(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 105740904 A
(43)申请公布日 2016.07.06

(21)申请号 201610066709.0

(22)申请日 2016.01.29

(71)申请人 东南大学
地址 210096 江苏省南京市四牌楼2号

(72)发明人 叶智锐 施晓蒙 汤斗南 赵鑫玮
陆加健 吴运腾 吴丽霞

(74)专利代理机构 南京苏高专利商标事务所（普通合伙）32204
代理人 孟红梅

(51)Int.Cl.
G06K 9/62(2006.01)
G08G 1/01(2006.01)

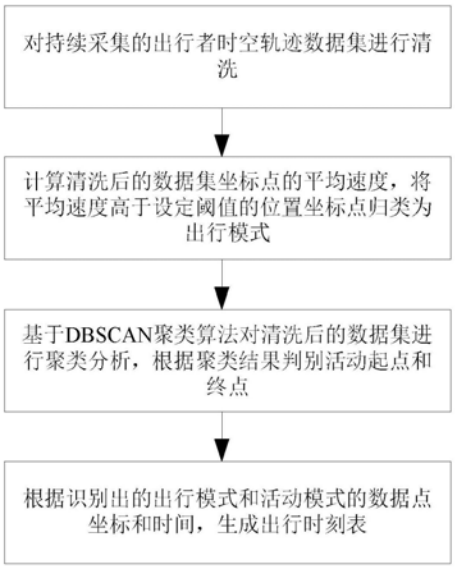
权利要求书1页 说明书4页 附图1页

(54)发明名称

一种基于DBSCAN聚类算法的出行与活动模式识别方法

(57)摘要

本发明公开了一种基于DBSCAN聚类算法的出行与活动模式识别方法，包括如下步骤：对持续采集的出行者时空轨迹数据集进行清洗；计算清洗后的数据集坐标点的平均速度，将平均速度高于设定阈值的位置坐标点归类为出行模式；基于DBSCAN聚类算法对清洗后的数据集进行聚类分析，根据聚类结果判别活动起点和终点；根据识别出的出行模式和活动模式的数据点坐标和时间，生成出行时刻表。本发明方法基于采集到的出行者时空轨迹序列集合，通过基于密度的聚类算法(DBSCAN)，将出行者的行为模式分为出行模式和活动模式。本发明方法便于计算与实际操作，实用性强，可以比较准确地判定出行者的行为模式，为后续的研究提供便捷，具有重要的现实意义。



CN 105740904 A

1.一种基于DBSCAN聚类算法的出行与活动模式识别方法,其特征在于,该方法包括如下步骤:

(1)数据清洗:对持续采集的出行者时空轨迹数据集进行清洗;

(2)出行模式识别:计算清洗后的数据集坐标点的平均速度,将平均速度高于设定阈值的位置坐标点归类为出行模式;

(3)活动模式识别,包括活动起点识别和活动终点识别,具体为:基于DBSCAN聚类算法对清洗后的数据集进行聚类分析,根据聚类结果判别活动起点和终点,在指定的时间间隔T1内的有大于指定的最小包含点数N1的数据点与数据点A的距离均小于指定的距离D1,则数据点A判别为活动的起点;若存在在指定的时间间隔T2内且不属于半径为D2的临界区域中的一个连续的数据点集,则该数据点集的第一个点判别为活动的终点;

(4)生成出行时刻表:根据识别出的出行模式和活动模式的数据点坐标和时间,生成出行时刻表。

2.根据权利要求1所述的基于DBSCAN聚类算法的出行与活动模式识别方法,其特征在于,所述数据清洗步骤中包括移除边界点的步骤,所述边界点为连续具有高于设定速度阈值的点坐标序列的起始坐标点。

3.根据权利要求1所述的基于DBSCAN聚类算法的出行与活动模式识别方法,其特征在于,所述出行模式识别步骤中包括:

选定一时间间隔,计算每个时间间隔内数据点的平均速度;

绘制平均速度随时间变化的曲线,设置一速度阈值;

将平均速度高于所设速度阈值的位置坐标点归类为出行模式。

4.根据权利要求1所述的基于DBSCAN聚类算法的出行与活动模式识别方法,其特征在于,活动起点的识别步骤包括:

设置临界区域的半径为D1,时间间隔为T1,最小包含点数N1,其中N1根据时间间隔与时间中值的比值确定;

对聚类结果的每个分组数据进行如下运算:从第一个数据点开始,计算在时间间隔T1区间内的所有坐标点与第一个数据点的距离,若所有距离均小于D1,或者,距离小于D1的点数大于N1,则该第一个数据点为活动的起点;带入下一个数据点进行同样的运算,直至遍历完成一个分组中的所有数据点。

