS技术的迅猛发展使得海量的时空数据因运而生. 这些数据给人们的生活和学者的科研带来了新的机遇与挑战. 通过对用户的海量的时空数据分析, 我们可以得到用户在日常生活中感兴趣的地点, 从而在许多基于位置的服务中更好地为用户推荐贴心的个性化服务[1,2]. 而如何高效准确地从海量的时空数据中得到用户感兴趣的地点是研究用户移动行为的可预见性和独一性[3,4]进而实现用户个性化推荐的关键问题之一[5,6]. 该问题已经引起了众多学者的关注和思考.

Palma 等人[7]基于改进的 DBSCAN 算法提出了一种根据用户移动速度的快慢来识别用户停留点的算法. 该算法将传统的 DBSCAN 算法中的邻域改为基于路程的 Eps-linear-neighborhood, 结合时间 MinTime, 对空间数据进行聚类, 所得到的簇作为该用户的停留点. 此外, 该算法还给出了聚类时所需半径的计算方法. Zhao Xiuli 等人[8]在 Palma 等人的基础上对聚类时所需半径的计算方法做了改进. Zhao 等人[8]认为聚类时所需的半径计算方法应该根据用户移动的速度快慢分为两种情况, 而不是 Palma 等人所认为的只有一种情况. Jose Antonio M.R.Rocha 等人[9]提出了一种根据方向变化来识别停留点的方法并且成功地应用于识别渔船的停留点. 上述文献只是对移动对象的单条轨迹进行了分析, 从而发现一些具有重要意义的位置[10]. 而实际生活中, 用户某一天的活动规律并不能代表该用户长期以来一直遵循这条活动规律. 例如: 某用户有一天

① 基金项目:陕西省教育厅科学研究计划(14JK1307);陕西省自然科学基金(2015JQ5157);西安工程大学研究生创新基金(CX201630)
收稿时间:2016-04-07;收到修改稿时间:2016-05-16 [doi: 10.15888/j.cnki.csa.005525]

© 中国科学院软件研究所 http://www.c-s-a.org.cn

去某个商店购物, 然而在接下来的 3 个月没有再去该商店购物. 此外, 许多研究旨在找出大量移动对象在一个较长时间段内的相似行为, 用来指导决策[11]. 然而这些方法取决于多个移动对象, 只能用于对热门区域的分析, 如: 预测到某条路线交通状况[12,13].

日常生活中, 由于移动通讯网络使用传统工作模式(如: 短信服务、广播服务)[14], 用户总会随时随地收到大量的垃圾信息, 而不是在相应的时间段和相应的地点得到自己感兴趣的信息. 这样, 一方面致使自己所关心的信息被淹没, 另一方面容易引起用户对该服务强烈的厌恶和不满. 因此, 如何从单用户已有的移动轨迹数据中得到其在特定时间段的兴趣点成为解决上述问题的关键. 目前, 虽然学者们就如何分析时空数据已经提出了许多改进的基于密度的聚类方法, 但是只采用其中一种算法难以发掘出用户在较长时间内(如: 3 个月)的同一特定时间段(如: 每天的 8:00~10:00)的兴趣点. 为此, 本文采用过滤--精炼策略, 分两个阶段运用基于密度的聚类, 较好地解决了该问题. 在过滤阶段, 对某用户的连续若干天中的同一特定时间段的轨迹数据聚类, 从而得出该用户在每一天的该特定时段内的停留点. 在精炼阶段, 对这些停留点聚类, 进而得到该用户在该特定时间段的兴趣点.

1　问题描述

定义 1 (GPS 轨迹). GPS 轨迹 T 是由一组连续的 GPS 点组成的有序序列, 可表示为 $T=\{(x_1,y_1,t_1),...,(x_n,y_n,t_n)\}$. 其中$(x_i,y_i,t_i)$表示用户的第 i 个 GPS 点. 在(x_i,y_i,t_i)中, x_i、y_i 分别表示在 t_i 时刻用户的经度和纬度.

