基于轨迹多特性的隐私保护算法

许华杰[1,2] 吴青华[1] 胡小明[3]
(广西大学计算机与电子信息学院 南宁 530004)[1]
(广西多媒体通信与网络技术重点实验室(广西大学) 南宁 530004)[2]
(上海第二工业大学计算机与信息工程学院 上海 201209)[3]

摘 要 现有基于聚类的轨迹隐私保护算法在衡量轨迹间的相似性时大多以空间特征为标准,忽略了轨迹蕴含的其他方面的特性对轨迹相似性的影响。针对这一情况可能导致的匿名后数据可用性较低的问题,提出了一种基于轨迹多特性的隐私保护算法。该算法考虑了轨迹数据的不确定性,综合方向、速度、时间和空间4个特性的差异作为轨迹相似性度量的依据,以提高轨迹聚类过程中同一聚类集合中轨迹之间的相似度;在此基础上,通过空间平移的方式实现同一聚类集合中轨迹的k-匿名。实验结果表明,与经典隐私保护算法相比,在满足一定隐私保护需求的前提下,采用所提算法实施隐私保护之后的轨迹数据整体具有较高的数据可用性。

关键词 轨迹隐私保护,隐私保护度,轨迹聚类,不确定性
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.01.029

Privacy Protection Algorithm Based on Multi-characteristics of Trajectory

XU Hua-jie[1,2] WU Qing-hua[1] HU Xiao-ming[3]
(School of Computer and Electronic Information, Guangxi University, Nanning 530004, China)[1]
(Guangxi Key Laboratory of Multimedia Communications and Network Technology, Guangxi University, Nanning 530004, China)[2]
(School of Computer and Information Engineering, Shanghai Second Polytechnic University, Shanghai 201209, China)[3]

Abstract Most of existing trajectory privacy protection algorithms based on trajectory clustering use spatial features as the standard when measuring the similarity between trajectories, ignoring the influence of other temporal and spatial characteristics of trajectories on trajectory similarity. In view of the fact that this situation may lead to the problem of low availability of anonymous data, a protection algorithm based on integrated spatiotemporal characteristics of trajectory was proposed. The proposed algorithm combines the uncertainty of trajectory data, and uses the difference of 4 aspects of direction, speed, time and space to measure similarity between trajectories, in order to improve the similarity between the trajectories in the same cluster set. And then the trajectories of the same clustering set are spatially shifted to achieve the k-anonymization of the trajectories in the same clustering set. The experimental results show that compared with the classical privacy protection algorithm, the trajectory data protected by proposed algorithm as a whole has higher data availability under certain privacy protection requirements.

Keywords Trajectory privacy protection, Degree of privacy protection, Trajectory clustering, Uncertainty

1 引言

定位技术的发展以及智能移动终端的普及,催生了大量的轨迹数据。轨迹数据语义丰富,对轨迹数据进行分析和挖掘可以支持多种与移动对象相关的应用,如优化道路网络及交通管理策略,分析用户行为模式以支持商业决策等[1]。轨迹数据采集之后可直接发布供用户使用,但直接发布很可能导致移动对象的个人敏感信息被泄漏,因此轨迹数据发布中的隐私保护问题逐渐成为研究热点[1]。

隐私保护的目的是保证攻击者不能以高置信度推测目标个体的敏感信息,对于轨迹隐私保护而言,既要保证轨迹本身的敏感信息不泄露,又要防止攻击者通过轨迹推导出其他的个人信息。目前关于轨迹隐私保护的研究工作主要是解决轨迹信息在数据发布和基于位置的服务(Location Based Service,LBS)中可能存在的隐私泄露问题:为了进行数据挖掘,数据所有者需要发布某些包含个人信息的数据,大量数据挖掘工具的使用要求数据所有者在发布轨迹数据时保证数据中的敏感信息不泄露,同时还要兼顾所发布轨迹数据的可用性,

到稿日期:2017-12-28 返修日期:2018-03-09 本文受广西自然科学基金项目(2014GXNSFAA118382),崇左市科技计划项目(崇科FB2018001),广西科技计划项目(2017AB15008),上海市教育委员会科研创新项目(14ZZ167),国家自然科学基金项目(71463003)资助。
许华杰(1974-),男,博士,副教授,CCF高级会员,主要研究方向为无线网络、网络安全、智能算法,E-mail:hjxu2009@163.com(通信作者);**吴青华**(1992-),女,硕士生,主要研究方向为信息安全;**胡小明**(1978-),女,博士,副教授,主要研究方向为密码学、信息安全。