5.根据权利要求1所述的基于DBSCAN聚类算法的出行与活动模式识别方法,其特征在于,活动终点的识别步骤包括:

设置临界区域的半径为D2,时间间隔为T2;

对聚类结果的每个分组数据进行如下运算:找到时间间隔T2内且不属于临界区域中的一个连续的序列点数据集,这个序列点数据集的第一个点将作为下一个活动行为的判定起点。

6.根据权利要求1所述的基于DBSCAN聚类算法的出行与活动模式识别方法,其特征在于,所述时间间隔T1为3分钟,距离D1为20米,时间间隔T2为10分钟,距离D2为100米。

一种基于DBSCAN聚类算法的出行与活动模式识别方法

技术领域

[0001] 本发明涉及交通出行信息技术领域，特别是涉及一种基于DBSCAN聚类算法的出行与活动模式识别信息采集方法。

背景技术

[0002] 居民出行数据是交通规划与管理的基础。交通需求建模理论发展至今，大致可归为两类理论体系：基于出行、基于活动的需求建模。基于出行的需求建模被广泛应用于传统“四阶段法”交通规划的实践中。基于出行的需求建模方法，从宏观角度，以独立的出行单元为对象，整体分析各个交通小区的出行需求。然而，该方法没有考虑到这些个体出行之间的联系，主要表现在两方面，一是缺乏对个体出行行为的考虑，二是没有考虑如何组织出行过程(出行时刻表)。而基于活动的出行需求分析将出行视为一种既得需求—从空间中分布的活动进行来获取需求，通过考虑这些活动与出行行为之间复杂的交互影响，分析得到出行者意图与需求，从而预测与识别群体的交通需求。

[0003] 现阶段我国主要采用人工调查法获取居民的出行信息，该方法既繁琐又耗费人力、财力。而且人工调查的结果受调查员的水平、居民参与的积极性、表格的回收率、意外事件等多方面因素的影响，得到的数据也往往精确性与真实性不足，常常是耗费了巨大人力物力却并没有得到很好的调查结果。随着科技的进步与发展，尤其是各种传感器的应用与发展，出现了如车载GPS、手机、公交卡、银行卡等可以记录人类的活动轨迹数据的技术。尤其是智能手机的广泛普及应用，为居民出行数据的采集提供了新思路。

[0004] 大数据时代下的多源数据，为基于活动的交通规划的实施提供了数据输入支撑，使更加精细、实时的交通规划成为可能。同时，对于出行者本身来说，出行与活动模式的划分也有助于自身的交通出行决策。出行模式是指出行者参与交通过程的状态，即各种通过交通方式进行交通出行；活动模式为交通参与者在出行过程中进行的一些活动，例如购物、休闲、娱乐。

[0005] 本发明基于DBSCAN聚类算法，该方法基于采集到的出行者时空轨迹序列集合，通过基于密度的聚类算法DBSCAN，精确识别出行者的行为模式，将其分为出行模式和活动模式。进行DBSCAN聚类分析，不需要对输入数据的分布做任何假设，且得到的结果与数据记录输入到算法中的顺序无关，给研究带来很大方便；同时，它能较好地处理高维数据表格对象，可以让我们获得出行者的时间、经纬度等多维信息；该方法能够发现任意形状的聚类，聚类挖掘的结果对异常数据具有非敏感性，有助于提高获取信息的精度，更为精确地识别出行者的模式。

发明内容

[0006] 发明目的：为了克服上述现有技术的不足，本发明提供了一种基于DBSCAN聚类算法的出行与活动模式识别方法，根据采集到的出行者时空轨迹序列集合比较准确地判定出行者的行为模式。

[0007] 技术方案:为实现上述发明目的,本发明采用如下技术方案:

[0008] 一种基于DBSCAN聚类算法的出行与活动模式识别方法,包括如下步骤:

[0009] (1)数据清洗:对持续采集的出行者时空轨迹数据集进行清洗;

[0010] (2)出行模式识别:计算清洗后的数据集坐标点的平均速度,将平均速度高于设定阈值的位置坐标点归类为出行模式;

[0011] (3)活动模式识别,包括活动起点识别和活动终点识别,具体为:基于DBSCAN聚类算法对清洗后的数据集进行聚类分析,根据聚类结果判别活动起点和终点,在指定的时间间隔T1内的有大于指定的最小包含点数N1的数据点与数据点A的距离均小于指定的距离D1,则数据点A判别为活动的起点;若存在在指定的时间间隔T2内且不属于半径为D2的临界区域中的一个连续的数据点集,则该数据点集的第一个点判别为活动的终点;

