

基于加权最小二乘的社交网络用户定位方法

时文旗^{1,2}, 罗向阳^{1,2}, 郭家山^{1,3}

(1. 网络空间态势感知河南省重点实验室, 河南 郑州 450001;

2. 数学工程与先进计算国家重点实验室, 河南 郑州 450001; 3. 郑州大学网络空间安全学院, 河南 郑州 450001)

摘 要: 社交网络在提供位置交友等服务时, 会对展示的用户距离文本进行混淆处理, 以保护用户位置隐私。为了验证当前社交网络采用的位置混淆机制能否有效保护用户的精确位置不被泄露, 提出了一种基于加权最小二乘的社交网络用户定位方法。该方法构造测试环境, 对位置交友服务中混淆后的距离文本进行大量搜集和统计, 结合真实距离数据识别报告距离对应的真实距离边界; 基于对目标用户所处坐标系象限的判别, 优化探针位置部署, 并利用三边测量定位模型得到目标用户的多个初步位置估计; 基于估计位置与探针的距离关系, 分别确定目标用户相较各探针的最远距离和最近距离的权重, 从而构造目标函数并基于加权最小二乘求目标函数的最优解, 该最优解即目标用户的最终定位结果。该方法基于距离边界约束推断社交用户位置, 避免了对位置服务的频繁查询, 保证了对社交用户的定位效率。基于微信平台开展了社交用户定位实验, 对 500 个微信用户的实际定位结果表明, 该方法能够实现对微信“附近的人”用户的准确定位, 与现有基于空间划分、基于启发式数论等典型定位方法相比, 定位精度和效率均有更好的性能表现, 平均定位误差降低了 10%以上, 定位过程中的位置服务访问次数减少了 50%以上。

关键词: 社交网络; 位置隐私; 用户定位; 加权最小二乘

中图分类号: TP309

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2022031

Social network user geolocating method based on weighted least squares

SHI Wenqi^{1,2}, LUO Xiangyang^{1,2}, GUO Jiashan^{1,3}

1. Key Laboratory of Cyberspace Situation Awareness of Henan Province, Zhengzhou 450001, China

2. State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China

3. School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou 450001, China

Abstract: When providing location-based dating and other location-based services, social networks will confuse the

收稿日期: 2022-03-04; 修回日期: 2022-05-05

通信作者: 罗向阳, luox_y_ieu@sina.com

基金项目: 国家自然科学基金 (1804263, 1736214, 62172435, 62002386); 中原科技创新领军人才计划 (214200510019)

Foundation Items: The National Natural Science Foundation of China (1804263, 1736214, 62172435, 62002386), Zhongyuan Science and Technology Innovation Leading Talent Project (214200510019)

引用格式: 时文旗, 罗向阳, 郭家山. 基于加权最小二乘的社交网络用户定位方法[J]. 网络与信息安全学报, 2022, 8(3): 41-52.

Citation Format: SHI W Q, LUO X Y, GUO J S. Social network user geolocating method based on weighted least squares[J]. Chinese Journal of Network and Information Security, 2022, 8(3): 41-52.

displayed user distance text to protect the user's location privacy. In order to verify whether the current location confusion mechanism adopted by social networks can effectively protect user's accurate location, a social network user geolocating method based on weighted least squares was proposed. The method constructed real-world tests to collect a large number of confused distance texts in location dating service, and identified the real distance boundary of reported distance combined with real distance data. Then, based on the discrimination of the quadrant of the coordinate system where the target user was located, the position of probes was optimally deployed, and multiple preliminary position estimations of the target user were obtained by using the trilateration model. The weights of the longest and the shortest distance of the target user were determined by the estimated position, and the related objective function was constructed. The optimal solution of the objective function was calculated based on the weighted least squares, which was the final geolocating result of the target user. The proposed method inferred the location of social network user based on the distance boundary constraint, which avoided the frequent query of location services and ensured the geolocating efficiency. The actual geolocating results based on 500 WeChat users showed that, the proposed method can accurately geolocate the "eople nearby" users of WeChat. Compared with the existing typical geolocating methods based on space partition and heuristic number theory, the proposed method have better performance in geolocating accuracy and efficiency. The average geolocating error is reduced by more than 10%, and the number of location service quarries in the geolocating process is reduced by more than 50%.

Keywords: social network, location privacy, user positioning, weighted least squares

0 引言

随着智能移动设备的普及和社交媒体的广泛流行, 社交网络中的多媒体数据得到爆发式增长^[1]. 位置文本作为多媒体数据的一种, 广泛存在于各类社交平台. 在基于位置的社交网络(LBSN, location-based social network)中, 这类位置文本往往用于提供基于位置的服务, 如导航、定向广告、好友发现等^[2-5], 为用户带来了巨大便利. 然而, 位置信息是一种重要的隐私信息, 用户在享受社交平台带来便利的同时, 存在自身位置隐私被社交平台泄露的风险. 以基于位置的交友服务为例, 该服务利用用户的位置信息, 为用户提供基于位置的附近用户发现功能, 并展示附近用户的距离文本, 如微信“附近的人”服务. 为了保护用户位置隐私, 社交平台往往会对距离信息进行混淆处理, 但展示的距离文本依然存在暴露用户位置的可能^[6-7]. 开展社交网络用户定位研究, 能够验证社交平台对用户位置隐私保护的有效性, 进而有助于用户享受到更安全便捷的服务, 具有重要的现实意义和研究价值.

现有关于 LBSN 用户定位的研究主要分为两类: 基于文本和社交关系的用户定位, 基于交友服务的用户定位. 前者利用用户在社交平台中发布的文本、签到数据、社交关系等信息推断用户

的真实位置^[8-10]; 后者研究如何利用社交网络的位置交友服务中公开展示的距离文本定位用户位置, 是本文重点讨论的问题.

基于位置的交友服务在展示用户间距离时, 会对其进行混淆处理, 使恶意用户难以基于展示的距离直接确定用户位置, 一定程度上起到保护用户隐私的作用. 然而, 现有研究表明, 在距离文本被混淆的情况下, 基于不同位置获取的用户混淆后的距离文本及其统计特征, 依然能够准确推断出用户的真实距离范围. 为此, 越来越多的社交平台逐渐加强对服务使用次数、用户实名认证等限制, 以防止恶意用户对服务的滥用.

为了验证当前社交网络中基于位置的交友服务采用的位置隐私保护机制的有效性, 本文提出基于加权最小二乘的社交网络用户定位方法. 该方法基于区域覆盖原则确定目标的大致位置, 并采用加权最小二乘法实现对目标的精确定位; 利用微信社交平台进行了实际定位实验, 实验结果证明所提方法能以很高的效率实现对微信用户的精确定位.

