基于轨迹特征及动态邻近性的轨迹匿名方法研究

王 超 [1] 杨 静 [1] 张健沛 [1]

摘 要 移动社会网络的兴起以及移动智能终端的发展产生了大量的时空轨迹数据, 发布并分析这样的时空数据有助于改善智能交通, 研究商圈的动态变化等. 然而, 如果攻击者能够识别出轨迹对应的用户身份, 将会严重威胁到用户的隐私信息. 现有的轨迹匿名算法在度量相似性时仅考虑轨迹在采样点位置的邻近性, 忽略轨迹位置的动态邻近性, 因此产生的匿名轨迹集合可用性相对较低. 针对这一问题, 本文提出了邻域扭曲密度和邻域相似性的概念, 充分考虑轨迹位置的动态邻近性, 并分别提出了基于邻域相似性和邻域扭曲密度的轨迹匿名算法; 前者仅考虑了轨迹位置的动态邻近性, 后者不仅能衡量轨迹位置的动态邻近性, 而且在聚类过程中通过最小化邻域扭曲密度来减少匿名集合的信息损失. 最后, 在合成轨迹数据集和真实轨迹数据集上的实验结果表明, 本文提出的算法具有更高的数据可用性.

关键词 隐私保护, 轨迹匿名, 动态邻近性, 邻域相似性, 邻域扭曲密度

引用格式 王超, 杨静, 张健沛. 基于轨迹特征及动态邻近性的轨迹匿名方法研究. 自动化学报, 2015, **41**(2): 330−341

DOI 10.16383/j.aas.2015.c140139

Research on Trajectory Privacy Preserving Method Based on Trajectory Characteristics and Dynamic Proximity

WANG Chao[1] YANG Jing[1] ZHANG Jian-Pei[1]

Abstract The rise of mobile social networks as well as mobile intelligent terminal has generated a lot of spatial-temporal trajectory data, publishing and analyzing such data is essential to improve transportation, to understand the dynamics of the economy in a region, etc. However, it will be a serious threat to the user′s privacy, if adversary is able to identify user′s identity corresponding to the trajectory. While calculating similarity of trajectories, the existing methods consider only locations proximity of the sampling point in the trajectory, and ignore the dynamic proximity of locations in the trajectory. So the produced trajectory anonymity set has a low utility. To solve this problem, we first present the concept of neighborhood similarity and neighborhood distortion density to fully consider the dynamics proximity of locations in the trajectory, and then propose two algorithms, i.e., trajectory anonymity algorithm based on neighborhood similarity and trajectory anonymity algorithm based on trajectory neighborhood distortion density. The former one only considers the dynamics proximity of locations in the trajectory, while the latter one also reduces information loss of anonymous collection by minimizing neighborhood distortion density during the clustering process. Finally, experimental results on a synthetic data set and a real-life data set demonstrate that our method offers better utility than comparable previous proposals in the literature.

Key words Privacy preserving, trajectory anonymity, dynamic proximity, neighborhood similarity, neighborhood distortion density

Citation Wang Chao, Yang Jing, Zhang Jian-Pei. Research on trajectory privacy preserving method based on trajectory characteristics and dynamic proximity. *Acta Automatica Sinica*, 2015, **41**(2): 330−341

移动社会网络的兴起及移动智能终端的发展带来了海量的新数据[1], 尤其是个人位置和轨迹数据. 由于位置和轨迹数据含有丰富的时空信息, 对其进行分析和挖掘可以支持多种与移动对象相关的应用[2], 比如: 智能交通、城市及道路规划、供应链管理等. 因此, 发布并分析这样的时空数据是非常有必要的.

然而, 只要有数据, 就必然存在安全与隐私的问题, 个人的隐私信息很可能会随着轨迹数据的发布而受到威胁. 简单地去掉标识符属性不能有效地保护个人的隐私信息. 在轨迹数据中, 最大的隐私威胁就是 "敏感位置泄露", 在这种情况下, 如果攻击者能够了解某人在哪些时间访问了哪些位置, 那么攻击者就能够确定此人在发布数据库中的真实记录, 并

收稿日期 2014-03-14 录用日期 2014-10-13
Manuscript received March 14, 2014; accepted October 13, 2014
国家自然科学基金 (61370083, 61073041, 61073043, 61402126), 高等学校博士学科点专项科研基金 (20112304110011, 20122304110012) 资助
Supported by National Natural Science Foundation of China (61370083, 61073041, 61073043, 61402126), and Research Fund for the Doctoral Program of Higher Education of China (20112304110011, 20122304110012)
本文责任编委 徐昕
Recommended by Associate Editor XU Xin
1. 哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001
1. College of Computer Science and Technology, Harbin Engineering University, Harbin 150001

且能够获取此人的其他轨迹信息, 进而推理得到此人的兴趣爱好、行为模式、社会习惯等隐私信息, 造成个人隐私信息的泄露. 因此, 面向移动社会网络轨迹发布的隐私保护方法是一个亟待解决的问题.

对于传统的关系型数据的隐私保护, 学术界已经进行了广泛的研究[3−9], 但是传统的关系型数据的隐私保护不能直接应用到包含时空数据的轨迹隐私保护方法中, 因为轨迹数据具有时间、位置相关性, 很难界定准标识符和敏感属性, 任何时空点或者时空点的组合都可以被视为准标识符属性. 在轨迹数据上直接进行 k-匿名[10−11] 操作, 对于任意一个轨迹, 都需要至少 k - 1 个其他原始轨迹被转换成完全相同的匿名轨迹来构成一个匿名轨迹集合, 而这样显然会造成巨大的信息损失.

为了降低轨迹数据发布产生的隐私泄露风险, 研究人员提出了多种轨迹匿名算法[12−15], 使发布的轨迹数据集合满足轨迹 k-匿名. 轨迹 k-匿名是位置 k-匿名的扩展[16], 在匿名过程中, 将用户的轨迹与至少 $k-1$ 条其他轨迹组成匿名集合, 然后进行发布, 这样攻击者在没有背景知识的情况下只有 $1/k$ 的概率得到用户的真实轨迹, 从而实现用户轨迹的隐私保护. 然而, 不合适的轨迹相似性度量会导致匿名过程中不必要的信息损失, 降低轨迹数据的可用性. 因此, 如何准确地衡量轨迹相似性, 构造满足最小信息损失的匿名集合是一个关键的问题.

Abul 等[14]、Huo 等[17] 在计算轨迹相似性时使用欧氏距离作为度量标准, Abul 等[15]、Chen 等[18] 在计算轨迹相似性时使用编辑距离作为度量标准, Tiakas 等[19] 在计算轨迹相似性时使用线性时空距离作为度量标准, Gao 等[20] 在计算轨迹相似性时, 用轨迹角度来衡量轨迹的相似性和方向, 文献 [21] 在研究 LBS (Location-based services) 中连续查询的隐私保护时, 考虑了速度因素.

移动对象形成的轨迹信息是包含采样点位置、时间、速度 (大小和方向) 等多种因素的, 在计算轨迹间相似时, 仅仅考虑其中的一个或者两个因素是不能够完整地衡量轨迹信息的, 只有综合考虑轨迹的时间、位置、速度 (大小和方向) 等内外在特征信息, 才能更加准确地度量轨迹间的相似性. 然而, 以上方法在计算轨迹相似性时大都仅仅考虑了当前轨迹点的位置和时间因素, 即轨迹在采样点位置的邻近性, 忽略了移动对象的速度因素对轨迹位置的影响, 而轨迹位置的邻近性会随着移动对象的运动而改变. 如果匿名集合内的对象速度 (大小和方向) 不同甚至差距过大的话, 会形成过大的匿名集合, 严重降低匿名数据的可用性. 图 1 模拟了移动对象的移动和匿名的形成.

如图 1 所示, 开始时刻邻近的对象 (C, D, E) 在最后时刻随着对象的散开而距离很远, 而开始时刻距离很远的对象 (A, B, F) 在最后时刻随着对象的移动而逐渐靠近. 因此, 在匿名轨迹时, 不仅仅应该考虑用户在采样点的位置, 还需要考虑移动对象在运动中的动态邻近性. 如果忽略轨迹位置的动态邻近性, 极有可能导致在形成移动对象的轨迹 k-匿名集合时, 集合内的移动对象在整个移动轨迹上并不相近, 因此, 在匿名过程中产生较大的信息损失, 降低匿名轨迹数据的可用性.

