ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 33,No. 3,May 2018,pp. 487－495
DOI:10.16337/j.1004-9037.2018.03.012

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于时空轨迹的移动对象汇聚模式挖掘算法

张逸凡　赵　斌　孙鸿艳　谈　超　吉根林

(南京师范大学计算机科学与技术学院，南京，210023)

摘　要：移动对象的聚集模式是时空轨迹模式挖掘中的重要课题，它研究移动对象群体在多个连续时刻中的空间聚集问题。现有的聚集模式基于共现模式进行定义，挖掘结果中夹杂大量非运动的聚集群体，严重影响模式挖掘的效果。为了解决此问题，本文提出了基于群体运动过程建模的汇聚模式。该模式定义从群体运动形态出发进行设计，准确识别向心运动的移动群体，有效排除非聚集类型运动群体的干扰。本文设计并实现了汇聚模式挖掘(Converging pattern mining，CPM)算法，该算法首先定位密度峰值点，确定候选的汇聚中心区域，然后依次识别每个时刻的汇聚群体，按照群体汇聚的持续性要求识别汇聚模式。基于真实轨迹数据进行实验，结果验证了本文提出的CPM算法在挖掘效果和算法效率的有效性。

关键词：轨迹数据挖掘；汇聚模式；聚集模式

中图分类号：TP391　　**文献标志码**：A

Algorithm for Mining Converging Patterns of Moving Objects from Spatiotemporal Trajectories

Zhang Yifan, Zhao Bin, Sun Hongyan, Tan Chao, Ji Genlin

(School of Computer Science and Technology, Nanjing Normal University, Nanjing, 210023, China)

Abstract: Gathering pattern is an important research topic in the field of trajectory pattern mining. It focus on collective gathering problem on consecutive time period. Traditional models of gathering patterns are based on co-concurrence patterns. Mining methods based on such models generate a lot of stationary gathering groups. In order to deal with such problems, we propose a converging pattern based on modelling of group moving objects, which accurately identifies gathering group instead of other types of moving group. A moving objects converging pattern mining (CPM) algorithm is presented and implemented. First, the algorithm locates all high density peak points and converges central zones. Second, the algorithm identifies converging groups on consecutive timestamps, and then detects converging patterns according to the durability of group patterns. Experimental results show the effectiveness and efficiency of the algorithm.

Key words: trajectory data mining; converging pattern; gathering pattern

---

基金项目：国家自然科学基金(41471371)资助项目。
收稿日期：2016-09-13；**修订日期**：2016-10-12

引 言

卫星定位和移动互联技术的日趋成熟催生了海量的时空轨迹数据。它们真实记录了移动对象的运动行为特征,包括位置、时间、方向和速度等属性。采集并分析移动对象群体(简称群体)的轨迹数据,可以有效地揭示群体运动的行为规律和常见模式。所以时空轨迹的群体模式挖掘是一个具有理论意义和应用价值的研究课题。通常,时空轨迹模式分为共现模式、伴随模式、聚集模式、频繁模式和异常模式等[1,2]。本文主要研究时空轨迹的聚集模式挖掘问题。2005 年,由 Laube 等首次提出了定义移动对象运动模式的 REMO 模型[3],其中定义了基于时空维度的聚集模式,即 Aggregation 模式。该模式规定在单个时间片中存在足够多的面向同一圆形区域运动的移动对象,且运动的方向向量都与该圆形区域相交,最终通过计算各方向向量延长线的交点来识别聚集模式。2013 年,郑凯等提出了聚集模式 Gathering[4,5],并总结了聚集模式的 5 种特征,分别是规模性、密集性、持久性、静态性、专一性。Gathering 模式由至少个连续时刻的密集快照构成,同时要求至少个参与者(Participator)必须在不低于个密集快照(可以不连续)中出现。Gathering 模式的优点是放松了参与者必须在连续密集快照中出现的要求,在保证足够参与性的同时也兼顾了每个时刻的对象密集性,这更适合实际的应用场景。2015 年,郑宇等研究基于 STG 图 (Spatio-temporal graph)的聚集模式。他们通过分析人群的移动行为发现城市中的黑洞模式[6]。该研究与众不同之处在于研究拓扑结构上的聚集模式问题。时空轨迹的伴随模式研究在建模方法上与聚集模式的比较接近。伴随模式是指群体在时间维度上连续出现,在空间维度上密集存在。典型的研究工作主要有 Flock[7-9],Moving cluster[10],Convoy[11,12],Swarm[13] 以及 Traveling companions[14] 等伴随模式。Flock 模式是指群体在连续的 k 个时刻中都在空间中维持圆形的聚集形状,并且这些时刻的聚集共享一定数量的移动对象。Convoy 模式放松了群体的聚集形状必须为圆形的要求,采用了密度相连的方式识别伴随运动的群体。Swarm 模式考虑到部分移动对象在某些时刻可能临时离开,然后又重回聚集群体。因而,在时间维度上放松为 k 个非连续的时刻要求。而 Traveling companions 模式则在处理方式采用在线处理的方式从时空轨迹流中挖掘伴随模式。不难发现,早期伴随模式的研究方法被借鉴到后续聚集模式研究中。例如,聚集模式的定义依然沿用了伴随模式定义的基本思路。经过分析可以发现,虽然聚集模式和伴随模式在群体形状、密度以及时间连续性方面各有不同,但是它们都是基于多时刻的共现模式所构成。这种建模方法的优点是可以从空间位置关系上准确识别“在一起”的群体,从而满足模式中“聚集存在”的需要,但是缺点也十分明显。由于缺乏群体在运动形态上的判断,因而无法避免“停车场”问题的出现。“停车场”问题是指聚集在一起的群体在连续时刻内不发生空间位置的改变,如同停车场一样。虽然处于“停车场”中的群体在连续时刻内都在空间邻域中聚集存在,但显然与聚集模式和伴随模式的基本思想不符合。

