一种基于最优轨迹的假查询隐私保护机制

刘燕妮[1,2], 叶阿勇[1,2], 张 强[1,2], 赵云涛[1,2]

[1]福建师范大学计算机与网络空间安全学院 福州 中国 350117
[2]福建省网络安全与密码技术重点实验室 福州 中国 350117

摘要 随着移动通信技术和无线传感器的发展, 基于位置服务的应用给我们的生活带来极大的便利。在实际使用中, 用户需要向不可信的 LBS 服务提供商发送自己的实时位置和相关的查询信息, 这可能会导致用户的个人隐私信息遭到泄露, 特别是在使用连续位置查询服务时, 服务提供商可以利用位置的时空相关性来构建用户的轨迹信息, 进而推断出用户的居住地址、公司位置等敏感信息。传统的位置隐私保护方法通常只考虑到当前位置, 在解决连续位置查询时存在挑战, 因此, 为了解决连续位置查询中难以权衡轨迹可用性与隐私性的问题, 提出一种基于最优位置轨迹的假查询隐私保护机制。首先, 通过真实轨迹和假轨迹间的互信息来度量轨迹的隐私, 解决轨迹隐私难以量化的问题。在此基础上, 提出一种基于马尔科夫链的轨迹互信息计算方法, 简化了轨迹互信息的计算过程,并使用两条轨迹上对应位置点间的欧几里距离来量化位置轨迹的可用性。其次, 考虑到生成的假轨迹可能并不符合用户的通行习惯, 容易被识别出来, 我们选择历史轨迹作为假轨迹。为了减少轨迹上位置点的数量, 使用四叉树法对路网区域进行划分, 将轨迹划分为不同的片段, 在相关约束条件下寻找最优的历史轨迹作为假轨迹, 从而保证使用的假轨迹更加真实、合理。最后, 实验结果表明, 本文的方案可以最大程度的实现位置数据隐私性和可用性平衡, 与其他方案相比, 安全性更高、系统计算开销更少。

关键词 隐私轨迹; 假查询; 互信息; 马尔科夫链
中图法分类号 TP309 DOI 号 10.19363/J.cnki.cn10-1380/tn.2023.11.09

A Privacy Preserving Mechanism with False Queries Based on Optimal Trajectories

LIU Yanni[1,2], YE Ayong[1,2], ZHANG Qiang[1,2], ZHAO Yuntao[1,2]

[1]College of Computer and Cyber Security, Fujian Normal University, Fuzhou 350117, China
[2]Fujian Provincial Key Laboratory of Network Security and Cryptology, Fuzhou 350117, China

Abstract With the development of mobile communication technology and wireless sensor, the application of location-based services has brought great convenience to our life. In actual use, the user needs to send untrusted LBS service provider own real-time position and related query information, this can lead to a user's personal privacy information leakage, especially when using continuous location query service, service providers can take advantage of the location of the spatial and temporal correlation to build a track of user information. In turn, sensitive information such as the user's residential address and company location can be inferred. Traditional location privacy protection methods have challenges in solving continuous location query. Therefore, in order to solve the problem that it is difficult to balance trajectory availability and privacy in continuous location query, an optimal location trajectory based fake query privacy protection mechanism is proposed. Firstly, the privacy of trajectory is measured through the mutual information between real trajectory and false trajectory to solve the problem that trajectory privacy is difficult to quantify. On this basis, a method of track mutual information calculation based on Markov chain is proposed to simplify the calculation process of track mutual information, and the usability of the track is quantified by the Euclidean distance between the corresponding position points on two tracks. Secondly, considering that the generated false track may not conform to the user's traffic habits and be easily identified, we choose the historical track as the false track. In order to reduce the number of position points on the track, the quadtree method is used to divide the road network area and divide the track into different segments. Under relevant constraints, the optimal historical track is found as the false track, so as to ensure that the false track used is more realistic and reasonable. Finally, experimental results show that the proposed scheme can maximize the balance between privacy and availability of location data, with higher security and less system computing overhead compared with other schemes.

Key words trajectory privacy; fake query; mutual information; markov chain

通讯作者: 叶阿勇, 博士, 教授, Email: yay@fjnu.edu.cn。
本课题得到国家自然科学基金(No. 61972096, No. 61771140, No. 61872088, No. 61872090)、福建省高校产学合作项目(No. 2022H6025)资助。
收稿日期: 2022-03-27; 修改日期: 2022-06-08; 定稿日期: 2023-09-02

1 引言

近年来, 随着移动通信技术的快速发展, 诞生了一系列基于位置服务(Location-Based Service, LBS)的应用, 渗透到我们生活的各个方面[1], 给我们带来极大的便利, 如紧急救援、智能交通和目标跟踪等服务。用户在使用连续 LBS 服务时, 需要连续向服务提供商提供自己的位置信息, 由于位置具有时空相关性, 服务提供商可以根据用户上传的多个位置构建出用户的时空轨迹, 通过进一步对生成的轨迹数据进行分析和挖掘, 可以实现交通路线推荐等服务和决策。然而, 轨迹数据中包含用户的隐私信息, 服务提供商或攻击者可以根据用户的时空轨迹判断出用户的家庭地址和工作地点等敏感信息, 进而可以通过数据挖掘等手段推断出用户的兴趣爱好、健康状况、消费水平和宗教信仰等隐私信息[2], 甚至威胁到用户的人身安全。

当前 LBS 所面临的关键问题就是如何保护用户的位置隐私, 为此, 许多隐私保护方法被相继提出[3]。其中, 基于假位置的方法[4-6]是常见的解决方案之一。其主要思想是生成一个或多个与用户真实位置相关联的虚假位置, 然后将虚假位置代替用户的实际位置提交给服务提供商, 这样, 服务提供商就无法获取用户的真实位置。与数据泛化和抑制等方法相比, 基于假位置的方法有以下优点: (1) 不依赖于第三方; (2) 提供精确的查询结果。因此, 基于假位置的方法被广泛应用于保护用户的位置隐私[7-11]。但是在实际应用中, 用户通常需要使用连续 LBS 服务而不是单点 LBS 服务[6]。例如, 在自驾旅游的路上, 用户需要不断查询周围的道路交通情况来决定旅行路线。如果直接采用现有的基于假位置的方案, 服务提供商可以根据位置的相关性, 识别出一些可信度高的假位置, 甚至可以直接获得用户的真实位置。为此, 文献[12]考虑了位置间的时间可达性和方向相似性等因素, 提出一种时空相关感知的隐私保护方案。然而, 该方案没有考虑轨迹的整体性, 其生成的轨迹容易被识别。文献[13]提出了一种基于马尔科夫链的在线轨迹隐私保护方法, 考虑前后位置点间的时空相关性, 实现每个位置的最优选择。文献[14]提出了一种最大化条件熵的位置扰动机制, 结合地理不可分辨性和重映射以提高效用。然而, 文献[13-14]都是基于用户当前的实际位置来生成扰动位置以实现隐私保护, 生成的轨迹只能实现每个位置最优, 无法实现轨迹最优。

因此, 本文提出一种基于最优轨迹的假查询隐私保护机制。首先, 从信息论的角度出发, 通过真实轨迹和假轨迹间的互信息来量化轨迹的隐私。与其他相似性度量方法相比, 互信息可以捕获到变量间非线性的统计相关性, 并且具有空间转换的不变性, 能有效度量变量间的真实依赖性。在此基础上, 提出一种基于马尔科夫链的互信息计算方法, 简化了轨迹互信息的计算。其次, 通过四叉树法对区域进行划分, 将轨迹划分为不同的片段, 在相关约束条件下寻找最优的历史轨迹作为假轨迹, 保证使用的假轨迹更加真实合理。

