基于指数机制的轨迹差分隐私保护方法

焦荟聪，刘文菊，王赜
天津工业大学计算机科学与技术学院，天津 300384

摘要
针对传统轨迹数据保护中忽略位置点携带的语义信息带来的隐私泄露问题，提出一种基于指数机制的轨迹差分隐私保护方法。针对位置空间属性及位置语义特征双重属性信息导致的隐私泄露，根据差分隐私中指数机制的特性，为位置点设计可用的打分函数后随机化输出，对轨迹进行了有效的隐私保护。该方法在保证位置隐私的同时减小数据集规模，并防止语义背景推断攻击，提高数据可用性。在真实轨迹数据集上进行实验，实验结果表明，该方法可以保证隐私强度，有效保护了用户的停留区域位置隐私，同时有效提高了数据可用性。

关键词
差分隐私；时空轨迹；语义位置；指数机制

中图分类号：TP391 文献标志码：A doi: 10.11959/j.issn.2096-0271.2022042

Trajectory differential privacy protection method based on exponential mechanism

JIAO Huicong, LIU Wenju, WANG Ze
College of Computer Science and Technology, Tiangong University, Tianjin 300384, China

Abstract
A trajectory differential privacy protection method based on exponential mechanism was proposed, aiming at the problem of privacy disclosure caused by ignoring semantic information carried by location points in traditional trajectory data protection. For the privacy disclosure caused by the dual attribute information of geographic features and semantic features of location, an available scoring function for location points was designed according to the characteristics of the index mechanism in differential privacy. And the function randomized the output to protect the trajectory effectively privacy. This scheme could reduce the size of data sets while ensure location privacy, prevent semantic background inference attacks and improve data availability. Experiments were carried out on real trajectory data sets, and the experimental results showed that the proposed method not only effectively protected the privacy of the user's stay area location, but also effectively improved the data availability while ensured the privacy intensity.

Key words
differential privacy, space-time trajectory, semantic location, index mechanism

0 引言

基于位置的服务近年来不断地出现在各个社交软件中，随着大数据时代的到来，多种软件带有全球定位系统（global positioning system，GPS），便于数据提供商收集用户信息进行数据分析和数据挖掘，以便为用户提供更优质的服务。连续的基于位置的服务（location-based service，LBS）签到形成一系列带有时间属性的轨迹数据，这些轨迹信息携带很多可通过相关性及某些属性信息带来推断攻击的信息，造成用户敏感信息隐私泄露，带来很大的安全隐患。为了达到保护隐私轨迹数据的目的，相比于仅对单个位置点进行保护，对整条轨迹进行保护才是当下更重要的问题。目前针对轨迹的隐私保护方法可以大致分为3类：泛化（匿名区域）、抑制（抑制敏感位置）、扰动（差分隐私，生成假位置）。参考文献[1]通过保护停留点，使用k匿名算法构建匿名区域的方式保护了敏感信息，同时降低了整个轨迹暴露的概率。参考文献[2]提出基于k-means++的轨迹(k，l,δ)-匿名算法，能有效抵抗轨迹相似性攻击。空间匿名技术具有较高的隐私保护水平，部署简单，计算开销较小，但是要以牺牲服务质量为代价。参考文献[3]依据语义位置流行度、用户设定的敏感语义位置类型及语义安全阈值对轨迹停留位置进行空间匿名，构建语义安全匿名区域，防止语义推断攻击。参考文献[4]以局部抑制代替全局抑制的方式实现轨迹数据的隐私保护，降低数据损失率并提高轨迹数据的可用性。参考文献[5]通过离散度控制假位置的分布情况，生成语义安全且分布稀疏的匿名集。然而，这些方法虽然可以保护数据的隐私，但它们都需要特殊的攻击假设和背景知识，而且无法提供一种有效且严格的方法来证明其隐私保护水平。

2008 年Dwork C[6]提出了一种更加严格的可证明隐私定义，即差分隐私保护方法。差分隐私由于其严格的可证明的安全性被广泛应用于数据隐私保护领域，目前逐渐成为主流的位置隐私保护方法。它主要有3个优点：一是用户的隐私泄露风险与攻击者掌握的背景知识量无关；二是它基于严格的数学证明，提供的隐私保护水平可以定量分析；三是隐私保护需要添加的噪声大小可以通过调整差分隐私预算来控制，不会随着数据集的变化而变化。参考文献[7]综合考虑轨迹上位置点的分布规律，利用希尔伯特曲线的性质对空间上的位置点进行降维映射，对空间点离散模式进行研究，提出了一种空间划分方法，划分后的区域全部使用其区域中心点来替换，对聚合后的轨迹进行发布，但是未考虑语义位置对聚类产生的语义推断攻击。参考文献[8]利用最小描述长度（minimum description length，MDL）算法对整条轨迹进行精简处理，利用前缀树存储轨迹段信息，对轨迹段计数进行差分隐私保护。但是由于前缀树结构假设数据有很多相同的前缀，而实际轨迹中位置点大多是分散的，因此该方法并不理想。参考文献[9]根据空间稀疏性以及最大运行速度提出两种空间攻击模型，以用户最大运行速度为约束提出了一种有效的一致性处理方法，使用四叉树对空间进行数据索引以及使用R树对路网进行数据索引发布。参考文献[10]提出了隐式位置的概念，建立了推演泄露模式的模型，通过位置替换和抑制对敏感位置进行匿名，同时考虑了用户行为模式和轨迹特征提出了两种约束模型。基于此，为了避免轨迹发布过程中语义信息带来的隐私泄露，