如何平衡隐私保护与匿名数据可用性之间的矛盾成为轨迹数据发布中的隐私保护研究需要考虑的一个重要问题。用户在获取 LBS 时需要提供自己的位置信息，然而通过位置隐私保护技术实现的移动对象位置隐私保护并不一定能对移动对象实时运行轨迹隐私进行有效保护，如何在保证服务质量的前提下保护移动对象运行轨迹的隐私是 LBS 中的轨迹隐私保护研究的关键。本文主要针对轨迹数据发布中的隐私保护问题进行研究。

目前，轨迹数据发布中采用的隐私保护方法主要有：假数据法[2-4]、抑制法[5]、泛化法[6-8]以及基于聚类的轨迹隐私保护方法[9-15]等。基于聚类的隐私保护方法通过定义轨迹间的相似性度量，将相似度较高的 k 条轨迹聚类到同一聚类集合中，然后基于该聚类集合完成轨迹 k-匿名，使攻击者在没有其他背景知识的情况下识别移动对象的概率变为原来的 $1/k$。大量研究表明，基于聚类的轨迹隐私保护方法在隐私保护程度和数据可用性上取得了较好的平衡，是目前主流的轨迹隐私保护方法之一[1]。目前国内外陆续出现了一些基于聚类的轨迹隐私保护方法的研究成果。Domingo 等人提出了一种基于轨迹微聚集和位置排列的匿名算法，即轨迹发布过程中先采用微聚集算法对轨迹进行聚类，然后使用位置排列算法对轨迹数据进行重构，从而实现 k-匿名[9-10]。Abul 等人首次提出了轨迹数据发布时的 k-匿名问题和轨迹数据的不确定性，并基于轨迹间的欧氏距离进行轨迹聚类以构建匿名轨迹集合，进而提出了基于聚类的经典轨迹隐私保护方法——NWA 算法[11]。为了解决 NWA 算法中采用的欧氏距离不能有效度量不同运行时间段轨迹之间的距离这一问题，Abul 等人使用编辑距离 EDR 代替欧氏距离，对 NWA 算法进行了改进，进而提出了 W4M 算法[12]；之后为了解决 EDR 计算量较大且需要事先设定轨迹点匹配时的时空阈值的问题，其又提出用线性时空距离 LSTD 代替 EDR 来对 W4M 算法进行改进，改进后的算法称为 W4M_L 算法[12]。王超等人提出的 GC-DM 算法在衡量轨迹间相似性的同时考虑了轨迹形状这一因素[13]。郭旭东等人提出的轨迹 l-差异性隐私保护算法和 SP 算法在轨迹聚类时都增加了 l-差异性的条件限制，以解决由于聚类集合中轨迹间的差异性不足而可能导致的用户隐私泄漏问题[14]。

基于聚类的轨迹隐私保护方法中聚类结果对匿名轨迹数据可用性的影响较大，而轨迹间相似性的度量方式是轨迹聚类的关键。现有的轨迹隐私保护算法往往以轨迹间的空间特性差异为主进行轨迹间相似性的计算，然而轨迹不仅具有空间特性，还蕴含了时间、方向以及移动对象速度等方面的信息，综合考虑轨迹多特性将有助于轨迹聚类时更加准确地度量轨迹之间的相似性，进而提高隐私保护后轨迹数据的可用性。基于上述分析，本文提出了一种基于轨迹多特性的隐私保护(Privacy Protection based on Multi-Characteristics of Trajectory，PPMCT)算法。

2　相关概念

1)轨迹数据不确定性

受到资源和设备方面的限制，轨迹数据的采样只能以离散的方式进行，离散采样获得的数据与移动对象连续的运动轨迹之间的矛盾导致采集到的轨迹数据难以准确地反映出移动对象的实时位置，因此轨迹数据通常具有一定的不确定性[11-12]。轨迹通常用采样获取的轨迹点序列表示，即 $T=\{(x_1,y_1,t_1),(x_2,y_2,t_2),\cdots,(x_n,y_n,t_n)\}$，其中 n 为轨迹 T 中包含的轨迹点数。轨迹数据的不确定性程度通常用其所对应的不确定性区域来反映。

2)两轨迹之间的夹角

本文中轨迹对应的向量用从轨迹起点指向轨迹终点的有向线段来表示；两轨迹之间的夹角定义为两轨迹对应的向量之间的夹角。如图 1 所示，轨迹 T_1 和轨迹 T_2 之间的夹角为两轨迹对应的向量 $\vec{T}_1$ 和 $\vec{T}_2$ 的夹角 θ。若 $\vec{T}_1=ai+bj$，$\vec{T}_2=ci+dj$，其中 a,b,c,d 均为实数，则：

$$\theta=\arccos\left(\frac{a\times c+b\times d}{\sqrt{a^2+b^2}\times\sqrt{c^2+d^2}}\right) \tag{1}$$