2 相关工作

近年来, 针对连续 LBS 查询中的隐私保护问题受到了国内外学者的广泛关注。基于连续 LBS 查询的位置隐私保护方法主要有数据泛化[15-18]、数据抑制[19-22]和假数据[4-6,7-11]。

基于数据泛化的隐私保护技术是指用某一隐藏区域来代替轨迹上的位置点进行 LBS 查询, 以达到隐私保护的目的。文献[15]最早提出了泛化区域的概念, 它保证在不同的时刻, 泛化的区域中始终包含 k 个位置, 满足 k-匿名, 并且在该泛化区域内 k 个位置信息无法区分, 进而保护用户的真实位置。但是, 该方法的泛化区域太大, 每一次查询造成的损失较大。因此, 文献[16]在文献[15]的基础上, 通过考虑用户的移动方向和移动速度, 进行最小化泛化区域, 但是该方法并不能保证每次泛化区域内的用户是相同的。Wang 等[17]为了确保用户在每次提交 LBS 请求时构造的匿名区中包含的匿名用户是相同的, 提出了一种基于贪心算法的匿名区构造方法, 为每个位置请求找到最小的泛化区域, 但是这样会使泛化区域的面积随着查询次数的增加而增加, 造成较大的计算开销。Latha 等[18]提出了一种 KRUPTO 算法, 它给出了邻居在地理区域存在的信息, 并且用户的敏感信息也被认为是隐藏区域, 它减少了与用户位置相关的处理延迟和匿名化成本。尽管轨迹数据泛化方法在一定程度上可以保护用户的隐私信息, 但是该隐私保护方法需要引入一个可信的第三方作为位置匿名服务器, 负责将用户的位置转换到隐藏区域中。

基于数据抑制的隐私保护技术是指在用户的轨迹数据中除去用户的敏感位置信息。Gruteser 等[19]研究了隐藏用户在敏感区域位置的信息披露控制算法, 将地图划分为敏感区域和非敏感区域, 通过抑制敏感区域中用户的位置信息来保护用户的信息。Terrovitis 等[20]设计了一种数据抑制技术, 在发布用户轨迹时抑制处于敏感区域的位置信息, 在满足用

户隐私的同时保证所公布的数据尽可能准确、有效。Goetz 等[21]提出了一种 MASKIT 系统, 限制了攻击者可以从过滤后的数据流中了解用户的敏感位置信息。赵婧等[22]提出一种基于轨迹频率抑制的方法对轨迹数据进行匿名处理。保证数据在经抑制处理后维持可用性。然而, 使用抑制法对轨迹中的位置信息进行保护处理虽然简单, 但是如何找到合理抑制的位置信息和降低信息损失是其面临的关键问题。

基于假数据法的隐私保护技术是指在不需要可信第三方的情况下, 用户选择一个虚拟位置代替自己的真实位置发送给 LBS 提供商, 达到位置隐私保护的效果。该思想最早由 Kido 等[7]提出, 用户使用虚假位置生成虚假轨迹, 进而保护用户的真实轨迹。Gao 等[8]从图论的角度出发, 提出了一种用于参与式感知的弹道隐私保护框架 TrPF, 对理论上的混合区模型进行了改进。有效地保护参与者的轨迹隐私, 降低信息损失和成本。Wu 等[9]利用 *k*-匿名技术向用户的位置轨迹中添加虚假数据来生成 *k*–1 条假轨迹, 达到匿名效果。雷凯跃等[10]从轨迹的整体方向、轨迹中相邻位置之间可达及移动距离对单条轨迹中相邻位置间的时空关联性和轨迹间的相似性进行分析, 提出一种假轨迹隐私保护方案, 有效保护轨迹发布中用户的轨迹隐私。Pan 等[11]对轨迹相似度进行研究, 提出了两种新的轨迹相似度度量, 即最大轨迹距离和最小轨迹距离和, 并分析了时空相似度和文本相似度之间的相关性。Zhang 等[4]将差分隐私机制与 *k*-means 算法相结合, 利用 *l*-分集的思想从 *k* 个簇中选择 *l* 个簇中心进行噪声处理, 将噪声位置点作为虚假位置, 避免了恶意攻击者拦截用户隐私信息的问题。Guo 等[5]利用用户的历史邻近位置代替用户的真实位置向 LBS 查询, 但无法为用户的移动轨迹实现足够的匿名保护。Soma 等[6]针对旅行计划, 将用户的虚假位置发送给 LBS, 并从 LBS 中增量检索关于虚假位置的最近兴趣点, 并利用几何特性将搜索空间细化为椭圆区域, 实现最小化行程距离。

3 预备知识

在本节中, 我们介绍了位置假查询机制, 给出了位置轨迹的隐私和可用性度量的相关定义。

3.1 位置假查询机制

本节我们主要介绍了位置假查询机制的过程。我们将用户的位置 L 表示为一个位置点序列 $L=\{l_1, l_2, ..., l_n\}$, 其中, n 为位置的个数, l_i 为位置点。一个隐私意识强的用户向不可信的服务提供商请求服务时, 发布一个代替其真实位置 L 的假位置 L'。具体而言, 用户在进行假查询时, 考虑位置间的时空相关性, 通过位置轨迹隐私保护机制(LTPPM)对真实位置进行隐私保护处理, 然后上传假位置进行 LBS 查询服务。为了简单起见, 我们假设$|L|=|L'|$, 即真实轨迹中的位置点数量和发布的假轨迹中的位置点数量相同。

3.2 位置轨迹的隐私和可用性度量

我们使用随机变量 L 和 L'分别表示真实的位置轨迹和发布的假轨迹, l 和 l'是这两个随机变量的可能取值。

定义 1. 位置轨迹隐私度量 给定真实的位置轨迹 $L=(l_1, l_2, ..., l_n)$和发布的假轨迹 $L'=(l_1', l_2', ..., l_n')$, 则 L 和 L'的隐私度量定义为:

$$I(L,L')=\sum_{l_i\in L}\sum_{l_i'\in L'} p(l_i,l_i')\log\left(\frac{p(l_i,l_i')}{p(l_i)p(l_i')}\right) \tag{1}$$

其中, $I(L, L')$是 L 和 L'的互信息; $p(l, l')$是 l 和 l'的联合概率分布; $p(l)$和$p(l')$分别表示 l 和 l'的边缘概率分布。

我们注意到, 在计算轨迹间的互信息时, 需要遍历真实位置轨迹和历史轨迹上位置点的所有组合$(l, l')\in(L, L')$, 指数复杂度高。这是因为目标函数中的变量数量取决于位置轨迹的长度, 想要消除优化问题中位置轨迹长度的影响, 使得最优解的计算更加简便, 互信息的计算需要独立于位置轨迹的长度。因此, 为了实现这一目标, 我们通过马尔科夫链[13]对互信息进行限制, 实现了互信息的计算只和当前的位置有关, 进而简化了互信息计算方法。定义 2 给出了基于马尔科夫链的位置轨迹隐私度量的具体定义。