定义 2 (停留点). 在某段 GPS 轨迹中, 如果用户在该段 GPS 轨迹的子轨迹 $T'=\{(x_i,y_i,t_i),...,(x_{i+k},y_{i+k},t_{i+k})\}$ 中的速度 $v \le \Delta v$, 则 T'为停留点. 其中, Δv 为阈值. 由此, 可得出 T'的总路程 s 不能超过阈值 Eps, 且该段子轨迹的持续时间 $|t_{i+k}-t_i| \ge \Delta T$.

在实际生活中, 定位系统以恒定的、短暂的时间间隔(通常为 2-5 秒)采集移动用户的 GPS 数据, 所采集数据的位置属性精度可以达到 1m. 通过聚类所获取的停留点 T'中任意两个相邻时间的 GPS 点之间的距离非常小, 即 T'所包含的 GPS 点在二维坐标下不仅是密集的, 而且具有一定形状. 因此, 为了便于计算, 通过计算 T'中 x、y 的平均值以获得 T'的中心位置坐标, 从而将 T'压缩, 以减少计算量.

$$\begin{cases} \overline{x}=\dfrac{1}{k+1}\sum\limits_{j=i}^{i+k}x_j \\ \overline{y}=\dfrac{1}{k+1}\sum\limits_{j=i}^{i+k}y_j \end{cases} \tag{1}$$

此时, 停留点 T'可表示为 $(\overline{x},\overline{y},t_s,t_e)$. 其中 t_s、t_e 分别为该停留点的开始时间和结束时间. 图 1 所示为某用户连续三天 8:00~10:00 的停留点, 其中每个平面分别代表相应的一天中 8:00~10:00 这一特定时间段.

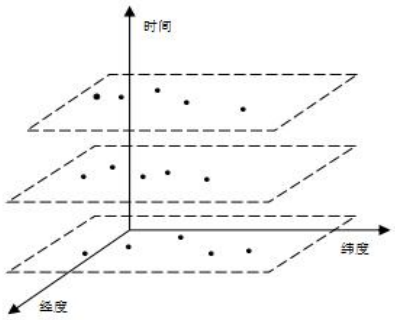

图 1　某用户连续三天 8:00~10:00 的停留点

定义 3 (移动点). 在某段 GPS 轨迹中, 如果出现下列情况, 则视为移动点: (1)两个时间相邻的停留点之间所夹着的那段连续的 GPS 轨迹点; (2)GPS 轨迹的起始点和第一个停留点之间所夹着的那段连续的 GPS 轨迹点; (3)GPS 轨迹的终止点和最后一个停留点之间所夹着的那段连续的 GPS 轨迹点; (4)如果整条 GPS 轨迹没有一个停留点, 则这条 GPS 轨迹中所有的点都是移动点.

定义 4 (兴趣点). 将集合 $\{T_{11}',T_{12}',...,T_{1k}';T_{21}',T_{22}',...,T_{2m}';...;T_{i1}',T_{i2}',...,T_{in}'\}$ 中的停留点聚类, 所形成的每一个簇代表用户在这连续的若干天内同一特定时间段经常访问的地点, 即用户在该特定时间段的兴趣点. 其中, $(T_{11}',T_{12}',...,T_{1k})$表示在第一天中用户在该特定时间段共有 k 个停留点. 类似的, $(T_{i1}',T_{i2}',...,T_{in}')$表示在第 i 天中用户在该特定时间段共有 n 个停留点.

2　特定时间段轨迹特征提取

与其它聚类方法相比, 基于密度的聚类可以发现任意形状的簇. 这更符合用户造访某地的实际情况. 然而, 只采用一种现有的基于密度的聚类算法难以准确地从单用户的大量轨迹数据中发现其较长时间内同一特定时间段的兴趣点. CB-SMoT[7]算法只适合提取用户单条移动轨迹的停留点. 而 ST-DBSCAN[15]算法

© 中国科学院软件研究所 http://www.c-s-a.org.cn

在处理单条移动轨迹时, 由于无法在时间维度得到适合的半径, 所以不能发现用户单条移动轨迹的停留点, 但是它可以对多条移动轨迹进行聚类分析. 综上, 本文采用过滤--精炼策略, 分为两个阶段(过滤阶段和精炼阶段)来有效地解决上述问题. 在过滤阶段, 从某用户连续若干天中同一特定时间段的 GPS 数据中过滤掉用户每一天该特定时段的移动点, 进而筛选出每一天该特定时段的停留点. 在精炼阶段, 从压缩后的所有停留点中提取出用户在这些天该特定时间段的兴趣点.