同时提高轨迹发布的可用性，本文将位置空间属性和位置语义特征属性相结合，提出一种基于指数机制的轨迹差分隐私保护方法。

1 相关定义

1.1 语义轨迹数据库

语义是具有坐标位置的一个很重要的属性，本文通过高德地图公开的应用程序接口（application program interface，API）获取每一个位置的地理标签及语义类别编码，将语义分为如图1所示的3个类别，以餐饮服务为例，餐饮服务属于大类，餐饮服务包括快餐厅，快餐厅包含肯德基、麦当劳等具体快餐品牌。本文共有23个大类、800多个小类。语义轨迹数据代表空间中一个移动对象的运行轨迹，包含多个时空位置点。单个位置的数据结构l={userid，lon，lat，time，sen_type，sen_typecode}，其中userid代表用户ID，lon代表经度，lat代表纬度，time代表时间，sen_type代表语义类型，sen_typecode代表语义类型编码。语义时空轨迹数据库是由$|D|$条语义轨迹构成的数据集，$D=\{T_1,T_2,\cdots,T_n\}$。

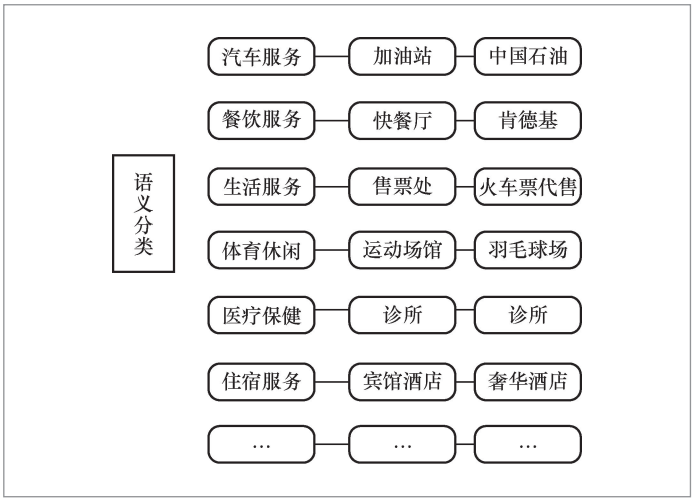

图1 语义分类

1.2 语义位置敏感度

信息熵的概念起源于香农信息论[11]，是一种可应用在很多领域的隐私度量方法，熵值越大，说明该位置的不确定性越高[12]，基本计算式如式（1）所示：

$$H(x)=-\sum_{i=1}^{n}p(x_i)\log_2 p(x_i) \quad (1)$$

设sloc是一个语义位置，用户u_i对位置sloc的访问总次数记为n_i，对所有位置的访问总次数记为N。因此，用户u_i对位置sloc访问总次数占所有位置被访问总次数的比例可表示为式（2）：

$$p(u_i.\text{sloc})=\frac{n_i}{N} \quad (2)$$

根据式（1）可以计算出位置信息熵，如式（3）所示：

$$H(\text{sloc})=-\sum_{i=1}^{n}p(u_i.\text{sloc})\log_2 p(u_i.\text{sloc}) \quad (3)$$

不同于传统概念中针对所有用户进行计算，本文根据不同用户需求，设计语义位置敏感度计算式，如式（4）所示：

$$\text{Sen}(u_i.\text{sloc})=\frac{p(u_i.\text{sloc})}{H(\text{sloc})} \quad (4)$$

Sen(u_i.sloc)的值越大，该位置对于当前用户越重要，隐私保护水平越高。

1.3 差分隐私

定义1 差分隐私[12-13]给定随机算法M及相邻数据集D_1和D_2，其中D_1和D_2相差

1条数据。对于算法M在数据集D_1和D_2上的任意输出O，若满足式(5)，则称随机算法 M 满足ε-差分隐私。

$$\Pr[M(D_1)\in O]\leqslant \Pr[M(D_2)\in O]\times e^{\varepsilon} \quad (5)$$

其中，Pr[·]由随机算法M控制，表示隐私被披露的风险。ε是一个可以调节的参数，称为隐私预算，它决定隐私保护的程度和发布数据集的精度，ε越接近0，算法M在D_1 和 D_2上输出相同结果的概率越接近1，隐私保护程度越高，相应的发布数据集的精度也就越低，可用性越低。