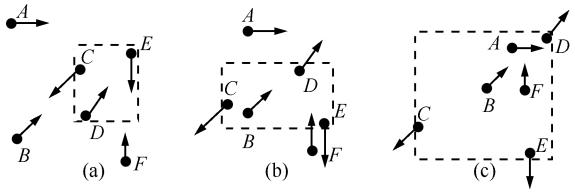


图 1　移动对象的移动和匿名形成示意图

Fig. 1　The diagram of the movement of move objects and the form of anonymous sets

本文研究轨迹数据发布的隐私保护, 在前人研究的基础上, 综合考虑轨迹的时间、位置、速度和方向等内外在特征信息对轨迹相似性的影响, 并提出相应的轨迹匿名算法, 在有效地保护轨迹数据的同时, 提高轨迹数据的可用性. 本文的主要贡献如下:

1) 提出邻域扭曲密度和邻域相似性的概念, 充分考虑轨迹位置的动态邻近性;

2) 分别提出基于邻域相似性和邻域扭曲密度的轨迹匿名算法来实现轨迹 k-匿名, 前者仅考虑了轨迹位置的动态邻近性, 后者不仅能衡量轨迹位置的动态邻近性, 而且在聚类过程中通过最小化邻域扭曲密度来减小匿名集合的信息损失;

3) 使用合成轨迹数据集和真实数据集对实验进行了评估, 实验结果表明, 本文提出的邻域扭曲密度和邻域相似性能准确地衡量轨迹间的相似性, 降低了匿名集合的信息损失.

本文的组织结构如下: 第 1 节介绍相关工作; 第 2 节介绍研究基础; 第 3 节提出轨迹隐私保护算法; 第 4 节给出算法的隐私保障和可用性分析; 第 5 节对实验及结果进行分析; 第 6 节给出结论.

1　相关工作

1.1　轨迹数据隐私保护技术

现有的轨迹隐私保护技术可以被分成以下几种类型[22].

1.1.1　基于假数据的轨迹隐私保护技术

该技术是指通过添加假轨迹对原始数据进行干扰, 同时又要保证被干扰的轨迹数据的某些统计属性不发生严重失真. 为了使攻击者不能有效地区分真实轨迹, You 等[23] 提出了生成假轨迹的方法, 该方法按照以下原则生成假轨迹: 1) 假轨迹的运动模

式与真实轨迹相似; 2) 轨迹间的交点尽可能的多. Gao 等[24] 提出生成假轨迹的方法, 来保护移动对象的位置和轨迹隐私, 并且得到较高的服务质量. 假轨迹隐私保护方法简单、计算量小, 但易造成假数据的存储量大及数据可用性降低等缺点. 因此, 该方法适用于实时性较高并且对数据可用性要求不高的系统.

1.1.2 基于抑制法的轨迹隐私保护技术

该技术是指根据具体情况有条件地发布轨迹数据, 不发布轨迹上的某些敏感位置或频繁访问的位置以实现隐私保护. Terrovitis 等[25] 假设不同的攻击者掌握不同的、不相交的移动对象轨迹信息, 并提出了一种基于抑制的方法来降低泄露整条轨迹的概率, 该方法迭代地抑制一些轨迹片段, 直到轨迹的泄露概率小于给定的阈值; Chen 等[26] 提出了 $(k,c)_L$ 隐私模型, 并使用局部抑制的方法来实现轨迹数据的匿名, 提高匿名数据的可用性. 抑制法不发布轨迹上的某些敏感位置或频繁访问的位置, 因此能保证用户较高的隐私保护度; 然而, 太多的轨迹片段被抑制会造成巨大的信息损失. 因此, 该方法适用于隐私保护度较高并且对数据可用性要求不高的系统.

1.1.3 基于划分法的轨迹隐私保护技术

该技术是指将轨迹上所有的采样点都泛化为对应的匿名区域, 以达到隐私保护的目的. 基于泛化法的轨迹 k-匿名技术在隐私保护度和数据可用性上取得了较好的平衡, 是目前轨迹隐私保护的主流方法.

Nergiz 等[27−28] 提出了一个基于泛化的算法. 该方法首先通过轨迹点匹配的方式将轨迹聚集到一系列聚类中, 然后将同一类中的轨迹的对应点泛化为最小边界矩形, 不匹配的点和轨迹被抑制. 最后, 对匿名的轨迹数据进行重构, 并发布重构后的原子轨迹.

Huo 等[17] 将轨迹 k-匿名集合的选择问题转化为图划分的问题, 通过最小化划分代价, 来降低匿名过程中的信息损失.

为了进一步降低信息损失, Huo 等[29] 提出了 You can walk alone (YCWA) 方法, 该方法提取轨迹中重要的停留点, 并使用基于网格和基于聚类的方法对停留点进行匿名.

Domingo-Ferrer 等[30] 提出了一个基于微聚集和排列的匿名算法, 该算法首先提出了一个轨迹相似性度量标准, 同时考虑了轨迹的时间和空间因素; 然后使用微聚集算法对轨迹进行聚类操作; 最后使用位置排列算法, 对轨迹数据进行重构.

Gao 等[20] 在计算轨迹相似性时, 用轨迹角度来衡量轨迹的相似性和方向, 将轨迹 k-匿名集合的选择问题转换为约束的最小生成树问题, 构造轨迹图模型来衡量轨迹的隐私度和数据可用性, 并通过在轨迹图中使用贪婪策略, 找到近似最优的轨迹 k-匿名集合.

1.1.4 其他轨迹隐私保护技术

除了以上主流的轨迹隐私保护算法, 差分隐私保护也逐渐地被用来进行轨迹隐私保护. 差分隐私是一种新的隐私保护框架, 它通过对原始数据及其转换或者统计结果添加噪音来达到隐私保护效果, 能够防止攻击者拥有任意背景知识下的攻击. 文献 [31−32] 对差分隐私及其在数据发布领域的应用进行了综述, Chen 等[33] 将差分隐私用于对运输信息的保护, 通过对数据加入 Laplace 噪声保证挖掘结果满足差分隐私需求, 从而实现对运输信息的轨迹隐私保护.

2 研究基础

2.1 基本概念

定义 1 (轨迹). 移动对象的轨迹指的是时间位置点的有序集合[30], 表示如下:

$$T=\{(t_1,x_1,y_1),(t_2,x_2,y_2),\cdots,(t_n,x_n,y_n)\} \quad (1)$$

定义 2 (子轨迹). 轨迹 $S=\{(t'_1,x'_1,y'_1),(t'_2,x'_2,y'_2),\cdots,(t'_m,x'_m,y'_m)\}$ 是式 (1) 中轨迹 T 的子轨迹, 如果存在整数 $1\le i_1<\cdots<i_m\le n$ 满足 $(t'_j,x'_j,y'_j)=(t_{i_j},x_{i_j},y_{i_j})$, $1\le j\le m$, 表示为 $S\le T$.

定义 3 (相交轨迹). 两条轨迹 $T_i=\{(t^i{}_1,x^i{}_1,y^i{}_1),(t^i{}_2,x^i{}_2,y^i{}_2),\cdots,(t^i{}_m,x^i{}_m,y^i{}_m)\}$ 和 $T_j=\{(t^j{}_1,x^j{}_1,y^j{}_1),(t^j{}_2,x^j{}_2,y^j{}_2),\cdots,(t^j{}_m,x^j{}_m,y^j{}_m)\}$, $I=\max(\min(t^i{}_n,\ t^j{}_m)-\max(t^i{}_1,\ t^j{}_1),\ 0)$. 如果 $I>0$, 那么轨迹 T_i 和 T_j 是相交轨迹; 否则, 轨迹 T_i 和 T_j 不是相交轨迹.

定义 4 (轨迹 p%-相交[30]). 两条轨迹 $T_i=\{(t^i{}_1,x^i{}_1,y^i{}_1),(t^i{}_2,x^i{}_2,y^i{}_2),\cdots,(t^i{}_m,x^i{}_m,y^i{}_m)\}$ 和 $T_j=\{(t^j{}_1,x^j{}_1,y^j{}_1),(t^j{}_2,x^j{}_2,y^j{}_2),\cdots,(t^j{}_m,x^j{}_m,y^j{}_m)\}$ 相交, $I=\max(\min(t^i{}_n,\ t^j{}_m)-\max(t^i{}_1,\ t^j{}_1),0)$, 如果满足

$$p=100\times\frac{I}{\max(t^i{}_n,t^j{}_m)-\min(t^i{}_1,t^j{}_1)} \quad (2)$$

那么 T_i 和 T_j 轨迹 p%-相交.

