基于位置服务的多因素假位置选择算法

李勇军, 祝跃飞, 吴魏, 白利芳

引用本文

李勇军, 祝跃飞, 吴魏, 白利芳. 基于位置服务的多因素假位置选择算法[J]. 计算机科学, 2025, 52(5): 357-365.

LI Yongjun, ZHU Yuefei, WU Wei, BAI Lifang. Multi-factor Dummy Location Selection Algorithm in Location-based Service [J]. Computer Science, 2025, 52(5): 357-365.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240200067

基于位置服务的多因素假位置选择算法

李勇军[1,2]　祝跃飞[1]　吴　魏[1]　白利芳[1,3]
1 信息工程大学网络空间安全学院　郑州 450007
2 中原工学院软件学院　郑州 450000
3 中国软件评测中心网络安全测评工程技术中心　北京 100048
(106449285@qq.com)

摘　要　针对现有假位置在进行基于位置服务的快照位置隐私保护时,忽略位置本身时间因素引发的背景知识攻击,以及对敏感位置同等对待等问题,提出一种多因素的假位置选取算法(Multi-Factor Dummy Location Selection Algorithm,MFDLS)。该算法综合考虑了影响隐私泄露的因素,包括位置的地理属性、语义属性、时间属性,以及查询概率等背景知识和用户敏感偏好,确保所选假位置不仅能有效抵御位置同质攻击、位置语义攻击和查询概率分布攻击,还能应对位置分布攻击、敏感同质攻击和链接攻击等多种威胁。算法选取满足与当前请求时间段内查询概率接近,语义多样化、匿名空间大且时间相对一致,非离群点和中心点要求的假位置。安全性分析和仿真实验结果表明:与已有的假位置选取算法相比,所提算法在敌手错误方面提升16%以上,质量损失方面降低30%以上,能更有效地抵御背景知识攻击,满足用户隐私需求。

关键词:假位置选择;多因素;地理位置;查询概率;位置语义;位置时间属性;敏感语义
中图分类号　TP309

Multi-factor Dummy Location Selection Algorithm in Location-based Service

LI Yongjun[1,2],ZHU Yuefei[1],WU Wei[1] and BAI Lifang[1,3]
1 School of Cyberspace Security,PLA Information Engineering University,Zhengzhou 450007,China
2 School of Software,Zhongyuan University of Technology,Zhengzhou 450000,China
3 Cybersecurity Testing Engineering Technology Center,China Software Testing Center,Beijing 100048,China

Abstract　In view of the existing dummy location selection methods in LBS snapshot location privacy protection,the background knowledge attack caused by the time factor of the location itself is ignored,and the sensitive locations are treated equally. Based on this,a multi-factor dummy location selection algorithm(MFDLS) is proposed,which comprehensively considers the factors that affect privacy leakage,including background knowledge such as geographical attributes,semantic attributes,time attributes of the location and query probability as well as the users' sensitive preferences. To ensure that the selected dummy locations can not only effectively resist location homogeneity attack,location semantic attack and query probability distribution attack,but also deal with multiple threats such as location distribution attack,sensitive homogeneity attack and link attack. The algorithm selects the dummy locations that meet the requirements of query probability close to the initiating time,semantic diversification,large anonymous space and relatively consistent time,non-outlier and central point. Compared with the existing dummy location selection algorithm,the security analysis and simulation results show that the proposed algorithm improves the adversary error by at least 16% and reduces the quality loss by at least 30%,which can more effectively resist the background knowledge attack and meet the users' privacy requirements.

Keywords　Dummy location selection,Multi-factor,Geographical location,Query probability,Location semantic,Location time attribute,Sensitive semantic

1　引言

随着定位技术及移动互联网的发展,基于位置的服务(Location-Based Service,LBS)作为智能终端的典型应用,已被广泛应用到社会生活、公共安全、交通、医疗等各领域[1]。目前,LBS已成为最有前途的移动服务之一,并在社会和商业领域取得了巨大的成功[2]。LBS虽然为用户提供了多样化的便捷服务,但也带来了严重的隐私泄露问题,引发了人们的担

到稿日期:2024-02-20　返修日期:2024-06-28
基金项目:科技委基础加强项目(2020-JCJQ-ZD-021);国家自然科学基金青年基金项目(62102447)
This work was supported by the Foundation Strengthening Project of Science and Technology Commission(2020-JCJQ-ZD-021) and National Natural Science Foundation of China(62102447).
通信作者:祝跃飞(yfzhu17@sina.com)

