第45卷第10期 中国科学技术大学学报 Vol. 45, No. 10
2015年10月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Oct. 2015

文章编号:0253-2778(2015)10-0829-07

# 基于移动数据的人群活动热点区域的发现

班雷雨,霍 欢,徐 彪

(上海理工大学光电信息与计算机工程学院,上海 200093)

**摘要**:移动数据记录了人群活动位置和时间的变化,反映了关于人类移动的语义知识.从区域语义分析移动人群频繁访问的热点区域,对理解区域特色在智慧城市方面的应用,具有重要意义.对于如何发现热点区域以及如何限制其覆盖范围,首先对人群位置序列进行分析,用网格作为空间数据索引结构并结合 Top-$k$ 排序;然后提出了基于核函数的热点区域发现方法,并给出了热点区域发现算法;最后在真实数据集上,验证了该方法的可行性和有效性.

**关键词**:移动数据;人群活动;热点区域;网格索引;密度; 核函数

**中图分类号**:TP18 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2015.10.005

**引用格式**: BAN Leiyu, HUO Huan, XU Biao. Discovery of hot regions about crowd activities based on mobility data[J]. Journal of University of Science and Technology of China, 2015,45(10):829-835,863.
班雷雨,霍 欢,徐 彪. 基于移动数据的人群活动热点区域的发现[J]. 中国科学技术大学学报, 2015,45(10):829-835,863.

# Discovery of hot regions about crowd activities based on mobility data

BAN Leiyu, HUO Huan, XU Biao

(*School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and technology, Shanghai* 200093, *China*)

**Abstract**: Mobility data records the change of location and time about crowd activities, showing semantic knowledge about human mobility. From the perspective of regional semantic knowledge, mining the hot regions visited frequently by moving crowds is essential to understand regional characteristics in the smart city applications. This paper studied how to discover hot regions and how to constraint their coverage size. Based on an analysis of the location sequence of moving crowd, a discovery method for discovering hot regions based on kernel function was proposed. This discovery method uses the grid as a spatial data indexing structure and the Top-$k$ sorting method. A discovery algorithm of hot regions was presented based on the discovery method. Finally, experimental results validate accurately the feasibility and effectiveness of the method on practical datasets.

**Key words**: mobility data; crowd activities; hot region; grid indexing; density; kernel function

## 0 引言

随着信息与通讯技术的快速发展和位置感知设备(如智能手机、车载 GPS 等)的应用普及,基于位置的服务、智能城市区域规划、城市交通监控等领域的研究应用成为当前的分析热点.其中,对移动数据

**收稿日期**:2015-08-27;**修回日期**:2015-09-29

**基金项目**:国家自然科学基金(61473001,71071045,71131002);安徽大学青年科学研究基金(33050054).

**作者简介**:班雷雨,女,1989 年生,硕士生,研究方向:数据挖掘. E-mail:1060891990@qq.com

**通讯作者**:霍欢,博士/副教授. E-mail:huo_huan@yahoo.com

方面的研究大部分集中在轨迹模式发现上[1-7]，虽然有些轨迹模式发现的研究工作涉及热点区域发现方法[6-7]，但是相关研究工作多将轨迹模式用于预测和分类，而对移动对象活动的运动模式研究较少[8]. 为此，本文将移动对象活动的热点区域引入到运动模式中，重点研究运动模式中移动对象运动轨迹的空间属性信息. 将移动对象空间信息的位置序列和区域特征方面的知识相关联，从而发现被移动对象频繁访问的热点区域，这是挖掘移动对象运动模式的关键前提，同时也是理解智慧城市区域特色的一种方式[9].

热点区域从一定程度上反映出活动事件具有较高的可能性，可以为智能城市土地价值评估、公共停车场规划等提供有效的参考信息，但是如何发现热点区域以及热点区域的覆盖范围如何约束，目前还缺乏深入研究. 本文为了有效地解决人群活动的热点区域，发现相关问题，首先提出一个基于核函数的热点区域发现方法，将人群活动空间离散化成 $w \times h$ 大小的网格，通过核方法和启发式方法获取网格中的单元格密度值. 其中，网格作为空间数据索引结构，与数据对象规模无关[8]，可有效降低热点区域发现的复杂度，提高热点区域发现算法的性能. 同时借鉴 Top-$k$ 排序选出密度阈值，作为界定热点区域的边界，解决了热点区域的覆盖范围约束问题；接着，提出了热点区域发现算法，热点区域发现问题的实验结果表明，本文提出的发现算法具有相对较好的算法性能和聚类效果.