为了解决“停车场”问题对聚集模式挖掘的干扰,本文提出了兼顾群体运动方向的聚集模式挖掘问题,也就是对群体运动过程进行建模,识别持续朝向中心区域聚集的群体汇聚模式,简称汇聚模式。与传统的聚集模式相比,汇聚模式挖掘中增加了对群体运动形态的甄别,提高了识别群体聚集行为的准确性,但也为算法设计增加了难度,主要表现在如下两方面:(1)汇聚模式的定义。已有的聚集模式和伴随模式都是基于群体的“共现”思想来设计并定义的。但是在汇聚模式中只有群体聚集到中心区域时才会表现出共现模式,在此之前无法采用共现模式的挖掘算法跟踪并识别移动群体。所以,本文提出的汇聚模式无法完全采用共现模式进行定义。(2)群体运动形态具有复杂性。在现实应用场景中,群体的运动形态具有多样性。例如,有向心运动的群体,有随机运动的群体,也有静止不动的群体。多种类型群体在空间维度上相互重合交织,这为汇聚模式识别增加了难度。

为了应对上述挑战,本文提出的汇聚模式针对移动群体的聚集过程进行建模。在模式定义中引入运动方向,从群体运动形态出发进行设计。这样可以有效识别向心运动的群体,而不受其他运动类型移

动对象的干扰。基于此思路,本文设计并实现了汇聚模式的挖掘算法。该算法从汇聚模式的汇聚中心点出发,首先使用基于密度的聚类算法定位高密度点,并以此作为候选汇聚的中心点,然后根据候选汇聚中心点邻域中的移动群体的运动形态识别向心汇聚模式。最后在连续时刻上,挖掘移动群体的汇聚模式。

1 问题描述

分析现实中的汇聚运动行为可以总结出如下 3 个特性:(1)规模性。移动群体在数量上应该具有规模性;(2)持续性。在时间维度上,群体汇聚行为应该持续一段时间,形成稳定的群体运动形态;(3)方向性。在空间维度上,移动群体从不同方向朝向同一中心区域汇聚。

根据以上特性,本文对移动群体的汇聚运动行为进行形式化。设移动对象集合 $O_{DB}=\{o_1,\cdots,o_n\}$,时间区间 $T=\langle t_1,\cdots,t_m\rangle$,其中移动对象 o 的轨迹定义为 $o.traj=\langle(p_{t_1},t_1),\cdots,(p_{t_m},t_m)\rangle$,$p_{t_i}=(x_{t_i},y_{t_i})\in\mathbf{R}^2$,$t_i\in T$,$o.p_{t_i}$ 为 o 在 t_i 时刻的空间位置。在 t_i 时刻所有对象的位置点集合定义为 $S_{t_i}=\{o.p_{t_i}\mid o\in O_{DB}\}$。

1.1 相关定义

定义 1(邻域) 给定距离阈值 ε 和点集 S,点 p 的 ε- 邻域定义为 $N_\varepsilon(p)=\{q\in S\mid D(p,q)\leqslant\varepsilon\}$,其中 $D(\cdot)$ 表示两点间的欧氏距离。