定义 2. 基于马尔科夫链的位置轨迹隐私度量 给定真实的位置轨迹 $L=(l_1, l_2, ..., l_n)$和发布的假轨迹 $L'=(l_1', l_2', ..., l_n')$, 则基于马尔科夫链的 L 和 L'的隐私度量定义为:

$$I_{Markov}(L,L')=\sum_{i=1}^{n} p(l_i,l_i')\log\left(\frac{p(l_i,l_i')}{p(l_i)p(l_i')}\right) \tag{2}$$

其中, $I_{Markov}(L, L')$是基于马尔科夫链的 L 和 L'的互信息。

然而, 为了从 LBS 中获得效用, 进行假查询所带来的信息失真应该限制在一定的阈值之内。因此, 定义 3 给出可用性度量的定义。

定义 3. 位置轨迹可用性度量 给定真实的位置轨迹 $L=(l_1, l_2, ..., l_n)$和发布的假轨迹 $L'=(l_1', l_2', ..., l_n')$, 则 L 和 L'的可用性度量定义为:

$$Dis(L,L')=\frac{1}{n}\sum_{i=1}^{n}\left\|l_i-l_i'\right\| \tag{3}$$

其中, $Dis(L, L')$表示轨迹 L 和 L'的距离; $\|l_i–l_i'\|$表示两条位置轨迹上位置点之间的欧几里得距离, n 表示轨

迹中位置点的数量。

4 本文方案

在本节中, 我们设计了基于最优轨迹的假查询隐私保护机制, 解决了隐私机制的可用性最优化问题。

4.1 基于最优轨迹的假查询隐私保护机制

在本节中, 我们提出一种基于最优轨迹的假查询隐私保护机制。其主要思想是: 在可用性的约束条件下, 通过计算真实轨迹和已有历史轨迹集中轨迹之间的互信息, 寻找出互信息最小的历史轨迹。然后, 在最优的历史轨迹上找出与真实位置相对应的位置进行假查询。具体而言, 如图 1 所示, 对于真实轨迹 L, 首先需要从历史轨迹集中找出最优的历史轨迹 L'。其次, 在历史轨迹 L'上, 找出与真实轨迹上的位置 l_1, l_2, ..., l_n 对应的假位置 l_1', l_2', ..., l_n'。最后, 通过假位置 l_1', l_2', ..., l_n'向 LBS 发送查询信息来获取服务。在整个过程中, 我们在满足可用性的条件下寻找隐私泄漏最小的轨迹。因此, 可以实现位置轨迹的隐私-可用性权衡。同时, 我们基于历史轨迹集选取发布的假轨迹, 可以保证筛选出来的轨迹更加真实、合理。

从直观上看, 用户需要的隐私保护级别越高, 其获得的效用就越少, 反之亦然。所以如图 1 所示, 在基于我们的可用性约束来设计最优轨迹的假查询隐私保护机制时, 存在着隐私-可用性平衡问题。从数据可用性来看, 受可用性约束的最小隐私泄漏是多少, 以及如何设计最优轨迹的假查询隐私保护机制来实现最小的隐私泄漏, 是我们最需要考虑的问题。因此, 我们给出以下定义来表述这个问题。

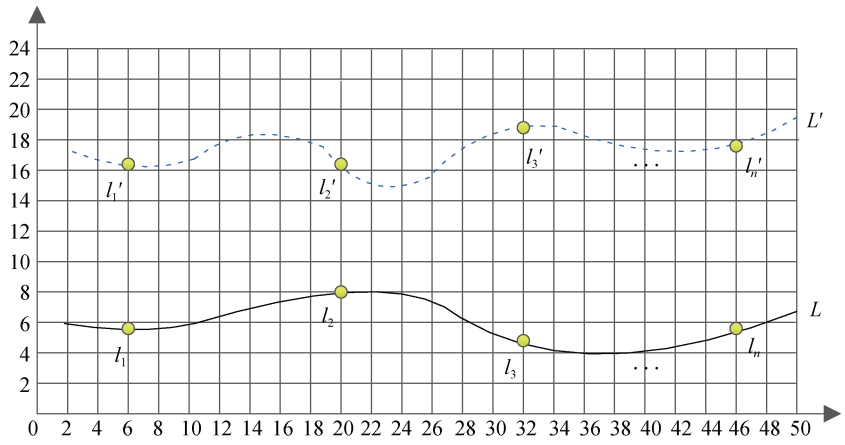


图 1 最优轨迹的假查询

Figure 1 False query for optimal trajectory

定义 4. 位置轨迹的隐私性-可用性平衡 对于真实的位置轨迹 $L=(l_1, l_2, ..., l_n)$和发布的假轨迹 $L'=(l_1', l_2', ..., l_n')$, L 和 L'的可用性约束为: $Dis\leqslant d$。一个最优轨迹的假查询隐私保护机制表示在可用性约束下, 位置轨迹的隐私泄漏最小, 即求解以下问题的最优解:

$$\begin{aligned} Q_R &= \underset{\substack{Dis(L,L')\leqslant d \\ L'\in R}}{\arg\min}\ I_{Markov}(L,L') \\ &= \underset{\substack{\{Dis(l_i,l_i')\leqslant d_i\}_{i=1}^{n} \\ l_i\in L;l_i'\notin L';L'\in R}}{\arg\min} \sum_{i=1}^{n} I_{Markov}(l_i,l_i') \end{aligned} \tag{4}$$

其中, R 表示整个路网区域的历史轨迹集; $I_{Markov}(L, L')$ 表示基于马尔科夫链的位置轨迹隐私性度量; d_i 表示区域 R 内第 i 个位置点的失真; n 表示轨迹中位置点的数量。

直接求解定义 4 中的优化问题, 即寻找最优历史轨迹 L'会导致较大的计算复杂度。这是因为我们不仅需要计算真实轨迹 L 和历史轨迹集 R 中任一 L' 的可能组合, 而且也需要考虑轨迹上的位置点的数量。也就是说, 我们在解决真实轨迹 L 和历史轨迹 L'上位置个数的优化问题的同时, 也要考虑历史轨迹集 R 中位置轨迹的数量。具体而言, 如果我们考虑一个区域内有 M 条历史轨迹, 假轨迹的选取总是在这 M 条历史轨迹上。那么当用户想查询的位置点个数为 N 时, 由于 L 和 L'上的变量个数$|L|=|L'|=N$, 则一条历史轨迹的组合数为 N, 则计算 M 条历史轨迹组合数的时间复杂度为 $O(MN^2)$。并且随着位置轨迹中位置个数和历史轨迹的不断增加, 变量的数量也在不断增长, 计算复杂度较高。因此, 在 4.2 中给出了基于子区域的假查询机制。

4.2 基于子区域的假查询机制

在本节中, 我们利用子区域对假查询机制进行优化。通过将原始区域划分为不同的子区域, 可以缩小所选区域 R 的面积, 从而减少区域轨迹片段上位置点的数据, 使得计算更加简便。

定义 5. 基于子区域的位置轨迹隐私泄露 在任意的子区域 R'中, 用户的真实位置轨迹和发布的假轨迹分别为 $L=(l_1, l_2, ..., l_m)$, $L'=(l_1', l_2', ..., l_m')$。则在子区域 R'中位置轨迹实际的隐私泄漏被定义为:

$$Q_{R'} = I_{Markov}(L,L') = \sum_{i=1}^{m} I_{Markov}(l_i,l_i') \tag{5}$$

我们使用 $Q_{R'}$作为隐私度量来评估基于子区域的位置轨迹的实际隐私泄漏情况。其中 $R'\subseteq R$, 即子区域 R'为区域 R 的子集, m 为位置点的个数, 且 $m<n$。