忧，具体包括位置隐私和查询隐私。如何保障用户在享受高质量的 LBS 服务的同时，确保自身的隐私安全，受到了学术界和工业界的广泛关注。

针对 LBS 位置隐私保护问题，国内外学者提出了许多隐私保护方法[1,3-7]，其有效保护了用户的位置隐私。假位置方法易于操作实现，因此受到学者们的青睐。通常，假位置方法与 k-匿名结合使用，而 k-匿名易受到如历史查询概率导致的概率分布攻击[8]，为此，学者们在生成或选取假位置时，考虑了查询概率、位置语义及物理分散度等因素，以抵御拥有背景知识的攻击者。目前的研究方案在保护位置隐私时，忽略了用户敏感位置偏好及位置本身的时间属性影响。由于每个用户对敏感位置的定义并不完全相同，如职业是医生的用户对医院的敏感度没有其他职业用户的敏感度高，因此假位置选取在抵御敏感同质攻击[9]的同时，需考虑用户个性化的敏感需求。在 LBS 应用中，除上述因素外，位置本身的时间因素也会导致推理攻击。例如，用户晚上 10 点在商业街发起最近酒店的查询请求，若产生的虚假位置是银行，而攻击者又知用户非银行工作人员，那么用户真实位置被暴露的风险就会增加。本文基于所述情况，综合考虑位置的地理属性、语义属性、时间属性、用户敏感偏好、分时段的查询概率等因素，提出了一种考虑多因素的假位置选择算法。本文的主要贡献如下：

1)为保护位置隐私，考虑位置时间属性的影响，提出多因素的假位置选取算法。假位置选取时综合考虑造成隐私泄露的多方因素，保证所选假位置的查询概率接近、语义信息多样化、时间属性一致，同时保障所选位置分散且不是离群点和假位置集的中心。

2)考虑了用户敏感语义偏好，安全性分析及实验结果表明，所提方法能够抵御更多攻击，如位置分布攻击、敏感同质攻击等。

3)在公开的真实数据集上进行的评估表明，相比其他算法，本文算法在抵御敌手敏感语义攻击和基于时间的链接攻击方面的保护效果更好，同时质量损失也有所降低。

2 相关工作

2.1 基于位置的方法

LBS 中的位置通常被赋予了城市功能，位置本身包括地理属性、语义属性和时间属性。

位置的地理属性在假位置方法中是不可忽略的因素，只是应用方式不同。2005 年 Kido 等[10]最早提出基于虚假位置的方法，直接使用位置的地理属性，采用随机游走模型生成虚拟位置；Lu 等[11]指出随机游走模型生成的假位置最终会变得极为密集，这仍然直接依赖于地理属性，他们提出的 Grid-Dummy 和 CirDummy 生成算法将虚拟网格或圆形区域划分成子区域，选择不同的顶点/半径作为虚假位置，从而保证较大隐私区域中合理的查询处理代价；Niu 等[12]提出的增强型熵度量的假位置选择(enhanced-Dummy Location Selection，enhanced-DLS)算法，为使文中设计的基于熵度量的假位置选择(Dummy Location Selection，DLS)形成的匿名区较大，使用位置的乘积距离来衡量位置分散度；Zhang 等[13]利用海伦公式筛选假位置；Zhang 等[14]使用位置的地名信息进行近似匹配，选取合适的假位置生成匿名集；Wang 等[15]使用曼哈顿距离来进行初次假位置生成；Song 等[16]通过欧氏距离均匀生成假位置；Jiao 等[6]使用鞋带定理计算假位置集中所有位置形成的多边形面积，以此衡量位置离散度；Zhang 等[7]计算假位置集中所有位置到位置集最小覆盖圆的圆心的欧氏距离，然后利用这些距离计算出距离熵，以此来衡量位置离散度。