根据邻域概念引申出另外两个定义:(1)汇聚模式中心点 p_a 的 ε- 邻域,记作 $N_\varepsilon(p_a)$;(2)以汇聚点 p_a 为中心,以半径 r 为距离阈值的邻域,记作 $N_r(p_a)$。$N_r(p_a)$ 是群体汇聚后的停留区域,而 $N_r(p_a)$ 是指移动对象从不同方向朝向汇聚点 p_a 运动的区域,如图 1 所示。

定义 2(邻域快照) 给定移动对象集合 O_{DB} 和距离阈值 ε,点 p 在 t 时刻的 ε- 邻域定义为 $N_\varepsilon(p,t)=\{q\in S_t\mid D(p,q)\leqslant\varepsilon\}$。

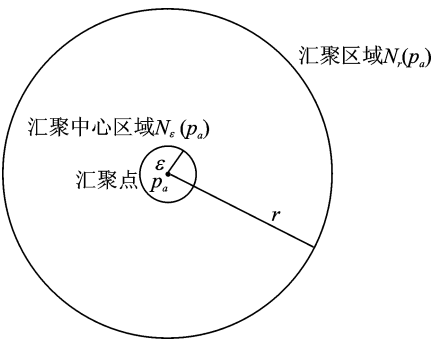


图 1 汇聚模式中的 P 点邻域

Fig. 1 Neighborhood of node P of converging pattern

定义 3(方向区域) 给定极坐标系 O_x,极轴与地理坐标的正东方向同向。将以极点为中心、以 r 为半径的圆形划分成夹角相等的 n 个方向区域,第 i 个方向区域定义为 $D_i=\left\{(\rho,\theta)\mid 0\leqslant\rho\leqslant r,\frac{(2i-3)\pi}{n}\leqslant\theta<\frac{(2i-1)\pi}{n}\right\}$,$i\in[1,\cdots,n]$。通常每个区间对应于地理空间上的一个方向,若设 $n=4$,分别对应于东、南、西和北 4 个方向,如图 2 所示。

定义 4(同向点集) 给定移动对象集合 O_{DB} 和方向区域 D,t 时刻方向区域 D 中的同向点集定义 $C_D(S_t)=\{(x,y)\in S_t\mid(\sqrt{x^2+y^2},\arctan(y/x))\in D\}$

定义 5(汇聚群体) 给定移动对象集合 O_{DB} 和时间阈值 k_1,在 t_e 时刻以 p_a 为汇聚点参与汇聚的移动对象集合记作 A_{t_e},必须满足以下要求:

(1) $A_{t_e}\subseteq O_{DB}$;

(2)设 $[t_{e-k_1},t_e]\subseteq T$,$\forall o\in A_{t_e}$,满足 $o.p_{t_e}\in N_\in(p_a,t_e)$,且 $o.p_{t_{e-k_1}}\notin N_r(p_a,t_{e-k_1})$。

定义 6(向心汇聚群体) 给定移动对象集合 O_{DB},方向区域同向点集的阈值 s,时间阈值 k_1,在 t_e 时刻以 p_a 为汇聚点参与向心汇聚的移动对象集合记作记作 A'_{t_e},必须满足 $\forall\ t_i\in[t_{e-k_1+1},t_{e-1}]$,$|C_{D_i}(S'_{t_i})|>s$,其中,$S'_{t_i}=\{o.p_{t_i}\mid o\in A_{t_e}-A_{t_{e-1}}\}$。该定义反映了汇聚模式的方向性和规模性。

定义 7(汇聚模式) 给定移动对象集合 O_{DB},时间阈值 k_2,以 p_a 为汇聚点的汇聚模式定义为 $G=$

$\langle A'_{t_a}, A'_{t_{a+1}}, \cdots, A'_{t_b} \rangle$，必须满足 $b-a+1 > k_2$。该定义反映了汇聚模式在时间维度上的持续性。

1.2 问题定义

给定移动对象集合 O_{DB}，汇聚区域半径 r，方向区域同向点集的阈值 s，时间阈值 k_1 和 k_2，时空轨迹汇聚模式挖掘是在时间区间 T 范围内发现所有的汇聚模式 G。