## 1 相关工作

移动数据记录了人群活动位置和时间的变化情况，将相应的数据映射到现实的地图区域位置上，就可以还原出对象的运动轨迹，对运动轨迹分析获得人群空间位置序列分布与个体活动特征信息. 本文对移动数据进行挖掘，分析与人群相关的活动区域问题，并涉及空间聚类、索引方法、区域范围发现方法这三个主要方面，下面给出这三个方面的相关工作.

近年来，数据挖掘领域已有的空间数据聚类算法中，比较典型的聚类方法有：划分聚类方法，如 $k$-means，$k$-medoids[10]，能比较好地发现数据规模不大的球状簇，但不能检测到任意形状的簇；层次聚类方法，如 Wu 等[11]提出了一种新的基于分段的聚类聚合方法，作为一种随机搜索方法，可以相当有效地找出最优解. 该方法虽然对传统的层次聚类方法进行了改善，然而从高效角度来说，聚类过程很耗时；基于密度的聚类方法，如 DBSCAN[12] 可发现任意形状簇，较好地发现空间中密集的区域和稀疏的区域. 此外，还有一些聚类算法，如 Liu 等[13]对移动对象聚类，提出基于移动的聚类算法，用以识别移动车辆潜在的热点区域和 Dai[14] 通过对轨迹集聚类研究移动对象轨迹的相似性，但这些方法引入了复杂的参数，增加了参数估计的难度，而且文献[14]是基于距离的计算，复杂的参数增加了计算的复杂性. 本文的算法是基于密度的计算，引入 Top-$k$ 排序，用 $k$ 值来确定密度阈值的大小，可以比较灵活地调整 $k$ 值，有利于分析热点区域发现的相关工作.

数据挖掘领域中，一些有代表性的索引方法，如 STRIPES[15]，用于解决轨迹预测问题，但以算法的空间复杂度为代价获取结果；$R$ 树，被 Mamoulis 等[16]证实其在聚类任意形状时不是一个有效的索引结构. 本文的索引方法用网格作为索引结构，提出基于核函数的热点区域发现方法，该索引方法与数据对象规模无关，并且可以有效率地管理移动对象，提高发现算法的效率.

对于如何约束区域的覆盖范围，Yuan 等[6]根据人类移动和兴趣点(POIs)，用主干道对城市划分成互不相连的区域，从而找到具有不同功能的区域，该方法可以帮助人们快速地了解复杂的城市等，但没有给出密集区域；Jensen 等[17]在区域密度概念的基础上，提出了二维的密度直方图，改善了密度获取的效率，虽然该方法要求用一对面积阈值限定密集区域面积的大小，但是未能给出有效地解决方法；文献[8-18]在问题描述中设定热门区域有一个紧凑的约束边界，即热门区域大小受限在一对约束半径之间，但涉及大区域的问题，问题变得复杂化. 本文采用比较经典的核方法[19]并结合 Top-$k$ 的排序思想，通过调整 $k$ 值来给出密度阈值，从而限定热点区域的范围，发现热点区域.

## 2 相关定义和描述

**定义 2.1**(运动轨迹)运动轨迹可由组成移动数据的时空点序列构成：$<TS_1, TS_2, \cdots, TS_n>$，其中，$TS_i$ 是一个具有经度、纬度和时间的三元组 $(lon_i, lat_i, t_i)$，表示移动对象在 $t_i$ 时刻的位置是 $(lon_i, lat_i)$，可反映出运动过程中的空间和时间信息属性. 图 1 反映出空间信息特性是运动轨迹的特征之一. 图 2 分别

从二维和三维空间角度分析运动轨迹.