两条轨迹 100%-相交, 当且仅当这两条轨迹有相同的开始时间和相同的结束时间; 两条轨迹 0%-相交, 即两条轨迹不相交, 当且仅当两条轨迹没有重叠的时间间隔. 本文用 $ot(T_i,T_j)$ 来表示任意两条相交轨迹 T_i 和 T_j 的重叠时间间隔.

定义 5 (同步轨迹). 两条轨迹 T_i 和 T_j 轨迹 p%-相交, $p>0$, 如果在重叠时间间隔 $ot(T_i,T_j)$ 内, 轨迹 T_i 和 T_j 有相同数量的位置点, 并且位置点

对应的时间点对应相同, 那么轨迹 T_i 和 T_j 是同步轨迹.

定义 6 (同步轨迹集合). 轨迹集合为同步轨迹集合当且仅当集合内任意两条 $p\,\%$-相交 $(p>0)$ 的轨迹为同步轨迹.

为了计算轨迹相似性, 必须设法将不满足同步轨迹条件的轨迹转换为同步轨迹. 本文假设移动对象在任意 2 个相邻采样点之间都做匀速直线运动, 如果轨迹 T_i 和 T_j 轨迹 $p\,\%$-相交, $p>0$, 轨迹 T_i 在 t 时刻 $(t\in ot(T_i, T_j))$ 有一个采样点 l_i, 轨迹 T_j 在 t 时刻没有采样点, 那么可以直接在轨迹 T_j 中 t 时刻插入一个采样点 l_j, 并满足假设, 反之亦然. 通过重复以上操作, 即可将轨迹 T_i 和 T_j 转换为同步轨迹, 具体的算法见文献 [30] 中的算法 1.

定义 7 (采样点 l 在 t 时刻的 k-距离). 存在 t 时刻的采样点 $o_t\in D$ (D 表示同步轨迹集合), 使得至少有 k 个 t 时刻的采样点 $o'_t\in D$, 满足 $dist\ (l_t,\ o'_t)\le dist\ (l_t, o_t)$, 并且最多有 k-1 个数据点 $o'_t\in D$, 满足 $dist\ (l_t,\ o'_t)<dist\ (l_t,\ o_t)$, 则 l 在时刻 t 的 k 距离 $k-t-dist\ (l)=dist\ (l_t,\ o_t)$.

定义 8 (采样点 l 在 t 时刻的矩形邻域). t 时刻的矩形邻域 $N_t(l)=(L_{x-,t},\ L_{y-,t},\ L_{x+,t},\ L_{y+,t})$, 其中, $(L_{x-,t},\ L_{y-,t}$ 和 $L_{x+,t},\ L_{y+,t})$ 分别是矩形邻域的左下角和右上角, $l_t\in N_t(l)$.

定义 9 (采样点 l 在 t 时刻的 k-矩形邻域). $N_{k,t}(l)=N_t(l)|\forall o_t\in D$ (D 表示同步轨迹集合) $o_t\in N_t(l)$, $dist(l_t,\ o_t)\le k-t-dist\ (l)$.

为了衡量轨迹中位置邻近性的变化, 需要了解任意时刻采样点的 k-矩形邻域, 即 k-矩形邻域的运动情况, 本文使用 $v_{x+,t}$, $v_{y+,t}$, $v_{x-,t}$, $v_{y-,t}$ 表示矩形邻域各个边界的运动速度. 采样点被概化为 k-矩形邻域会产生信息损失, 本文提出邻域位置扭曲度的概念来定量地描述概化产生的信息损失.

定义 10 (邻域位置扭曲度). 假设在初始时刻 s, 采样点 l 被概化为 l 的 k-矩形邻域 $N_{k,s}(l)=(L_{x-,s},\ L_{y-,s},\ L_{x+,s},\ L_{y+,s})$, 则该采样点概化产生的位置扭曲度为[21]

$$Loss(N_{k,s}(l))=(L_{x+,s}-L_{x-,s})\times(L_{y+,s}-L_{y-,s}) \tag{3}$$

那么, 在任意时刻 t, 该采样点概化产生的位置扭曲度为

$$\begin{aligned}Loss(N_{k,t}(l))=&(L_{x+,s}+v_{x+,t}\times(t-s)-L_{x-,s}-\\&v_{x-,t}\times(t-s))\times(L_{y+,s}+v_{y+,t}\times(t-s)-\\&L_{y-,s}-v_{y-,t}\times(t-s))\end{aligned} \tag{4}$$

那么, 对于同步轨迹集合 D, 轨迹中对应采样点被概化为 k-矩形邻域, 产生的轨迹邻域位置扭曲度为

$$Loss(D)=\int_s^e Loss(N_{k,t}(l))\mathrm{d}t \tag{5}$$

其中, s 和 e 分别为同步轨迹采样的开始和结束时间.

定义 11 (邻域扭曲密度). 轨迹的邻域扭曲密度为轨迹的邻域位置扭曲度与该邻域内轨迹对应的采样点个数的比值, 表示为

$$\varrho(N_{k,t}(l))=\frac{\int_s^e Loss(N_{k,t}(l))\mathrm{d}t}{k+1} \tag{6}$$

其中, $N_{k,t}(l)$ 为采样点 l 在 t 时刻概化形成的 k-矩形邻域, s 和 e 分别为同步轨迹采样的开始时间和结束时间.

邻域扭曲密度用来衡量每个采样点造成的邻域位置扭曲度的大小, 对于同样满足轨迹 k-匿名的不同轨迹集合, 较小的邻域扭曲密度意味着每个采样点产生较小的位置扭曲度, 即较小的信息损失.

既然较小的邻域扭曲度意味着较小的信息损失, 那么相似的两条轨迹匿名产生的邻域扭曲度一定小于不相似的轨迹产生的邻域扭曲度, 因此, 本文使用邻域位置扭曲度作为衡量任意两个同步轨迹相似性的标准, 并提出邻域相似性的概念.

定义 12 (邻域相似性). 轨迹的邻域相似性为任意两个同步轨迹从起始时间到结束时间对应采样点构成的邻域的位置扭曲度之和, 表示为

$$Sim(T_i,T_i)=\int_s^e Loss(\{T_i,T_i\})\mathrm{d}t \tag{7}$$

其中, T_i,T_i 为任意同步轨迹, $\{T_i,T_i\}$ 表示将轨迹 T_i,T_i 匿名化形成的轨迹匿名区域, s 和 e 分别为同步轨迹采样的开始和结束时间.

本文提出的轨迹邻域相似性是关于时间和速度的函数, 很明显, 同步轨迹的状态 (包括初始时间、位置和速度 (包含方向)) 越相似, 其匿名后的邻域位置扭曲度则越低; 其邻域扭曲密度越低, 邻域相似性越高; 状态差异越大, 匿名后的邻域位置扭曲度则越高; 其邻域扭曲密度越高, 邻域相似性越低.

本文提出的轨迹邻域相似性不仅考虑了时间、位置、速度和方向等内外在的轨迹特征信息, 还考虑了轨迹中位置邻近性的变化, 能较准确地衡量轨迹间的相似性. 在上文中, 本文假设移动对象在任意 2 个相邻采样点之间都做匀速直线运动, 因此, 在面向轨迹数据发布的隐私保护研究中, 任意轨迹点的速度 (包括方向) 都可以由相邻两个轨迹点的位置和时间关系求得.

2.2 轨迹隐私模型

2.2.1 攻击模型

众所周知, 轨迹数据发布时个人隐私泄露的风险与攻击者所掌握的背景知识有关, 攻击者掌握的背景知识越多, 个人的隐私泄露风险越大, 反之亦然. 本文假设攻击者掌握以下的背景知识:

1) 攻击者可以访问发布的匿名轨迹集合 D^*;

2) 攻击者知道原始的目标轨迹 T ($T \in D$, D 表示原始轨迹集合) 的子轨迹 S;

定义 13 (攻击模型). 对于已知的子轨迹 S ($S \leq T$) 和匿名轨迹集合 D^*, 如果攻击者能够确定匿名轨迹 T^* ($T^* \in D^*$) 是 T 的匿名轨迹, 则认为用户的个人隐私遭到攻击.

2.2.2 轨迹隐私模型

为了能准确地衡量轨迹隐私保护程度, 本文引入以下轨迹隐私模型[30].

给定子轨迹 S ($S \leq T$), $Pr_{T^*}[T|S]$ 表示攻击者能够确定匿名轨迹 T^* ($T^* \in D^*$) 是 T 的匿名轨迹的概率.