为保证假位置集能够抵御位置语义攻击，需保证假位置集语义的多样化。Jiao 等[6]在选择假位置时，考虑到语义距离，由百度地图的逆地理编码 API 获取语义信息；Zhang 等[7]采用 Wi-Fi AP 收集覆盖范围内的位置语义信息，使用调整的余弦相似度构建具有语义差异性的假位置集；Zhang 等[13]基于 WordNet 结构计算语义相似度；Zhang 等[14]使用编辑距离计算侯选集中假位置的语义相似度；Hara 等[17]考虑到位置语义，提出了一种控制用户访问地点偏好保留程度的匿名方法，但忽略了假位置离散度对隐私保护的影响。Wang 等[18]利用位置语义树描述位置集的语义差异，该算法需要提前构建位置语义树，适用于 POI(Point of Interest)类型较多的区域，因此在某些区域匿名效果可能较差；Wang 等[15]利用语义加权有向图描述语义的时间分布及转移关系；Tu 等[19]将地名作为语义信息。

2.2 基于查询概率的方法

Niu 等[12]设计的 DLS 算法基于 Borlange 数据集产生查询概率，选取与用户所处真实位置查询概率尽可能相似的位置作为假位置；Zhang 等[13]采用了网格划分方法，将每个网格的历史查询次数与所有网格历史查询次数总和的比值作为历史查询概率，在选择假位置时，优先选取历史查询概率与当前查询概率接近的位置；Wang 等[15]使用位置的查询次数与用户的历史查询总次数的比值作为位置历史概率，综合位置历史概率和大众评价信息衡量位置的可信度，以构造匿名集；Wang 等[18]利用 Wi-Fi AP 覆盖范围划分网格，计算每个网格内地理位置的历史查询概率，仍然选择查询概率接近的位置作为假位置；Tu 等[19]使用数据集内每个位置访问时间占比作为历史查询概率，选取历史查询概率与真实位置查询概率尽可能接近的位置；Xia 等[20]根据用户的服务请求概率分布情况，提出了一种基于半可信第三方的假位置匿名算法，并结合 Stackelberg 博弈模型对其改进。

以上研究未考虑到不同时段查询概率的差异，从而可能引发用户隐私泄露问题。基于此，Jiao 等[6]对位置查询概率进行时间分片，使用同一时间段内每个位置的出现次数占比作为分时段位置查询概率，优先选取分时段查询概率最为接近的位置作为假位置。Zhang 等[7]使用 Wi-Fi AP 收集其覆盖范围内的历史查询概率，使用过去一段时间内位置单元被查询的次数与所有位置单元在过去一段时间内被查询的总次数的比值作为位置单元的历史查询概率，仍选择优先查询概率接近的位置作为假位置。

2.3 基于用户偏好的方法

针对用户对隐私的需求不尽相同的问题，Wang 等[15]使用语义转移关系描述用户偏好；Hara 等[17]将用户偏好定义为访问地点的一系列类别；Li 等[21]根据位置和速度选择假位置，用户自定义敏感兴趣点，选择不同类别的敏感信息点进行

匿名，通过删除敏感兴趣点的方式保护用户位置隐私。但这些方案主要针对的是 LBS 轨迹情况。

以上文献在关注位置隐私时，能够充分考虑到位置的地理属性、语义属性和查询概率等，而目前考虑用户偏好的保护方案主要聚焦在 LBS 轨迹，且位置时间属性作为背景知识，当被攻击者获取后易受到链接攻击。例如生成的假位置是博物馆，通常它的开放时间为 9:00－17:00，而用户 8:00 已在博物馆出现，若已知用户身份并不是博物馆工作人员，很明显用户此时出现在博物馆肯定是不妥当的。因此，考虑位置时间属性是很有必要的。

基于此，本文提出一种综合用户敏感偏好和影响 LBS 位置隐私因素的假位置选取方案，从位置距离、位置语义、位置时间属性、查询概率、用户敏感偏好等方面进行筛选，保证假位置具有最大的真实性，能够更好地抵御背景知识攻击。

3 预备知识

3.1 系统架构

本文系统架构主要由卫星定位系统、Wi-Fi 接入点（Access Point，AP）、移动终端和 LBS 服务器组成，如图 1 所示。该架构中，移动终端被认为是可信的；卫星定位系统提供用户当前位置信息，Wi-Fi AP 收集地图信息、位置语义信息和不同时间段内的历史查询概率等，LBS 服务器为用户提供服务，被认为是不可信的。