接下来, 我们在下面的定义中给出了基于子区域的位置轨迹的隐私-可用性权衡问题。

定义 6. 基于子区域的位置轨迹隐私性-可用性平衡 基于子区域位置轨迹中隐私性与可用性的平衡问题如下:

$$\begin{aligned} Q_{R'} &= \arg\min_{L'\in R'} I_{Markov}(L,L') \\ &= \arg\min_{\substack{\{Dis(l_i,l_i')\leqslant d\}_{i=1}^{m} \\ l_i\in L, l_i'\in L', L'\in R'}} \sum_{i=1}^{m} I_{Markov}(l_i,l_i') \end{aligned} \tag{6}$$

其中, d_i 表示子区域内第 i 个位置点的失真。

因此, 为了实现子区域的位置轨迹隐私性-可用性平衡, 需要对路网进行划分。然而, 由于路网中的轨迹是不均匀的, 所以我们利用四叉树将路网划分为不同的网格区域, 进而轨迹将会划分在不同的区域中。图 2 是一个基于四叉树的路网区域划分示意图, 从图 2 中可以看出, 在轨迹越密集的区域, 分区的面积越小; 反之, 在轨迹越稀疏的区域, 分区的面积越大。

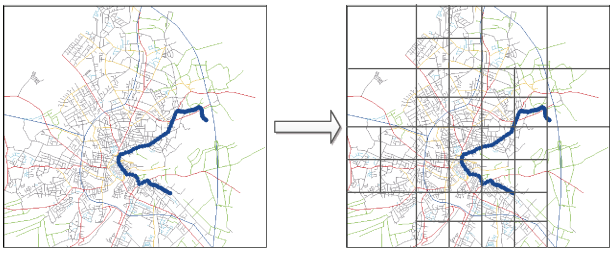


图 2 区域划分图

Figure 2 Zoning map

通过四叉树法将路网划分为不同的区域, 轨迹 L 被划分为不同的轨迹片段 L_i 且分布在不同的子区域 R'中, 这样可以缩短轨迹的长度, 从而减少轨迹上位置点的数量, 简化了计算过程。因此, 我们只需要对于每个子区域中的轨迹片段 L_i 进行处理, 如图 3 所示。

从图 3 可以看出, 由于我们将真实轨迹划分为不同的轨迹片段, 对于每一轨迹片段, 只需要从历史轨迹中找出代替它的轨迹片段。并且我们通过定理 1 证明了我们提出的基于子区域的假查询机制优化方法, 可以保证划分后的各个子区域中的轨迹片段的隐私泄漏大小之和等于原始轨迹的隐私泄露量。

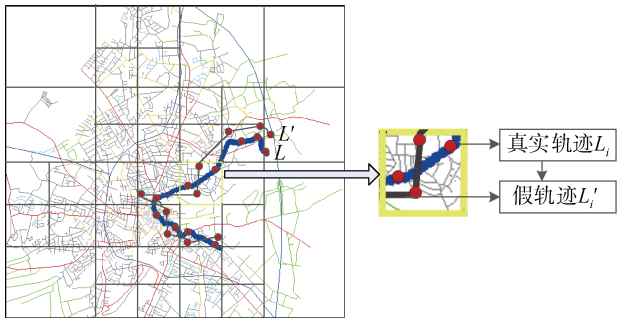


图 3 轨迹片段处理

Figure 3 Track fragment processing

定理 1 原始轨迹与发布假轨迹间的互信息等于各个原始轨迹片段和发布假轨迹间的互信息之和。

证明 假设原始轨迹为 $L=(l_1, l_2, ..., l_m)$, 发布的假轨迹为 $L'=(l_1', l_2', ..., l_m')$。

我们考虑了将原始轨迹划分为两个轨迹片段的情况。假设原始轨迹 L 划分为两个轨迹片段 $L_1=(l_1, l_2, ..., l_m)$和 $L_2=(l_{m+1}, l_{m+2}, ..., l_n)$。则有:

$$\begin{aligned} I_{Markov}(L,L') &= \sum_{i=1}^{n} I_{Markov}(l_i,l_i') \\ &= I_{Markov}(l_1,l_1') + ... + I_{Markov}(l_m,l_m') \\ &\quad + I_{Markov}(l_{m+1},l_{m+1}') + ... + I_{Markov}(l_n,l_n') \\ &= \sum_{i=1}^{m} I_{Markov}(l_i,l_i') + \sum_{i=m+1}^{m} I_{Markov}(l_i,l_i') \\ &= I_{Markov}(L_1,L_1') + I_{Markov}(L_2,L_2') \end{aligned}$$

同理, 将原始轨迹 L 划分为 k 个轨迹片段, 其中 k 取任意的大于 2 的正整数, 以上等式亦成立。因此, 可得原始轨迹与发布假轨迹间的互信息等于各个原始轨迹片段和发布假轨迹间的互信息之和, 证毕。

4.3 基于轨迹偏离角度的子区域优化机制

基于子区域的位置轨迹虽然保证了位置轨迹中位置的数量变少, 有效的降低了计算复杂度。然而, 由于历史轨迹数量繁多, 在每一次计算时, 需要遍历所有的历史轨迹, 计算复杂度还是较为复杂。因此, 为了减少历史轨迹的数量, 进一步优化计算过程, 我们考虑了各个子区域上轨迹位置点之间的位置关系, 通过定义子区域内位置轨迹间的偏离角度对历史轨迹进行约束。

定义 7. 轨迹偏离角度 假设用户的原始轨迹为 $L_i=\{l_{i1}, l_{i2}, \ldots, l_{iN}\}$, $l_{i1}=\{(x_{i1}, y_{i1}), (x_{i2}, y_{i2}), \ldots, (x_{im}, y_{im})\}$

表示用户在第 i 个区域内的轨迹, 其起点为(x_{i1}, y_{i1}), 终点为(x_{im}, y_{im})。则轨迹的偏离向量为 $\vec{s}_i=(x_{im}-x_1, y_{im}-y_1)$, 同样地, 发布的假轨迹为 $L_i'=\{l_{i1}', l_{i2}', \ldots, l_{iN}'\}$, $l_{i1}'=\{(x_{i1}', y_{i1}'), (x_{i2}', y_{i2}'), \ldots, (x_{im}', y_{im}')\}$ 的轨迹偏离向量为 $\vec{s}_i'=(x_{im}'-x_1', y_{im}'-y_1')$, 则在第 i 个区域内, 原始轨迹 l_{i1} 和假轨迹 l_{i1}'的偏离角度为:

$$\arccos\theta_i=\frac{\vec{s}_i\times\vec{s}_i'}{|\vec{s}_i||\vec{s}_i'|} \tag{7}$$

其中, θ_i 表示 $\vec{s}_i$ 与 $\vec{s}_i'$ 之间的夹角;

$\vec{s}_i\times\vec{s}_i'=(x_{im}-x_1)(x_{im}'-x_1')+(y_{im}-y_1)(y_{im}'-y_1')$;