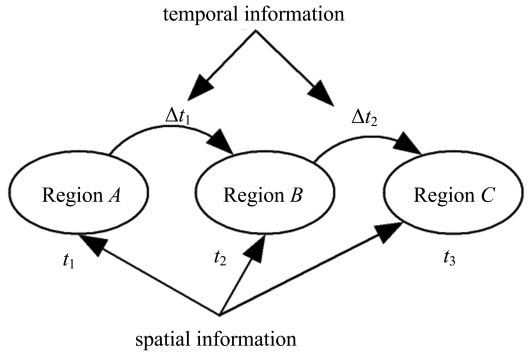


**图 1　运动轨迹属性图**

**Fig. 1　Motion trajectory properties**

从图 1 中可以看出,移动对象从 Region $A$ 经过时间 $\Delta t_1$ 运动到 Region $B$,经过时间 $\Delta t_2$ 从 Region $B$ 运动到 Region $C$,所以该移动对象的运动轨迹就可以体现出空间和时间两个信息属性,并且通常用定义 2.1 中的时空点序列给出定义,在图 2 中可以反映出来.

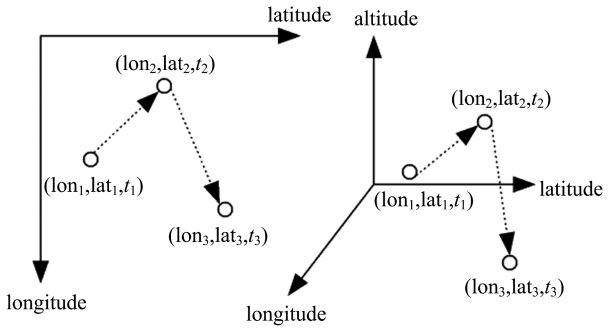


**图 2　从二维和三维角度分析运动轨迹**

**Fig. 2　Motion trajectories in 2, 3 dimensions respectively**

图 2 从空间信息属性分析移动对象的运动轨迹,图 2 中的左侧图是二维空间上移动对象的运动轨迹,即从时空点($lon_1$, $lat_1$, $t_1$)经过($lon_2$, $lat_2$, $t_2$)运动到($lon_3$, $lat_3$, $t_3$). 同样地,图 2 中的右图是三维空间上移动对象的运动轨迹,在左图中经度和纬度的二维坐标的基础上考虑了空间中的高度坐标,形成空间中的三维坐标.

根据运动轨迹的定义知,运动轨迹包含空间和时间两种信息属性,而本文的研究重点是从空间信息属性角度去挖掘与移动对象活动相关的热点区域,下面给出位置序列的定义.

**定义 2.2**　(位置序列)一个移动对象的位置序列可由反映移动对象运动过程的原始移动数据按一定的时间间隔进行线性插值后得到的序列 $p$ 组成,可记为 $Loc=<p_1, p_2, \cdots, p_n>$,其中 $p_i$ 为一空间位置点($p_i.x$, $p_i.y$).

发现被移动对象频繁访问的热点区域,对于挖掘移动对象运动模式非常关健. 本文给定一个位置序列,目的是挖掘热点区域,下面给出热点、热点区域的概念以及如何发现热点区域.

**定义 2.3**　(热点)活动空间里的单元格被称为热点(hot point),当且仅当满足密度大于等于一个密度阈值 $\delta$.

**定义 2.4**　(密度阈值 $\delta$)借鉴 Top-$k$ 的排序思想,从移动对象位置序列的所有单元格密度值中,选取第 $k$ 大的密度值作为密度阈值 $\delta$.

**定义 2.5**　(热点区域)热点区域是指由一系列热点聚集成的区域,即移动对象频繁访问的相对密集的区域. 在位置序列中的热点区域集合可记作 $HR=\{HR_1, HR_2, \cdots, HR_n\}$, $HR_i=\{\text{hot points}\}$, $n$ 为热点区域的数目. 图 3 给出了某个移动对象的热点区域.