1.3 总体框架

本文提出的移动对象汇聚模式挖掘的总体框架主要包括以下 4 个阶段：

(1)数据预处理。真实的 GPS 轨迹长度往往不等，采样率各不相同。预处理阶段的主要任务是采用线性插值的方法将不等长的轨迹基于相同的时间序列 T 进行“对齐”。

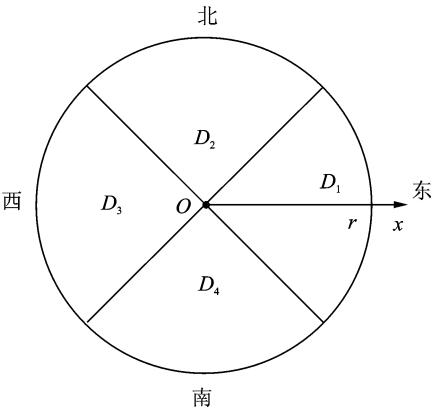


图 2 汇聚模式方向区域

Fig. 2 Direction regions of converging pattern

(2)定位候选的汇聚中心区域。汇聚模式的中心区域往往具有高密度的特点，因而通过识别密度峰值区域可以确定候选的汇聚中心区域。

(3)识别汇聚模式。在阶段 2 的基础上，分析汇聚中心区域中的移动群体在历史时间区间中的运动形态，进而识别汇聚模式及其中心区域。

(4)展示汇聚模式。通过可视化技术展示时空轨迹的汇聚模式，包括汇聚模式的移动对象、汇聚区域及其中心和汇聚模式的生命周期。

2 算法设计

基于汇聚模式定义，本文提出了汇聚模式挖掘算法 CPM。该算法包含定位密度峰值点、识别单时刻汇聚群体和挖掘汇聚模式 3 个阶段。由于算法在每个时刻都要进行多次区域搜索，因此对每个时刻建立 R 树索引。不作说明，下文算法中涉及到区域搜索的步骤，均使用 R 树索引提升效率。

2.1 定位密度峰值点

通常，密度峰值点具有以下两个特点：(1)密度峰值点的密度大于其邻域中其他点的密度；(2)密度峰值点与密度更高的点之间的距离相对较大。根据这两个特点，Rodriguez 等提出了基于密度峰值的聚类算法[15]。

对于点 p_i，分别计算其密度 ρ_i 和距离 δ_i，即

$$\rho_i = \sum_j f(d_{ij} - d_c) \tag{1}$$

$$\delta_i = \min_{j:\rho_j > \rho_i}(d_{ij}) \tag{2}$$

式中：d_{ij} 表示 p_i 与 p_j 的距离；d_c 为搜索半径，当 $d_{ij}-d_c<0$ 时，$f(d_{ij}-d_c)=1$，否则 $f(d_{ij}-d_c)=0$。δ_i 表示 p_i 距离密度更大的点的最小距离。特别地，当 ρ_i 为最大时，$\delta_i = \max_j(d_{ij})$。

如果 ρ_i 大于密度阈值 ρ，δ_i 大于距离阈值 δ，则 p_i 为密度峰值点。算法 1 简要描述了这一过程。首先计算每个点的密度(行 1～5)，然后计算每个点与密度更高点的距离(行 6～10)，最后判断每个点是否满足阈值要求(行 11～13)。算法时间复杂度为 $O(n^2)$，使用 R 树索引后，时间复杂度为 $O(n \cdot \log n)$。

算法 1 密度峰值点查找算法(Density peak query, DPQ)

输入：移动对象点集 S，搜索半径 d_c，密度阈值 ρ，距离阈值 δ

输出：密度峰值点集合 P

1. for each point $p_i \in S$

2. $\rho_i \leftarrow 0$;//密度初始化为 0
3. for each point $p_j \in S$
4. if($d_{ij} < d_c$)
5. $\rho_i \leftarrow \rho_i + 1$;//密度更新
6. for each point $p_i \in S$
7. $\delta_i \leftarrow \max(d_{ij})$;
8. for each point $p_j \in S$
9. if($\rho_j > \rho_i$)
10. $\delta_i \leftarrow \min(\delta_i, d_{ij})$;//计算 p_i 距离密度更大的点的最小距离
11. for each point $p_i \in S$
12. if($\rho_i > \rho$ and $\delta_i > \delta$)
13. $P \leftarrow P \cup p$;//输出满足阈值 ρ,δ 的点
14. return P

2.2 挖掘单时刻向心汇聚群体

向心汇聚的汇聚中心通常是密度峰值点,因此可以将密度峰值点作为候选汇聚中心点,然后识别该点邻域中的移动对象运动形态是否满足向心汇聚群体要求。如果满足要求,则构成候选汇聚模式。由于汇聚具有渐变性,因而可以将候选汇聚模式所在位置作为候选汇聚中心点,从而减少密度峰值点的计算。根据定义 7 可以看出,汇聚模式满足向下闭合属性。即,给定汇聚模式 $G=\langle A'_{t_s}, A'_{t_{s+1}}, \cdots, A'_{t_e}\rangle$,如果$\exists A'_{t_{s-1}}$,使得 $A'_{t_{s-1}}$ 附加到 G 会产生一个新的汇聚模式,则 G 为闭合汇聚模式。挖掘闭合汇聚模式可以避免冗余结果产生,这不仅减少了不必要的结果输出,还减轻了计算量。