2.1 获取停留点

用户在运动时, 当其速度在某地变小时, 说明该用户在此地有停留, 对此地周围的景物感兴趣. 反之, 当用户的速度变大时, 说明该用户正忙着赶路, 无暇顾及周围的景物. 因此, 当用户一条 GPS 轨迹中的某一段子轨迹的总路程 s 不能超过阈值 Eps, 且该段子轨迹的持续时间大于等于阈值 ΔT 时, 则该段子轨迹被认为是该用户的停留点.

根据上述理论, 本文在获取停留点时, 采用 Eps-linear-neighborhood 邻域, 即: 在某段子轨迹 $\{p_1,.....p_{k-1},p_{k+1},.....,p_n\}$ 中, 点 p_k 的邻域为

$$(\sum_{i=m}^{k-1} dist(p_i,p_{i+1})) \le Eps \cup (\sum_{i=k+1}^{n} dist(p_{i-1},p_i)) \le Eps \quad (2)$$

其中 $t_0 \le t_m \le t_k \le t_n \le t_N$, $dist()$ 表示两点之间的距离.

由新邻域的定义可知, 如果用户在时间 ΔT 内移动的路程为 $2Eps$, 那么 $2Eps/\Delta T$ 相当于对用户在该段子轨迹中的行走速度做了限制. 由于 GPS 数据的采样时间的时间间隔是一定的, 所以 ΔT 可以认为是聚类时所形成的簇的最小密度.

2.2 获取兴趣点

将上一小节所得的所有停留点压缩, 然后对其聚类, 从而得到用户在这些天该特定时间段的兴趣点. 本次聚类需要三个参数, 它们分别是 Eps_1、Eps_2、$MinPts$. 其中, Eps_1 用来判断某一停留点在经纬度维度是否属于另一停留点的邻域; 而 Eps_2 用来判断某一停留点在时间维度是否属于另一停留点的邻域. Eps_1、Eps_2 的计算公式如下:

$$Eps_1 = \sqrt{(\overline{x}_1-\overline{x}_2)^2+(\overline{y}_1-\overline{y}_2)^2} \quad (3)$$

$$Eps_2 = \sqrt{(t_{s1}-t_{s2})^2+(t_{e1}-t_{e2})^2} \quad (4)$$

其中, $(\overline{x}_1,\overline{y}_1,t_{s1},t_{e1})$、$(\overline{x}_2,\overline{y}_2,t_{s2},t_{e2})$ 分别代表两个任意的压缩后的停留点. $MinPts$ 表示聚类时所形成的簇的最小密度.

3 实验与分析

本文采用微软(亚洲研究所)对外公开的 GeoLife 数据集作为实验数据集. 随机选取 1 名用户 A 的连续 3 个月中每天 8:00~10:00 的 GPS 数据作为研究对象.

由于在实际生活中人们造访某地时的速度是在一定范围内的, 所以在获取用户停留点时所需的参数对所有用户具有通用性. 因此, 本文在获取用户停留点时, 根据已有的文献和先验知识, 选取了具有通用性的实验所需的参数. 由于每个用户的生活习惯不同, 他们在特定时间段内的停留点个数、停留时间的长短不同, 而这些都会直接地对获取用户兴趣点造成影响, 因此在获取某用户兴趣点时, 需要寻求适合该用户的最佳参数, 而不是一概而论. 为此, 本文通过实验得到适合用户 A 在获取兴趣点时所需的最佳参数, 进而验证了所提出的方法的有效性.