1 相关工作

1.1 LBSD 服务

随着移动社交网络的普及, 基于位置的交友

(LBSD, location-based social discovery) 服务^[11]快速发展并得以流行, Facebook、Telegram、微信、QQ 等全球流行的社交平台均提供此类服务。该服务通过获取用户手机等移动设备的位置信息, 向用户推荐其自身位置附近的其他用户, 并展示附近用户的相对距离(称为报告距离)。为了保护用户位置隐私, LBSD 服务器会采用对报告距离进行混淆处理, 起到一定的位置模糊效果。

常见的报告距离混淆策略有设置最小粒度、添加随机噪声、次序映射等。通过设置最小粒度, 将用户间的精确距离向上取整, 以防止三边测量定位攻击; 通过添加随机噪声, 在保证服务质量的前提下, 使报告距离呈现一定的随机性, 以应对基于统计特征的定位攻击; 通过隐藏距离信息, 并以次序映射的形式体现用户距离, 能够增加定位用户位置的成本。

为了避免服务被恶意用户滥用, LBSD 服务提供商会采用检测虚拟位置、限制查询次数、限定查询距离范围、设置用户推荐时间域等方式, 对定位用户所需的频繁查询行为进行限制。

1.2 LBSD 用户定位相关工作

LBSD 用户定位技术主要基于 LBSD 服务公开展示的附近用户报告距离、相对次序等特征, 通过部署位置已知的探针在不同位置获取目标用户被混淆后的报告距离等信息, 进而推断目标用户的位置。

Ding 等^[12]利用报告距离的分段特性, 采用距离外扩的方式确定目标用户相距探针的真实距离范围, 并基于经典的三边测量定位方法, 实现对目标用户的粗粒度定位和追踪; 为了突破最小报告距离对定位精度的限制, Li 等^[13]提出了基于空间划分的定位 (SPBG, space partition-based geolocation) 方法, 该方法通过对最小报告距离对应地理区间的二分查找, 实现对微信用户更高精度的定位, 后续工作进一步以更大的开销实现了定位精度的提升^[14-15]; 针对报告距离随机化问题, Peng 等^[17]在文献[16]所提的二维数论定位方法的基础上, 设计了基于启发式数论的定位 (HNBG, heuristic number theory-based geolocation) 方法, 提高了实际环境下对 LBSD 用户的定位精度。Polakis 等^[18]将用户定位问题进行模型化

描述, 提出了一个用户定位框架, 并实现了对 Facebook Messenger、Skout 等多款社交平台用户的定位。

1.3 问题描述

如上述相关工作所述, 虽然当前 LBSD 服务采用了多种策略保护用户位置隐私, 现有方法依然可以通过频繁获取目标用户的报告距离, 基于其统计特征实现对目标真实位置的逼近和定位。然而, 为了实现对目标的精确定位, 现有方法在定位目标过程中均需要部署大量探针, 并利用探针在不同位置对目标用户的报告距离进行频繁搜集, 实现对目标真实位置的精确定位。在当前 LBSD 服务加强对用户的实名认证、限制用户的服务使用次数等措施下, 现有方法定位目标所需的开销过大, 其实用性有待提高。

2 基于加权最小二乘的社交用户定位

本文针对现有 LBSD 用户精确定位方法定位开销大、定位时间长的不足, 提出了一种基于加权最小二乘的社交网络用户高效定位方法。该方法主要分为两个部分: 距离边界识别、目标用户定位。其中, 第一部分为定位前的准备工作, 通过搜集大量的真实 LBSD 服务数据, 识别不同报告距离对应的真实距离边界, 为定位目标时提供用户距离划定依据。当定位目标用户时, 本文方法首先通过优化探针部署方式, 并基于区域覆盖初步确定目标用户的多个估计位置; 然后, 基于估计位置对目标用户的真实距离边界进行权值分配; 最后, 基于分配的权值构建目标函数, 利用最小二乘法求目标函数的最优解, 即目标用户的最终定位结果。该方法的原理示意如图 1 所示, 主要步骤如下。

步骤 1 真实距离边界识别

利用已知位置的社交账户, 搜集大量真实环境下的报告距离, 从而确定不同报告距离对应的真实距离范围。通过控制并调整已知位置的用户间的真实距离, 记录不同真实距离下用户间的报告距离, 建立包含<真实距离, 报告距离>数据对的数据集; 将数据集中报告距离相同的真实距离进行归类并排序, 得到该报告距离对应的真实距离序列, 序列的首尾值即该报告距离对应的最远

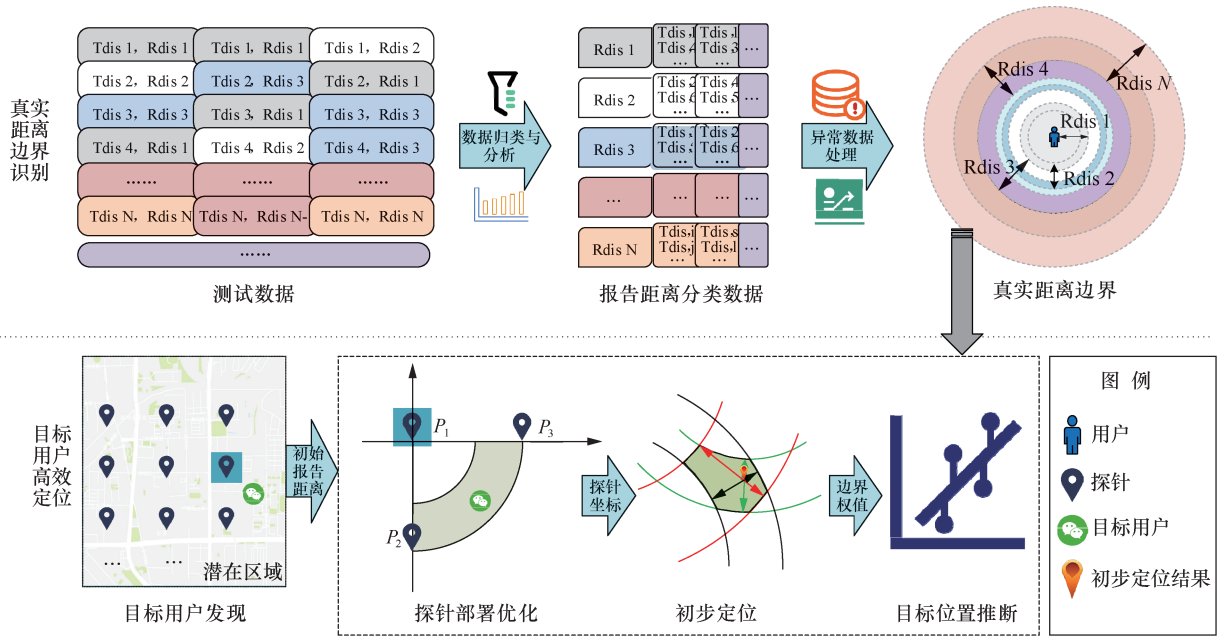


图 1 基于加权最小二乘的社交网络用户定位方法原理示意

Figure 1 Diagram of social network user geolocating method based on weighted least squares