定义2 全局敏感度。对于任意一个查询函数$f: D\rightarrow R_d$ ，D表示数据集，R表示所映射的实数空间，d表示函数f的查询维度，该函数作用于数据集D后返回一个d维向量，其全局敏感度如式(6)所示：

$$\Delta f = \max_{D_1,D_2} \| f(D_1)-f(D_2)\|_1 \quad (6)$$

其中，D_1 和D_2 是任意两个相邻数据集。

定义3 指数机制。给定数据集D及随机算法M，M的输出为实体对象$r(r\in R)$，用$u(D,r)$表示打分函数，用来评估输出值r的优劣程度，Δf表示打分函数的全局敏感度，如果M以正比于P的概率从R中选择并输出r，那么M满足ε-差分隐私，如式(7)所示：

$$P = e^{\frac{\varepsilon\times u(D,r)}{2\Delta f}} \quad (7)$$

1.4 停留区域

GPS愈来愈精确。在海量的轨迹数据中，每条轨迹位置点拥有很大的体量，但并非所有位置点都是有意义的，用户频繁且长时间停留的区域容易暴露用户的行为模式[14]，只有能代表用户行为模式的位置点具有代表性。本文利用已有文献中提出的兴趣区域的概念，设置时间阈值和距离阈值，将用户长时间停留的相近位置点集合定义为停留区域，对停留区域包含的实际位置点进行隐私保护。

1.5 基于语义背景知识的攻击

一个位置，除了地理上的形状还有一些语义信息(如学校、医院、电影院等)，这些语义信息依赖于这个位置上的实体。攻击者将地理空间信息和语义信息相结合，可以有效推测出用户的位置。基于位置语义背景知识的攻击中最常见的是语义推断攻击[15]，当发布的轨迹数据集中包含大量的用户敏感语义位置时，攻击者就可以根据此语义进一步推测出用户的其他敏感信息。例如，用户A的简单轨迹记录示例见**表1**。用户A一个月的轨迹记录中频繁出现某小区到某医院的轨迹，攻击者结合地理信息等其他公开信息可以推测出该用户的家庭地址，并知晓该用户是患者或者职业是医生；再结合轨迹中的时间信息，前往医院的时间符合上下班的规律，可以初步推断用户A是一名在职医生；再通过其他信息得知用户A的家庭信息、个人爱好等，进而窃取用户敏感信息。

2 基于指数机制的轨迹差分隐私保护方法

差分隐私的指数机制是面向非数值性数据的一种隐私保护机制。指数机制以一定的概率值返回一个数值，从而实现差分隐私，概率值由打分函数确定。笔者基于此提出一种基于位置空间属性和位置语义的轨迹隐私保护方法，并根据指数机制特性进行设计，减少位置语义信息带来的隐私泄露问题。

为了抵抗语义推断攻击，本文通过抑制发布和差分隐私技术相结合来设计算法。指数机制核心概念是设计打分函数，对于数据集合中的语义位置来说，对每个位置进行打分，打分越高的位置输出概率越高，同时加入差分隐私技术的随机化来达到隐私保护的目的。本文设计打分函数，使用户位置敏感度较低的位置分数高，这样在随机输出时，敏感语义位置大概率不被发布，同时结合地理位置属性保持空间特征，保护用户隐私。

表1 用户A的简单轨迹记录示例

时间	语义
2021-01-02 08:10	XX小区
2021-01-02 08:20	便利店
2021-01-02 09:10	XX医院
2021-01-02 17:10	饭店
2021-01-02 18:30	XX小区
…	…

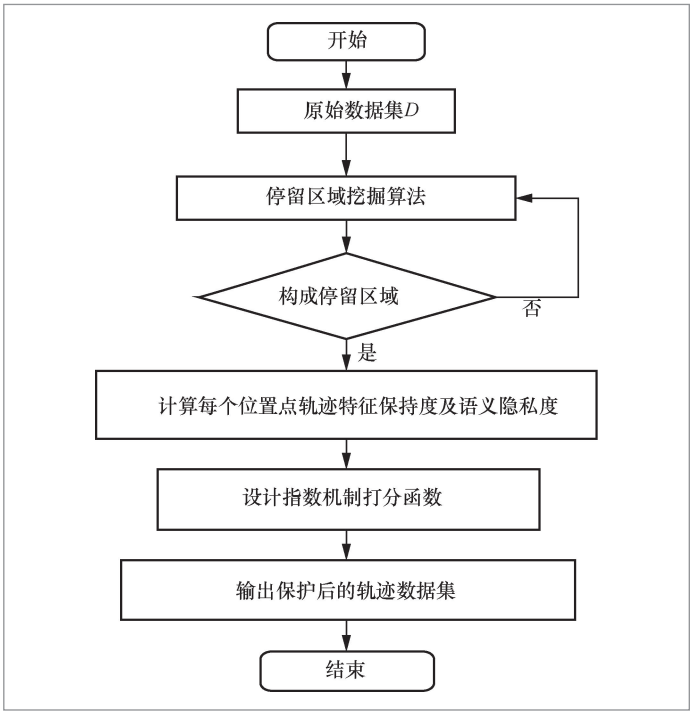

图2 基于指数机制的轨迹差分隐私保护方法框架

2.1 设计思路

本文提出一种基于指数机制的轨迹差分隐私保护方法，主要设计思路如下。

- 根据停留区域挖掘算法得到整条轨迹中的停留区域集合，即需保护隐私的轨迹位置区域。
- 根据改进的MDL算法计算待保护区域中包含所有位置点的轨迹特征保持度，根据语义位置敏感度计算待保护区域中包含所有位置点的语义隐私度。
- 结合语义隐私度和位置轨迹特征保持度设计打分函数，利用指数机制将待保护区域内位置点进行随机输出，删除其他位置点，发布隐私保护后的轨迹。