[0012] (4)生成出行时刻表:根据识别出的出行模式和活动模式的数据点坐标和时间,生成出行时刻表。

[0013] 进一步地,所述数据清洗步骤中包括移除边界点的步骤,所述边界点为连续具有高于设定速度阈值的点坐标序列的起始坐标点。

[0014] 进一步地,所述出行模式识别步骤中包括:

[0015] 选定一时间间隔,计算每个时间间隔内数据点的平均速度;

[0016] 绘制平均速度随时间变化的曲线,设置一速度阈值;

[0017] 将平均速度高于所设速度阈值的位置坐标点归类为出行模式。

[0018] 进一步地,活动起点的识别步骤包括:

[0019] 设置临界区域的半径为D1,时间间隔为T1,最小包含点数N1,其中N1根据时间间隔与时间中值的比值确定;

[0020] 对聚类结果的每个分组数据进行如下运算:从第一个数据点开始,计算在时间间隔T1区间内的所有坐标点与第一个数据点的距离,若所有距离均小于D1,或者,距离小于D1的点数大于N1,则该第一个数据点为活动的起点;带入下一个数据点进行同样的运算,直至遍历完成一个分组中的所有数据点。

[0021] 进一步地,活动终点的识别步骤包括:

[0022] 设置临界区域的半径为D2,时间间隔为T2;

[0023] 对聚类结果的每个分组数据进行如下运算:找到时间间隔T2内且不属于临界区域中的一个连续的序列点数据集,这个序列点数据集的第一个点将作为下一个活动行为的判定起点。

[0024] 有益效果:本发明方法通过简单地数据清洗和聚类分析,以平均速度为指标,简单而巧妙的判断人的出行活动并记录轨迹,并且可以全天不间断地记录数据,大大节省了相关调查的人力、财力和物力,对出行活动数据的精确采集和判断具有积极作用。针对模式识别和轨迹记录,可以揭示人类活动轨迹在时间、空间的从聚模式、周期性等特点,进而为用户进行轨迹预测、城市动态景观、城市交通等方面的研究提供支持。

附图说明

[0025] 图1为本发明方法的总体流程图。

具体实施方式

[0026] 下面结合附图和实施例对本发明作更进一步的说明。应理解这些实施例仅用于说明本发明而不用于限制本发明的范围，在阅读了本发明之后，本领域技术人员对本发明的各种等价形式的修改均落于本申请所附权利要求所限定的范围。

[0027] 如图1所示，本发明实施例公开的一种基于DBSCAN聚类算法的出行与活动模式识别方法，该方法依次包括数据清洗、出行模式识别、活动模式识别和生成出行时刻表的步骤。

[0028] 数据清洗步骤，即对24小时持续采集的出行者时空轨迹数据进行清洗。由于WIFI信号缺失或者Google Play Service的错误，空间轨迹数据是存在间隙的。因此，为了方便研究，这些间隙被视为标记的序列，有间隙的数据被定义为分段数据。然后，我们统计边界点的空间位置坐标，通过计算每个点的经纬度坐标，筛选出具有不寻常的较高速度的点坐标(速度阀值可根据数据和需求自行定义)。若存在连续具有较高速度的点坐标序列，则定义该坐标序列的起始坐标点为边界点。为了提升聚类与活动探测的效率，我们将所有边界点从数据集中移除。

[0029] 出行模式识别步骤，即计算完成清洗过程后的数据点集的平均移动速度，基于平均速度进行出行模式归类。选定一个时间间隔(如10分钟)，计算出每个时间间隔内数据点的平均速度。绘制出平均速度随时间变化的曲线，然后设置一个速度阀值(该速度阈值可根据数据和需求自行定义)，则我们可以识别出高于阀值的位置坐标点，然后将其归类为出行模式。虽然低于阀值的数据点之中也许会包含一部分出行模式，但也同时包含了所有活动模式的数据点。经过这个步骤，我们可以将出行模式与活动模式的分类结果通过可视化的方式展现出来。

[0030] 活动模式识别步骤，包括活动起点识别和活动终点识别。首先基于DBSCAN聚类算法对清洗后的数据集进行聚类分析，然后根据聚类结果来判别活动起点和终点。经过数据清洗得到个一系列数据点，每个数据点包括三个参数:UTC时间t、经度lon、纬度lat。对DBSCAN算法输入三个参数:数据点、最小活动时间min_time(如3min)、搜索领域半径Eps(如20米)，最小包含点数可由时间间隔(最小活动时间)与时间中值的比值求出，并四舍五入取整，其中时间中值为相邻数据点的时间差值的中值。识别活动模式的具体过程为:

[0031] 选取第一个数据点为例，先找到第一个数据点与其余数据点的距离小于20米的点，然后判断它在最小活动时间内包含的数据点小于最小包含点数，故输出第一个数据点的类型为外部点。依次循环，找到某个数据点在最小活动时间内包含的数据点数大于临界包含点数时，输出这个数据点的点类别为中心点，行为模式类型为活动模式。

[0032] 其中，判别活动起点步骤，即可使用聚类结果的分组数据，来进行活动起点判别测试。具体过程为:我们设置一个半径为D1的临界区域(本算法中设为20米)，选定一个特定的时间间隔T1(本算法中建议选用3分钟)。最小包含点数N1由时间间隔与时间中值的比值确定，并四舍五入取整，其中时间中值为相邻数据点的时间差值的中值。然后从一个分组里的第一个数据点开始，计算在此时间间隔区间内的所有坐标点与第一个数据点的距离，距离公式为:

[0033] $$D = \sqrt{(x_1 - x_i)^2 + (y_1 - y_i)^2}$$

[0034] 其中，x，y为数据点的经纬度，D为两数据点的距离。

[0035] 如果所有距离都小于20米，或者，距离小于20米的点数大于最小包含点数，则可以判断为活动开始。如果任何一个距离大于20米，则带入下一个数据点重复上述计算过程。

[0036] 同样，基于聚类分析的结果，我们可以进行活动终点的判别。为了判别活动的终点，我们设置一个半径为D2(本算法中设为100米)的临界区域，选定一个时间间隔T2(本算法中建议选用10分钟)，找到该时间间隔内且不属于临界区域中的一个连续的序列点数据集。此时活动终止，这个序列点数据集的第一个点将作为下一个活动行为的判定起点。

[0037] 重复判别活动起点、终点过程，进行直到所有数据集都被遍历。最后根据识别出的出行模式和活动模式的数据点坐标和时间，生成出行时刻表。

[0038] 本发明方法充分考虑了活动的特征，通过对数据的筛选和聚类分析以及以平均速度为指标，对人的出行活动进行了准确判断和识别，并生成出行时刻表，方便了数据采集。用本发明方法获取的数据中蕴含着人类行为的时空分布模式，通过对这些轨迹的研究可以挖掘个体轨迹模式，进而为轨迹预测、城市规划、交通监测等方面的研究提供技术支持。

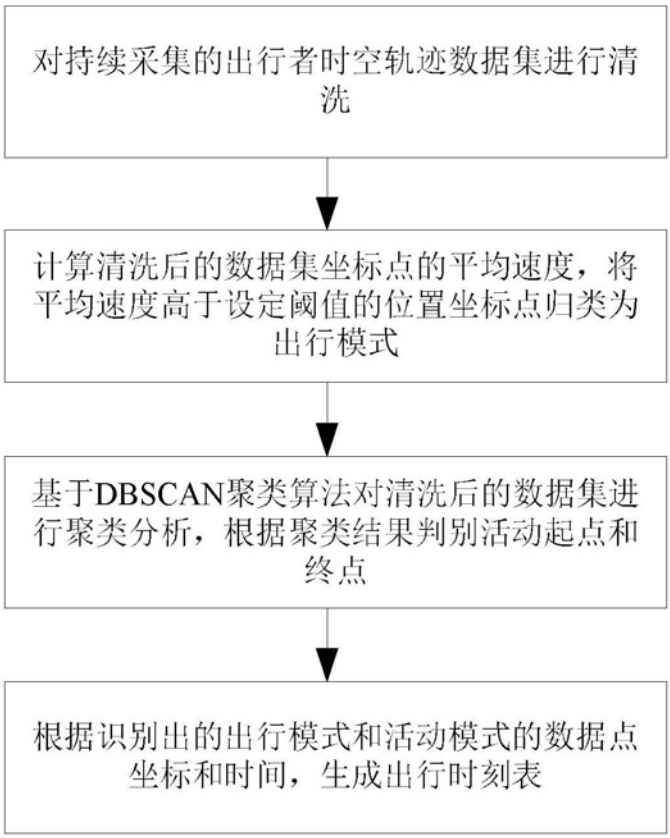

对持续采集的出行者时空轨迹数据集进行清洗
计算清洗后的数据集坐标点的平均速度，将平均速度高于设定阈值的位置坐标点归类为出行模式
基于DBSCAN聚类算法对清洗后的数据集进行聚类分析，根据聚类结果判别活动起点和终点
根据识别出的出行模式和活动模式的数据点坐标和时间，生成出行时刻表

图1