$|\vec{s}_i|=\sqrt{(x_{im}-x_1)^2+(y_{im}-y_1)^2}$ 。

在整个区域内, 假轨迹集合 L'相对于原始轨迹 L 的偏离角度序列为 $S=\{\theta_1, \theta_2, ..., \theta_N\}$。所以, 原始轨迹 L 和假轨迹 L'的轨迹偏离角度 sim_θ为:

$$sim_\theta=\frac{1}{N}\sum_{i=1}^{N}\arccos\theta_i \tag{8}$$

图 4 给出了区域优化处理示意图。由图 4 可知, 通过轨迹的偏离角度对历史轨迹进行约束, 会缩小子区域的范围, 减少历史轨迹的数量, 进一步对基于子区域的位置轨迹进行计算优化。

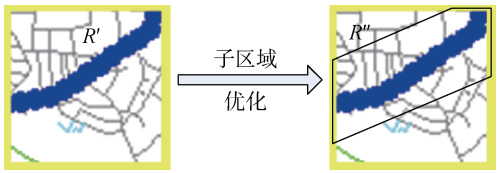


图 4 子区域优化图

Figure 4 Subregion optimization diagram

因此, 我们通过定义 8 给出基于优化子区域的位置轨迹隐私泄露的定义。

定义 8. 基于优化子区域的位置轨迹隐私泄露 在经过优化的任意优化子区域 R''中, 用户的真实位置轨迹为 $L=(l_1, l_2, \ldots, l_m)$和发布的假轨迹为 $L'=(l_1', l_2', \ldots, l_m')$。则在优化子区域 R''中位置轨迹实际的隐私泄露被定义为:

$$Q_{R''}=I_{Markov}(L,L')=\sum_{i=1}^{m}I_{Markov}(l_i,l_i') \tag{9}$$

我们使用 $Q_{R''}$作为隐私度量来评估基于子区域的位置轨迹的实际隐私泄漏情况, 其中 $R''\subseteq R'\subseteq R$, 即优化子区域 R''为子区域 R'的子集, 且 R''中的历史轨迹要远小于 R'中历史轨迹的数量。

接下来, 我们在下面的定义中提出了基于优化子区域的位置轨迹的隐私-可用性权衡问题。

定义 9. 基于优化子区域的位置轨迹隐私性-可用性平衡 基于优化子区域位置轨迹中隐私性与可用性的平衡问题如下:

$$\begin{aligned}Q_{R''}&=\arg\min_{L'\in R''}I_{Markov}(L,L')\\&=\arg\min_{\substack{\{Dis(l_i,l_i')\le d_i\}_{i=1}^m\\ l_i\in L, l_i'\in L', L'\in R''}}\sum_{i=1}^{m}I_{Markov}(l_i,l_i')\end{aligned} \tag{10}$$

因此, 我们只需要解决在优化子区域 R''中, 在可用性的约束下, 通过计算位置轨迹间的互信息, 寻找出互信息最小的历史轨迹。我们通过定理 2 证明了我们提出的基于优化子区域的假查询机制优化方法可以进一步简化计算方法。

定理 2 优化子区域可以进一步简化互信息的计算复杂度。

证明 假设子区域为 R', R'中包含 M 条历史轨迹; 优化子区域为 R'', R''中包含 S 条历史轨迹, 其中 $R''\subseteq R'$且 $S\ll M$。每条轨迹上包含 m 个位置点。因为:

$$\begin{aligned}Q_{R'}&=\arg\min_{L'\in R'}I_{Markov}(L,L')\\&=\arg\min_{\substack{\{Dis(l_i,l_i')\le d_i\}_{i=1}^m\\ l_i\in L, l_i'\in L', L'\in R'}}\sum_{i=1}^{m}I_{Markov}(l_i,l_i')\end{aligned}$$

则在子区域为 R'中选取互信息最小的轨迹时需要计算所有的历史轨迹和真实轨迹之间的互信息。假设遍历一条轨迹的计算复杂度为 $O(m)$, 当 L'遍历 R'中所有的历史轨迹, 计算复杂度为 $MO(m)$。又因为:

$$\begin{aligned}Q_{R''}&=\arg\min_{L'\in R''}I_{Markov}(L,L')\\&=\arg\min_{\substack{\{Dis(l_i,l_i')\le d_i\}_{i=1}^m\\ l_i\in L, l_i'\in L', L'\in R''}}\sum_{i=1}^{m}I_{Markov}(l_i,l_i')\end{aligned}$$

则在优化子区域为 R''中, 当 L'遍历 R''中所有的历史轨迹时, 其计算复杂度为 $SO(m)$。

由于 $S\ll M$, 所以计算复杂度 $SO(m)$远小于 $MO(m)$, 即有 $SO(m)\ll MO(m)$。

因此, 优化子区域可以进一步简化互信息的计算复杂度, 证毕。

4.4 最优位置轨迹选择算法

基于 4.1、4.2 和 4.3 的分析, 本节提出一种最优位置轨迹选择算法, 具体的伪代码见算法 1。

在算法 1 中, 为了计算方便, 我们假设每条历史轨迹上对应的位置数目与真实轨迹位置个数相等。在每一个区域内的真实轨迹有 n 个位置, 有 m 条历史轨迹。首先, 我们通过计算轨迹间偏离角度, 对历史轨迹进行约束, 筛选出 $sim_\theta<\alpha$的历史轨迹。其次, 我们计算真实位置轨迹 $L_i=(l_{i1}, l_{i2}, \ldots, l_{in})$与历史轨迹 $H_i^h=(l_{i1}^h, l_{i2}^h, ..., l_{in}^h)$ 的距离, 筛选出 $Dis(L_i, H_i^h)$小于阈值β的历史轨迹。最后, 在可用性的约束下, 通过计

算轨迹间的互信息, 选出互信息最小的历史轨迹作为假轨迹 $L'=(l_1', l_2', \ldots, l_n')$, 然后以假轨迹上的位置点 $l_1', l_2', \ldots, l_n'$进行查询。

表 1 OSLTSA 算法实现步骤

Table 1 OSLTSA algorithm implementation steps

算法 1: 最优位置轨迹选择算法(OSLTSA)

输入: 真实位置轨迹 L, 轨迹偏离角度α, 轨迹距离阈值β

输出: 最优位置轨迹 L'

1.初始化 L'

2.**Loop**:

3.j=1 to n, h=1 to m

4.**for** $l_{ij}\in L_i$, $l_{ij}^h\in H_{ij}^h$

5. calculate sim_θ;

6. **if** $sim_\theta<\alpha$ && $|Dis(L_i,H_i^h)|<\beta$

7. calculate $Dis(L_i,H_i^h)$;

8. calculate $I_{Markov}(L_i,H_i^h)$;

9. select argmin $I_{Markov}(L_i,H_i^h)$;

10. **else**

11. goto **Loop**

12. **end if**

13. **end for**

14. **return** $L'=(l_1',l_2', \ldots, l_n')$.