真实距离和最近真实距离。对于不同的报告距离 Rd_i ，其对应的最远真实距离和最近真实距离分别记为 $Rd[i]_{\max}$ 和 $Rd[i]_{\min}$ 。

步骤 2 潜在区域目标发现

在目标所处潜在区域中等距部署探针搜寻目标用户，直到目标用户出现在探针的附近用户列表中，记录此时目标用户的报告距离为 Rd_1 ，将发现目标用户的探针称为初始探针 P_1 。

步骤 3 探针位置部署优化

探针间位置分布会对定位精度及成功率产生重要影响。为了避免不合理探针分布对定位的影响，本步骤通过判断目标用户在以初始探针为原点的坐标系中所处的象限，优化探针的部署。

首先，以初始探针 P_1 为原点建立坐标系，如图 2 所示。在坐标轴上坐标为 $(Rd[1]_{\max}, 0)$ ， $(-Rd[1]_{\max}, 0)$ ， $(0, Rd[1]_{\max})$ 和 $(0, -Rd[1]_{\max})$ 处分别部署探针；然后，各探针分别查询附近的人以获取目标用户的报告距离；最后，选取探针中获取的目标用户的报告距离最小且不在同一坐标轴上的探针（记为 P_2 和 P_3 ，对应的目标用户的报告距离为 Rd_2 和 Rd_3 ），结合初始探针 P_1 作为三边测量定位的探针。若各探针获取的目标报告距离相同，为便于处理，选取 $(Rd[1]_{\max}, 0)$ 和 $(0, Rd[1]_{\max})$ 作为 P_2 和

P_3 ，此时认为目标用户在以 P_1 为原点的坐标系中，位于 P_2 、 P_1 和 P_3 所对应的象限，如图 2 中阴影部分所示， P'_2 、 P 和 P'_1 分别对应 P_2 、 P_1 和 P_3 。

步骤 4 目标用户初步定位

步骤 3 中确定了定位目标所需的探针 (P_1 、 P_2 和 P_3) 及其对应的目标用户的报告距离 (Rd_1 、 Rd_2 和 Rd_3)，此时利用步骤 1 中确定的报告距离对应的真实距离边界，分别以各探针位置 (P_1 、 P_2 和 P_3) 为中心做圆环，圆环的内径和外径分别为该探针获取的目标的报告距离所对应的真实距离边界，目标用户的位置就在 3 个圆环的交叉区域内。取圆环交叉区域的重心点作为对目标的初步定位结果 L_1 。

更新探针位置，将所有探针沿坐标轴方向平移 σ 距离，得到的新的探针位置记为 P_4 、 P_5 和 P_6 。采用上述方法得到目标用户的另一个初步定位结果 L_2 。

步骤 5 加权最小二乘求最优解

由于系统噪声的影响，步骤 1 中建立的报告距离对应的真实距离边界较为宽泛，从而导致步骤 4 中圆环的交叉区域的范围较大，初步定位结果很大概率与目标真实位置存在较大的误差。而若以报告距离为观测数据建立误差目标方程，得

到的目标用户位置的估计值与真实位置偏差较大,进而造成定位结果并不可靠。为此,本文方法以初步定位结果为参考,通过对目标用户的真实距离赋不同权值,进而执行对目标用户的下一轮定位。

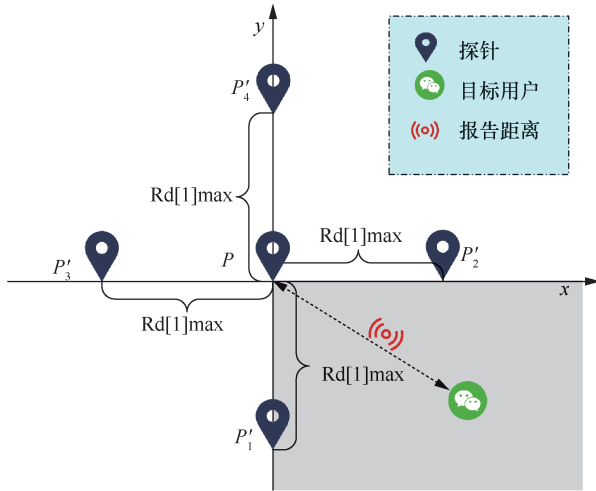


图2 探针部署示意
Figure 2 Diagram of probe deployment

计算 L_1 和 L_2 的中点,记为 L_e ,计算 L_e 与探针 P_i 的真实距离,记为 d'_i ,设目标用户的真实位置为 L_T ,对各探针对应的目标用户最远真实距离和最近真实距离分别赋予不同权重 inw_i 和 ouw_i ,并构建如下目标函数。

$$f(X) = \sum_{i=1}^m \left[\text{inw}_i \cdot (\text{Tdis}(P_i, X) - \text{Rd}[i]_{\min})^2 + \text{ouw}_i \cdot (\text{Rd}[i]_{\max} - \text{Tdis}(P_i, X))^2 \right] \quad (1)$$

其中, m 表示用于定位目标的探针数量, $\text{Tdis}(P_i, X)$ 表示探针 P_i 与位置 X 之间的真实距离,且

$$\text{inw}_i = 1 - \frac{d'_i - \text{Rd}[i]_{\min}}{\text{Rd}[i]_{\max} - \text{Rd}[i]_{\min}} \quad (2)$$

$$\text{ouw}_i = 1 - \frac{\text{Rd}[i]_{\max} - d'_i}{\text{Rd}[i]_{\max} - \text{Rd}[i]_{\min}} \quad (3)$$