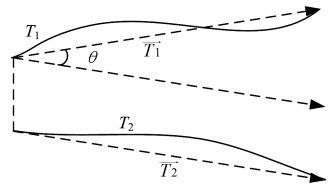

图 1　两轨迹之间的夹角示意图

Fig. 1　Schematic diagram of angle between two trajectories

3　基于轨迹多特性的隐私保护算法

为了提高隐私保护后轨迹数据的可用性，本文提出基于轨迹多特性的隐私保护(PPMCT)算法，如算法 1 所示。算法主要包含 2 个模块：1)轨迹聚类算法模块 TraClu(S,k,t_{tol},ω)，其根据隐私保护需求和轨迹间的相似性在原始轨迹集合上进行聚类操作，将相似性较高的轨迹聚类到同一集合中；2)空间平移算法模块 TraTrans(S_{clu},δ)，其在轨迹聚类的基础上将聚类中心轨迹所对应的不确定性区域作为匿名区域，并通过空间平移的方式完成同一聚类集合中轨迹的 k-匿名。

算法 1　PPMCT 算法

输入：原始轨迹数据集合 S，隐私保护度 k，可容忍时间误差 t_{tol}，时空信息权重分配情况 ω，不确定性区域半径 δ

输出：匿名轨迹集合 S_{anon}

Begin

$S_{clu}\leftarrow$TraClu(S,k,t_{tol},ω)；

$S_{anon}\leftarrow$TraTrans(S_{clu},δ)；

End

算法 1 中 S_{clu} 为聚类后的轨迹集合。下面将介绍 PPMCT 算法中轨迹相似性的度量方式以及该算法中轨迹聚类和空间平移两个模块的执行过程。

3.1　轨迹的相似性度量

轨迹数据蕴含了轨迹多方面的特性信息，为了使两轨迹之间的相似性度量更为合理，PPMCT 算法首先计算出两轨迹在方向、速度、时间以及空间 4 个特性上的差异，然后对各特性的差异值进行 Z-score 标准化[15]，最后通过设置 4 个特性的权重来综合计算出轨迹之间的相似度，并将该相似度值作为轨迹数据聚类的依据。

1)方向差异的度量

本文将轨迹方向定义为由轨迹起点指向终点,用角度距离[15]反映轨迹之间的方向差异。轨迹 T_i 与 T_j 之间的角度距离如式(2)所示[15]:

$$dirD(T_i,T_j)=\begin{cases}|T_j|\times\sin\theta, & 0°\leqslant\theta\leqslant90°\\ |T_j|, & 90°<\theta\leqslant180°\end{cases} \tag{2}$$

其中,$|T_j|$表示 T_j 的长度,θ 表示轨迹 T_i 和 T_j 的夹角,$\sin\theta$ 的值可结合式(1)求得。

2)速度差异的度量

本文用速度差值反映轨迹之间的速度差异。轨迹 T_i 与 T_j 之间的速度差值用式(3)计算[15]:

$$speD(T_i,T_j)=\frac{(|V_{\max T_i}-V_{\max T_j}|)+(|V_{\min T_i}-V_{\min T_j}|)+(|V_{\mathrm{avg}T_i}-V_{\mathrm{avg}T_j}|)}{3} \tag{3}$$

根据轨迹 T_i 中轨迹点的位置和时间信息可计算出移动对象在 T_i 各轨迹段内的平均速度,$V_{\max T_i}$ 和 $V_{\min T_i}$ 分别表示该移动对象在 T_i 上各轨迹段平均速度的最大值和最小值,$V_{\mathrm{avg}T_i}$ 表示该移动对象在 T_i 上各轨迹段平均速度的平均值。采用相同的方式对轨迹 T_j 进行计算,结果用于计算式(3)。

3)空间差异的度量

空间差异用于衡量轨迹之间的位置差距,反映了两轨迹在空间上的相似程度。本文用两轨迹之间的空间距离来表示其空间差异程度,计算方式为:依次以聚类中心轨迹 T_p 上的各轨迹点为基准,在可接受的时间误差范围(t_{tol})内计算其与轨迹 T_i 上对应时间段内的轨迹点之间的欧氏距离,取其中的最小值除以不确定性区域半径 δ,并将其结果取整后进行累加,以累加之和表示两轨迹间的空间距离,如式(4)所示:

$$spaD(T_p,T_i)=\sum_{T_p}\left|\frac{\min(dis(P_p,P_{i_1}),dis(P_p,P_{i_2}),\cdots,dis(P_p,P_{i_x}))}{\delta}\right| \tag{4}$$