2.2 算法设计

基于指数机制的轨迹差分隐私保护方法框架如**图2**所示，其中核心部分在于指数机制打分函数设计。

（1）停留区域挖掘

假设某条轨迹$T=\{L_1, L_2, L_3, \cdots, L_n\}$，其中$L_i$代表一个位置点，该位置点携带语义属性及地理属性经纬度信息。设置时间阈值ΔT和距离阈值ΔS。从轨迹第一个位置点开始，将该位置点记作start，并将其加入停留区域area，向后计算下一个位置点与该位置点的时间差Δt。如果Δt大于时间阈值，则计算两点之间的距离差Δs（使用真实经纬度距离计算实际位置距离）；如果Δs小于距离阈值，则将该位置点加入停留区域area，继续遍历下一个位置点，直至到达该条轨迹最后一个位置点，形成一个以start位置点为基础点的停留区域area。接着遍历除已形成的停留区域外的第一个位置点，将其作为start继续挖掘停留区

域，直至整个轨迹的停留区域全部形成。

（2）轨迹特征保持度计算

根据MDL算法计算待保护区域中包含所有位置点的轨迹特征保持程度权重值。在将MDL算法应用于轨迹数据时，D是运动物体的轨迹数据集，H是运动物体轨迹的轨迹分割。本文方法中对停留区域和前后两个位置点构成的局部轨迹进行轨迹特征保持，可利用MDL算法精确度高的特点，计算每个位置点的轨迹特征保持度。单个停留区域示意如**图3**所示，该停留区域中n个位置点和前后两个位置点构成局部特征区域，根据实际坐标计算出这n个位置点分别与前后两个位置点的位置几何权重，将平均值作为该位置的轨迹特征保持度。

对于停留区域内的位置点c_i，有式（8）所示关系：

$$c_0 = s, c_{\text{last}} = e \tag{8}$$

某位置点的模型精度值等于该位置点与前后两个端点的欧氏距离的平均值，如式（9）所示：

$$L(H) = \frac{\left(\text{dis}(s,c_i) + \text{dis}(e,c_i)\right)}{2} \tag{9}$$

本文分两部分计算模型复杂度。取该位置点与区域外邻近的两个位置点（**图3**中s点和e点）分别计算复杂度值，根据MDL算法的定义求出垂直距离和角距离，计算后求平均值，如式（10）~式（13）所示：

$$L_1(D|H) = \sum_{j=0}^{i-1} \log_2\left(\text{dis}_{\perp}(sc_i, c_j c_{j+1}) + 1\right) + \log_2\left(\text{dis}_{\angle}(sc_i, c_j c_{j+1}) + 1\right) \tag{10}$$

$$L_2(D|H) = \sum_{j=0}^{\text{last}-1} \log_2\left(\text{dis}_{\perp}(ec_i, c_j c_{j+1}) + 1\right) + \log_2\left(\text{dis}_{\angle}(ec_i, c_j c_{j+1}) + 1\right) \tag{11}$$

$$L(D|H) = \frac{\left(L_1(D|H) + L_2(D|H)\right)}{2} \tag{12}$$

$$W = L(H) + L(D|H) \tag{13}$$

其中，W为总描述长度，将每个位置点的W作为轨迹特征保持度输出。

（3）语义隐私度计算

为了防止语义推断攻击，将位置语义属性作为保护区域内位置点的隐私度量。原始轨迹数据集中不包括语义属性，为了保证方案的真实性，使用高德地图的API，根据位置点对应实际地图上的经纬度坐标进行语义爬取，采取距离该位置点最近的语义地点标签定义语义属性，保证整个爬取过程距离范围在200 m以内。高德地图将语义分为三大类别，据此对位置点语义属性进行语义编码。挖掘到停留区域时，对包含的所有位置点提取语义属性，根据式（4），基于位置熵的概念计算每个语义位置的语义敏感度，并将其作为语义特征权重，得出每个位置点的语义隐私度。语义隐私度的值越大，代表语义位置隐私需求越高。

（4）指数机制打分函数设计

差分隐私的指数机制关键在于打分函数的设计，打分函数设计的目的是可以以较大的概率输出泄露用户隐私概率最小的位置语义，同时最大限度保持轨迹特征变

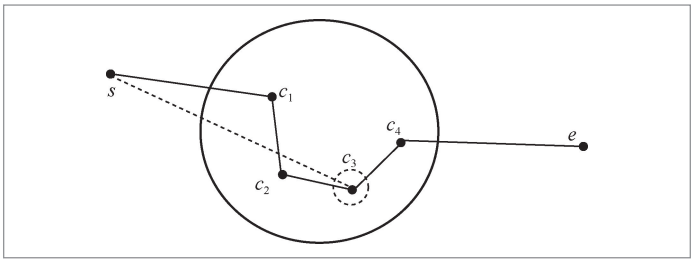

图 3 单个停留区域示意

化，以保证隐私。本文打分函数如式（14）所示：

$$u(D,r)=1-\frac{\text{sen}(u_i.\text{sloc})}{\text{sen}(u_i.\text{sloc})+\text{geo}(u_i.\text{sloc})} \quad (14)$$