利用最小二乘法求使目标函数 f 最小的最优解 X , X 为对目标用户位置的最终定位结果,即

$$L_T \approx \min_X f(X) \quad (4)$$

在上述步骤中,步骤3和步骤5是本文方法

的关键。

3 基于象限判定优化探针部署

三边测量定位用户时的探针位置分布对定位结果有着重要的影响。当探针间夹角过大时,容易导致3个圆环之间形成狭长的交叉区域,在该交叉区域内取得目标定位结果时,定位精度难以保证。因此,所提方法通过设置探针间最大夹角,避免不合理的探针位置分布;同时通过判定目标用户在坐标系中的象限,避免因探针位置与目标用户距离过远导致定位误差增大。

3.1 距离边界划定

探针的位置部署依赖于报告距离 Rd_i 对应的真实距离边界,下面介绍真实距离边界的划定方法。

大多数 LBSD 服务以报告距离的方式展示附近用户的位置信息,且会设置最小报告距离,如微信、探探以 100 m 为单位,陌陌以 10 m 为单位展示附近用户的距离。在没有系统混淆的情况下,该报告距离反映了用户的真实距离边界,即

$$\text{Rd} - K \leq d < \text{Rd} \quad (5)$$

其中, K 表示报告距离的变化单位。然而,系统误差的存在使当真实距离小于 $\text{Rd} - K$ 或大于 Rd 时,报告距离依然可能为 Rd ,即实际环境下的真实距离边界会更为宽泛。

为了寻找实际环境下报告距离对应的真实距离边界,需要获取大量的真实数据。通过控制多个不同账户的位置,能够控制用户间的真实距离使之逐渐增大,每次距离变化的增量记为 Δd 。用户变化位置后,分别查询附近其他用户的报告距离,从而得到<真实距离,报告距离>数据对。通过大量测试,能够获取实际环境下的真实距离边界数据集(BTD, boundary of true distance)。BTD 中一轮测试过程可表示如下。

$$\text{BTD} = \{\text{append}(\langle d_i, \text{Rd} \rangle) \mid \text{Rd} \in A, d_i = d_0 + i \cdot \Delta d, 0 \leq i < M\} \quad (6)$$

其中,数据集 A 表示所有可能出现的报告距离, M 表示一轮测试过程中的测试次数, d_0 表示该轮测试开始时用户间的初始真实距离, d_i 表示第 i

次测试时用户间的真实距离。获得数据集后，依据报告距离不同，将数据集中数据对分成若干子集。每个子集的数据对中的真实距离 d_i 表示在某次测试中，当用户间的真实距离为 d_i 时，两者之间的报告距离为该子集对应的报告距离。最后，剔除各子集中真实距离 d_i 的异常值，并对其进行排序，取排序后的首位值分别作为该报告距离对应的最远真实距离和最近真实距离。需要注意的是，最小报告距离的存在，使最小报告距离对应的最小真实距离为 0，此时需要综合考虑社交平台的位置混淆特性和实际定位效果，选择定位过程的触发阈值。

由于位置更新时间、设备定位误差等因素的影响，各子集中可能存在一些异常值，异常值最明显的特征为远大于或远小于其他真实距离。这些异常值并不能反映实际条件下该报告距离对应的真实距离范围，且会导致定位目标时划定的目标距离边界过大，影响定位精度。因此，需要剔除各子集中的异常值，使确定的真实距离边界能够有效反映实际条件下 LBSD 服务的距离特征，从而为目标用户的准确定位提供保障。

3.2 探针优化部署

三边测量定位方法具有很高的效率，也是本文方法的理论基础^[19]。当利用三边测量定位方法定位目标时，合理的探针分布会更有利于得到集中的交叉区域，从而有助于提高定位精度。

在定位目标过程中，当初始探针发现目标用户时，可根据目标用户的报告距离（记为 Rd_1 ）及其对应的真实距离边界，划定目标用户相较于初始探针的距离范围，即 $[Rd[1]_{\min}, Rd[1]_{\max}]$ 。此时，在以初始探针位置为中心的坐标轴上选择其余探针的位置时，其与初始探针的距离对判定目标所处象限的准确性影响很大。

不失一般性，假设目标用户位于坐标系的第四象限，且 P'_1 至 P'_4 距离初始探针的真实距离分别为 d'_1 至 d'_4 ，以 P'_3 为例进行讨论，如图 3 所示。

由于 $\text{dis}(P, T) \leq Rd[1]_{\max}$ ，且 $\angle P'_3PT$ 为钝角，则

$$\angle P'_3PT > 90^\circ > \angle PP'_3T \quad (7)$$

$$\text{且 } \angle P'_3PT > 90^\circ > \angle TPP'_3 \quad (8)$$

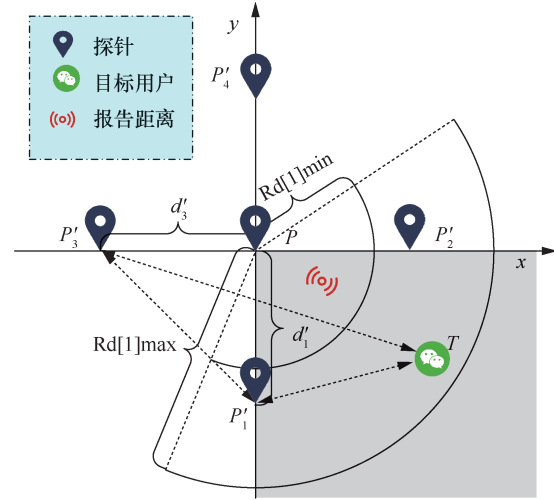


图 3 探针间距离关系示意

Figure 3 Diagram of distance relationship between probes

由三角形性质可知

$$\text{dis}(P'_3, T) > \max(\text{dis}(P, T), \text{dis}(P, P'_3)) \quad (9)$$

若探针 P'_3 与初始探针的距离 $\text{dis}(P, P'_3)$ 过小，即图中 d'_3 过小时，可能存在 $Rd[1]_{\max} > \text{dis}(P'_3, T) > d'_3$ ，此时，在探针 P'_3 处得到的目标用户的报告距离依然可能是 Rd_1 ，基于报告距离变化将无法判断目标所处的象限。

本文方法依据初始探针获取的目标用户报告距离的最大真实距离，来确定其他探针的位置分布。此时， $\text{dis}(P, P'_3) = Rd[1]_{\max}$ ， $\text{dis}(T, P'_3) > Rd[1]_{\max}$ 。即当目标用户位于坐标系的第四象限时，探针 P'_3 得到的目标用户的报告距离必然大于 Rd_1 ，因此，当 P'_3 得到的目标用户的报告距离不大于 Rd_1 时，目标用户必然不在第四象限。同理，当目标用户位于第一象限时，上述关系依然存在。