其中,$P_p\in T_p$,$P_{i_1},P_{i_2},\cdots,P_{i_x}\in T_i$,且 $P_{i_1},P_{i_2},\cdots,P_{i_x}$ 与 P_i 的时间差异不超过 t_{tol}。

4)时间差异的度量

时间差异反映了两轨迹在运行时间方面的相似程度,本文用两轨迹对应起止时间段的不重合时长来反映两轨迹的时间差异。若 T_i 对应的移动对象的运行时间段为$[t_{i1},t_{im}]$,T_j 对应的移动对象运行的时间段为$[t_{j1},t_{jn}]$,则两轨迹间对应的起止时间段的不重合时长为:

$$timeD(T_i,T_j)=|t_{j1}-t_{i1}|+|t_{jn}-t_{im}| \tag{5}$$

3.2 Z-score 标准化及轨迹相似度计算

3.2.1 Z-score 标准化

由于轨迹相似性度量所采用的 $dirD,speD,spaD,timeD$ 等的量级不同,因此本文使用 Z-score 标准化[15]分别对上述度量值进行标准化处理;在此基础上结合权重设置对轨迹相似度进行计算,用于综合衡量轨迹间的相似性。设轨迹集合中有 $n+1$ 条轨迹,当前的聚类中心轨迹为 T_{n+1},则进行轨迹聚类时需要分别计算集合中其他轨迹(记为 $T_i,i\in[1,n]$)与 T_{n+1} 之间的轨迹相似度。首先计算出 T_i 与 T_{n+1} 在方向、速度、空间以及时间方面的差异度量值,然后对这 4 个特性的度量值进行 Z-score 标准化处理。以时间差异的度量值 $timeD$ 的标准化为例,假设 T_i 与 T_{n+1} 对应的起止时间段的不重合时长 $timeD$ 为 $x_i,i\in[1,n]$,则对 x_i 的 Z-score 标准化处理过程如下[15]。

1)计算绝对偏差的平均值

$$S=\frac{1}{n}(|x_1-m|+|x_2-m|+\cdots+|x_n-m|) \tag{6}$$

其中,m 为 x_i 的均值:

$$m=\frac{1}{n}(x_1+x_2+\cdots+x_n) \tag{7}$$

2)计算标准度量值 Z_i

$$Z_i=(x_i-m)/S \tag{8}$$

其中,Z_i 为标准化后的度量值,表示原始度量值 x_i 偏离平均值的程度,服从正态分布。

3.2.2 轨迹相似度计算

根据上述轨迹之间在方向、速度、时间、空间 4 个方面的差异度量以及标准化结果,给出轨迹 T_i 和轨迹 T_j 之间的相似度计算方式:

$$Sim(T_i,T_j)=\frac{1}{\omega_1*Z_{dirD}+\omega_2*Z_{speD}+\omega_3*Z_{spaD}+\omega_4*Z_{timeD}} \tag{9}$$

其中,$\omega_1,\omega_2,\omega_3,\omega_4$ 分别表示轨迹间的差异在方向、速度、时间和空间上的权重,权重的设置可依据用户对轨迹 4 个方面特性的隐私敏感度来决定,各权重之和为 1;Z_{dirD},Z_{speD},Z_{spaD},Z_{timeD} 分别为 T_i 和 T_j 之间在上述 4 个特性的度量值 $dirD,speD,spaD,timeD$ 标准化后的结果。

3.3 轨迹聚类

轨迹聚类的目的是将轨迹集合中相似度较高的轨迹聚类到同一集合中,聚类集合中轨迹间的相似度越高,空间平移所带来的信息损失就越小,匿名后的轨迹数据的可用性就越高。轨迹聚类的过程为:每次迭代从未聚类的轨迹集合(S_{unclu})中随机选择一条轨迹作为聚类中心轨迹 T_p,根据轨迹间的相似度从 S_{unclu} 中选出与 T_p 相似度较高的 $k-1$ 条轨迹组成一个大小为 k 的轨迹集合 S_{now},并将其添加到聚类集合 S_{clu} 中,重复上述聚类操作直到 S_{unclu} 中轨迹数($|S_{\mathrm{unclu}}|$)不足 k,即无法达到 k 聚类的条件为止。轨迹聚类算法的具体过程如算法 2 所示。

算法 2 轨迹聚类算法 TraClu($S,k,t_{\mathrm{tol}},\omega$)