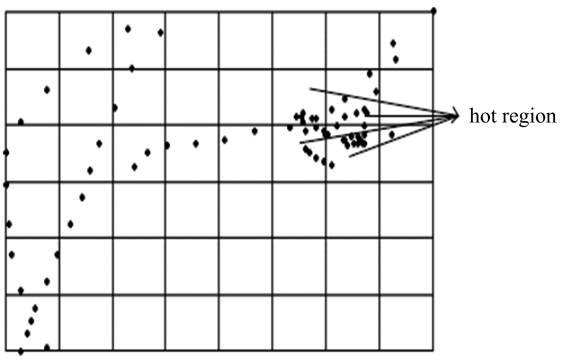


**图 3　移动对象的位置序列和热点区域图**

**Fig. 3　Location sequence and hot areas of moving objects**

图 3 给出了某个移动对象活动的位置序列,对位置序列运用本文提出的基于核函数的热点区域发现方法,可以得到图 3 中由单元格聚类成的热点区域. 下面具体给出基于核函数的热点区域发现方法.

## 3　基于核函数的热点区域发现方法

### 3.1　网格结构划分设计

在分析基于核函数的热点区域发现方法的过程中,活动空间进行网格划分后,就可以根据核方法和启发式公式计算含有位置点的单元格的密度值. 网格的引入加快了热点的获取,提高了热点区域发现算法的效率,因此网格在整个热点区域发现过程中起到了重要作用,但是网格如何划分,即如何将移动对象的活动空间进行网格化,这是本节研究的重点.

本文使用经度和纬度的划分方法将移动对象的活动位置空间离散化成 $w\times h$ 大小的网格. 网格划分后，移动对象活动的位置点序列被映射到相应的单元格中. 同时，为了保证发现的任意热点区域之间不重叠，即保证热点区域之间的无冗余性，对已经属于某一个热点区域的单元格要进行标记处理.

为了更好地理解网格结构划分的设计，结合图4的移动对象位置序列进行分析. 首先，将移动对象的活动空间用经度和纬度的方法划分成 $w\times h$ 大小的网格，$w\times h$ 大小可根据所观察的数据期望的分辨率适当调整(在图4中 $w\times h$ 取为 $3\times 3$). 同时，为了尽可能使单元格的长度 $\Delta w$ 和高度 $\Delta h$ 具有较清晰的分辨率，下面给出网格划分的步长参考值.

$$\Delta w=\frac{\max|p_i.\text{latitude}-p_j.\text{latitude}|}{w}$$
$$\forall\ p_i,p_j\in \text{Loc},i\neq j \tag{1}$$

和

$$\Delta h=\frac{\max|p_i.\text{longitude}-p_j.\text{longitude}|}{h}$$
$$\forall\ p_i,p_j\in \text{Loc},i\neq j \tag{2}$$

式中，Loc 为移动对象的位置序列，$p_i,p_j$ 为活动空间内的位置点，$p_i.$ latitude，$p_i.$ longitude，$p_j.$ latitude，$p_j.$ longitude 分别为 $p_i,p_j$ 在纬度和经度方向上的坐标值，所以根据上面的划分方法，得到移动对象位置序列被网格划分后的图，如图4所示.

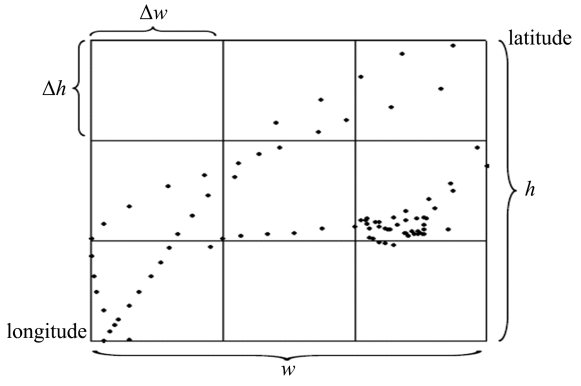


**图4 移动对象的活动位置空间网格化图**

**Fig. 4 Moving object location space grids**

### 3.2 发现方法

空间信息特性是移动对象运动轨迹的特征之一，反映了移动对象的位置变化信息. 本文从空间信息属性的角度分析移动数据以便获取移动对象的频繁访问的活动热点区域. 在本文中，为了提高热点区域的发现效率，采用网格结构划分设计的划分方法，将整个移动对象活动位置空间离散化成 $w\times h$ 大小的网格，即由划分线垂直于坐标抽划分而成的若干个等大的单元格(cell)组成，在此基础上计算每个单元格的密度. 为了便于在网格空间内对移动对象的位置序列进行表述，对每个移动对象的活动位置点取其在空间单元格坐标的所有位置点的平均坐标值，可表示为$(x,\ y)$，这样原来由聚类在同一单元格的多个位置点，现在只需要用单个特征点坐标值表示. 结合前面对运动轨迹和位置序列的介绍，则 $x$ 和 $y$ 的计算公式如下：