其中，D是兴趣区域集合，r是想要输出的抽样的语义位置点。以正比于$p(D,r)$的概率将n种结果随机输出，如式（15）所示：

$$p(D,r)=\frac{e^{\frac{\varepsilon\times u(D,r)}{2\Delta f}}}{\sum_{r\in R}e^{\frac{\varepsilon\times u(D,r)}{2\Delta f}}} \quad (15)$$

其中，ε是隐私预算，Δf是打分函数的全局敏感度，对兴趣区域执行隐私保护算法，输出隐私泄露最小的语义位置，删除其他语义位置。本文将所提方法命名为Index_DP算法，算法伪代码如下。

算法1 Index_DP算法

输入：原始轨迹$T=\{L_1,L_2,\cdots,L_n\}$，轨迹长度N，距离阈值ΔS，时间阈值ΔT

输出：隐私保护后的轨迹T_1

1：for i=1;i<N; i++ do

2：　if $\Delta t(L_i,L_{i+1})<\Delta T$ and $\Delta s(L_i,L_i+1)>\Delta S$ then

3：　将L_i,L_{i+1}加入停留区域A

4：　对于$A=\{L_j,L_{j+1},\cdots,L_{j+n}\}$

5：　if L_{j+n}!=L_n then

6：　将A加入邻近位置，形成带边界区域E

7：　g=MDL(A,E)

8：　s=$-p\times np$.log2(p)

9：　设计打分函数score=1-sen_w/2geo_w

10：　out=index_mechanism(A,pr)

11：　end if

12：　end if

13：end for

14：end

3 实验分析

3.1 实验环境与数据集

本文的实验环境为Windows 10版本操作系统，编程语言使用Python，采用的数据集是真实用户轨迹采样Geolife数据集。Geolife数据集是微软亚洲研究院项目组历时5年多收集的182位用户的GPS轨迹数据集，GPS轨迹由一系列时间戳点表示，每个点包含纬度、经度和时间信息。该数据集包含17 621条轨迹，经过初步筛选，最短轨迹包含300个携带语义属性的位置点，最长轨迹包含2 000个轨迹点。在原始轨迹数据集上，通过高德地图API爬取每个位置点在地图上对应的语义信息，将语义类型编码作为不同语义的象征，并添加到轨迹每个位置点的属性当中，作为轨迹隐私保护的一个属性。

3.2 衡量标准

（1）数据可用性

本文采用动态时间弯曲（dynamic time warping，DTW）[14,16]距离来衡量隐私保护前后轨迹的失真度。DTW是一个动态迭代的过程，原始轨迹$T_1=\{L_1,L_2,L_3,\cdots,L_m\}$和处理后的可发布轨迹$T_2=\{C_1,C_2,C_3,\cdots,C_n\}$的长度分别为$m$和$n$，则两条轨迹之间的DTW距离计算式如式（16）所示：

$$D(T_1,T_2)=D(i,j)+\min\begin{cases}D(i-1,j)+D(i,j)\\D(i-1,j-1)+2D(i,j)\\D(i,j-1)+D(i,j)\end{cases} \quad (16)$$

(2)隐私保护度

使用差分隐私中的指数机制对轨迹进行隐私保护,分析不同的隐私预算对轨迹可用性的影响来衡量差分隐私方法的隐私保护度。从差分隐私指数机制的定义可以得出,隐私预算与可用性成正比,与隐私保护成反比。

3.3 实验结果

为了分析算法效果,将本文提出的Index_DP算法分别与参考文献[17]提出的ANoise算法以及参考文献[18]提出的GNoise算法进行对比。

(1)不同参数设定下算法性能对比

分析在不同的时间阈值及不同的距离阈值下,Index_DP算法对数据可用性产生的影响。首先分析在时间阈值固定的情况下,距离阈值不同时轨迹的失真度变化趋势。时间阈值为10 min时,距离阈值分别为200 m、400 m、800 m时的失真度变化趋势如**图4**所示。随着距离阈值的增大,轨迹失真度增加,这是由于在停留区域挖掘过程中,距离范围越大,停留区域中包含的轨迹位置点越多,保护的隐私区域越大,导致处理之后的轨迹长度越短,与原始轨迹差异性越大。

分析在距离阈值固定的情况下,时间阈值不同时轨迹的失真度变化趋势。距离阈值为400 m时,时间阈值分别为5 min、10 min、15 min时3种算法的失真度变化趋势如**图5**所示。随着时间阈值的增大,轨迹失真度降低,这是由于搜索的时间越短,保护的隐私区域越小,处理后轨迹失真度越低。