定义 14 (轨迹 p-隐私). 如果 $Pr_{T^*}[T|S] \leq p$, ($\forall T \in D, S \leq T$), 则称匿名轨迹满足轨迹 p-隐私.

定义 15 (轨迹 k-匿名). 给定子轨迹 S ($S \leq T$), 满足轨迹 k-匿名当且仅当满足轨迹 $1/k$-隐私.

3 轨迹匿名算法

在本节中, 基于邻域相似性和邻域扭曲密度等概念, 本文分别提出基于邻域相似性的轨迹匿名算法 (Trajectory anonymity algorithm based on neighborhood similarity, TAA-NS) 和基于邻域扭曲密度的轨迹匿名算法 (Trajectory anonymity algorithm based on trajectory neighborhood distortion density, TAA-NDD) 来实现轨迹 k-匿名. 前者仅考虑轨迹位置的动态邻近性, 不能保证匿名集合最小的信息损失; 后者通过轨迹的邻域扭曲密度来度量轨迹相似性, 不仅能衡量轨迹位置的动态邻近性, 而且通过最小化邻域扭曲密度来减小匿名集合的信息损失.

3.1 基于邻域相似性的轨迹匿名算法

基于邻域相似性的轨迹匿名算法主要依靠轨迹的邻域相似性来衡量轨迹的相似性和轨迹位置的邻近性. 通过将邻域相似的轨迹聚集到一起, 达到隐私保护的效果, 具体过程见算法 1.

算法 1 (TAA-NS).

输入. 原始同步轨迹集合 D, 轨迹 k-匿名的参数 k.

输出. 轨迹等价类集合 Q.

BEGIN

1) $Q = \emptyset$;

2) $V = D$; //V 表示没有形成匿名集合的轨迹集合

3) If ($|D| < k$)

4) Return;

5) While ($|D| \geq k$) do

6) $C = \emptyset$; //C 表示某个轨迹等价类

7) 在集合 D 中随机选择一条轨迹 T_i 加入集合 C;

8) While ($|C| < k$) do

9) If ($\exists$ T_w, 满足: T_i 和 T_w 邻域最相似, 并且 $T_w \notin C$, $T_w \in V$)

10) $C = C \cup \{T_w\}$;

11) Else

12) Break;

13) End If

14) End While

15) If $|C| \geq k$ do

16) $Q = Q \cup C$;

17) $D = D - C$;

18) $V = V - C$;

19) Else

20) $D = D - \{T_i\}$;

21) End If

22) End While

23) $V = V \cup D$; //将没有形成匿名集合的轨迹合并

24) While ($V \neq \emptyset$) do

25) 在集合 V 中随机选择一条轨迹 T_i;

26) If ($\exists$ T_w, 满足: T_i 和 T_w 邻域最相似, $T_w \in W, W \in Q$) do

27) $W = W \cup \{T_i\}$;

28) $V = V - \{T_i\}$;

29) End If

30) End While

31) Return Q;

32) End If

END

算法 TAA-NS 首先判断输入轨迹集合 D 中的轨迹数量是否小于 k (步骤 3)), 如果条件满足, 那么算法肯定不能产生满足轨迹 k-匿名的等价类集合算法, 则返回 (步骤 4)).

步骤 5) ~ 22) 是本算法的关键. 当 D 中轨迹数量不小于 k 时, 算法重复执行步骤 5) ~ 22), 每次最多得到一个类 C, 使类 C 中至少有 k 个轨迹. 具体的方法是: 首先初始化集合 C, 随机地从 D 中选择轨迹 T_i 加入到 C 中; 然后在步骤 8) ~ 14) 中贪婪地在 V 中选择与 C 中轨迹邻域最相似的轨迹, 并加入到 C 中. 如果形成的集合 C 中含有足够的

轨迹 (不小于 k 个), 将 C 加入到 Q 中, 并分别从 D 和 V 中删除 C; 否则, 将 T_i 从 D 中移除 (步骤 15)～20)). 重复上述策略, 直到 D 中轨迹数量小于 k 为止 (步骤 22)).

如果仍有剩余的轨迹, 算法在步骤 23)～31) 将剩余轨迹逐一加入到与其轨迹邻域相似性最大的类中. 最后返回轨迹等价类集合 Q.

3.2　基于邻域扭曲密度的轨迹匿名算法

算法 TAA-NDD 通过轨迹的邻域扭曲密度来衡量轨迹的相似性和轨迹位置的邻近性. 通过将邻域扭曲密度最小的轨迹聚集到一起, 达到隐私保护的效果, 具体过程见算法 2.

算法 2 (TAA-NDD).

输入. 原始同步轨迹集合 D, 轨迹 k-匿名的参数 k.

输出. 轨迹等价类集合 Q.

BEGIN

1) $Q=\emptyset$;

2) $V=D$; //V 表示没有形成匿名集合的轨迹集合

3) If ($|D|<k$)

4)　Return;

5) While ($|D|\geq k$) do

6)　$C=\emptyset$; //C 表示某个轨迹等价类

7)　在集合 D 中随机选择一条轨迹 T_i 加入集合 C;

8)　While ($|C|<k$) do

9)　If ($\exists\ T_w$, 满足: $T_w\notin C$, $T_w\in V$, T_w 加入 C 中产生的邻域扭曲密度最小)

10)　$C=C\cup\{T_w\}$;

11)　Else

12)　Break;

13)　End If

14)　End While

15)　If $|C|\geq k$ do

16)　$Q=Q\cup C$;

17)　$D=D-C$;

18)　$V=V-C$;

19)　Else

20)　$D=D-\{T_i\}$;

21)　End If

22) End While

23) $V=V\ \cup\ D$; //将没有形成匿名集合的轨迹合并

24) While ($V\neq\emptyset$) do

25)　在集合 V 中随机选择一条轨迹 T_i;

26)　If ($\exists W$, 满足:$W\in Q$, T_i 加入 W 中产生的邻域扭曲密度最小) do

27)　$W=W\cup\{T_i\}$;

28)　$V=V\ -\ \{T_i\}$;

29)　End If

30) End While

31) Return Q;

32) End If

END

算法 TAA-NDD 与算法 TAA-NS 有近似的实现策略, 不同之处在于: 算法 TAA-NDD 通过轨迹的邻域扭曲密度来度量轨迹的相似性, 不仅能衡量轨迹位置的动态邻近性, 而且在聚类过程中通过最小化邻域扭曲密度来减小匿名集合的信息损失 (步骤 9) 和步骤 26)).

在算法 1 和算法 2 的实现过程中, 由于考虑轨迹位置的动态邻近性, 需要了解任意时刻 t 轨迹采样点的 k-矩形邻域. 在算法 1 中, 轨迹的邻域相似性只需要计算任意 t 时刻 2 条轨迹采样点的矩形邻域, 可以由 t 时刻采样点的相对位置确定, 比较简单; 在算法 2 中, 需要确定任意时刻 t 时, 至少 k 条轨迹的采样点形成的矩形邻域. 考虑到跟踪所有时刻矩形邻域的边界状态代价较大, 本文只考虑矩形邻域边界改变时的状态.

由于矩形邻域内各个采样点速度不一致, 在某一时刻矩形邻域边界的速度可能发生变化, 如图 2 所示.

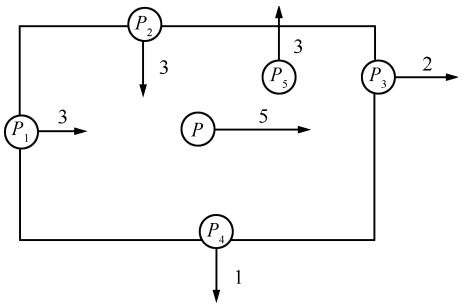


图 2　矩形邻域边界速度变化示意图

Fig. 2　The velocity change diagram of the rectangular neighborhood boundary

如图 2 所示, 在目前时刻, 该矩形邻域的边界速度分别由 P_1、P_2、P_3、P_4 四个采样点确定, 但是由于 P 和 P_5 的存在, 边界速度会在不久发生改变, 由于 P 的速度大于 P_3 的速度, P 会代替 P_3 成为新的右边界, 同理, P_5 会代替 P_2 成为新的上边界, P_2 代替 P_4 成为新的下边界等. 因此, 矩形邻域边界的速度是一个关于时间 t 的分段函数, 问题的关键是确定矩形邻域边界改变的时间. 以 P 和 P_3 为例, 已知 P 和 P_3 的位置和在 x 轴方向的速度, 很容易求出 P 追上 P_3 的时间. 由于算法比较直观, 具体的算法省略.