$$x=\frac{\sum_{i=1}^{n}p_i.x}{n} \tag{3}$$

和

$$y=\frac{\sum_{i=1}^{n}p_i.y}{n} \tag{4}$$

移动对象的活动空间被网格划分后，位置点的密度就可以转化为计算单元格的密度. 为了估算每个单元格的密度，本文使用比较经典的核方法(kernel method)[19]，用二次核密度函数来计算移动对象活动所在的单元格的密度. 对于任一单元格 c 和活动空间任一位置点 $p_i$，则每个单元格的密度值估算 f(c)可根据二元正态密度核函数计算得出，即

$$f(c)=\frac{1}{n\gamma^2}\sum_{i=1}^{n}\frac{1}{2\pi}\exp\left(-\frac{|c-p_i|^2}{2\gamma^2}\right) \tag{5}$$

式中，$|c\text{-}p_i|$ 是任一单元格 c 和活动空间任一位置点 $p_i$ 之间的距离.

此外，由启发式方法(heuristic method)[19]给出运动轨迹平稳光滑参数 $r$，则 $r$ 的计算公式如下所示：

$$r=\frac{1}{2}n^{-\frac{1}{6}}(\sigma_x{}^2+\sigma_y{}^2)^{\frac{1}{2}} \tag{6}$$

式中，$\sigma_x$ 和 $\sigma_y$ 分别是整条位置序列 Loc 中所有位置点相应的 $x$ 和 $y$ 坐标的标准差.

在估算出单元格密度值后，移动对象较为密集的位置点通过一定的条件聚类聚集为一个热点区域，其中热点区域可以借鉴地图上的等值线定义的方法给出. 由于任意单元格的密度值是不完全相等的，甚至密度值差别很小，所以等值线可以定义为由相邻密度值相同或相近的单元格连接构成，其判断条件公式为：

$$|f(c_1)-f(c_2)|\leqslant\varepsilon \tag{7}$$

式中，本文选取的 $\varepsilon$ 值为 0.05，作为热点区域等值线的判断条件.

密度阈值 $\delta$ 的选择影响着发现的热点区域占活动空间的比例,该密度阈值可以用 Top-$k$ 排序思想的方法给出,即从所有单元格的所有密度值中选取第 $k$ 大的密度值作为密度阈值 $\delta$. 本文选取的 $k$ 值为 15,作为挖掘移动对象位置序列得出的热点区域的条件.

## 4　热点区域发现算法

### 4.1　算法设计

为了提高热点区域发现方法的性能,本文引入网格结构以解决空间数据聚类中的数据索引问题,主要优势是因为其快速的处理时间,该时间只与活动空间划分的单元格数目有关,与移动对象的规模无关. 在本文的聚类算法中,由于单元格的密度值会被频繁访问,并且单元格的数目较多,本文使用数组 A 存放所有的单元格密度值,用队列 $Q$ 结构存储满足大于等于密度阈值 $\delta$ 的单元格密度值.

算法 4.1 给出了热点区域发现算法的伪代码,分为三个步骤实现. 首先,该算法将移动对象的活动空间离散化成 $w\times h$ 大小的网格. 其中,活动空间是由记录移动对象运动轨迹的移动数据在空间上的坐标投影形成的位置空间;其次,引入核方法和启发式公式获取单元格密度值. 该算法需要对单元格执行遍历,利用二元正态核密度函数估算出每一个含有位置点且未被处理的单元格的密度值,并对遍历过的单元格进行标记. 特别地,对于每个包含位置点且未被处理的单元格,通过函数 CompCellValue(算法 4.2)获取单元格密度值,在函数 GetTop(算法 4.3)中使用 Top-$k$ 排序方法得到密度最大的 $k$ 条记录;最后,结合等值线的判断条件通过聚类将满足条件的单元格聚集成热点区域.