(2)不同阈值下算法性能对比

选取距离阈值为400 m,分别取时间阈值为5 min、10 min、15 min和20 min,在轨迹数据集中随意抽取10条轨迹计算出4个时间阈值下的轨迹失真度后取平均值,3种算法分别进行实验得出结果如**图6**所示。**图6**中横轴对应的是本文实验中选取的4个时间阈值,纵轴对应的是轨迹失真度,失真度的值越高,代表轨迹的可用性越差。由**图6**中可以看出,Index_DP算法在不同时间阈值下的平均失真度均小于其他两种算法。随着时间阈值的增大,3种算法的失真度都大致呈现减小趋势,符合停留区域挖掘算法的阈值影响。在时间阈值为5 min时,Index_DP算法失真度为0.0158;ANoise算法失真度为0.0608,比Index_DP算法高0.045;GNoise算法采用纯加噪声的方法使其失真度最高达到0.209。由此可见Index_DP算法在对轨迹进行隐私保护后,轨迹的信息损失远低于GNoise算法和ANoise算法,大大提高了轨迹可用性。

时间阈值为10 min,距离阈值分别为200 m、300 m、500 m和700 m时3种算法的失真度变化趋势如**图7**所示。由**图7**可知,在距离阈值变化下Index_DP算法在对轨迹进行隐私保护后轨迹的信息损失仍然最小,且远低于GNoise算法和ANoise算法。随着时间阈值和距离阈值的不断变化,Index_DP算法的平均失真度并未产生较大波动,均小于0.05,而ANoise算法的失真度基本为0.05以上。由实验可知,Index_DP算法不仅大大提高了轨迹可用性,且具有很好的延展性,在当前大数据情况下,轨迹长度越来越长,轨迹数量越来越多,本文算法仍有良好的普适性。

(3)不同隐私保护度下算法性能对比

差分隐私的隐私保护程度与隐私预算ε有关,分析不同的ε取值对平均失真度的影响。由于本文方法设计目的是在越来越长的轨迹趋势下保护轨迹隐私,且每个

用户轨迹长度各有不同，本文将用户轨迹大致分为3组进行对比试验，分别为500~1 000 m、1 000~1 500 m以及1 500~2 000 m。在本文数据集中，轨迹长度在1 000~1 500 m的轨迹数量最多，轨迹长度在500 m以下的轨迹较少，因此不考虑该部分轨迹。由**表2**可以看出，不同的轨迹长度在轨迹失真度上的变化没有逐渐增大，1 500~2 000 m分段的轨迹失真度小于500~1 000 m的轨迹失真度，说明Index_DP算法可以适用在更长的轨迹上，具有很好的延展性。

由**表2**可以看出，随着隐私预算ε增大，数据失真度逐渐降低，这是由Index_DP算法设计的打分函数的目标及指数机制的性质决定的。ε越大，选择出期望结果的可能性也越大，轨迹保持更好。打分函数设计的目的首先是保证语义位置的隐私，其次是尽可能地保持轨迹特征，轨迹保持得更好的情况下也会更大概率地泄露敏感位置，隐私预算与可用性成正比，与隐私保护成反比，这个结果符合差分隐私指数机制的规律。由**表2**可以看到大部分的轨迹受隐私预算影响不大，这是因为在该机制中加入的语义属性对打分函数进行了一部分约束，本文主要目的是在保护隐私的情况下尽可能提高轨迹可用性，因此为了同时兼顾语义安全和保持轨迹特征，牺牲了一部分可用性。

4 结束语

本文提出了一种基于指数机制的轨迹差分隐私保护方法。此方法针对轨迹位置隐私保护中往往忽略位置点的语义属性带来的隐私泄露问题，在对轨迹进行保护时不仅考虑其隐私保护，也要保证轨迹本身的特征尽可能不发生较大的改变，保证隐