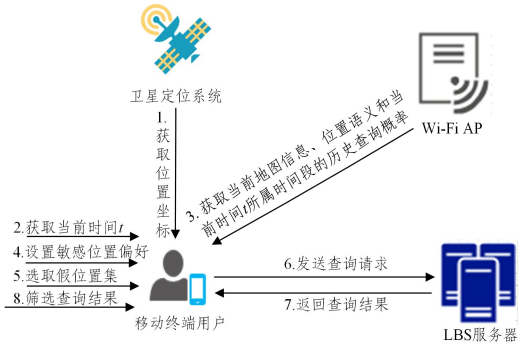

图 1 系统架构

Fig. 1 System architecture

从图 1 中可看出，用户在发起请求前，通过卫星定位系统获取当前位置信息，通过移动终端设置获取当前时间 t，接着从 Wi-Fi AP 获取覆盖范围内的地图信息、位置语义信息和当前时间段内的历史查询概率。用户自行设置敏感位置集，使用假位置选择算法，得到 $k-1$ 个假位置，其与真实位置共同组成 k 匿名集向 LBS 服务器发起查询请求，LBS 服务器解析请求内容，查询出相应的结果集返回到移动终端，移动终端对返回的结果集进行处理后，为用户提供最终结果。

3.2 攻击模型

按攻击方式，将攻击者分为被动攻击和主动攻击。被动攻击是通过截获、窃取用户所发送的信息来达到其目的，通常涉及合谋攻击。主动攻击的敌手会直接攻击 LBS 服务器，以获取服务器所存储的全部信息，包括但不限于用户发送的假位置集，进一步结合背景知识推断用户的隐私信息，实施推理攻击。两种方式存在的攻击包括以下类别。

位置分布攻击[22]：攻击者利用用户在匿名区域中的位置分布特点，推断出服务请求者的身份。例如，离群点攻击[23]中，攻击者根据用户位置分布不均匀的特点，推断出密度稀疏区域用户（离群点）的查询隐私信息。

位置同质攻击[24]：产生的虚假位置与真实位置过近，很容易推理出用户真实位置。例如，产生的虚假位置与真实位置之间的距离不超过 10 m，用户真实位置就很容易被暴露。

位置语义攻击[25]：产生的假位置与真实位置语义类型一致，可推理出用户真实位置的语义信息。若用户在医院发起请求，而产生的虚假位置也在医院，则敌手很容易推理出用户就在医院。

敏感同质攻击[9]：即使匿名集同时满足 k 匿名与 l 多样性，用户的隐私仍有可能被泄露。例如，匿名区域中的请求用户个数大于 k，且类型数大于 l，但其中每个请求都包含酒吧、医院等敏感语义，攻击者就可推测出用户提出的请求包含敏感内容。

查询概率分布攻击[24]：攻击者具有当前地图的历史查询概率，针对假位置集的查询概率进行分析，认为查询概率较高的位置大概率为用户真实位置。

聚类攻击：假位置集以真实位置为中心，易引起聚类攻击。若选取的假位置与真实位置较近，则通过中心聚类，很容易推理出用户的真实位置。

基于时间的位置链接攻击：假位置具有的时间属性与真实位置具有的时间属性不同，根据用户发起查询的时间或用户特征推理出假位置是不合理的。

3.3 最分散的假位置选取

3.3.1 假位置分散思想

图 2 给出了假位置选取时的 3 种情况。假设 $k=3$，图 2 实线区域 A_1 选择的假位置单元距离用户真实位置单元较近，且具有与用户真实位置单元相同的当前时间段查询概率；虚线区域 A_2 和点线区域 A_3 选择的假位置单元距离两个用户真实位置单元较远，同样具有与用户真实单元相同的当前时间段查询概率。虽然 3 种方式下假位置数量满足匿名度要求，但由于 A_1 中的假位置单元彼此之间距离很近，攻击者很容易将真实用户定位在一个很小的区域中，易遭受位置同质攻击。A_2 和 A_3 选取的假位置单元更加分散，用户被成功攻击的可能性大大降低，但从区域面积大小角度出发，A_3 被成功攻击的概率是最小的。