综上，本文方法基于探针获取的目标用户的报告距离大小来判断目标用户所处的象限，从而选定三边测量定位的探针位置，有助于确定更合理的探针位置分布。

4 基于估计位置确定距离边界权重

通过设置不同社交账号的位置，能够利用 LBSD 服务获取不同真实距离下的报告距离，从而获得不同报告距离对应的真实距离范围。假设报告距离 Rd 对应的真实距离边界为 $[150 \text{ m}, 350 \text{ m}]$ ，则当两个用户间的真实距离在此范围内时，其查询

服务可能得到对方的报告距离为 R_d 。需要说明的是，由于系统误差的影响，不同报告距离的真实距离边界存在重叠部分，即用户在某个真实距离条件下，其能够得到的对方的报告距离可能不止一种。

最小二乘法作为一种数学优化方法，通过对真实数据与多组观测数据的误差（即目标函数）进行最小化，来估计真实数据^[20-21]。该方法应用于 LBSD 目标用户定位时，如何选取观测数据及如何建立目标函数对于估计结果的准确性至关重要。

4.1 观测数据选取

在定位目标过程中，探针能够获取目标用户的报告距离。然而，位置混淆的存在导致报告距离并不准确，目标用户的真实距离可能远大于或远小于该报告距离。如图4所示，目标用户的报告距离为 R_d ，且 R_d 的真实距离边界分别为 $R_{d_{\max}}$ 和 $R_{d_{\min}}$ 。在定位目标用户时，探针仅能获取目标用户被混淆后的报告距离 R_d ，若 $R_{d_{\max}}$ 与 R_d 差距较大，且目标用户的真实距离略小于 $R_{d_{\max}}$ 时，此时目标的真实距离与 R_d 之间存在较大的差距，以 R_d 作为观测数据无法准确反映目标用户相对于探针的真实距离状态。以 R_d 为观测数据构建目标函数，并利用最小二乘法估算与其误差最小的目标用户位置时，定位结果并不可靠。

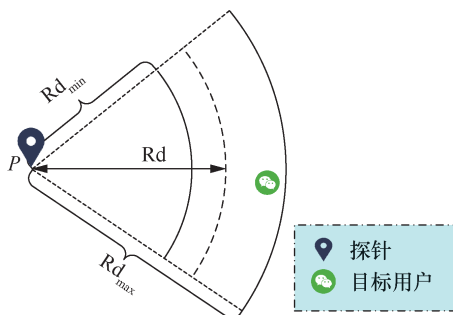


图4 目标用户真实距离与距离边界的关系示意
Figure 4 Diagram of relationship between real distance and distance boundary

在定位目标时，基于探针获取的目标用户的报告距离可以划定目标用户的真实距离范围，这个距离范围可以认为是准确的，即当目标用户的报告距离为 R_d 时，其相对于探针位置的真实距离必然在 $R_{d_{\max}}$ 和 $R_{d_{\min}}$ 之间。因此，本文方法选择

报告距离对应的真实距离边界作为观测数据，并利用其构建目标函数，能够较为准确地反映目标用户的真实距离状态。

4.2 目标函数构建

目标函数应相对准确地反映目标用户的真实距离状态。本文方法选择报告距离对应的真实距离边界作为观测数据，然而，由于观测数据包含两个值（最远真实距离和最近真实距离），在构建目标函数时，需要对两个距离边界赋予不同的权重。

基于直观的想法，在以探针为圆心、以最远真实距离和最近真实距离分别作为外径和内径的环形区域中，当目标用户更靠近内径时，说明目标的真实距离与最近真实距离之差更小，此时最近真实距离应获得更高的权重；否则，最远真实距离应获得更高的权重。因此，所提方法采用式(2)和式(3)所述方式，确定最近真实距离和最远真实距离的权值。

通过初步定位过程，能够得到不同圆环的交叉区域，该交叉区域为一个不规则曲边图形。由于目标的真实距离会影响报告距离，直接影响圆环的内径和外径，进而对交叉区域的形状产生影响。因目标的真实距离范围由圆环严格限定，确定目标的初步定位结果时要保证定位结果在交叉区域内，本文方法选择重心作为初步定位结果。通过计算该交叉区域的重心，能近似推断目标用户相较于探针的真实距离。此时，如不更新探针位置，将无法构建有效的目标函数实现对目标用户更高精度的定位。因此，得到初步定位结果后，更新探针位置是必要的。通过对探针位置进行更新，得到的新的探针位置依然满足步骤2中对目标用户的象限判定。此时依据初步定位结果，能够确定构建目标函数时最远距离和最近距离的权重。

交叉区域重心位置的确定方法如图5所示。首先，计算多个圆环交叉区域的顶点坐标，为了便于计算，基于这些顶点坐标做凸多边形，设凸多边形的边数为 n ，将该 n 多边形的重心作为不规则交叉区域的重心；然后，从凸多边形的一个顶点出发，连接与其不相邻的其余 $n-3$ 个顶点，将图形分割成 $n-2$ 个三角形；接着，分别求各三

角形的重心和面积, 记为 L 和 S 。

$$L = \{(x_i, y_i) | 0 < i \leq n-2\} \quad (10)$$

$$S = \{s_i | 0 < i \leq n-2\} \quad (11)$$

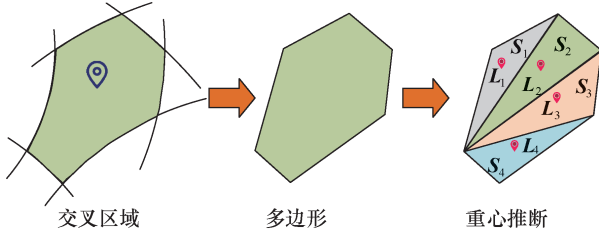


图 5 交叉区域重心位置示意

Figure 5 Diagram of center of gravity position in intersection area

最后, 根据各三角形面积占多边形总面积的比重, 确定该三角形重心坐标的权值, 将各重心的加权平均值作为该交叉区域的重心 L' 。设 $L' = (x, y)$, 则

$$x = \frac{\sum_{i=1}^{n-2} s_i \cdot x_i}{\sum_{i=1}^{n-2} s_i} \quad (12)$$

$$y = \frac{\sum_{i=1}^{n-2} s_i \cdot y_i}{\sum_{i=1}^{n-2} s_i} \quad (13)$$

上述过程能够求得交叉区域的重心, 将其作为中间定位结果, 该结果一定程度上能够反映出目标用户在圆环中的大致位置。因此, 在构建目标函数时, 基于中间定位结果确定最远距离和最近距离的权重, 相比对其设置同等权重, 更具合理性, 更有助于提高真实环境下的定位精度。