在获取停留点的实验中, ΔT 根据文献[7]设为 180s. 根据人们的生活常识, 一个正常的成年人步行的速度约为 1.2m/s. 当移动用户的运动速度不大于 1.2m/s 时, 表示该用户对周围的景物感兴趣, 因此, 本文将 Eps 设为 216m. 图 2 所示为用户 A 在某天该时间段的轨迹. 图 3 所示为该用户在这一天 8:00~10:00 的停留点.

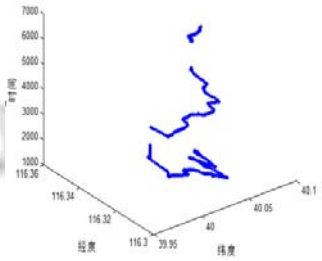

图 2 用户 A 某天 8:00~10:00 的轨迹

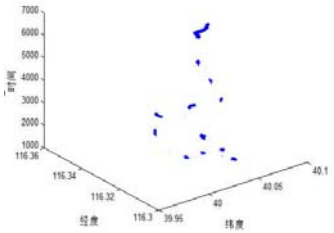

图 3 用户 A 某天 8:00~10:00 的停留点

在获取兴趣点的实验中, 本文将 Eps_1 分别取 50m、100m、150m、200m、250m, Eps_2 分别取 400s、450s、500s、550s、600s, $MinPts$ 分别取 2、3、4、5. 通

© 中国科学院软件研究所 http://www.c-s-a.org.cn

过组合 Eps_1、Eps_2 和 $MinPts$ 的不同值，观察用户 A 在每组参数下所形成簇的个数，从而得到最佳的参数组合. 图 4 所示为用户 A 在每组参数下所形成簇的个数.