算法 1 简要描述了单时刻向心汇聚群体的挖掘过程。这里,O' 表示候选汇聚中心邻域中的移动对象集合。对于时刻 t_e,首先找到候选汇聚模式对应的汇聚中心以外的点集(行 2~4),然后计算这些点集的密度峰值点,并将其作为候选向心汇聚中心(行 5),最后根据向心汇聚群体的定义判断每个候选向心汇聚在该时刻是否满足定义要求(行 6~18)。如果满足要求,则更新其持续时间(行 13)。如果不满足要求且该候选向心汇聚的持续时间满足阈值 k_2,则输出为闭合汇集模式(行 16)。

算法 2 单时刻汇聚群体挖掘算法(Converging group mining,CGM)

输入:移动对象集合 O_{DB},候选汇聚模式 C,半径阈值 ε,γ, 同向点集的阈值 s,时刻 t_e,时间阈值 k_1,k_2

输出:当前时刻汇聚模式 R

1. $C' \leftarrow \Phi, O' \leftarrow \Phi, R \leftarrow \Phi$;
2. for each c $\in C$
3. $O' \leftarrow O' \cup \{o \mid o \in N_\varepsilon(c._{\text{center}})\}$;
4. $S_{t_e} \leftarrow \{o.p_{t_e} \mid o \in (O_{DB} - O')\}$;
5. C $\leftarrow C \cup \{c \mid c_{<p_s>} \in \text{DPQ}(S_{t_e})$;//计算当前密度峰值点,并将其作为候选向心汇聚中心
6. for each $c \in C$ do
7. if($c.$ is Converge = false)
8. continue;
9. flag $\leftarrow$ true;
10. for each $t_i \in [t_{e-k_1+1}, t_{e-1}]$ do
11. if ($C_{D_i}(S'_{t_i}) < s$) then
12. flag $\leftarrow$ false;

13. break;
14. if(flag = true) then//如果每个分区的点集规模满足阈值 s,则为汇聚
15. c.updateTime;//更新候选汇聚模式时间
16. $C' \leftarrow C' \cup c$;
17. else if(c.lifetime $\geq k_2$)
18. $R \leftarrow R \cup c$;
19. else
20. C.remove(c) ;//从候选汇聚模式中删除闭合汇集模式
21. return R ;

2.3 挖掘连续时刻汇聚模式

向心汇聚反映了移动对象从四周向中心汇聚的运动形态,是一个汇聚的过程。如果只是单时刻或者持续时间很短的向心汇聚则不能形成一个大规模群体事件,这样的汇聚模式也是没有意义的。而当汇聚模式持续了较长的一段时间,汇聚中心的规模会越来越大,直到达到一个峰值后,再趋于稳定,然后又慢慢消失,这个过程恰恰反映了一个大规模群体事件从产生到消失的过程。根据汇聚模式的时间连续性,算法按时间推进,使用单时刻汇聚模式挖掘算法 CGM 计算每个时刻的闭合汇聚模式,同时计算当前时刻的候选汇聚模式并更新候选汇聚模式集合。

算法 3 简要描述了这一过程。R 表示汇聚模式集合,C 表示当前候选汇聚模式集合。算法迭代地调用单时刻汇聚模式挖掘算法 CGM,并更新当前候选集合 C(行 2~4),最后输出所有汇聚模式 R。

算法 3 汇聚模式挖掘算法(Converging patterns mining,CPM)

输入:移动对象集合 O_{DB} ,半径阈值 $\in$、γ ,同向点集的阈值 s ,时间阈值 k_1 ,k_2 ,时间域 T

输出:汇聚模式集合 R

1. $R \leftarrow \Phi, C \leftarrow \Phi$;
2. for each $t_e \in T$ (in ascending order)
3. $R \leftarrow R \cup$ CGM(O_{DB} ,C, $\in$,γ ,s ,t_e ,k_1 ,k_2) ;
4. C. update;
5. return R;

3 实验与分析

3.1 实验设置

为了验证汇聚模式及其算法的有效性和高效性,本文采用真实的 GPS 轨迹数据(http://www.ccf.org.cn/sites/ccf/dashuju.jsp? contentId=2756825351305 # 大赛赛题)进行实验。该数据集是北京市 12 408 辆出租车在 2012 年 11 月的 GPS 数据。本文实验选取了其中 11 月 1 日全天的 GPS 数据,共 12 408 条轨迹,大小为 1.8 GB。经过轨迹预处理后,出租车轨迹的时间区间以分钟为单位,共 1440(24×60)个时刻。本文实验所有程序采用 Java 语言编写,运行在 CentOS 6.4 操作系统上,硬件平台的配置情况为 2 个 6 核 Intel(R) Xeon(R) CPU,主频为 2.40 GHz,内存为 32 GB。