5 实验分析

本节通过实验分析验证本文方案的性能, 并且将本文的实验结果与文献[13]提出的 DLTP 机制和文献[14]中提出的 ABCP 机制进行比较。

5.1 实验设置

本实验将采用真实的轨迹数据集 Geolife 数据集[23]评估本文方法的实际隐私泄露。其中, Geolife 数据集采集了 182 个用户 2007 年 4 月至 2012 年 8 月五年期间在北京的运动轨迹。同时, 实验环境为: Intel(R) Core(TM) i5-3470U CPU @3.20GHz 2.40 GHz; 4GB(RAM)内存; Windows 10 专业版 64 位操作系统。考虑到实验误差的影响, 每组实验重复进行 3 次, 结果取平均值。

5.2 信息泄露度

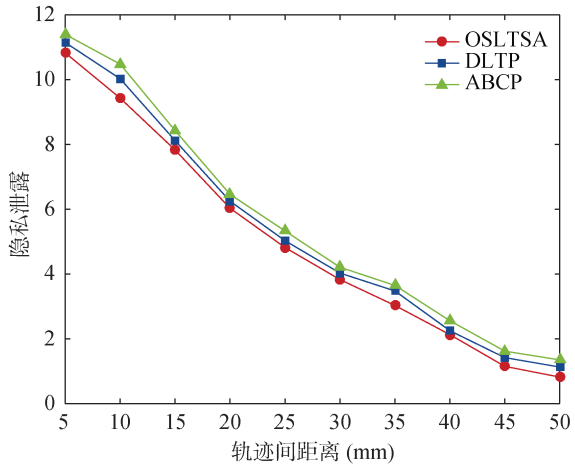
为了分析数据的信息泄露度随轨迹偏离角度α和轨迹距离阈值β值的变化规律, 我们进行了以下实验并且对实验结果进行分析。

首先, 我们研究轨迹信息泄露随着不同的轨迹偏离角度α值的变化情况。图 5(a～d)说明了当α=15°、α=30°、α=45°和α=60°时, OSLTSA、DLTP 和 ABCP 三种算法随着轨迹距离阈值β的变化而变化的情况。从图 5 中可以看出, 随着轨迹偏离角度α值的增大, 3 种机制下轨迹数据的隐私泄露程度也在不断减少。这是因为随着轨迹偏离角度α值的增大, 区域中历史轨迹的数量也在增加, 进而轨迹集中位置数增多, 造成的隐私泄露程度减少。并且从图 5 可知, 从整体上来看, 随着轨迹偏离角度α值的增大, 本文方案的数据隐私泄露量均小于 DLTP 机制和 ABCP 机制的隐私泄露量, 这也进一步说明本文方法更能减少数据的隐私泄露, 可以更好的保护数据的隐私安全。

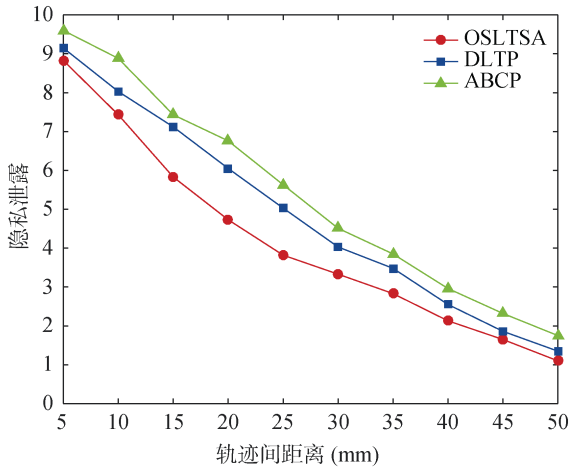
其次, 我们研究不同的轨迹距离阈值β值对数据信息损失度的影响。图 6(a～d)说明了当β=10、β=20、β=30 和β=40mm 时, OSLTSA、DLTP 和 ABCP 机制的隐私泄露随着轨迹偏离角度α的变化而变化的情况。从图 6 中可以看出, 随着轨迹距离阈值β值的增大, 3 种机制下轨迹数据的隐私泄露程度也在不断增大。这是因为随着轨迹距离阈值β值的增大, 历史轨迹可选区域变大, 可选轨迹数量增加, 进而轨迹集中位置增加, 造成的泄露程度也在不断增加。并且整体而言, 在不断增大轨迹距离阈值β的条件下, 本文的方案所造成的隐私泄露均要小于其他两种机制, 安全性高。