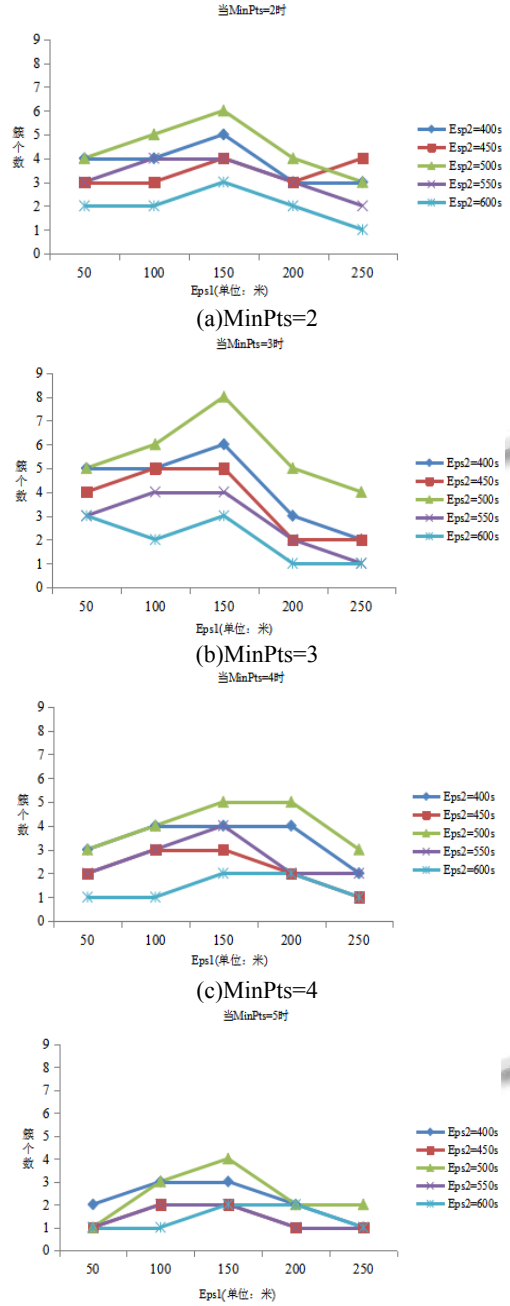

图 4　用户 A 在不同参数下聚类的平均值

从图 4 中可以得到：当 Eps_1=150、Eps_2=500、$MinPts$=3 时，所形成的簇最多. 因此，该参数组合为用户 A 的最佳参数组合. 此外当 Eps_1、Eps_2 和 $MinPts$ 三者中任意两个参数一定时，与其相对应的实验结果的趋势均随着另一个参数的增大而先增大后减小. 这是因为在基于密度的聚类过程中，只有在一定范围内增大某一个参数时，其形成的簇的个数会随着该参数的增大而增大. 当该参数的值超过某一数值时，会出现几个簇的合并，从而使得簇的个数减少.

使用上述最佳参数组合，对已经得到的用户 A 所有停留点聚类，得到如图 5 所示的结果. 从图 5 中可以看出来，该用户在这 90 天中这一时间段内的兴趣点共有 8 个.

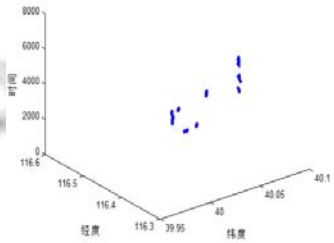

图 5　用户 A90 天 8:00~10:00 经常停留的地点

4　结语

本文采用过滤—精炼策略，运用基于密度的聚类对单用户在特定时间段的移动轨迹数据进行分析，从而得到该用户在指定时间段的兴趣点. 这为实现基于位置的个性化推荐服务提供了扎实的基础. 在实际应用中，服务商可以根据用户在特定时间段所处的位置为其提供更准确的个性化推荐服务.

参考文献

1 宋国杰,唐世渭,杨冬青,等.一种无线通信环境中的用户移动模式的挖掘算法.软件学报,2002,13(8):1465–1471.

2 孟祥武,胡勋,王立才,等.移动推荐系统及其应用.软件学报,2013,24(1):91–108.

3 郭迟,刘经南,方媛,等.位置大数据的价值提取与协同挖掘方法.软件学报,2014,25(4):713–730.

4 Song C, Qu Z, Blumm N, et al. Limits of predictability in human mobility. Science, 2010, 327(5968): 1018–1021.

5 刘树栋,孟祥武.一种基于移动用户位置的网络服务推荐方法.软件学报,2014,25(11):2556–2574.

6 刘树栋,孟祥武.基于位置的社会化网络推荐系统.计算机学报,2015,38(2):322–336.

7 Palma T, Bogorny V, Kuijpers B, et al. A clustering-based approach for discovering interesting places in trajectories. Proc. of the 2008 ACM Symp. on Applied Computing. New

© 中国科学院软件研究所 http://www.c-s-a.org.cn

York. ACM. 2008. 863–868.

8 Zhao XL, Xu WX. A clustering-based approach for discovering interesting places in a single trajectory. 2009 Second International Conference on Intelligent Computation Technology and Automation. New York. IEEE. 2009. 429–432.

9 Rocha JAMR, Times VC, Oliveira G, et al. DB-SMOT: A direction-based spatio-temporal clusteringmethod. IEEE Conference of Intelligent Systems. New York. IEEE. 2010. 114–119.

10 刘大有,陈慧灵,齐红,等.时空数据挖掘研究进展.计算机研究与发展,2013,50(2):225–239.

11 刘奎恩,肖俊超,治明,等.轨迹数据库中热门区域的发现.软件学报,2013,24(8):1816–1835.

12 乔少杰,金琨,韩楠,等.一种基于高斯混合模型的轨迹预测算法.软件学报,2015,26(5):1048–1063.

13 李国徽,钟细亚.一种基于固定网格的移动对象运动轨迹索引模型.计算机研究与发展,2006,43(5):828–833.

14 孟祥武,王凡,史艳翠,等.移动用户需求获取技术及其应用.软件学报,2014,25(3):439–456.

15 Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatial-temporal data. Data & Knowledge Engineering, 2007, 60(1): 208–221.

www.c-s-a.org.cn

© 中国科学院软件研究所 http://www.c-s-a.org.cn