输入:原始轨迹数据集合 S,隐私保护度 k,可容忍时间误差 t_{tol},时空信息权重分配情况 ω

输出:轨迹聚类集合 S_{clu}

```
Begin
1.  S_unclu ← S;
2.  While(|S_unclu| ≥ k)
3.      T_p ← select one Trajectory form S_unclu;
4.      S_now ← T_p;
5.      S_unclu ← S_unclu \ T_p;
6.      For each T_i in S_unclu
7.          Calculate Sim(T_i, T_p);
8.      End for
9.      While(temp ≤ k-1)
10.         T_s ← Select max Sim(T_i, T_p);
```

```
11.      S_now←T_s;
12.      S_unclu←S_unclu\T_s;
13.      temp++;
14.    End while
15. S_clu←S_now;
16. End while
End
```

3.4 空间平移

空间平移的目的是实现轨迹聚类集合中轨迹的 k-匿名，使匿名轨迹聚类集合中的任意一条轨迹与该聚类集合中的其他 $k-1$ 条轨迹无法区分。空间平移算法如算法 3 所示。根据轨迹聚类结果 S_{clu} 对每个轨迹聚类集合中位于匿名区域外的轨迹进行空间平移操作，以保证匿名后的轨迹满足：1）同一聚类集合中的轨迹在同一个匿名区域内；2）同一聚类集合中的轨迹所处的时间段相同。

算法 3　空间平移算法 TraTrans(S_{clu}，δ)

```
输入：轨迹聚类集合 S_clu，不确定区域半径 δ
输出：匿名轨迹集合 S_anon
Begin
1. For each clustering in S_clu
2.   T_p←get the center trajectory of clustering set;
3.   For each Trajectory T in S_clu
4.     If(t_s!=t_ps ‖ t_e!=t_pe)
5.       add or del point;
6.     End If
7.     For each point in T
8.       If(dis(P_i,P_pi)>δ)
9.         move P_i into anonymous region
10.      End If
11.    End For
12.  End For
13. End For
End
```

如算法 3 所示，空间平移算法 TraTrans(S_{clu}，δ)首先以 T_p 为参照，针对聚类集合中的其他轨迹，通过增删其轨迹点使这些轨迹与 T_p 所处的时间段相匹配（步骤 4－步骤 6），然后将该聚类集合中位于 T_p 匿名区域之外的轨迹平移至区域内，所采用的平移方式如图 2 所示。

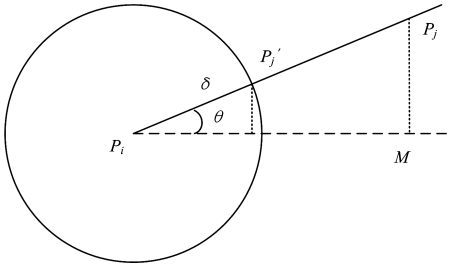

图 2　轨迹点空间平移示意图

Fig. 2　Schematic diagram of trajectory point space translation

如图 2 所示，P_i 为聚类中心轨迹 T_p 上 t 时刻对应的轨迹点位置，P_j 为该聚类集合中一条非聚类中心轨迹在 t 时刻的轨迹点位置。根据平移距离最短原则，P_j 应平移至图 2 中 P_j'的位置。若 P_i 的位置坐标为(x_i, y_i)，P_j 的位置坐标为(x_j, y_j)，则 P_j'的坐标(x_j', y_j')可通过式(10)和式(11)计算得到。

$$x_j' = x_i + \delta \times \frac{|P_iM|}{|P_iP_j|} = x_i + \delta \times \frac{|x_j - x_i|}{\sqrt{(x_j - x_i)^2} + \sqrt{(y_j - y_i)^2}} \tag{10}$$

$$y_j' = y_i + \delta \times \frac{|P_jM|}{|P_iP_j|} = y_i + \delta \times \frac{|y_j - y_i|}{\sqrt{(x_j - x_i)^2} + \sqrt{(y_j - y_i)^2}} \tag{11}$$

4 实验与分析

为验证 PPMCT 算法的有效性，从轨迹数据发布和应用的角度出发，将时空查询错误率和频繁序列匹配度量作为评判匿名轨迹数据可用性的标准，在此基础上进行轨迹隐私保护的相关实验。

4.1 实验数据和环境

OLDEN 数据集由 Brinkhoff 移动对象产生器基于奥尔登堡地图产生[11-14]，是文献[11-12]采用的标准数据集，其轨迹中相邻轨迹点间的时间差为 10 min。本文从中选择前 500 条轨迹数据构成数据集进行实验，数据集的统计信息如表 1 所列。

表 1　实验数据集统计信息

Table 1　Statistical information of experimental data sets

参数	取值
轨迹点数	28439
轨迹数	500
数据大小/kB	1368
最长轨迹点数	145
最短轨迹点数	5

实验用 Java 语言在 Eclipse 开发平台上对 PPMCT 轨迹隐私保护算法加以实现，并对其性能进行测试。实验环境为：处理器 Intel(R) Core i7，主频 2.6 GHz，操作系统 Windows 7，物理内存 4 GB。