**算法 4.1**　热点区域发现算法

1. Discretize the space into a regular w×h grid
2. For each nonempty && unprocessed cell c do
3. den = compCellValue(c, $p_i$)　/* compute the cell density value */
4. δ = getTop(den,k)
5. HotCell = getTop(den, δ)　/* obtain the satisfied cells occupied by hot points */
6. HR=Clustering(HotCell,ε)
7. Return HR　/* the hot regions identified by contour line using top-k density value threshold. */

移动对象区域的密度反映人群在某一区域的聚集程度,而区域是由一系列单元格构成,可以通过分析位置序列中的位置点的分布来分析移动对象的聚集程度. 根据算法 4.1 的分析,在分析位置点的分布情况时,本文引入网格数据结构,使用网格索引作为空间数据聚类的数据索引方法,此时移动对象区域的密度问题就可以细化为计算单元格的密度值. 下面给出计算单元格密度值的算法,该算法对于任一单元格 $c$ 和任一位置点 $p_i$ 构成的 $(c,p_i)$ 点对,利用二元正态核密度函数估算出网格中任一单元格 $c$ 的密度,并将密度值存储在数组 $A$ 中. 另外,在算法中引入 $r$ 参数,可以根据上文中介绍的启发式公式计算得出.

**算法 4.2**　CompCellValue(c, p)　/* compute the cell density value */

1. Cell density array A　/* A is a double array */
2. For each pair(c,$p_i$) do　/* c is a cell and $p_i$ is a location point */
3. Using the heuristic method to obtain r
4. Using the kernel method to obtain f(c)
5. A. insert(f(c))

算法 4.2 给出了获取单元格密度值的方法,由该方法得到的存储在数组中的密度值,数据量较大. 这些数据并不都是本文所考虑的,此时就要将较小的密度值过滤掉,筛选出比较合适的值. 在热点和热点区域定义中,给出了热点区域的阈值下限,即密度阈值 δ,密度阈值 δ 的确定在算法 4.3 中涉及. 算法 4.3 用 Top-$k$ 排序的方法筛选出了第 $k$ 大单元格的密度值作为密度阈值 $\delta$,并将所有满足大于等于 $\delta$ 条件的单元格密度值添加到队列 $Q$ 中.

**算法 4.3**　GetTop(den,k)　/* obtain the satisfied cells */

1. Q=empty
2. For each den in A do
3. Select Top-k den from A　/* thekth den is δ */
4. Q. insert(den)

### 4.2　算法复杂度分析

对于热点区域发现算法的复杂度,根据前面介绍可知,该算法的复杂度和网格大小 $w\times h$、位置点数目 $n$ 有关,即为 $O(w\times h\times n)$.

## 5 实验

本文通过以下四部分来验证本文提出的发现方法的可行性和有效性. 第一,利用网格将人群的活动空间离散化;第二,利用核方法和启发式公式估算出单元格密度值,借鉴地图上的等值线方法发现人群活动热点区域;第三,分析不同密度阈值 $\delta$ 条件下,$k$ 值对热点区域发现算法的影响;最后,对算法的性能进行分析.

第一,在发现基于核函数的人群活动热点区域的过程中,具有划分作用的网格对于获取热点区域而言起到索引作用,所以为了获取人群活动的热点区域,本文首先把人群的活动空间进行离散化,将位置序列转换成点层图,根据前面网格结构划分设计的方法,将点层图划分成大小为 50×50 的网格,其中每一个单元格的面积统计约为 0.08 $km^2$. 本文使用的数据集来自北京市连续 5 年内的 182 个移动对象的移动数据,包括 18 670 条轨迹和 24 876 978 个位置点. 从数据集中取北京市四环以内 2012 年 11 月 10 日当天的 6 051 个数据点作为测试数据,结果如图 5 所示.