5 时间复杂度分析

现有方法在定位目标用户时往往需要对 LBSD 服务进行频繁访问, 以寻找报告距离的统计特征。由于社交平台加强对服务查询次数的限制, 现有方法的效率和实用性受到了极大制约。而本文方法在确定报告距离对应的真实距离边界时, 受服务查询次数限制的影响较小。实际上, 现有方法大多存在这种数据预处理过程。例如, 现有 SPBG 方法在定位开始前需要首先进行大量测试以确定报告距离的真实距离边界。因此, 本

文以定位目标过程中所需的探针对 LBSD 服务的访问次数为指标, 对比所提方法与现有典型代表性定位方法的时间复杂度。

现有 SPBG 方法首先通过查询目标的报告距离, 来确定目标对应的真实距离范围, 并通过对目标所位于的空间区域进行二分判断实现目标定位, 其整体时间复杂度为 $O(\log N)$ 。HNBG 方法首先通过在所构建的坐标系的坐标轴上等距部署大量探针, 推断初始估计坐标, 然后分别沿着初始估计坐标执行一维定位方法得到最终定位结果, 其整体时间复杂度为 $O(N)$ 。

与现有方法相比, 本文提出的基于加权最小二乘的社交用户高效定位方法采用了相同的目标发现策略, 但区别在于所提方法在定位目标之前, 基于大量的实际测试, 建立了各报告距离对应的真实距离范围, 从而避免了确定目标距离范围过程中, 对目标用户报告距离的频繁迭代查询或校验。在定位目标过程中, 仅需要对目标报告距离的数次查询, 即可利用最小二乘法推断目标位置, 给出定位结果, 其时间复杂度为 $O(1)$ 。

综上, 本文所提方法在时间复杂度方面, 相比现有典型的代表性方法具有明显优势。

6 实验结果与分析

为了验证所提方法在实际环境下的定位效率及定位精度, 本节基于微信社交平台, 设计并实施了用户定位实验。微信作为一款流行的社交软件, 截至 2021 年年底, 在全球范围内的月活跃用户超 12.6 亿, 且微信“附近的人”是一类典型的基于位置的交友服务^[22]。微信以 100 m 为单位展示附近用户的距离, 当报告距离大于 1 km 时, 其变化粒度发生变化 (1 km 为单位)。本文实验仅考虑报告距离小于 1 km 时的用户定位。实验结果与现有典型方法 HNBG^[17]、SPBG^[13]和 RRABG (relation between reported and actual distance based geolocation)^[15]做对比。

6.1 实验设置

本文实验包括两个部分: 真实距离边界识别实验及微信用户定位实验。实验设置如表 1 所示。

表1 实验设置
Table 1 Experiment Settings

项目	设置
报告距离范围	100 m 至 900 m
目标数量	500
使用微信账号数量	10
实验区域	郑州市市区 5 km×5 km 的区域
测试轮数	50 轮
数据条数	11 724
Δd	5 m
目标发现过程中相邻探针间隔	400 m
平移距离 σ	100 m
自动化工具	Appium
运行环境	Nox Player 模拟器
最小二乘实现函数	Python/scipy/optimize/bfgs
多边形中心点求解	Python/shapely/intersection

本文利用多个微信账号，实施真实距离边界确定实验，为了便于阐述该过程，仅以用户 A 和用户 B 来表示该过程。首先，将微信用户 A 位置固定，作为查询者；用户 B 作为附近用户。用户 B 的初始位置与 A 相同，之后调整其位置使其与用户 A 的距离以 5 m 为单位逐渐增大。每次用户 B 的位置变化后，用户 A 查询附近用户并记录用户 B 在该距离下的报告距离，直到用户 B 的报告距离出现 1 km。上述过程称为一轮测试，共计进行了 50 轮测试。基于得到的测试结果，划定 9 种不同报告距离（100 m 至 900 m，以 100 m 为单位变化）的真实距离边界。

目标用户定位实验部分，选取郑州市市区 5 km×5 km 的区域作为实验区域，利用安卓模拟器将多个微信账号的位置随机设置在实验区域中，将其作为待定位的目标用户。利用其他微信账号作为探针，探针在实验区域内每隔 400 m 查询一次附近用户，直到目标用户出现在探针的附近的人列表中，然后对其执行定位。成功定位所有目标后，重新按照上述方法更新目标的位置，直至得到 500 个定位结果。

所对比的 HNBG、SPBG 和 RRABG 方法采用文献[17]、文献[13]和文献[15]的实验设置。

6.2 真实距离边界识别

基于上述实验设置，共计得到 11 724 条<真

实距离，报告距离>记录，按照 2.1 节中介绍的距离边界判定方法，得到的报告距离（100 m 至 900 m）对应的真实距离范围如表 2 所示。

表2 报告距离对应的真实距离范围
Table 2 True distance range of the reported distance
单位：m

报告距离	最小真实距离	最大真实距离	距离范围
100	0	205	205
200	35	260	225
300	125	405	280
400	240	515	275
500	335	585	250
600	430	715	285
700	495	825	330
800	630	940	310
900	685	1085	400

由表 2 可知，微信的报告距离对应的真实距离范围较大，且该距离范围与报告距离大致呈正相关。随着用户间真实距离增大，报告距离的不确定性也会增大。由于用户更期望与附近的用户建立社交关系，因此这种位置混淆策略是合理的。基于获得的真实距离边界，利用所提方法对微信目标用户进行实际定位。

6.3 定位结果对比与分析

本文从定位精度和定位效率两个方面对实验结果进行分析与对比。通过计算各方法的定位结果（经纬度）与目标真实位置之间的偏差来计算定位误差；以在发现和定位目标用户的过程中，各探针对目标用户报告距离的查询次数为指标，对比各方法的定位效率。

（1）定位精度对比与分析

通过对 500 个微信目标用户的定位，分别计算所提方法和对比方法的定位误差，其结果如图 6 所示。

图 6 展示了所提方法与对比方法的平均定位误差和最小误差。由图 6 可知，在微信“附近的人”加强用户位置混淆的前提下，所提方法依然能够有效地定位微信用户，最小误差达到 13.6 m，平均误差为 53.7 m。与现有典型微信用户定位方法相比，平均定位误差分别降低了 23.6%、20.3%和 10.2%。