3.3 算法时间复杂性分析

设原始同步轨迹集合 D 中轨迹数为 n, 采样点数为 d, 算法在聚类完成后得到 m 个类.

算法在步骤 5)～22) 中每得到一个新类 C, 最多扫描 k 遍 D, 并且通过轨迹的对应采样点来计算轨迹的邻域相似性或者邻域扭曲密度. 因为在算法执行过程中 $|D| \leq n$, 所以每生成一个新类用时不超过$\mathrm{O}(dkn)$. 步骤 5)～22) 共生成 m 个类, 执行时间为$\mathrm{O}(dkmn)$.

步骤 22) 结束后得到 m 个类, 每个类至少 k 个元组, 所以至多剩余 $n-mk$ 个元组, 每次循环需要扫描 Q 一遍, 并且通过轨迹的对应采样点来计算轨迹的邻域相似性或者邻域扭曲密度, 执行时间为 $\mathrm{O}(dm)$. 所以, 步骤 24)～30) 的执行时间为 $\mathrm{O}(dm(n-mk))$.

因此, 算法 TAA-NDD 和算法 TAA-NS 总体执行时间为 $\mathrm{O}(dkmn) + \mathrm{O}(dm(n-mk))$, 因为 $km < n$, 所以在最坏情况下算法的时间复杂度为 $\mathrm{O}(dn^2)$.

4 轨迹隐私保障和可用性分析

4.1 轨迹隐私保障

定理 1 (TAA-NS 算法实现了轨迹 k-匿名).

证明. 假设子轨迹 S 是攻击者掌握的关于原始目标轨迹 T_s 的背景知识, 满足 $S \leq T_s$, 并且攻击者可以访问发布的匿名轨迹集合 D^*. 由于攻击者掌握子轨迹 S 的信息, 他可以推测出移动对象的部分速度信息, 但是不会对隐私效果产生影响: 1) 如果攻击者掌握的是部分相邻轨迹点的信息, 他能推测出相邻点间的速度, 但是速度信息是不断变化的, 攻击者难以根据推测的速度信息得出下一个轨迹点的具体位置; 2) 如果攻击者掌握的是部分不相邻的轨迹点信息, 由于速度信息是不断变化的, 攻击者难以推测出两个轨迹点间的速度信息, 因此也难以确定轨迹点的具体位置.

在最好的情况下, 攻击者不知道轨迹的隐私保护方法, 此时隐私泄露概率为 $1/area$, 其中 $area$ 为矩形邻域的面积; 在最坏的情况下, 如果攻击者知道轨迹的隐私保护方法, 并且知道该矩形邻域内轨迹采样点的数量 k, 那么, 隐私泄露概率最大为 $1/k$. 综上所述, 经过 TAA-NS 算法的匿名保护, 轨迹隐私泄露的概率小于等于 $1/k$, 根据定义 15 可知, TAA-NS 算法实现了轨迹 k-匿名. □

定理 2 (TAA-NDD 算法实现了轨迹 k-匿名).

证明. TAA-NDD 算法与 TAA-NS 算法的实现过程大体相同, 在同样的背景知识下, TAA-NDD 算法也能实现轨迹 k-匿名, 证明过程同上. □

4.2 轨迹可用性度量标准

为了更全面地衡量发布的匿名轨迹数据的可用性, 本文从匿名数据的质量和信息损失两方面进行分析.

4.2.1 匿名数据质量的度量

对于满足轨迹 k 匿名模型的匿名轨迹数据而言, 可以通过匿名组的平均规模以及匿名组的总量来度量匿名数据的质量. 一般来说, 匿名数据包含的匿名组越多, 平均匿名组规模越小, 匿名化的数据越接近原来的真实数据, 可用性也越高. 显然原始数据表包含最多的匿名组, 匿名组规模最小, 其可用性也最高.

本文使用 CAVG 作为数据可用性度量标准. CAVG 是 LeFevre 等[34] 提出的, 表示为

$$CAVG = \left(\frac{|D|}{k|Q|} \right) \tag{8}$$

其中, $|D|$ 表示原始轨迹集合中的轨迹数量, $|Q|$ 表示集合 Q 中的聚类个数. CAVG 值越小, 数据可用性越高.

4.2.2 匿名数据信息损失的度量

由于本文算法在聚类过程中, 能保证将所有轨迹聚集到某个聚类中, 即不存在整条轨迹的删除, 因此信息损失主要来源于匿名集构造过程中对于轨迹点的删除和泛化操作, 为了准确度量匿名轨迹的信息损失, 本文采用如下的信息损失计算公式:

$$Infoloss = \sum_{i=1}^{l} \sum_{j=1}^{m} Area(x_i, y_j, t_j) + n \cdot \Omega \tag{9}$$

其中, l 代表等价类的个数; m 表示该等价类中同步轨迹的位置数量; $Area(x_j, y_j, t_j)$ 表示采样点 (x_j, y_j) 在时刻 t_j 被泛化的区域面积; n 代表删除的位置数量. Ω[30] 是一个常数, 表示对删除位置的惩罚.

5 实验及结果分析

本文使用了 2 个数据集对算法的可用性进行分析.

1) 合成数据集. 本文使用 Brinkhoff 轨迹生成器生成了 1 000 条合成轨迹, 共包含德国奥尔登堡市的 46 906 个位置, 记为 OLDENBURG.

2) 真实数据集. 本文还使用了一个收集自美国旧金山市的出租车移动轨迹作为本实验的真实数据集[35], 经过过滤操作, 得到 2 393 条轨迹信息, 其中每条轨迹平均包含 74.9 个位置, 记为 TAXI. 使用该数据集的优势在于, 它包含更大量的轨迹信息, 而且这些轨迹信息都是真实的.

实验的硬件环境为: Intel (R) Core (TM) 2 Quad CPU Q8400 @2.66 GHz, 2.67 GHz, 2.00 GB

内存, 操作系统为 Microsoft Windows 7, 算法均在 Matlab R2012a 下实现.

考虑到实验结果的可再现性, 本文将生成合成轨迹数据的 Brinkhoff 轨迹生成器的参数公布如下: 6 个移动对象类, 3 个外部对象类; 每个时间点生成 10 个移动对象和 1 个外部对象; 共 100 个时间点; 移动速度 250; "probability" 为 1 000. 这种设置生成了包含 46 906 个位置的 1 000 条合成轨迹, 最长的轨迹包含 100 个位置信息, 每条轨迹平均包含 46.9 个位置.

本文使用的旧金山出租车数据集由若干个文件构成, 每个文件都包含一辆出租车在 2008 年 5 月和 6 月的 GPS 信息, 文件中的每行都包含着出租车在给定时间的坐标 (经度和纬度). 然而, 出租车在一个月内的移动轨迹不可能被看作一条轨迹, 因此, 本文设置最大时间间隔, 将一条移动轨迹划分成多个轨迹.

为了方便实验, 本文仅选取一天的轨迹数据作为数据集合, 考虑到 2008 年 5 月 25 日 12 : 04 到 5 月 26 日 12 : 04 这一天的轨迹信息最多, 本文将其作为数据集合. 在该数据集合中, 连续位置的平均时间间隔为 92 秒, 因此本文设置 3 分钟 (2 倍的平均时间间隔) 为最大时间间隔, 并选取采样点数量在 20 ∼ 200 之间的轨迹, 最后, 得到 2 393 条轨迹信息, 其中每条轨迹平均包含 74.9 个位置.

为了方便对本文提出算法的可用性进行比较, 本文实现了文献 [30] 提出的算法 (记为 MDAV) 和 Abul 等[14] 的 NWA 算法, 并对算法及其可用性进行了分析. 其中, NWA 算法是最早实现轨迹 k-匿名的经典隐私保护算法; MDAV 算法是一个高效的轨迹 k-匿名算法, 具有较快的执行效率, 匿名数据集可用性较高. 其中, NWA 算法参数 δ 取值为 1 500, Max-trash = 10, MDAV 算法参数 Time-threshold = 10 000, Max-radius = 178.95, Max-trash = 10, Space-threshold = 10^9.

5.1 可用性比较

在本文中, 最主要的目标就是在提供足够的隐私保护的基础上, 最大化轨迹数据的可用性. 本文从匿名数据的质量和信息损失两方面进行分析.

5.1.1 匿名数据质量的度量

图 3 和图 4 分别描述了 MDAV、NWA、TAA-NS 和 TAA-NDD 四个算法随隐私级别 k 变化的 CAVG 值, 其中, 图 3 表示在数据集 OLDENBURG 下的实验结果, 图 4 表示在数据集 TAXI 下的实验结果.