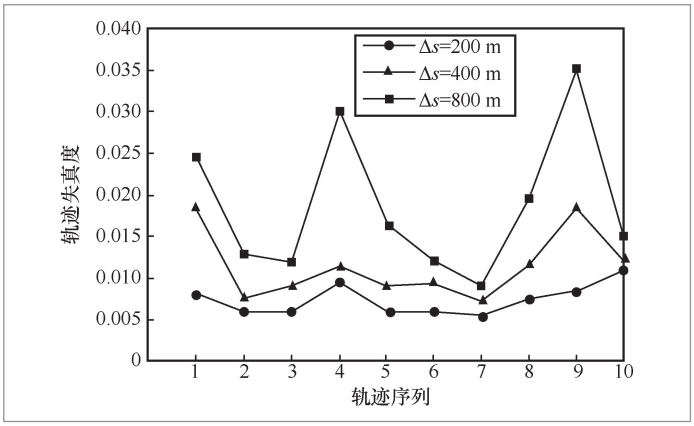

图 4 时间阈值为 10 min 时，距离阈值分别为 200 m、400 m、800 m 时的失真度变化趋势

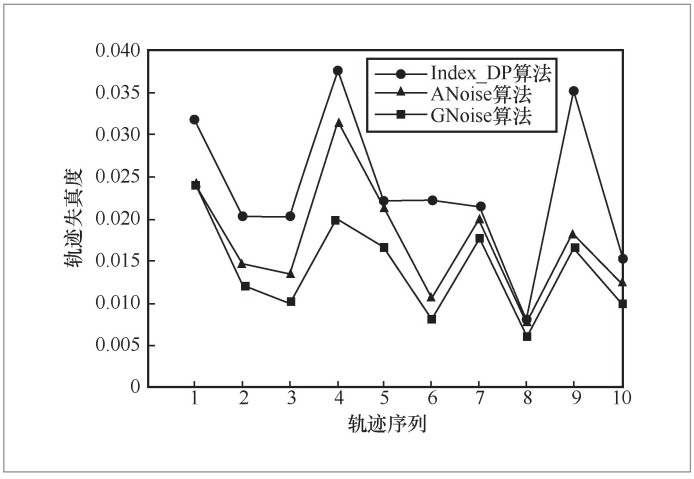

图 5 距离阈值为 400 m 时，时间阈值分别为 5 min、10 min、15 min 时 3 种算法的失真度变化趋势

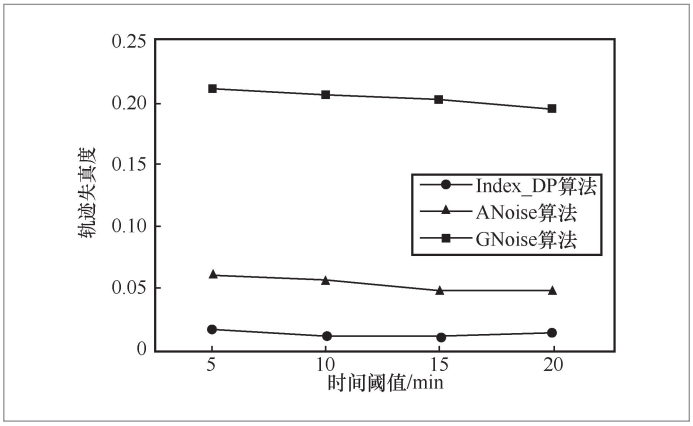

图 6 距离阈值为 400 m，时间阈值分别为 5 min、10 min、15 min 和 20 min 时 3 种算法的失真度变化趋势

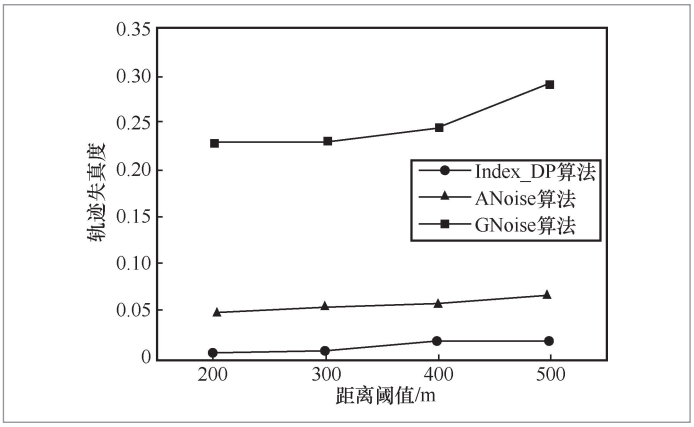

图 7　时间阈值为 10 min，距离阈值分别为 200 m、300 m、500 m 和 700 m 时 3 种算法的失真度变化趋势

表 2　不同隐私预算下的失真度分析

隐私预算 ε	500~ 1 000 m	1 000~ 1 500 m	1 500~ 2 000 m
0.05	0.025011	0.029588	0.015624
0.5	0.025010	0.027752	0.015473
1	0.018261	0.021688	0.012797
5	0.018109	0.022347	0.013085
100	0.016554	0.019989	0.012183
1 000	0.015691	0.019708	0.012155

私保护后轨迹的可用性。根据语义时空轨迹本身的特性，使用差分隐私方法中的指数机制，对停留区域中具有不同隐私水平的位置点进行概率随机化选择输出，由此对轨迹进行隐私保护。基于停留区域的概念，考虑位置点地理几何特征的同时，为了防止语义推断攻击在原始数据集上融入位置语义属性，使用指数机制在保证隐私条件下对区域内语义位置进行抽样选择。在真实数据集上进行的多次仿真实验证明了Index_DP算法具有良好的延展性和普适性，在轨迹保护中有效地提高了数据可用性。由于本文使用了比较复杂的数学模型，下一阶段将考虑如何降低算法的运行时间，提高运行效率。

参考文献:

[1] HUO Z, MENG X F, HU H B, et al. You can walk alone: trajectory privacy-preserving through significant stays protection[C]// Proceedings of International Conference on Database Systems for Advanced Applications. Heidelberg: Springer, 2012: 351-366.

[2] ZHANG X L, YANG W J. Trajectory anonymous algorithm based on k-means++ against similarity attack[J]. Computer Science and Application, 2020, 10(4): 610-618.