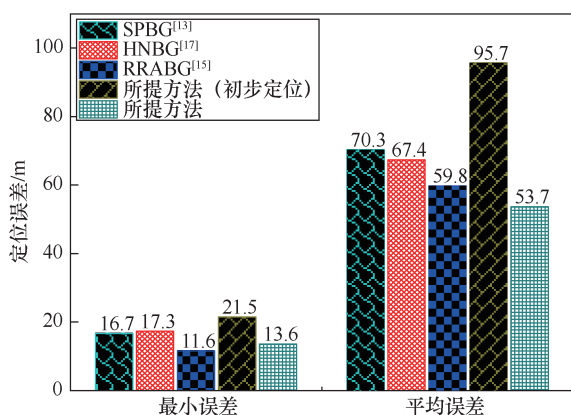


图 6 定位误差对比
Figure 6 Comparison of geolocation error

各方法的定位误差分布情况如图 7 所示，图中横坐标表示定位误差的分布区间，纵坐标表示不同定位误差分布区间内的定位结果在全部定位结果中所占的比例。由图 7 可知，所提方法能够将 27% 左右 ($0.059+0.213$) 的目标定位在 50 m 以内，将超过 70% ($0.059+0.213+0.452$) 的目标定位在 80 m 以内，高于对比方法。其中，所提方法的初步定位结果与目标真实位置的误差较大，这是由于微信的位置混淆机制使报告距离对应的真实距离范围较大，基于该距离范围的三边测量定位方法得到的交叉区域也较大，难以实现对目标的高精度定位。而本文方法在对目标实施两轮三边测量定位的基础上，通过采用加权最小二乘法求目标位置的最优解，能够进一步提高定位精度。

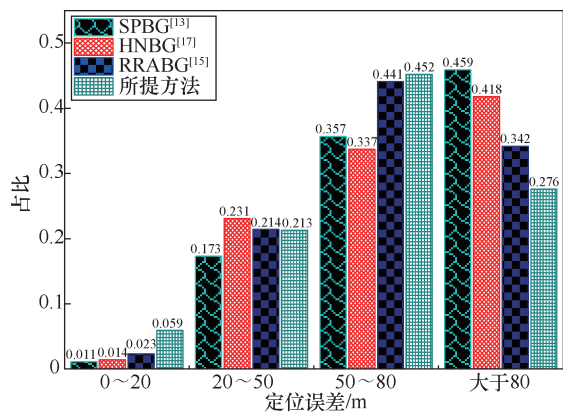


图 7 误差分布对比
Figure 7 Comparison of error distribution

以探针在测试区域内发现目标用户并开始定位时，目标用户的报告距离为依据，对实验结果

进行分类讨论，分别计算各初始报告距离对应的平均定位误差，其结果如图 8 所示。由图 8 可知，目标用户的初始报告距离越近，其定位平均误差相对越小。当目标用户的初始报告距离大于 700 m 时，所提方法对其的定位误差显著增大。这是由于此时报告距离对应的真实距离会更加离散，且变化幅度更大，因此所提方法在对目标进行初步定位和精确定位时均会存在较大的误差。在实际定位过程中，通过在目标用户发现阶段设置对目标用户报告距离的限制，如当探针发现目标用户且目标的报告距离小于 600 m 时执行定位，将进一步提高对目标的定位精度，但会带来更大的定位开销。

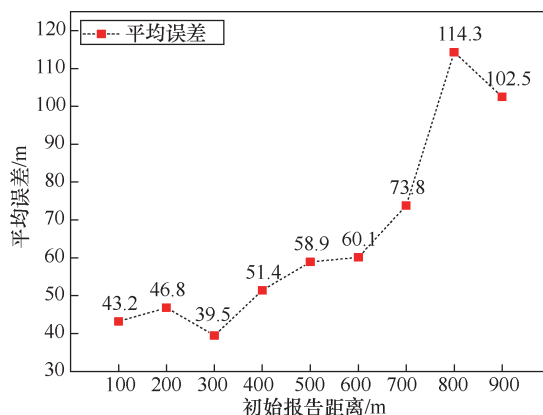


图 8 目标不同初始报告距离下的定位平均误差
Figure 8 Average geolocating error of target under different initial reporting distances

(2) 定位效率对比与分析

所提方法和对比方法对目标用户报告距离的平均查询次数如图 9 所示。RRABG 方法是在 SPBG 方法的基础上，通过多探针协作判定，提高空间划分的成功率，其所需查询次数大于 SPBG，因此本节选择效率更高的 SPBG 方法作为对比方法之一。图 9 中 AQT-F 表示方法在目标发现过程中的平均查询次数 (average query times of target finding)，AQT-G 表示在目标定位过程中的平均查询次数 (average query times of target geolocating)，AQT-T 表示两者之和。由于 HNBBG 方法讨论 $1\text{ km}\times 1\text{ km}$ 范围内的目标定位，因此本文将测试区域分为 25 个 $1\text{ km}\times 1\text{ km}$ 区域，并分别利用 HNBBG 方法对各区域内的目标进行定位，计算 AQT-G。

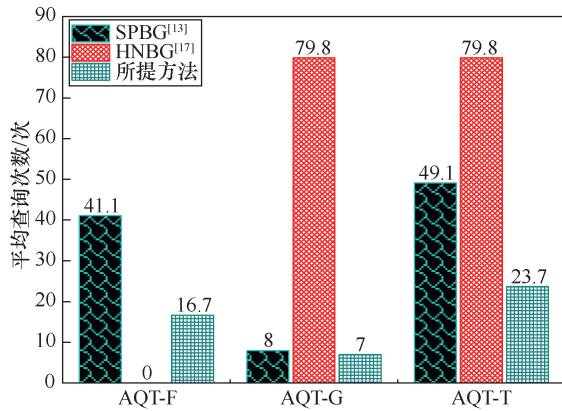


图9 各方法定位效率对比
Figure 9 Comparison of geolocating efficiency

由图9可知,所提方法在完整定位目标过程中对目标用户报告距离的平均查询次数为23.7次,远低于对比方法,具有更高的效率。SPBG方法在目标发现过程中,除了使目标出现在探针附近用户列表中,还需限制目标用户的报告距离,从而需要更多的查询次数。HNBBG方法不存在目标发现过程,因此其AQT-F为0,但其需要沿区域边界部署大量探针以确定目标用户的初始坐标,并在应用一维数论方法精确定位目标时,依然需要对目标用户报告距离进行频繁查询,所需查询次数最多。