由图 3 和图 4 发现, 随着 k 值增加, 四个算法的 CAVG 值波动不大. 对于同样的 k 值, TAA-NS、TAA-NDD 和 MDAV 算法的值相差不大, 而 NWA 的值最大. 这是因为, NWA 算法在预处理过程中, 由于轨迹时间段不匹配, 会删除许多轨迹; 而且, 在聚类过程中, 不仅删除位置信息, 还可能删除不合条件的轨迹, 因此降低了算法产生的匿名数据的可用性. 而其他 3 个算法, 在同步化过程中不会删除轨迹信息, 在聚类过程中, 也能保证聚类的大小, 因此数据可用性相对较高. 关于 k 值的优化问题, 文献 [36] 进行了详细的讨论说明, 本文在此不作详细讨论.

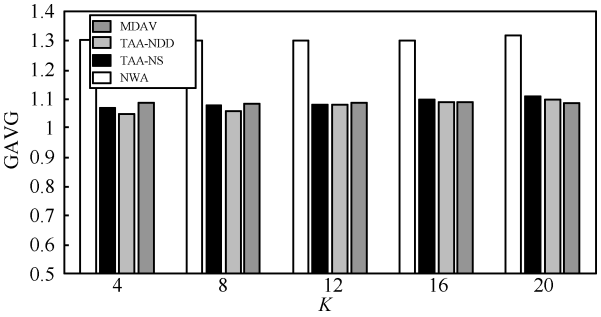


图 3　OLDENBURG 数据集下的 CAVG 值

Fig. 3　Value of CAVG in OLDENBURG dataset

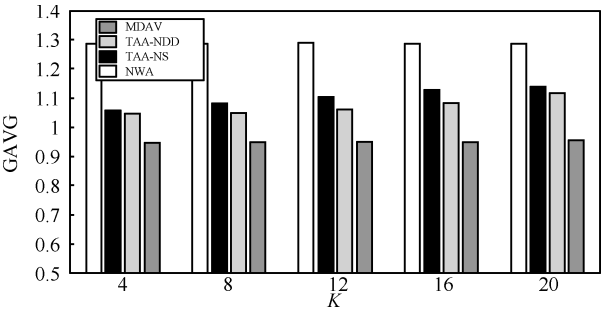


图 4　TAXI 数据集下 CAVG 值

Fig. 4　Value of CAVG in TAXI dataset

5.1.2 匿名数据信息损失的度量

根据本文提出的轨迹信息损失度量标准, 为了获得准确的信息损失, 需要选择合适的 Ω. 本文使用位置泛化产生的信息损失的最大值来近似的估计 Ω. 实验结果如图 5 和图 6 所示.

由图 5 和图 6 可知, 位置泛化产生的信息损失的最大值随着 k 值的增加波动比较大, 在 OLDENBURG 数据集下, 本文取 $k = 12$ 对应的信息损失的最大值作为 Ω 的取值, 即 $\Omega = 1.0810 \times 10^6$; 在 TAXI 数据集下, 本文取 $k = 18$ 对应的信息损失的最大值作为 Ω 的取值, 即 $\Omega = 2.6048 \times 10^9$.

图 7 和图 8 分别描述了 MDAV、NWA、TAA-NS 和 TAA-NDD 四个算法随隐私级别 k 变化的位置删除比例, 图 7 表示在数据集 OLDENBURG 下的实验结果, 图 8 表示在数据集 TAXI 下的实验结果.

由图 7 和图 8 发现, 随着 k 值增加, 四个算法的删除位置的比例都呈增加趋势, 因为随着 k 值的增加, 聚类过程中需要考虑的轨迹数量增加, 会删除更多的位置信息来满足轨迹 k-匿名; 在删除位置比例的大小方面, 对于同样的 k 值, TAA-NS 和 TAA-NDD 算法的值基本相同, 且最小, MDAV 其次, NWA 具有最大的删除位置比例. 这是由于 TAA-NS 和 TAA-NDD 算法的位置删除都发生在轨迹聚类时的轨迹同步化过程中; 而 MDAV 算法不仅在轨迹同步化过程中会删除位置, 而且在之后的位置置换过程中也会对位置进行删除, 因此它的位置删除比例大于 TAA-NS 和 TAA-NDD 算法; NWA 算法具有最大的位置删除比例, 因为相对于 MDAV 算法, 它有以下几个劣势: 1) 在预处理过程中, 由于轨迹时间段不匹配, 会删除许多轨迹; 2) 不能计算时间跨度不相交的轨迹的相似性; 3) 在聚类过程中, 不仅删除位置信息, 还可能删除不合条件的轨迹.

图 9 和图 10 分别描述了 MDAV、NWA、TAA-NS 和 TAA-NDD 四个算法随隐私级别 k 变化的信息损失, 其中, 图 9 表示在数据集 OLDENBURG 下的实验结果, 图 10 表示在数据集 TAXI 下的实验结果.

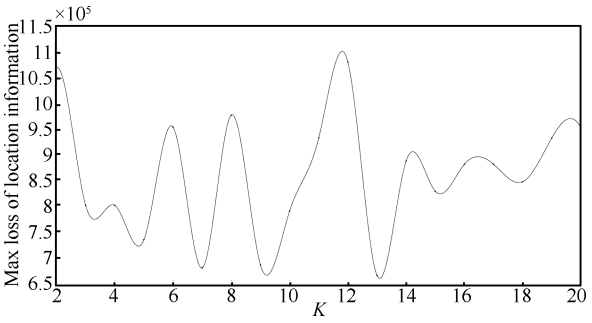


图 5 OLDENBURG 数据集下位置泛化的最大信息损失

Fig. 5 The maximum information loss of location generalization in OLDENBURG dataset