4.2 数据可用性的衡量标准

轨迹数据发布的目的主要是供研究人员或位置服务提供商进行查询，或用于进行用户行为模式方面的研究[2,12]。与文献[12]类似，本文将时空查询错误率和频繁序列匹配度量作为评判匿名轨迹数据可用性的标准。

1）时空查询错误率

本文所针对的时空查询指的是在轨迹所在的空间区域内随机生成一个查询区域，在轨迹所处的总运行时间段内随机生成一个时间段作为查询时间段，根据生成的查询时间段和查询区域信息在轨迹数据集合上进行查询，并统计满足条件的轨迹数量作为查询结果。在给定查询时间段内的任一时刻，可能经过查询区域 PSI 的轨迹均视为满足查询条件[12]。实验中根据生成的查询时间段和查询区域信息在原始轨迹数据集合和匿名后轨迹数据集合上分别进行 PSI 时空查询，并由查询统计结果计算匿名轨迹数据集合对应的 PSI 时空查询错误率。时空查询错误率越低，匿名数据的可用性越高。设 $RealNum$ 表示在真实轨迹数据集合上查询到满足查询条件的轨迹数，$AnonNum$ 表示在匿名轨迹数据集合上查询到满足

查询条件的轨迹数，则匿名轨迹数据集合对应的 PSI 时空查询错误率可通过式(12)求得[12]。

$$PSI\ error\ ratio=\frac{|RealNum-AnonNum|}{RealNum} \tag{12}$$

2)频繁序列匹配度量

将原始轨迹数据集合对应轨迹所在的空间区域划分成 $m\times n$ 的网格区域，则轨迹集合中的任何一条轨迹均可转化为一个区域序列，每个区域序列即为一种频繁序列模式[12]。假设从原始轨迹数据集合中提取出的频繁序列模式集合为 S，从匿名轨迹数据集合中提取出的频繁序列模式集合为 S'，$|S|$ 和 $|S'|$ 分别表示集合 S 和集合 S' 中频繁序列模式的个数，那么原始轨迹数据集合与匿名轨迹数据集合之间的序列模式匹配度可以用频繁序列模式匹配度 *F-Measure* 来衡量[12]。

$$F\text{-}Measure=\frac{2\alpha\beta}{\alpha+\beta} \tag{13}$$

其中，$\alpha=|S'\cap S|/|S'|$，$\beta=|S'\cap S|/|S|$。*F-Measure* 值越大，表示从匿名轨迹数据集合与原始轨迹数据集合中提取出的频繁序列模式的匹配度越高(两个数据集合所反映的用户行为模式的差异越小)，匿名数据集合的数据可用性越好。

4.3 实验结果与分析

文献[14]的算法针对的是差异性不足引起的用户隐私泄漏风险的问题，没有考虑到数据的可用性。文献[13]提出的 GC-DM 算法在进行数据隐私保护时没有考虑轨迹数据本身的不确定性；且对于数据可用性的定义及评价，其更多的是考虑隐私保护过程中删除轨迹数据的比例、删除位置信息的比例以及总的信息损失。本文所提出的算法在进行数据隐私保护时结合了轨迹数据本身的不确定性，且考虑到数据的价值在于其可被用户使用，因此本文算法主要从是否利于用户对数据进行使用(数据查询和数据挖掘)的角度出发将时空查询错误率和频繁序列匹配度量作为评判匿名轨迹数据可用性的标准。相对于文献[13]，本文更多地从数据隐私保护的结果而不是过程加以考虑。基于以上原因，在对算法进行性能评估时，主要将所提方法与文献[12]提出的 W4M_L 算法进行对比。

由于空间平移过程中出于对轨迹时间段匹配的考虑而进行的轨迹点增删等操作可能会对轨迹数据的可用性造成较大的影响，实验中在进行轨迹相似度计算时考虑为时间特性差异设置较高的权重；在此基础上，结合单特性实验和多特性实验对权重值进行设置。首先，通过每次只考虑单个轨迹特性的单特性实验评估算法中不同特性对轨迹数据可用性的影响程度，并根据实验结果将轨迹相似性在方向、速度、时间和空间 4 个方面的权重初始值设置为 0.1，0.1，0.5 和 0.3；在此基础上，通过多特性实验验证权重设置的合理性，分析和综合多次实验的结果，并据此对权重值不断进行调整，最终将方向、速度、时间和空间 4 个特性的权重分别设为 0.1，0.1，0.6 和 0.2，在后续实验中，都采用这一权重设置。实验中，本方算法与文献[12]的设置一致，将轨迹数据不确定性区域半径 δ 设置为 600 m；在进行频繁序列统计时，将原始轨迹数据集合对应的空间区域划分成 10×10 的网格区域；将时空查询的时长范围设置为 20～100 min。PPMCT 算法和 W4M_L 算法在不同隐私保护度 k 下的时空查询错误率如图 3 所示，其中时空查询错误率数据均是查询 1000 次所得的 *PSI error ratio* 结果的平均值。