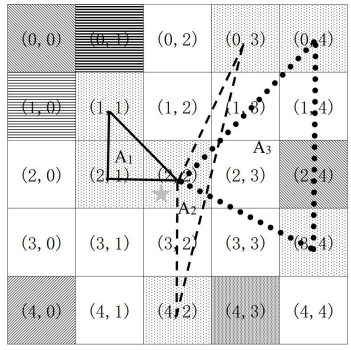

图 2 假位置分散思想

Fig. 2 Dispersion for dummy location selectio

3.3.2 寻找最分散位置

如何从图 2 中选取出 A_3 呢？本文按区域面积选取。假

位置集形成的区域通常是不规则区域，如图 3 所示。

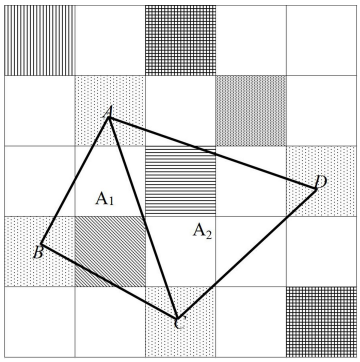

图 3 不规则区域示例

Fig. 3 Example of irregular region

假设 $k=4$，A 为用户真实位置所在位置单元的中心，B、C、D 为假位置所在位置单元的中心。以 4 个位置为顶点构成的多边形区域面积可通过以下方法进行计算：根据任意一个 n 边形总能被划分为 $(n-2)$ 个三角形定理，将图 3 中的四边形划分成 2 个三角形，每个三角形的面积分别记为 A_1 和 A_2，求出 AB、AC、AD、BC、BD 的距离。现利用海伦公式分别计算出 2 个三角形的面积，所得面积之和即为该区域的面积。海伦公式为：

$$A=\sqrt{r(r-a)(r-b)(r-c)}$$

其中，a、b、c 分别为三角形的边长，$r=(a+b+c)/2$ 是三角形的半周长，A 为三角形的面积。

为保证假位置集中位置不过于集中，须保证形成的不规则区域尽可能大，也即不规则区域面积最大。为抵御聚类攻击，最好保证用户真实位置不是假位置集中心点。

3.4 相关定义及问题描述

3.4.1 相关定义

设 Ω 为赋予城市功能的所有地理空间集，$D\in\Omega$ 为虚假位置集，$SP\in\Omega$ 为敏感位置集。位置 $p\in\Omega$，由地理属性 G、语义属性 S 和时间属性 T 构成，记作 $p=\langle G,S,T\rangle$。G 通常由经度 lat 和纬度 lon 坐标构成，记作 $G=(lat,lon)$；S 常用位置的 POI 来表示；T 为 p 的营业时间，通常由日期（周几）$\boldsymbol{w}$、时间区间 h 组成，记作 $T=(\boldsymbol{w},h)$。

定义 1（分时段位置查询概率，Time-Segment Location Query Probability，TSLQP）[6] 在给定时间段内，某个特定位置被查询的概率，如式(1)所示：

$$p_i^t=\frac{n_i}{\sum_{i=1}^{N^2}n_i} \tag{1}$$

其中，$i=1,2,\cdots,N^2$，t 表示当前所处的时间段，n_i 表示某个位置单元的出现次数，N^2 表示所有位置单元的总数。

定义 2（语义类别相似度） 设 c_p 为位置 p 的语义属性类别编码，p_1 和 p_2 的语义类别相似度如式(2)所示：

$$S(p_1,p_2)=\frac{L_1(c_{p_1},c_{p_2})+L_2(c_{p_1},c_{p_2})+L_3(c_{p_1},c_{p_2})}{|c_{p_1}|} \tag{2}$$

其中，$L_1(c_{p_1},c_{p_2})$，$L_2(c_{p_1},c_{p_2})$，$L_3(c_{p_1},c_{p_2})$ 分别表示 p_1 和 p_2 的大类、中类、小类是否相同。若某一类别相同，则相应的值为 2，否则为 0。

定义 3（时间相似度） 任意两位置 p_1 和 p_2 的时间相似度如式(3)所示：

$$T(p_1,p_2)=\frac{\sum_{i=1}^{7}w_1[i]\times w_2[i]}{\sqrt{\sum_{i=1}^{7}(w_1[i])^2}\times\sqrt{\sum_{i=1}^{7}(w_2[i])^2}}\times\frac{|h_1\cap h_2|}{|h_1\cup h_2|} \tag{3}$$

时间相似度由两部分乘积得出，第一部分是时间中 $\boldsymbol{w}$ 的相似度，采用 one-hot 编码方式，使用 7 维空间的余弦相似度进行计算；第二部分是时间中的 h 相似度，使用区间数相似度计算。