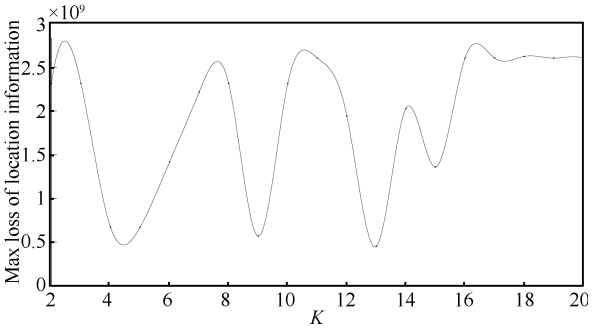


图 6 TAXI 数据集下位置泛化的最大信息损失

Fig. 6 The maximum information loss of location generalization in TAXI dataset

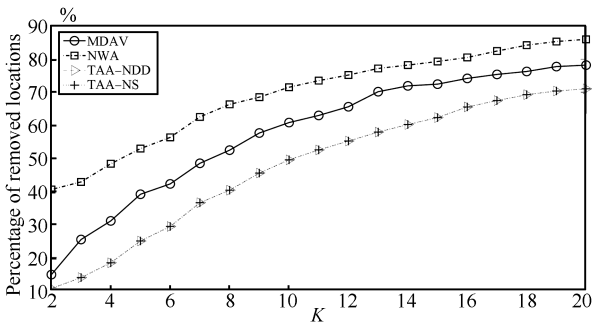


图 7 OLDENBURG 数据集下删除轨迹位置的百分比

Fig. 7 The percentage of the removed trajectory location in OLDENBURG dataset

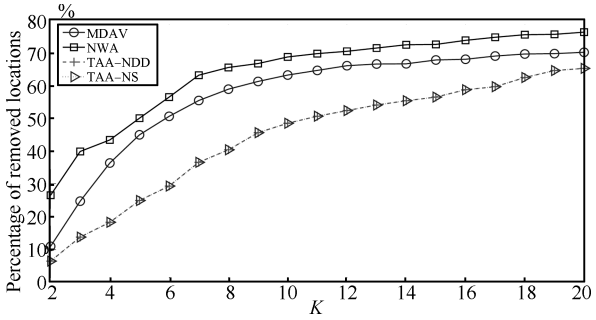


图 8 TAXI 数据集下删除轨迹位置的百分比

Fig. 8 The percentage of the removed trajectory location in TAXI dataset

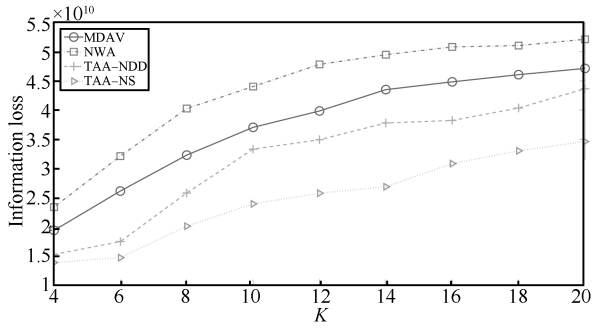


图 9 OLDENBURG 数据集下的信息损失

Fig. 9 The information loss in OLDENBURG dataset

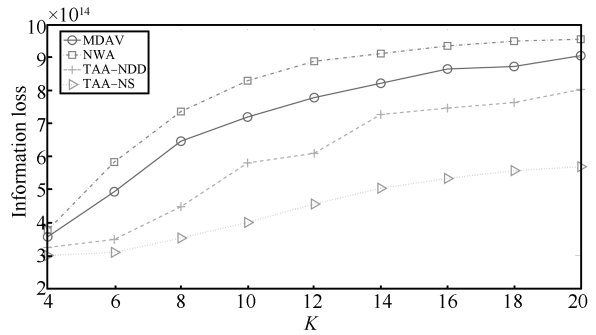


图 10 TAXI 数据集下的信息损失

Fig. 10 The information loss in TAXI dataset

由图 9 和图 10 发现, 随着 k 值增加, 四个算法的信息损失都呈增加趋势, 这是因为随着 k 值的增加, 聚类过程中需要考虑的轨迹数量增加, 会产生更大面积的泛化区域, 造成信息损失增加; 在信息损失的大小方面, 对于同样的 k 值, TAA-NS 和 TAA-NDD 算法比 MDAV 算法具有更小的信息损失, NWA 具有最大的信息损失. 这是由于 TAA-NS 和 TAA-NDD 算法在计算轨迹间相似性时考虑了轨迹中位置邻近性的变化, 能更好地衡量轨迹间的相似性; 而 TAA-NS 算法比 TAA-NDD 算法具有更大的信息损失, 这是因为 TAA-NDD 通过轨迹的邻域扭曲密度来度量轨迹的相似性, 不仅能衡量轨迹位置的动态邻近性, 而且在聚类过程中通过最小化邻域扭曲密度来减小匿名集合的信息损失. 通过对比图 9 和图 10 的信息损失可以发现, TAXI 数据集的信息损失要大于 OLDENBURG 数据集的信息损失, 这是因为相对于 OLDENBURG 数据集, TAXI 数据集更加稀疏, 在聚类过程中损失的信息会更多.

5.2　执行时间比较

图 11 和图 12 分别描述 MDAV、NWA、TAA-NS 和 TAA-NDD 四个算法随隐私级别 k 变化的执行时间, 其中, 图 11 表示在数据集 OLDENBURG 下的实验结果, 图 12 表示在数据集 TAXI 下的实验结果.

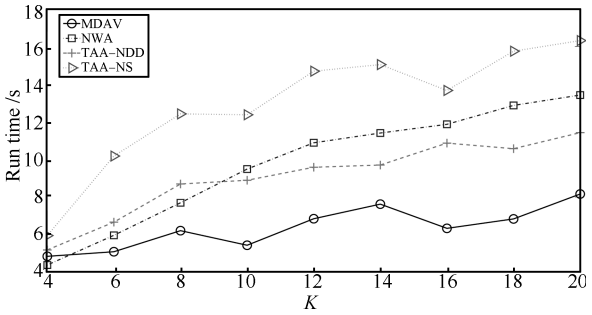


图 11　OLDENBURG 数据集下的执行时间

Fig. 11　The run time in OLDENBURG dataset

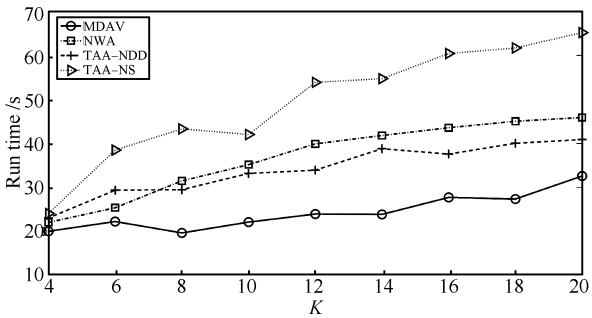


图 12　TAXI 数据集下的执行时间

Fig. 12　The run time in TAXI dataset

由图 11 和图 12 可以发现, 随着 k 值的增加, 四个算法的执行时间都呈增加趋势, 因为随着 k 值增加, 聚类过程中需要考虑的轨迹数量增加, 需要更多时间来选择相似的轨迹; 在执行时间的大小方面, 对于同样的 k 值, TAA-NS 和 TAA-NDD 算法比 MDAV 算法具有更多的执行时间, 这是由于 MDAV 算法只考虑了采样点的相似性, 而 TAA-NS 和 TAA-NDD 算法在计算轨迹间相似性时不仅考虑采样点的相似性, 还有轨迹中位置邻近性的变化, 因此增加了计算时间; 而 TAA-NDD 算法比 TAA-NS 算法具有更多的执行时间, 是因为 TAA-NDD 不仅考虑每对轨迹的邻域相似性, 还考虑匿名集合构成的矩形邻域的邻域扭曲密度是否最小, 因此增加了计算时间. 通过对比图 11 和图 12 可以发现, TAXI 数据集的执行时间大于 OLDENBURG 数据集的执行时间, 这是因为相对于 OLDENBURG 数据集, TAXI 数据集含有更多的轨迹, 在聚类过程中花费的时间更多.

6　结论

针对现有轨迹匿名算法在度量相似性时仅考虑轨迹在采样点位置的邻近性, 忽略轨迹位置的动态邻近性的问题, 本文提出了邻域扭曲密度和邻域相似性的概念, 充分考虑轨迹位置的动态邻近性, 并分别提出了基于邻域相似性和邻域扭曲密度的轨迹匿名算法; 前者仅考虑轨迹位置的动态邻近性, 后者不仅能衡量轨迹位置的动态邻近性, 而且在聚类过程中通过最小化邻域扭曲密度, 有效地降低了匿名集合的信息损失. 最后, 在合成轨迹数据集和真实轨迹数据集上的实验结果表明, 本文提出的算法具有更高的数据可用性.

References

1 Li Jian-Zhong, Liu Xian-Min. An impotant aspect of big data: data usability. *Journal of Computer Research and Development*, 2013, **50**(6): 1147−1162
(李建中, 刘显敏. 大数据的一个重要方面: 数据可用性. 计算机研究与发展, 2013, **50**(6): 1147−1162)

2 Liu Da-You, Chen Hui-Ling, Qi Hong, Yang Bo. Advances in spatiotemporal data mining. *Journal of Computer Research and Development*, 2013, **50**(2): 225−239
(刘大有, 陈慧灵, 齐红, 杨博. 时空数据挖掘研究进展. 计算机研究与发展, 2013, **50**(2): 225−239)

3 Han Jian-Min, Yu Juan, Yu Hui-Qun, Jia Dong. A multi-level l-diversity model for numerical sensitive attributes. *Journal of Computer Research and Development*, 2011, **48**(1): 147−158
(韩建民, 于娟, 虞慧群, 贾洞. 面向数值型敏感属性的分级 l-多样性模型. 计算机研究与发展, 2011, **48**(1): 147−158)

4 Han Jian-Min, Cen Ting-Ting, Yu Hui-Qun. Research in microaggregation algorithm for k-anonymization. *Acta Electronica Sinica*, 2008, **36**(10): 2021−2029
(韩建民, 岑婷婷, 虞慧群. 数据表 k-匿名化的微聚集算法研究. 电子学报, 2008, **36**(10): 2021−2029)

5 Ni Wei-Wei, Xu Li-Zhen, Chong Zhi-Hong, Wu Ying-Jie, Liu Teng-Teng, Sun Zhi-Hui. A privacy-preserving data perturbation algorithm based on neighborhood entropy. *Journal of Computer Research and Development*, 2009, **46**(3): 498−504
(倪巍伟, 徐立臻, 崇志宏, 吴英杰, 刘腾腾, 孙志挥. 基于邻域属性熵的隐私保护数据干扰方法. 计算机研究与发展, 2009, **46**(3): 498−504)

6 Yang Jing, Wang Bo. Personalized *l*-diversity algorithm for multiple sensitive attributes based on minimum selected degree first. *Journal of Computer Research and Development*, 2012, **49**(9): 2603−2610
(杨静, 王波. 一种基于最小选择度优先的多敏感属性个性化 *l*-多样性算法. 计算机研究与发展, 2012, **49**(9): 2603−2610)