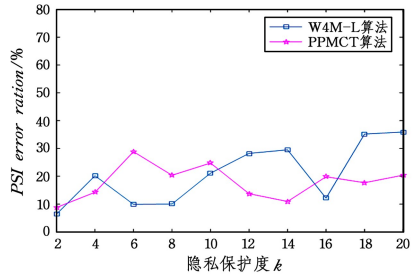

图 3 时空查询错误率

Fig. 3 Spatial temporal query error rate

由图 3 可知，PPMCT 算法与 W4M_L 算法在隐私保护度 k 处于 2～20 之间时对应的平均时空查询错误率相当，这说明在轨迹匿名过程中二者对原始数据的变更程度相当，匿名后的数据在查询时所体现出的数据可用性没有明显差别。其中，PPMCT 算法对应的时空查询错误率的波动幅度为 8.77%～28.84%，W4M_L 算法对应的时空查询错误率的波动幅度为 6.38%～35.9%，前者相较于后者所对应的时空查询错误率在不同隐私保护度 k 下的波动更小。查询错误率是体现轨迹数据可用性的一个重要方面，PPMCT 算法比 W4M_L 算法对应的查询错误率波动小，也就意味着经 PPMCT 算法匿名后的轨迹数据在数据可用性方面的性能更稳定。

图 4 为 PPMCT 算法和 W4M_L 算法在不同隐私保护度 k 下频繁序列匹配度的结果对比图。由图 4 可知，W4M_L 算法和 PPMCT 算法对应的 *F-Measure* 值均随着隐私保护度 k 的增大整体呈现下降趋势。因为随着隐私保护度 k 的增大，空间平移过程中被移动的轨迹点增多且移动距离增大，从匿名轨迹数据中提取的频繁序列的模式集合与从原始轨迹数据中提取的频繁序列的模式集合之间的交集变小。此外，在相同隐私保护度 k 下，PPMCT 算法对应的 *F-Measure* 值较 W4M_L 算法对应的 *F-Measure* 值高，说明相较于 W4M_L 算法，通过 PPMCT 算法匿名化的轨迹数据集合与从原始轨迹数据集合提取出的频繁序列的模式匹配度更高，其所反映的用户行为模式与原始数据集合对应的用户行为模式的差异更小，因此在用户行为模式挖掘方面的数据可用性也更高。

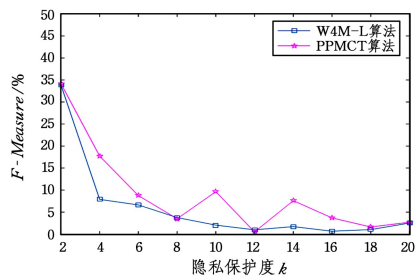

图 4 频繁序列模式匹配度 *F-Measure*

Fig. 4 Pattern matching ratio of frequent sequences *F-Measure*

此外，在算法的效率方面，由于数据发布过程中的轨迹隐

私保护问题通常针对的是离线轨迹，因此对完成轨迹隐私保护所需的时间并没有过高的要求。PPMCT 算法与 W4M_L 算法完成轨迹隐私保护所需的时间相差不多，在本实验中二者完成轨迹隐私保护所需的最长时间分别为 13 s 和 6 s。由于 PPMCT 算法和 W4M_L 算法均是采用随机选取聚类中心轨迹的方式进行匿名前的轨迹聚类操作，因此算法每次执行所获得的匿名结果会有一定的随机性，但从多次实验的结果及其分析可得出结论：PPMCT 算法能达到与 W4M_L 算法相当的时空查询错误率水平，且在与原始轨迹数据集合的频繁序列模式的匹配度方面具有更优的性能。

结束语　本文针对现有基于聚类的隐私保护算法在度量轨迹间的相似性进行聚类时没有充分考虑轨迹多方面的特性可能导致的匿名后轨迹数据可用性较低的问题，提出基于轨迹多特性的隐私保护（PPMCT）算法。PPMCT 算法考虑了轨迹在方向、速度、时间以及空间 4 个方面的差异，综合衡量了轨迹之间的相似性，并在此基础上进行轨迹聚类，以空间平移的方式对聚类后的轨迹进行匿名化操作。实验结果表明，在满足相同隐私保护度的情况下，PPMCT 算法的时空查询错误率与 W4M_L 算法相当，且 PPMCT 算法在与原始轨迹数据集合的频繁序列模式匹配度方面的性能更胜一筹，说明在满足隐私保护需求的前提下，PPMCT 算法能够使匿名后的轨迹数据整体具有更高的数据可用性。将数据匿名与数据可用性相结合是今后的一个研究方向。

参 考 文 献

[1] HUO Z，MENG X F. A Survey of Trajectory Privacy Preserving Techniques [J]. Chinese Journal of Computers，2011，34(10)：1820-1830. (in Chinese)
霍峥，孟小峰. 轨迹隐私保护技术研究[J]. 计算机学报，2011，34(10)：1820-1830.