**图 5 网格化点层图**

**Fig. 5 Grid point layer**

第二,从上面的点层图中可以初步分析出,人群活动在某种程度上呈现出一定的聚集性和离散性,位置点比较聚集的区域反映出与之相对应的区域被频繁访问,相应地,某些离散的位置点反映出与之相应的区域很少被访问. 为了进一步分析区域被活动人群访问的频率,本文引入"热点区域"的概念,并用二次核密度函数挖掘出热点,借鉴地图中的等值线的方法给出人群活动的热点区域. 本文的算法的计算平台为 Intel(R) Core(TM) CPU 2.67 GHz,2 GB内存. 本文从数据集中取出 8 000 个位置点作为该算法的测试数据,其中,热点区域边界由借鉴地图的等值线方法给出,根据前面介绍的定义等值线的方法,图 6 中有 1 和 2 两个热点区域,如图 6 所示.

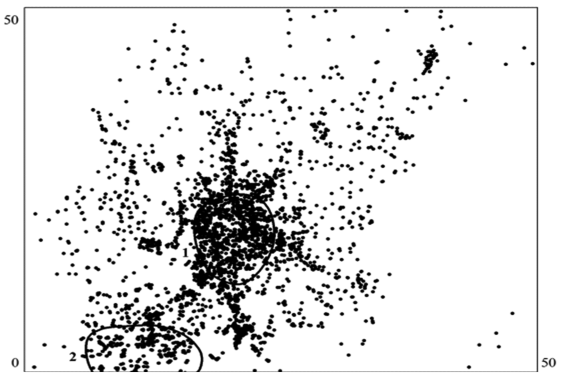


**图 6 结合地图发现的热点区域图**

**Fig. 6 Hot spots combined map discovering**

第三,分析不同 $k$ 值对热点区域发现算法的影响. 根据前面对热点区域发现方法分析的介绍,等值线上数据点表示附有某一密度阈值的单元格的密度值,并且数据点的密度值是相同的或接近的,其中,密度阈值 $\delta$ 的选取借鉴 Top-$k$ 的思想,即 $k$ 值大小影响着热点区域的范围,$k$ 值越大热点区域的范围就越大. 根据前面对算法复杂度的分析,该算法的复杂度和网格大小、数据规模有关,下面首先固定网格大小 50×50,从数据规模的角度分析 $k=5$,$k=15$,$k=25$对应的单元格密度值作为密度阈值时对人群活动热点区域发现算法时间开销的影响,如图 7 所示. 接着,固定数据规模 80 000,从网格大小分析 $k=5$,$k=15$,$k=25$ 对应的单元格密度值作为密度阈值时对人群活动热点区域发现算法时间开销的影响,如图 8 所示.

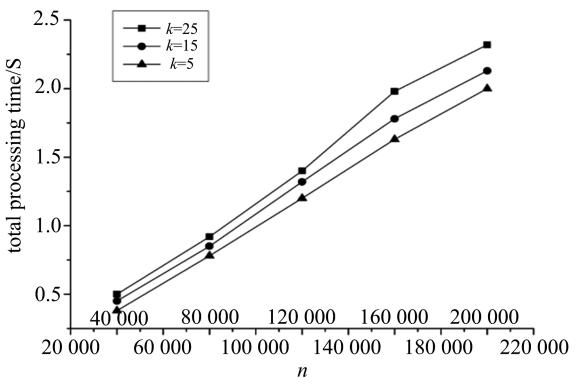


**图 7 固定网格大小不同数据规模下的算法时间开销**

**Fig. 7 Algorithm time expenses by fixed grid size in different data size**

分析图 7 可知,在加载运动轨迹和网格划分所需时间相同的情况下,随着 $k$ 值增大,发现热点区域

所需时间趋于上升. 这是因为 $k$ 值增大,热点区域的范围就会变大,算法排序时间呈上升趋势.