7 Wang Bo, Yang Jing. A personalized privacy anonymous method based on inverse clustering. *Acta Electronica Sinica*, 2012, **40**(5): 883−890
(王波, 杨静. 一种基于逆聚类的个性化隐私匿名方法. 电子学报, 2012, **40**(5): 883−890)

8 Zhou Shui-Geng, Li Feng, Tao Yu-Fei. Privacy preservation in database applications: a survey. *Chinese Journal of Computers*, 2009, **32**(5): 847−861
(周水庚, 李丰, 陶宇飞. 面向数据库应用的隐私保护研究综述. 计算机学报, 2009, **32**(5): 847−861)

9 Xiong Ping, Zhu Tian-Qing. A data anonymization approach based on impurity gain and hierarchical clustering. *Journal of Computer Research and Development*, 2012, **49**(7): 1545−1552
(熊平, 朱天清. 基于杂度增益与层次聚类的数据匿名方法. 计算机研究与发展, 2012, **49**(7): 1545−1552)

10 Samarati P, Sweeney L. Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression. In: Proceedings of the 1998 IEEE Symposium on Research in Security and Privacy. Palo alto, CA: IEEE, 1998. 1−19

11 Sweeney L. *k*-anonymity: a model for protecting privacy. *International Journal on Uncertainty Fuzziness and Knowledge-based Systems*, 2002, **10**(5): 557−570

12 Domingo-Ferrer J, Sramka M, Trujillo-Rasúa R. Privacy-preserving publication of trajectories using microaggregation. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS. California, USA: ACM, 2010. 26−33

13 Monreale A, Andrienko G L, Andrienko N V, Giannotti F, Pedreschi D, Rinzivillo S. Movement data anonymity through generalization. *Transactions on Data Privacy*, 2010, **3**(2): 91−121

14 Abul O, Bonchi F, Nanni M. Never walk alone: uncertainty for anonymity in moving objects databases. In: Proceedings of the 24th International Conference on Data Engineering. Cancun: IEEE, 2008. 376−385

15 Abul O, Bonchi F, Nanni M. Anonymization of moving objects databases by clustering and perturbation. *Information Systems*, 2010, **35**(8): 884−910

16 Gruteser M, Grunwald D. Anonymous usage of location-based services through spatial and temporal cloaking. In: Proceedings of the 1st International Conference on Mobile Systems, Applications and Services. San Francisco, USA: ACM, 2003. 31−42

17 Huo Z, Huang Y, Meng X F. History trajectory privacy-preserving through graph partition. In: Proceedings of the 1st International Workshop on Mobile Location-based Service. Beijing, China: ACM, 2011. 71−78

18 Chen L, Özsu M T, Oria V. Robust and fast similarity search for moving object trajectories. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data. Baltimore, Maryland: ACM, 2005. 491−502

19 Tiakas E, Papadopoulos A N, Djordjevic-Kajan S. Searching for similar trajectories in spatial networks. *Journal of Systems and Software*, 2009, **82**(5):772−788

20 Gao S, Ma J F, Sun C, Li X H. Balancing trajectory privacy and data utility using a personalized anonymization model. *Journal of Network and Computer Applications*, 2014, **38**(1):125−134

21 Pan Xiao, Hao Xing, Meng Xiao-Feng. Privacy preserving towards continuous query in location-based services. *Journal of Computer Research and Development*, 2010, **47**(1): 121−129
(潘晓, 郝兴, 孟小峰. 基于位置服务中的连续查询隐私保护研究. 计算机研究与发展, 2010, **47**(1): 121−129)

22 Huo Zheng, Meng Xiao-Feng. A survey of trajectory privacy-preserving techniques. *Chinese Journal of Computers*, 2011, **34**(10): 1820−1830
(霍峥, 孟小峰. 轨迹隐私保护技术研究. 计算机学报, 2011, **34**(10): 1820−1830)

23 You T H, Peng W C, Lee W C. Protecting moving trajectories with dummies. In: Proceedings of the 8th International Conference on Mobile Data Management. Mannheim, Germany: IEEE, 2007. 278−282

24 Gao S, Ma J F, Shi W S, Zhan G X. LTPPM: a location and trajectory privacy protection mechanism in participatory sensing. *Wireless Communications and Mobile Computing*, 2012, doi: 10.1002/wcm.2324

25 Terrovitis M, Mamoulis N. Privacy preservation in the publication of trajectories. In: Proceedings of the 9th International Conference on Mobile Data Management. Beijing, China: IEEE, 2008. 65−72

26 Chen R, Fung B C M, Mohammed N, Desai B C, Wang K. Privacy-preserving trajectory data publishing by local suppression. *Information Sciences*, 2013, **231**: 83−97

27 Nergiz M E, Atzori M, Saygin Y, Güç B. Towards trajectory anonymization: a generalization-based approach. In: Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS. Irvine, California, USA: ACM, 2008. 52−61

28 Nergiz M E, Atzori M, Saygin Y, Güç B. Towards trajectory anonymization: a generalization-based approach. *Transactions on Data Privacy*, 2009, **2**(1): 47−75

29 Huo Z, Meng X F, Hu H B, Huang Y. You can walk alone: trajectory privacy-preserving through significant stays protection. In: Proceedings of the 17th International Conference on Database Systems for Advanced Applications. Busan, South Korea: ACM, 2012. 351−366

30 Domingo-Ferrer J, Trujillo-Rasua R. Microaggregation- and permutation-based anonymization of movement data. *Information Sciences*, 2012, **208**: 55−80

31 Xiong Ping, Zhu Tian-Qing, Wang Xiao-Feng. A survey on differential privacy and application. *Chinese Journal of Computers*, 2014, **37**(1): 101−122
(熊平, 朱天清, 王晓峰. 差分隐私保护及其应用. 计算机学报, 2014, **37**(1): 101−122)

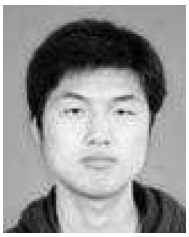
32 Zhang Xiao-Jian, Meng Xiao-Feng. Differential privacy in data publication and analysis. *Chinese Journal of Computers*, 2014, **37**(4): 927−949
(张啸剑, 孟小峰. 面向数据发布和分析的差分隐私保护. 计算机学报, 2014, **37**(4): 927−949)

33 Chen R, Desai B C, Sossou N M. Differentially private transit data publication: a case study on the montreal transportation system. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China: ACM, 2012. 213−221

34 LeFevre K, DeWitt D, Ramakrishnan R. Mondrian multidimensional k-anonymity. In: Proceedings of the 22nd International Conference on Data Engineering. Atlanta, Georgia USA: IEEE, 2006. 25−36

35 Piorkowski M, Sarafijanovoc-Djukic N, Grossglauser M. A parsimonious model of mobile partitioned networks with clustering. In: Proceedings of the 1st International Conference on Communication Systems and Networks. Bangalore, India: IEEE, 2009. 1−10

36 Song Jin-Ling, Liu Guo-Hua. Selection algorithm for optimized k-values in k-anonymity model. *Journal of Chinese Computer Systems*, 2011, **32**(10): 1987−1993
(宋金玲, 刘国华. k-匿名隐私保护模型中 k 值的优化选择算法. 小型微型计算机系统, 2011, **32**(10): 1987−1993)



王 超 哈尔滨工程大学计算机科学与技术学院博士研究生. 主要研究方向为数据库与知识工程, 数据挖掘, 隐私保护.
E-mail: wangchao0605@hrbeu.edu.cn
(**WANG Chao** Ph. D. candidate at the College of Computer Science and Technology, Harbin Engineering University. His research interest covers knowledge engineering, data mining, and privacy protection.)

杨 静 哈尔滨工程大学教授. 主要研究方向为企业智能计算, 数据库与知识工程, 隐私保护. 本文通信作者.
E-mail: yangjing@hrbeu.edu.cn
(**YANG Jing** Professor at Harbin Engineering University. Her research interest covers enterprise intelligence computing, database, knowledge engineering, and privacy protection. Corresponding author of this paper.)

张健沛 哈尔滨工程大学教授. 主要研究方向为企业智能计算, 数据库与知识工程, 社会网络.
E-mail: zhangjianpei@hrbeu.edu.cn
(**ZHANG Jian-Pei** Professor at Harbin Engineering University. His research interest covers enterprise intelligence computing, database, knowledge engineering, and social network.)