[2] LEI K Y，LI X H，LIU H，et al. Dummy Trajectory Privacy Protection Scheme for Trajectory Publishing based on the Spatio-temporal Correlation [J]. Journal on Communications，2016，37(12)：156-164. (in Chinese)
雷凯跃，李兴华，刘海，等. 轨迹发布中基于时空关联性的假轨迹隐私保护方案[J]. 通信学报，2016，37(12)：156-164.

[3] XU T，CAI Y. Exploring Historical Location Data for Anonymity Preservation in Location-Based Services[C]// Proceedings of IEEE Conference on Computer Communications. New York：IEEE Press，2008：547-555.

[4] HWANG R H，HSUEH Y L，CHUNG H W. A Novel Time-Obfuscated Algorithm for Trajectory Privacy Protection [J]. IEEE Transactions on Services Computing，2014，7(2)：126-139.

[5] ZHAO J，ZHANG Y，LI X H，et al. A Trajectory Privacy Protection Approach via Trajectory Frequency Suppression [J]. Chinese Journal of Computers，2014，37(10)：2096-2106. (in Chinese)
赵婧，张渊，李兴华，等. 基于轨迹频率抑制的轨迹隐私保护方法[J]. 计算机学报，2014，37(10)：2096-2106.

[6] NERGIZ M E，ATZORI M，SAYGIN Y，et al. Towards Trajectory Anonymization：A Generalization-based Approach [J]. Transactions on Data Privacy，2009，2(1)：47-75.

[7] YANG J，ZHANG B，ZHANG J P，et al. Personalized Trajectory Privacy Preserving Method based on Graph Partition[J]. Journal on Communications，2015，36(3)：1-11. (in Chinese)
杨静，张冰，张健沛，等. 基于图划分的个性化轨迹隐私保护方法[J]. 通信学报，2015，36(3)：1-11.

[8] NERGIZ M E，ATZORI M，SAYGIN Y. Towards Trajectory Anonymization：A Generalization-based Approach[C]// Proceedings of ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS. California：ACM，2008：52-61.

[9] DOMINGO F J，SRAMKA M，TRUJILLO-RASUA R. Privacy-preserving Publication of Trajectories Using Microaggregation [C]// Proceedings of ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS. San Jose：ACM，2010：26-33.

[10] DOMINGO F J，SRAMKA M，TRUJILLO-RASUA R. Micro-aggregation and Permutation-based Anonymization of Movement Data[J]. Information Sciences，2012，208(21)：55-80.

[11] ABUL O，BONCHI F，NANNI M. Never Walk Alone：Uncertainty for Anonymity in Moving Objects Databases[C]// Proceedings of the 24th IEEE International Conference on Data Engineering. Cancun：IEEE，2008：376-385.

[12] ABUL O，BONCHI F，NANNI M. Anonymization of Moving Objects Databases by Clustering and Perturbation [J]. Information Systems，2010，35(8)：884-910.

[13] WANG C，YANG J，ZHANG J P. Privacy Preserving Algorithm based on Trajectory Location and Shape Similarity [J]. Journal on Communications，2015，36(2)：144-157. (in Chinese)
王超，杨静，张健沛. 基于轨迹位置形状相似性的隐私保护算法[J]. 通信学报，2015，36(2)：144-157.

[14] GUO X D，WU Y J，YANG W J，et al. L-diversity Algorithm for Privacy Preserving Trajectory Data Publishing [J]. Computer Engineering and Applications，2015，51(2)：125-130. (in Chinese)
郭旭东，吴英杰，杨文进，等. 隐私保护轨迹数据发布的 l-差异性算法[J]. 计算机工程与应用，2015，51(2)：125-130.

[15] SHI L K，ZHANG Y R，ZHANG X. Trajectory Data Clustering Algorithm based on Spatio-temporal Pattern [J]. Journal of Computer Applications，2017，37(3)：854-859. (in Chinese)
石陆魁，张延茹，张欣. 基于时空模式的轨迹数据聚类算法[J]. 计算机应用，2017，37(3)：854-859.