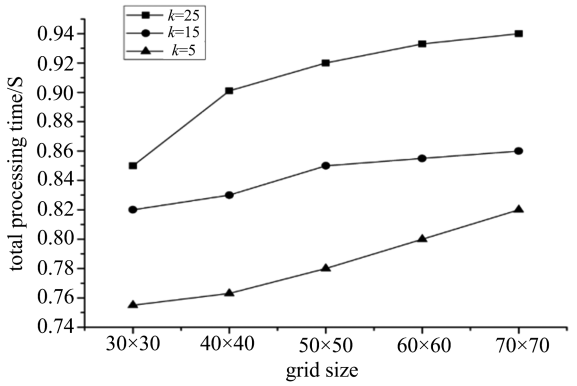


**图 8　固定数据规模不同网格大小下的算法时间开销图**

**Fig. 8　Algorithm time expenses by fixed data size in different grid size**

本文用 Top-$k$ 方法设定密度阈值并用核方法和启发式公式发现热点区域,为了进一步验证该方法的性能,与经典的基于密度的聚类算法 DBSCAN[12]进行比较. 固定网格大小 80×80 和密度阈值 0.8,分析在不同的移动对象数据规模下,两种聚类算法对发现的密集区域所用时间开销的比较,结果如图 9 所示.

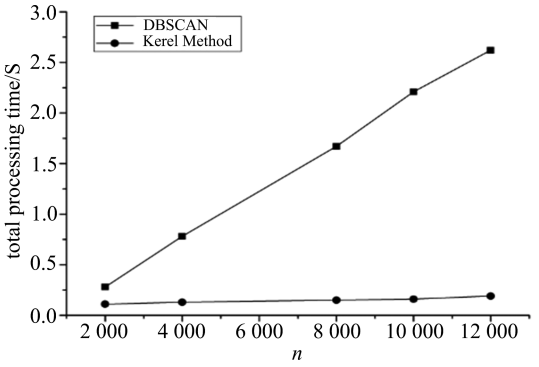


**图 9　两种方法性能比较图**

**Fig. 9　Performance comparison of two methods**

从图 9 可以看出 DBSCAN 和本文提出的 kernel method 这两种聚类算法在不同移动对象数据规模下的比较,本文提出的 kernel method 方法的性能优于 DBSCAN 聚类算法.

## 6　结论

本文通过对人群运动轨迹的空间信息特性进行分析,解决了移动人群位置序列的热点区域问题. 对于热点区域挖掘问题的解决,本文引入网格索引结构并通过设计二次核密度函数和启发式公式,解决了带有密度阈值的热点区域发现的问题. 本文的算法在真实的移动数据集上进行了大量实验,并且可以在短时间内完成对数百万的位置点的聚类,大量的实验结果表明了本文提出的算法的可行性、有效性. 在未来,我们会继续优化算法,实现对区域的实时监测,实现实时、动态监控区域的人群流量变化.

**参考文献(References)**

[1] Smith G, Wieser R, Goulding J, et al. A refined limit on the predictability of human mobility[C]// IEEE International Conference on Pervasive Computing and Communications. Budapest, Hungary: IEEE Press, 2014: 88-94.

[2] Lin M, Hsu W J, Lee Z Q. Predictability of individuals' mobility with high-resolution positioning data[C]// Proceedings of the ACM Conference on Ubiquitous Computing. London: ACM Press, 2012: 381-390.

[3] Qiao S J, Shen D Y, Wang X T, et al. A self-adaptive parameter selection trajectory prediction approach via hidden Markov models[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(1): 284-296.

[4] Qiao S J, Han N, Zhu W, et al. TraPlan: An effective three-in-one trajectory-prediction model in transportation networks[J]. IEEE Transactions on Intelligent Transportation Systems, 2014, 16(3): 1188-1198.

[5] Houenou A, Bonnifait P, Cherfaoui V, et al. Vehicle trajectory prediction based on motion model and maneuver recognition[C]// IEEE/RSJ International Conference on Intelligent Robots and Systems. Tokyo, Japan: IEEE Press, 2013: 4363-4369.

[6] Yuan J, Zheng Y, Xie X. Discovering regions of different functions in a city using human mobility and POIs[C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China: ACM Press, 2012: 186-194.

[7] Giannotti F, Nanni M, Pedreschi D, et al. Trajectory pattern mining[C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, USA: ACM Press, 2007: 330-339.

[8] 刘奎恩, 肖俊超, 丁治明, 等. 轨迹数据库中热门区域的发现[J]. 软件学报, 2013, 24(8): 1816-1835.
Liu K E, Xiao J C, Ding Z M, et al. Discovery of hot region in trajectory databases[J]. Journal of Software, 2013, 24(8): 1816-1835.

(下转第 863 页)