3.2 实验评价

与本文研究最接近的工作是文献[4]中提出的聚集模式及其算法,该算法是目前在聚集模式挖掘方面最新的研究成果,因而以此作为基准测试方法。实验将从模式定义有效性和挖掘算法效率两方面对汇聚模式和聚集模式进行比较。在模式定义有效性方面,通过人工检查的方式对模式挖掘的结果进行分类,分析有效模式所占比例,比例越高说明模式挖掘效果越好。这里的有效模式是指移动群体的运动

中心区域在地理空间中与(Point of interest,POI)重合的模式。而算法效率的比较相对简单,主要考虑不同数据量下的算法运行时间,时间越短则效率越高。通过前文的分析可以发现,两种模式最大的区别在于模式定义中是否考虑移动群体的运动方向。本文实验将证明考虑方向性的汇聚模式比聚集模式在挖掘效果和算法效率两方面表现得更出色。

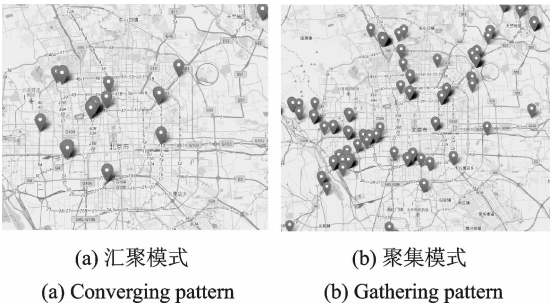
3.3 挖掘算法有效性

在分析两种模式的有效性之前,先介绍实验参数的设置情况。汇聚模式的参数设置为:汇聚中心区域定位算法的距离阈值 $d_c=1\ 000$ m,密度阈值 $\rho=20$,峰值点间的距离阈值 $\delta=1\ 500$ m,半径阈值 $\varepsilon=800$ m,$r=5\ 000$ m,方向区域同向点集阈值 $s=5$,时间阈值 $k_1=20$,$k_2=5$。

聚集模式的参数设置为:DBSCAN 数量阈值 $m=5$,半径阈值 $\varepsilon=200$ m,聚集快照中移动对象的个数阈值 $m_c=15$,相邻簇的豪斯多夫距离阈值 $\delta=300$ m,群体生命期阈值 $k_c=10$,参与者生命期阈值 $k_p=5$,参与者个数阈值 $m_p=10$。

首先通过可视化技术在地图上展示两种模式的挖掘结果,即地理空间中的分布情况。可以明显发现聚集模式在数量上多于汇聚模式,两者空间分布差异性较大。汇聚模式主要集中在北京市三环以内人流量较多的地方;而聚集模式分布较为广泛,不局限在城市的中心区域,如图 3 所示。

为了能够量化评估模式挖掘的有效性,结合群体运动的地理位置对挖掘结果进行分类,划分成 3 种类型:路口聚集、停车场聚集和 POI 聚集。其中,路口聚集是指移动群体在道路路口的停留;停车场聚集表明移动群体在某个固定区域长时间滞留,移动群体本身变动较少;POI 聚集是指地理空间中人们感兴趣的位置,如体育场、交通枢纽和娱乐场所等。不难发现,只有 POI 聚集可以体现模式挖掘的有效性。



(a) 汇聚模式 (b) 聚集模式

(a) Converging pattern (b) Gathering pattern

图 3 模式挖掘结果可视化

Fig. 3 Visualization of mining pattern

按照上述分类,作者将 2016 年 11 月 1 日一天中两种模式挖掘算法的结果统计在表 1 中。可以发现,聚集模式发现了 326 个结果,远远多于汇聚模式的 40 个结果。但是 73.3%的结果都属于停车场聚集。经过分析发现,大量出租车停留在交通枢纽的停车场,或者下午集中停留在出租车停靠点,这些位置大多是路边。而 POI 聚集一共 28 个,占 8.6%。这样的结果说明聚集模式的挖掘质量不高,有效的聚集比例较低。另一方面,汇聚模式中 POI 聚集共 33 个,占 82.5%,在绝对数量上超过聚集模式,并且汇聚模式挖掘算法的质量较高,可以发现更多有意义的模式。