[3] 俞望年, 宣占祥, 马小明, 等. 轨迹数据发布中基于敏感语义位置的隐私保护算法[J]. 现代计算机, 2020(27): 3-9.
YU W N, XUAN Z X, MA X M, et al. Privacy protection algorithm based on sensitive semantic location in trajectory data publishing[J]. Modern Computer, 2020(27): 3-9.

[4] 俞庆英, 王燕飞, 叶梓彤, 等. 基于优化局部抑制的轨迹数据发布隐私保护算法[J]. 计算机工程, 2020, 46(8): 112-118.
YU Q Y, WANG Y F, YE Z T, et al. Privacy protection algorithm based on optimized local suppression for trajectory data publication[J]. Computer Engineering, 2020, 46(8): 112-118.

[5] 王辉, 朱国宇, 申自浩, 等. 基于用户偏好和位置分布的假位置生成方法[J]. 计算机科学, 2021, 48(7): 164-171.
WANG H, ZHU G Y, SHEN Z H, et al. Dummy location generation method based on user preference and location distribution[J]. Computer Science, 2021, 48(7): 164-171.

[6] DWORK C. Calibrating noise to sensitivity in private data analysis[J]. Lecture Notes in

Computer Science, 2012, 3876(8): 265–284.

[7] HAN Q, XIONG Z, ZHANG K. Research on trajeetory data releasing method via differential privacy based on spetial partition[J]. Security and Communication Networks, 2018: 1–14.

[8] ZHAO X, PI D, CHEN J. Novel trajectory privacy–preserving method based on prefix tree using differential privacy[J]. Knowledge–based Systems, 2020, 198: 105940.

[9] 霍峥, 孟小峰. 一种满足差分隐私的轨迹数据发布方法[J]. 计算机学报, 2018, 41(2): 400–412.
HUO Z, MENG X F. A trajectory data publication method under differential privacy[J]. Chinese Journal of Computers, 2018, 41(2): 400–412.

[10] 刘向宇, 刘竹丰, 夏秀峰, 等. 一种保持轨迹数据高可用性的隐式位置访问隐私保护技术[J]. 沈阳航空航天大学学报, 2019, 36(2): 66–75.
LIU X Y, LIU Z F, XIA X F, et al. An algorithm of protecting hidden location visit privacy with high trajectory utility[J]. Journal of Shenyang Aerospace University, 2019, 36(2): 66–75.

[11] 彭长根, 丁红发, 朱义杰, 等. 隐私保护的信息熵模型及其度量方法[J]. 软件学报, 2016, 27(8): 1891–1903.
PENG C G, DING H F, ZHU Y J. et al. Information entropy models and privacy protection[J]. Joumal of Software. 2016, 27(8): 1891–1903.

[12] VISVALINGAM M, WHYATT J D. Line generalisation by repeated elimination of points[J]. The Cartographic Journal, 1993, 30(1): 46–51.

[13] MCSHERRY F, TALWAR K. Mechanism design via differential privacy[C]// Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science. Piscataway: IEEE Press, 2007: 94–103.

[14] MYERS C S, RABINER L R. A comparative study of several dynamic time warping algorithms for speech recognition[J]. Bell Labs Technical Journal, 2013, 60(7): 1389–1409.

[15] 马明杰, 杜跃进, 李凤华, 等. 基于语义的位置服务隐私保护综述[J]. 网络与信息安全学报, 2016, 2(12): 1–11.
MA M J, DU Y J, LI F H, et al. Review of semantic–based privacy–preserving approaches in LBS[J]. Chinese Journal of Network and Information Security, 2016, 2(12): 1–11.

[16] JIANG K F, SHAO D X, BRESSAN S, et al. Publishing trajectories with differential privacy guarantees[C]//Proceedings of the 25th International Conference on Scientific and Statistical Database Management. [S.l.:s.n.], 2013: 1–12.

[17] 兰微, 林英, 包聆言, 等. 融入兴趣区域的差分隐私轨迹数据保护方法[J]. 计算机科学与探索, 2020, 14(1): 59–72.
LAN W, LIN Y, BAO L Y, et al. Trajectory–differential privacy–protection method with interest region[J]. Journal of Frontiers of Computer Science and Technology, 2020, 14(1): 59–72.

[18] LEE J G, HAN J W, WHANG K Y. Trajectory clustering: a partition–and–group framework[C]//Proceedings of 2007 ACM SIGMOD international conference on Management of data. [S.l.:s.n.], 2007: 593–604.

作者简介

焦荟聪（1997– ），女，天津工业大学计算机科学与技术学院硕士生，主要研究方向为云计算、隐私保护。

刘文菊（1963– ），女，天津工业大学计算机科学与技术学院教授，主要研究方向为无线网络信息安全、互联网应用系统研发。

王赜（1976– ），男，博士，天津工业大学计算机科学与技术学院教授，主要研究方向为计算机网络与安全、互联网+应用、云网络安全。

收稿日期： 2021-10-21

通信作者： 王赜，wangze@tiangong.edu.cn