综上,相比现有典型的微信用户位置推断方法,本文所提方法具有更好的定位精度,且定位效率远高于现有方法。实验结果表明,当前微信社交平台对于用户距离文本的隐私保护并不能有效防止用户准确距离的泄露,对报告距离文本的混淆有待进一步加强。

7 结束语

为了验证当前社交网络对于用户位置隐私保护的有效性,本文提出了一种基于加权最小二乘的社交网络用户定位方法。利用该方法对微信用户开展了实际定位,结果表明,微信现有位置隐私保护机制不能有效防止用户位置隐私泄露,仍需在保证服务质量的前提下进一步加强对距离文本的混淆强度。本文工作期望能加强社交平台对于用户位置隐私保护的重视。为了抵御本文提出的定位方法,一种可行的方案是服务提供商对报告距离对应的真实距离边界进行定时更新,从而缩短定位参数的有效时间域,增大定位用户所需的代价。此外,基于差分隐私等的隐私增强方法^[23-24]的

尽快应用将进一步提升用户位置隐私的安全性。

参考文献:

- [1] 姚瑞欣,李晖,曹进,等. 社交网络中的隐私保护研究综述[J]. 网络与信息安全学报, 2016, 2(4): 33-43.
YAO R X, LI H, CAO J, et al. Overview of privacy preserving in social network. Chinese Journal of Network and Information Security, 2016, 2(4): 33-43.
- [2] 李宇溪,周福才,徐紫枫. 支持 K-近邻搜索的移动社交网络隐私保护方案[J]. 计算机学报, 2021, 44(7): 1481-1500.
LI Y X, ZHOU F C, XU Z F. Privacy-preserving K-nearest-neighbor search over mobile social network[J]. Chinese Journal of Computer, 2021, 44(7): 1481-1500.
- [3] ZHENG Y. Location-based social networks: users[M]. Springer, New York, 2011.
- [4] LI J, YAN H, LIU Z, et al. Location-sharing systems with enhanced privacy in mobile online social networks[J]. IEEE Systems Journal, 2017, 11(2): 439-448.
- [5] WANG H, LI Y, YANG C, et al. Co-location social networks: linking the physical world and cyberspace[J]. IEEE Transactions on Mobile Computing, 2016, 18(5): 1028-1041.
- [6] XIE R, CHEN Y, LIN S, et al. Understanding skout users' mobility patterns on a global scale: a data-driven study[J]. World Wide Web, 2018, 22(1): 2655-2673.
- [7] NOVAK E, LI Q. Near-pri: private, proximity based location sharing[C]//33rd IEEE Conference on Computer Communications. 2014: 37-45.
- [8] LUO X, QIAO Y, LI C, et al. An overview of microblog user geolocation methods[J]. Information Processing & Management, 2020, 57(6): 102375-102392.
- [9] ZHENG X, HAN J, SUN A. A survey of location prediction on Twitter[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 30(9): 1-11.
- [10] ABHINAV K, JYOTI P. Location reference identification from tweets during emergencies: a deep learning approach[J]. International Journal of Disaster Risk Reduction, 2019, 2(33): 365-375.
- [11] PUTTASWAMY K, WANG S, STEINBAUER T, et al. Preserving location privacy in geosocial applications[J]. IEEE Transactions on Mobile Computing, 2014, 13(1): 159-173.
- [12] DING, Y., PEDDINTI, S. T., ROSS, K. W. Stalking Beijing from Timbaktu: a generic measurement approach for exploiting location-based social discovery[C]//4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices. 2014: 75-80.
- [13] LI M, ZHU H, GAO Z, et al. All your location are belong to us: breaking mobile social networks for automated user location tracking[C]//15th ACM International Symposium on Mobile Ad Hoc Networking and Computing, 2014: 43-52.
- [14] CHENG H, MAO S, XUE M, et al. On the impact of location errors on localization attacks in location-based social network service

- es[C]/9th International Conference on Security, Privacy and Anonymity in Computation, Communication and Storage. 2016: 343-357.
- [15] SHI W, LUO X, ZHAO F, et al. Geolocating a WeChat user based on the relation between reported and actual distance[J]. International Journal of Distributed Sensor Networks, 2018, 14(4): 1-13.
- [16] XUE M, LIU Y, ROSS K, et al. I know where you are: thwarting privacy protection in location-based social discovery services[C]/IEEE Conference on Computer Communications Workshops. 2015: 179-184.
- [17] PENG J, MENG Y, XUE M, et al. Attacks and defenses in location-based social networks: a heuristic number theory approach[C]/36th IEEE International Symposium on Security and Privacy in Social Networks and Big Data. 2015: 64-71.
- [18] POLAKIS I, ARGYROS G, PETSIOS T, et al. Where's Wally?: precise user discovery attacks in location proximity services[C]/22nd ACM SIGSAC Conference on Computer and Communications Security. 2015: 817-828.
- [19] MORIYA K, FUJIMOTO M, ARAKAWA Y, et al. Effective trilateration-based in-door localization method utilizing active control of lighting devices[J]. Sensors Mater, 2020, 32(2): 625-650.
- [20] LIU J, ZHANG Y, ZHAO F. Robust distributed node localization with error management[C]/7th ACM International Symposium on Mobile Ad Hoc Networking and Computing. 2006: 250-261.
- [21] LUENBERGER D, YE Y. Linear and nonlinear programming[M]. MA: Addison-Wesley, 1984.
- [22] Tencent Homepage[EB].
- [23] SIDDULA M, LI Y, CHENG X, et al. Privacy-enhancing preferential LBS query for mobile social network users[J]. Wireless Communications and Mobile Computing, 2020(4): 1-13.
- [24] JIANG H, LI J, ZHAO P, et al. Location privacy-preserving mechanisms in location-based services: A comprehensive survey[J]. ACM Computing Surveys, 2021, 54(1): 1-36.

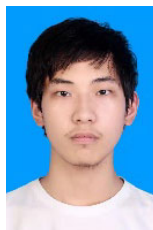
[作者简介]



时文旗（1993—），男，河南周口人，网络空间态势感知河南省重点实验室硕士生，主要研究方向为网络目标定位、网络空间资源测绘。



罗向阳（1978—），男，湖北荆门人，网络空间态势感知河南省重点实验室教授、博士生导师，主要研究方向为网络空间资源测绘、多媒体安全。



郭家山（1998—），男，河南郑州人，网络空间态势感知河南省重点实验室硕士生，主要研究方向为网络空间资源测绘。