3.4.2 问题描述

假位置选取时，在查询概率方面，应选取查询概率与用户真实位置 p_{real} 的请求时间 t 所在时间段的查询概率相近的位置，即：

$$\arg\min\{\max(p'_{p_{\text{real}}}-p'_p)\},p\in D,\ p\neq p_{\text{real}}$$

在位置语义方面，应确保整个假位置集 D 的语义具有多样性，以最大化假位置集的语义差异性。因此，在选择时应当从侯选集中选出与 D 中已有位置语义类别相似度最小的位置，这样可最大程度地保证任意两位置的语义相似度在一个较低水平，即：

$$\arg\min\{\text{sum}(S(p_1,p_2))\},p_1,p_2\in D,p_1\neq p_2$$

在位置离散度方面，应选择使假位置集 D 面积最大的位置，设 A 为 D 的面积，即 $\arg\min\{1/A\}$。

在时间属性方面，应使整个假位置集 D 的时间具有一致性，使得最后假位置集的时间差异性最小。因此，在选择时应当选择侯选集中与 D 中所有位置时间语义相似度最大的位置，这样可最大程度地保证两位置的时间相似度较大，即：

$$\arg\max\{\text{sum}(T(p_1,p_2))\},p_1,p_2\in D,p_1\neq p_2$$

在敏感位置保护方面，应选择使假位置集 D 中所有位置与敏感位置集 SP 的语义类别相似度最小的位置，这样可最大程度上保证 D 中位置与敏感位置集中的语义相似度最小，即：

$$\arg\min\{\text{sum}(S(p_s,p_d))\},p_s\in SP,p_d\in D$$

综合考虑上述 5 个因素，可将多目标问题转化为极大值求解问题，即：

$$\arg\max\left(\frac{\text{sum}(T(p_i,p_j))}{\max(p'_{p_{\text{real}}}-p'_p)+S(p_i,p_j)+\text{sum}(S(p_s,p)+1/A}\right),p_s\in SP,p_i,p_j,p\in D,p_i\neq p_j,p\neq p_{\text{real}}$$

和位置分散度较大、营业时间相似的位置，与用户真实位置共同构成满足 k-匿名的假位置集。算法 1 给出了假位置选择算法的伪代码。

4 假位置选取算法

4.1 算法描述

本文提出的假位置选择算法根据分时段历史查询概率、位置语义信息、位置时间属性和地理位置信息来选择 $k-1$ 个与用户所处真实位置的分时段查询概率相似、位置语义距离

算法 1 假位置选择算法

输入：用户真实位置 p_{real}，用户匿名度需求 k，用户所处范围内的当前时段历史查询概率 p^t，位置语义信息，位置时间属性，用户敏感

位置集 SP
输出：假位置集 D
1. 将用户所处范围内当前时段历史查询概率按升序排列
2. 分别从真实位置 p_{real} 左边和右边选取 k 个位置作为侯选集 C
3. 将用户真实位置 p_{real} 加入到假位置集 D 中
4. while |D| ＜ k do：
5. for p in C：
6. sumSem ← 0
7. for dp in D：
8. 计算 p 与 dp 的语义类别相似度 S(p,dp)
9. sumS ← sumS＋S(p,dp)//计算 p 与 D 的语义相似度之和
10. endfor
11. if(|D| ＜ 2)：
12. A＝0；//D 中位置少于 2 时，无法形成多边形，面积为 0
13. else：
14. 将 p 加入到 D 中
15. 计算多边形面积 A
16. 将 p 从 D 中移除
17. endif
18. for dp in D：
19. 计算 p 与 dp 的时间相似度 T(p,dp)
20. sumT ← sumT＋T(p,dp)//计算 p 与 D 的时间相似度之和
21. endfor
22. for sp in S：
23. 计算 p 与 sp 的语义类别相似度 S(p,sp)
24. sumN ← sumN＋S(p,sp)//计算 p 与 SP 语义相似度之和
25. endfor
26. d ← sumT/(sumS＋1/A＋sumN)
27. 用 max 记录 d 的最大值，用 cur 记录 max 对应的位置 p//假位置集应时间一致，语义多样，分散度大，与敏感语义不同
28. endfor
29. 计算 D 的中心点 O
30. 计算 D 中位置到 O 的距离 Dis
31. 使用 max 记录 Dis 中的最大值
32. 计算 p_{real} 到 O 的距离 dis
33. if p_{real}≠O 并且 dis≠Dis：//不是中心点和离群点
34. 将 cur 加入到 D
35. 将 cur 从 C 中移除
36. endif
37. endwhile
38. return D