表 1 2012 年 11 月 1 日北京市的汇聚模式和聚集模式挖掘情况

Tab. 1 Details of converging and gathering on November 1, 2012 in Beijing

模式类型	数量	路口聚集/%	停车场聚集/%	POI 聚集/%
聚集模式	326	18.10	73.30	8.60
汇聚模式	40	17.50	0	82.50

针对同一份轨迹数据,两种模式挖掘结果中存在 4 个相同的 POI 聚集结果。按照发生地点、时间和规模的不同展现在图 4 中。矩形框的粗细代表群体模式的规模大小,长短表示群体模式持续的时间。通过分析可以发现以下结果:(1)这 4 个相同的 POI 聚集主要发生在交通枢纽;(2)汇聚模式出现的时刻主要为早晚高峰,也就是交通枢纽最繁忙的时间段,但是有一半的聚集模式发生的时间段在凌晨,主要是交通枢纽的停车区域;(3)从群体模式规模上来看,汇聚模式普遍大于聚集模式。

之所以有这样的差异,和两种模式的定义有关。聚集模式主要考虑各个时刻移动群体的聚集状态,

而不考虑聚集的运动过程及方向，这就出现了聚集模式无法解决的“停车场”问题。移动群体在早晚高峰期的交通路口减速慢行或者等待红绿灯时，被识别成一个密集区域。按照聚集模式定义，持续一段时间这样的群体聚集行为被识别为聚集模式。而汇聚模式关注于聚集形成的运动过程，很自然地类似“停车场”这样的模式都被过滤掉，所以最终的挖掘结果质量比较高。

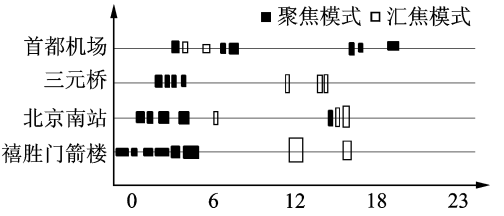


图 4 汇聚模式和聚集模式挖掘结果的交集

Fig. 4 Intersection of converging and gathering

3.4 算法效率

本节将比较两种模式挖掘算法的执行效率。分别比较 1 000 条、2 000 条、3 000 条、4 000 条和 5 000 条轨迹数据的测试结果。如图 5 所示，同样数据量情况下汇聚模式挖掘算法 CPM 比聚集模式挖掘算法 TAD[4] 效率更高，并且十分明显。因而，在算法整体运行时间上 CPM 算法优势明显。

为了进一步发现两种挖掘算法性能差异的原因，本文实验将挖掘算法分为两个阶段，分别是预处理阶段和模式挖掘阶段。R 树索引仅在两种模式的挖掘阶段被使用。如图 6 所示，首先可以发现随着数据量的增加两种算法各阶段的执行时间均随之增加。在两种模式的 4 个阶段中，聚集模式的预处理阶段最耗时，这是由于该阶段需要进行 DBSCAN 聚类计算，该计算的时间复杂度较高(具有索引支撑的 DBSCAN 时间复杂度为 $O(n\log n)$，否则为 $O(n^2)$)。其次，发现聚集模式的 TAD 算法对于数据集的增加比较敏感。这主要是 TAD 算法中包含了较为耗时的豪斯多夫距离计算，时间复杂度为 $O(mn)$。所以，在聚集模式挖掘中采用了聚类算法和豪斯多夫距离计算两种耗时的计算步骤，这是影响挖掘算法效率的主要因素。