5.3 运行时间

图 7 和图 8 分别给出了 OSLTSA、DLTP 和 ABCP 机制随不同的轨迹偏离角度α值以及不同轨迹距离阈值β下的运行时间的比较。



(a) α=15°



(b) α=30°

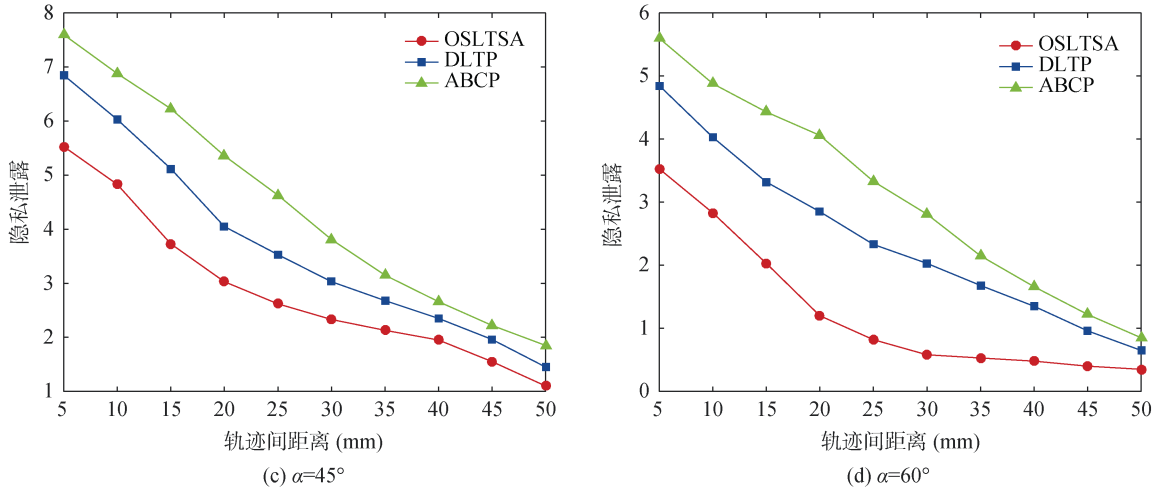


图 5　α值固定时，隐私泄露随β值的变化情况

Figure 5　The change of privacy disclosure with β value when α value is fixed

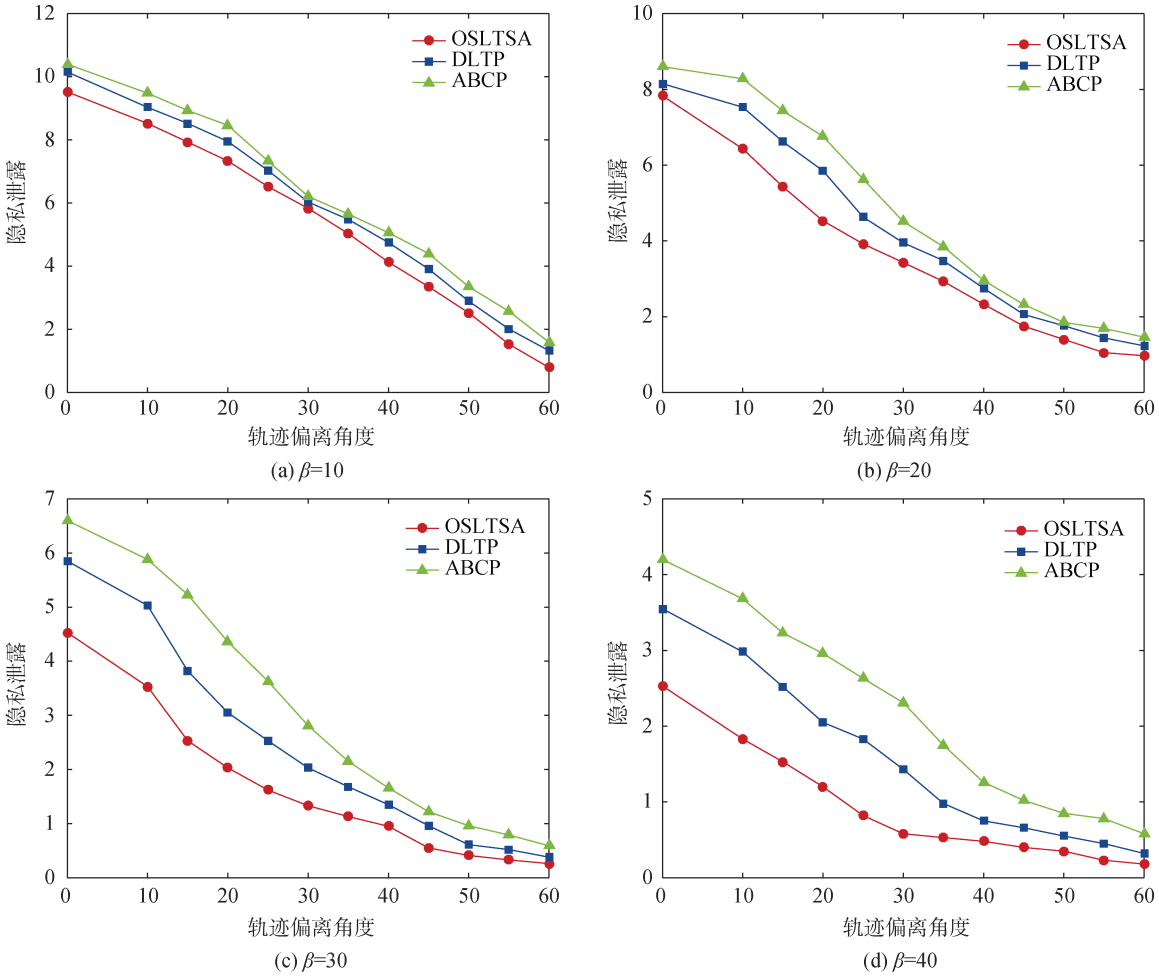


图 6　β值固定时，隐私泄露随α值的变化规律

Figure 6　The change rule of privacy disclosure with α value when β value is fixed

图 7(a～b)给出了当α=15°和α=30°时, OSLTSA、DLTP 和 ABCP 机制随着轨迹距离阈值的变化情况。从图 7 可以看出，当α值发生变化时，其执行时间也相应改变。随着α值的增大，执行时间也在不断增加。这是因为α值增大，区域中历史轨迹的个数增加，计算互信息时需要遍历所有的历史轨迹，进而执行时间变长。而且从图 7 中可以看出，本文提出的 OSLTSA 算法的执行时间相比 DLTP 和 ABCP 机制，

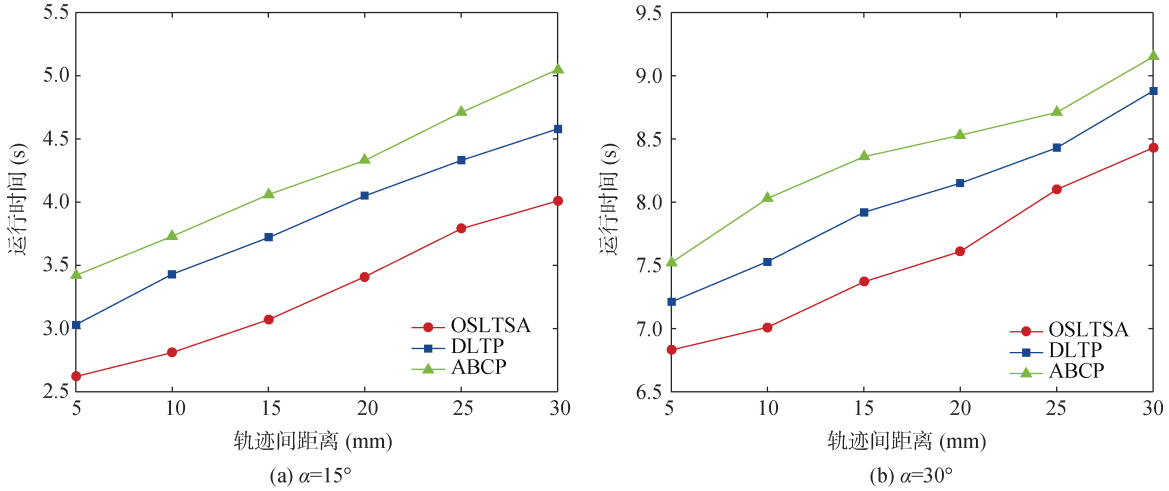


图 7　α值固定时，不同β值下的运行时间

Figure 7　The running time at different values of β when α is fixed

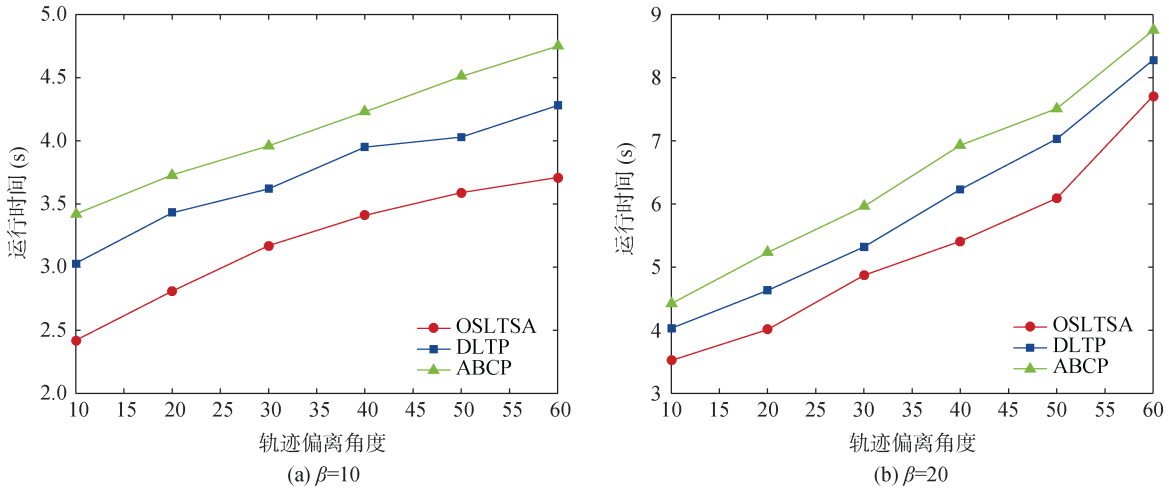


图 8　β值固定时，不同α值下的运行时间

Figure 8　The running time at different values of α when β is fixed

执行时间最短。这是因为，本文的方法是在可用性的约束下寻找互信息最小的历史轨迹，优化了计算过程，进而运行时间较短。

图 8(a～b)给出了当β=10mm 和β=20mm 时，OSLTSA、DLTP 和 ABCP 机制随着轨迹偏离角度α的变化情况。从图 8 可以看出，当β值发生变化时，其执行时间也相应改变。随着β值的增大，执行时间也在不断增加。这是因为β值增大，历史轨迹可选择的区域变大，其可选轨迹数增加，而计算互信息时需要遍历所有的历史轨迹，进而执行时间变长。而且从图 8 中可以看出，本文提出的 OSLTSA 算法的运行时间相比 DLTP 和 ABCP 机制而言，运行时间最短。这是因为，本文的方法是在可用性的约束下寻找互信息最小的历史轨迹，优化了计算过程，进而运行时间较短。