算法 1 输入的参数为用户真实位置 p_{real}、用户匿名度需求 k、用户所处范围内的当前时段历史查询概率 p^t、位置语义信息、位置时间属性、用户敏感位置集 SP。首先，根据分时段历史查询概率 p^t 得到与用户的分时段查询概率相近的 $2k$ 个位置形成侯选位置集 C（第 1－2 行），将真实位置 p_{real} 加入假位置集 D（第 3 行），遍历 C 中的每个位置（第 5 行），计算其与 D 中每个位置的语义类别相似度之和、匿名区域面积、时间相似度和（第 6－20 行），计算 C 中每个位置与敏感位置集中的语义类别相似度和（第 22－25 行），综合考虑 $d=sumT/(sumS+1/A+sumN)$（第 26 行），记录最大值及相应位置（第 27 行），判断不是离群点和中心点（第 29－33 行），将满足要求的位置 cur 加入到 D 中并从 C 中移除（第 34－35 行）。重复上述步骤（第 5－35 行），直到 D 中假位置的数量为 k，最后返回 D。

4.2 安全性分析

位置分散性分析：当假位置集都集中在很小的一块儿区域时，攻击者可锁定用户在地图上较为精确的位置，使得隐私信息暴露。本文算法在选择 p 加入 D 时，计算其与 D 所形成的多边形面积，选择面积最大的 p 加入 D，保证最终生成的 D 面积较大，使得攻击者无法确定用户位置；p_{real} 偏离假位置集中其他位置时，会成为离群点，易受到位置分布攻击，本算法在将 p 加入 D 前进行离群点判断，能够较好地抵御位置分布攻击；p_{real} 是假位置集中心时，易受到聚类攻击，本算法在将 p 加入 D 前进行中心点判定，能够较好地抵御聚类攻击。

语义安全性分析：当假位置集中的语义信息较少时，攻击者根据背景知识，易获取用户位置的语义信息，完成位置语义攻击。在本文算法中，对 D 中每个位置都计算与 p 的语义相似度，选择总体语义相似度加和最小的 p 加入，从而降低 D 中语义相似的位置数量，保证假位置集中语义的多样性，使攻击者无法确定用户真实位置的语义信息。

敏感语义安全性分析：即使假位置集中的语义信息较多，但语义均集中在用户敏感语义上，仍易受到敏感同质攻击。在本文算法中，对 D 中每个位置都计算与用户敏感位置集 SP 中所有位置的语义相似度，选择总体语义相似度加和最小的 p 加入，增加非敏感语义的位置数量，保证假位置集中非敏感语义的多样性，使攻击者无法确定用户真实位置的敏感语义信息。

查询概率安全性分析：本文算法在进行假位置选取时考虑用户发起查询请求的时刻，选取的假位置查询概率与用户真实位置的查询概率在该时间段内是最接近的，攻击者无法使用本文所述的概率分布攻击方式找到用户的真实位置，只能随机猜测，成功概率为 $1/k$。

时间安全性分析：若假位置中时间属性与用户真实位置 p_{real} 不同，如 p_{real} 的营业时间为晚 7：00－晚 12：00，而假位置集中除用户真实位置外的其他位置的营业时间均不在这个时间范围内，攻击者根据背景知识，易推理出用户真实位置，从而遭受基于时间的位置链接攻击。本文算法中，对 D 中每个位置都计算其与 p 之间的时间相似度，选择总体时间相似度加和最大的 p 加入，保证假位置集中位置的时间属性最为一致，使攻击者很难根据时间属性推理出用户真实位置。

5 实验结果与分析

5.1 实验设置

实验环境：Intel(R) Core(TM) i7-8650u CPU @ 1.90 GHz 2.11 GHz processor，16.0 GB RAM，Windows 10，PyCharm Edu 2020.3，Python 3.6.5。

实验数据使用 Geolife[26] 中数据最为密集的区域纬度[39.5，40]N，经度[116.30，116.35]E，此区域涉及 177 名用户，11 822 个位置，为 273.73 km^2 的方形区域，分成 200＊200 个网络，位置语义使用高德地图获取 POI 信息，分成大类、中类、小类。位置查询概率分为两种情况，一种是不考虑时间属

性的查询概率，另一种是考虑时间属性的查询概率。前者用当前网格中人数除以总人数 177，后者用当前时刻当前网格中人数除以总人数 177。所有网格的访问概率总和为 1。