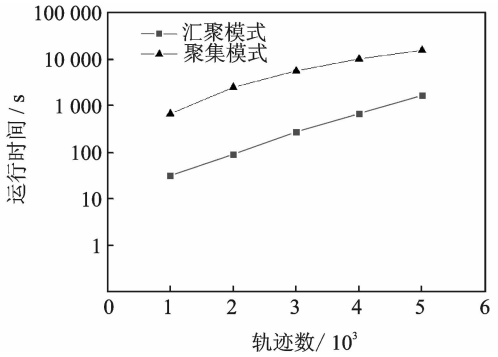


图 5 不同数据量的算法效率比较

Fig. 5 Performance comparison of algorithms under data size

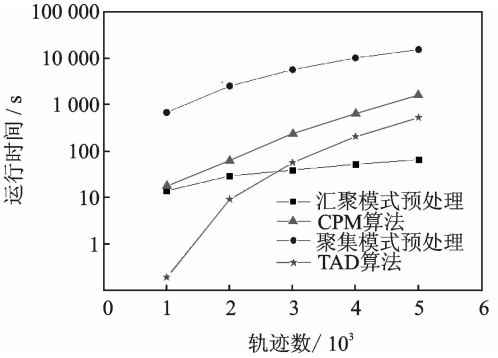


图 6 不同数据量的算法各阶段效率比较

Fig. 6 Performance comparison of stages under data size

4 结束语

本文提出了一种基于群体运动过程建模的时空轨迹汇聚模式。该模式定义可以有效地解决现有聚集模式和伴随模式挖掘中无法避免的“停车场”问题。基于此模式定义设计并实现了汇聚模式挖掘算法 CPM。该算法首先使用算法 DPQ 定位密度峰值点，并以此作为候选的汇聚中心区域，然后使用算法 CGM 识别单一时刻的向心汇聚群体。最后将 CGM 算法应用到连续时间片上，挖掘出所有满足规模性和持续性要求的汇聚模式。为了验证本文提出的汇聚模式及其算法的优越性，本文以真实的轨迹数据进行实验，实验表明本文提出的汇聚模式及其算法在挖掘效果和算法效率两方面都明显优于现有的聚集模式的挖掘方法。

参考文献:

[1] Zheng Y. Trajectory data mining: An overview[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2015, 6(3): 1-29,41.

[2] 吉根林, 赵斌. 时空轨迹大数据模式挖掘研究进展[J]. 数据采集与处理, 2015, 30(1): 47-58.
Ji Genlin, Zhao Bin. Research progress in pattern mining for big spatio-temporal trajectories[J]. Journal of Data Acquision and Processing,2015,30(1):47-58.

[3] Laube P, Marc V K, Stephan I. Finding REMO-detecting relative motion patterns in geospatial lifelines [M]. Berlin: Springer, 2005: 201-215.

[4] Zheng K, Zheng Y, Yuan N J, et al. On discovery of gathering patterns from trajectories[C]// IEEE 29th International Conference on Data Engineering. Brisbane, Australia:[s. n], 2013: 242-253.

[5] Zheng K, Zheng Y, Yuan N J, et al. Online discovery of gathering patterns over trajectories[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 8(26): 1974-1988.

[6] Hong L, Zheng Y, Yung D, et al. Detecting urban black holes based on human mobility data[C]//Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems. Bellevue, WA, USA:[s. n], 2015, 35:1-10.

[7] Benkert M, Gudmundsson J, Hübner F, et al. Reporting flock patterns[C]//European Symposium on Algorithms. Zurich, Switzerland:[s. n], 2006: 660-671.

[8] Vieira M R, Bakalov P, Tsotras V J. On-line discovery of flock patterns in spatio-temporal data[C]//Proceedings of the 17th ACM SIGSPATIAL International Conference On Advances In Geographic Information Systems. Seattle, Washington, USA: [s. n], 2009: 286-295.

[9] Gudmundsson J, Kreveld M J. Computing longest duration flocks in trajectory data[C]//Proceedings of the 14th Annual ACM International Symposium on Advances In Geographic Information Systems. Arlington, Virginia, USA:[s. n], 2006: 35-42.

[10] Kalnis P, Mamoulis N, Bakiras S. On discovering moving clusters in spatio-temporal data[C]//International Symposium on Spatial and Temporal Databases. Angra dos Reis, Brazil:[s. n], 2005: 364-381.

[11] Jeung H, Yiu M L, Zhou X, et al. Discovery of convoys in trajectory databases[J]. Proceedings of the VLDB Endowment, 2008, 1(1): 1068-1080.

[12] Jeung H, Shen H T, Zhou X. Convoy queries in spatio-temporal databases[C]//2008 IEEE 24th International Conference on Data Engineering. Cancun, Mexico:[s. n], 2008: 1457-1459.

[13] Li Z H, Ding B L, Han J W, et al. Swarm: Mining relaxed temporal moving object clusters[J]. Proceedings of the VLDB Endowment, 2010, 3(1/2): 723-734.

[14] Tang L A, Zheng Y, Yuan J, et al. On discovery of traveling companions from streaming trajectories[C]//2012 IEEE 28th International Conference on Data Engineering. Washington D C, USA:IEEE, 2012: 186-197.

[15] Rodriguez A, Laio A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.

作者简介:

张逸凡(1990-),男,硕士研究生,研究方向:轨迹数据挖掘,E-mail:zyfnnu@163.com。

赵斌(1978-),男,副教授,研究方向:数据挖掘、数据库及其应用,E-mail:zhaobin@njnu.edu.cn。

孙鸿艳(1992-),女,硕士研究生,研究方向:轨迹数据挖掘,E-mail:hysnnu@163.com。

谈超(1983-),女,讲师,研究方向:机器学习及其应用,E-mail:73022@nju.edu.cn。

吉根林(1964-),通信作者,男,教授,研究方向:数据挖掘及其应用,E-mail:glji@njnu.edu.cn。

(编辑:刘彦东)