6　结论

针对现有位置轨迹隐私保护机制难以实现可用性和隐私性的平衡问题，提出一种基于最优轨迹的假查询隐私保护机制。从信息论的角度出发，提出轨迹互信息的隐私度量方法，先通过计算真实轨迹与历史轨迹间的互信息来量化轨迹的隐私，再基于四叉树法对地图进行划分，将真实轨迹划分为不同的轨迹片段，将轨迹间的互信息转化为所有轨迹片段的互信息之和。特别地，我们基于马尔科夫链简化了互信息的计算方法，降低了系统的计算开销。并且我们选择的假轨迹是基于历史轨迹生成的，可以保证生成的假轨迹更加真实、合理。最后，通过实验分析，证明本文的方案可以在满足可用性的条件下保护用户的隐私。

参考文献

[1] Jiang J F, Han G J, Wang H, et al. A Survey on Location Privacy Protection in Wireless Sensor Networks[J]. *Journal of Network and Computer Applications*, 2019, 125: 93-114.

[2] Han M, Wang J B, Yan M Y, et al. Near-Complete Privacy Protection: Cognitive Optimal Strategy in Location-Based Services[J]. *Procedia Computer Science*, 2018, 129: 298-304.

[3] He Y, Chen J G. User Location Privacy Protection Mechanism for Location-Based Services[J]. *Digital Communications and Networks*, 2021, 7(2): 264-276.

[4] Zhang Q Y, Zhang X, Wang M Y, et al. DPLQ: Location-Based Service Privacy Protection Scheme Based on Differential Privacy[J]. *IET Information Security*, 2021, 15(6): 442-456.

[5] Guo X Y, Wang W M, Huang H P, et al. Location Privacy-Preserving Method Based on Historical Proximity Location[J]. *Wireless Communications and Mobile Computing*, 2020, 2020: 1-16.

[6] Soma S C, Hashem T, Cheema M A, et al. Trip Planning Queries with Location Privacy in Spatial Databases[J]. *World Wide Web*, 2017, 20(2): 205-236.

[7] Kido H, Yanagisawa Y, Satoh T. An Anonymous Communication Technique Using Dummies for Location-Based Services[C]. *ICPS '05. Proceedings. International Conference on Pervasive Services*, 2005: 88-97.

[8] Gao S, Ma J F, Shi W S, et al. TRPF: A Trajectory Privacy-Preserving Framework for Participatory Sensing[J]. *IEEE Transactions on Information Forensics and Security*, 2013, 8(6): 874-887.

[9] Wu X C, Sun G Z. A Novel Dummy-Based Mechanism to Protect Privacy on Trajectories[C]. *2014 IEEE International Conference on Data Mining Workshop*, 2015: 1120-1125.

[10] Lei K Y, Li X H, Liu H, et al. Dummy Trajectory Privacy Protection Scheme for Trajectory Publishing Based on the Spatiotemporal Correlation[J]. *Journal on Communications*, 2016, 37(12): 156-164.
(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.)

[11] Pan X, Ma A, Zhang J W, et al. Approximate Similarity Measurements on Multi-Attributes Trajectories Data[J]. *IEEE Access*, 2018, 7: 10905-10915.

[12] Li G S, Yin Y M, Wu J H, et al. Trajectory Privacy Protection Method Based on Location Service in Fog Computing[J]. *Procedia Computer Science*, 2019, 147: 463-467.

[13] Zhang W J, Li M, Tandon R, et al. Online Location Trace Privacy: An Information Theoretic Approach[J]. *IEEE Transactions on Information Forensics and Security*, 2019, 14(1): 235-250.

[14] Oya S, Troncoso C, Pérez-González F. Back to the Drawing Board: Revisiting the Design of Optimal Location Privacy-Preserving Mechanisms[C]. *The 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017: 1959-1972.

[15] Chow C Y, Mokbel M F. Enabling Private Continuous Queries for Revealed User Locations[C]. *The 10th international conference on Advances in spatial and temporal databases*, 2007: 258-273.

[16] Pan X, Meng X F, Xu J L. Distortion-Based Anonymity for Continuous Queries in Location-Based Mobile Services[C]. *The 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2009: 256-265.

[17] Yu W, Xu D B, He X, et al. L2P2: Location-Aware Location Privacy Protection for Location-Based Services[C]. *2012 Proceedings IEEE INFOCOM*, 2012: 1996-2004.

[18] Latha K, Jayanthi S, Elavenil V. KRUPTO: Supporting Privacy Against Location Dependent Attacks in Wireless Sensor Network[C]. *2013 International Conference on Communication and Signal Processing*, 2013: 908-912.

[19] Gruteser M, Liu X. Protecting Privacy, in Continuous Location-Tracking Applications[J]. *IEEE Security & Privacy*, 2004, 2(2): 28-34.

[20] Terrovitis M, Mamoulis N. Privacy Preservation in the Publication of Trajectories[C]. *The Ninth International Conference on Mobile Data Management*, 2008: 65-72.

[21] Götz M, Nath S, Gehrke J. MaskIt: Privately Releasing User Context Streams for Personalized Mobile Applications[C]. *The 2012 ACM SIGMOD International Conference on Management of Data*, 2012: 289-300.

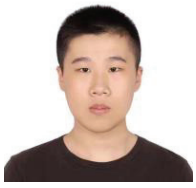
[22] Zhao J, Zhang Y, Li X H, et al. A Trajectory Privacy Protection Approach via Trajectory Frequency Suppression[J]. *Chinese Journal of Computers*, 2014, 37(10): 2096-2106.
(赵婧, 张渊, 李兴华, 等. 基于轨迹频率抑制的轨迹隐私保护方法[J]. *计算机学报*, 2014, 37(10): 2096-2106.)

[23] Zheng Y, Xie X, Ma W Y. GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory[J]. *IEEE Data Eng Bull*, 2010, 33: 32-39.

刘燕妮 于 2021 年在韩山师范学院计算机科学与技术专业获得学士学位。现在福建师范大学网络空间安全专业攻读硕士学位。研究领域为数据安全与机器学习。研究兴趣包括: 机器学习、数据共享、隐私保护。Email: 2426759858@qq.com

张强 于 2021 年在福建师范大学应用数学专业获得硕士学位。研究领域为位置隐私。研究兴趣包括: 隐私保护、数据安全。Email: denghn@163.com

叶阿勇 于 2009 年在西安电子科技大学计算机系统结构专业获得博士学位。现为福建师范大学计算机与网络空间安全学院的教授和博士生导师。研究领域为网络安全与隐私。研究兴趣包括: 区块链、网络安全、位置隐私。Email: yay@fjnu.edu.cn



赵云涛 现在福建师范大学攻读学士学位。研究领域为位置隐私。研究兴趣包括: 隐私保护、网络安全。Email: 49675983@qq.com