为验证算法能够抵御基于位置的物理属性、语义属性、时间属性以及查询概率发起的攻击，本文从不确定性、位置分散度、语义多样性、时间一致性、敏感语义多样性 5 个方面进行评价。为衡量方案效果，从敌手错误和质量损失 2 个角度进行评估，并与 DLS 算法[12]、enhanced-DLS 算法、TSDLS 算法、DLG_SI 算法[7]行比较。

DLS 算法使用随机选取方式，从与真实位置查询概率最接近的 $2k$ 个假位置中选择，确保形成的假位置集具有最大熵值；enhanced-DLS 算法使用 DLS 算法形成 $2k$ 个侯选假位置集，从中选取遮罩面积最大的假位置集作为最终结果；TSDLS 算法从与真实位置分时段查询概率最接近的 $2k$ 个假位置中，选取使位置离散度和语义距离尽可能大的假位置以形成匿名区；DLG_SI 算法生成假位置时，首先选取与真实位置分时段查询概率最接近的 $8k$ 个假位置，其次从 $8k$ 个假位置中随机选取 m 组 $6k$ 大小的假位置集，接着从具有最大位置熵的一组假位置集中选出与真实位置查询概率相近的 m 组 $4k$ 个假位置集，再从具有最大时间熵的一组假位置集中选取满足 $2k$ 个语义要求的位置，最后从中选取距离熵最小的 $k-1$ 个位置形成假位置集。

5.2 评价指标

1)信息熵

信息熵用来衡量保护方法抵御查询概率分布攻击的能力。本文使用分时段位置查询概率，计算方式同文献[6]。

2)位置分散度

位理分散度用于衡量保护方法抵御位置分布攻击和位置同质攻击的能力。本文采用面积来衡量。假位置集所围面积越大，所选出的假位置(包括假位置和真实位置)就越分散，这样攻击者就很难将用户位置缩小到一个很小的范围内，从而提高用户的隐私效果。

3)语义差异性

语义差异性用来度量保护方法抵御位置语义攻击的能力。本文使用语义类别相似度来衡量。语义差异性越大，表明抵御位置语义攻击的效果越好。语义差异性的计算如式(4)所示：

$$SD=1-\frac{\sum SEM}{k} \tag{4}$$

其中，$SEM=\{S(p_1,p_2)\}$，p_1 和 p_2 分别为 D 中的任意两个位置。SD 的取值范围为[0,1]。

4)时间一致性

时间一致性用来度量保护方法抵御基于时间的位置链接攻击的能力。本文使用时间相似度来衡量。时间一致性越大，表明抵御基于时间的位置链接攻击效果越好。时间一致性的计算如式(5)所示：

$$TD=1-\frac{\sum TS}{k} \tag{5}$$

其中，$TS=\{T(p_1,p_2)\}$，p_1 和 p_2 分别为 D 中的任意两个位置。TD 的取值范围为[0,1]。

5)敏感差异性

敏感差异性用来度量敏感位置集与假位置集之间的语义差异性，进而衡量保护方法抵御敏感同质攻击的能力。本文使用语义类别相似度来衡量。敏感差异性越大，表明抵御敏感同质攻击的效果越好。敏感差异性的计算如式(6)所示：

$$ND=1-\frac{\sum_{p\in D}\sum_{sp\in SP}S(p,sp)}{k} \tag{6}$$

6)敌手错误

敌手错误[27]是敌手推理得出的位置与用户真实位置之间的偏差。设 $p,p'\in D,\hat{p}\in\Omega$，$p$ 为用户真实位置，p' 为经保护机制 M 扰动后生成的扰动位置，$\hat{p}$ 为敌手推理得出的位置，$\pi_a(p)$ 为敌手对 p 的先验知识，H 为敌手的推理机制，敌手错误如式(7)所示：

$$AE(\pi_a,M,H)=\frac{\sum_{p,p',\hat{p}}\pi_a(p)Pr(p'\mid p'=M(p))Pr(\hat{p}\mid\hat{p}=H(p'))d(\hat{p},p)}{S(p,\hat{p})\times T(p,\hat{p})} \tag{7}$$

7)质量损失

质量损失[27]是保护方法得出的位置与用户真实位置之间的偏差。设 $p,p'\in D,p$ 为用户真实位置，p' 为经保护机制 M 扰动后生成的扰动位置，用户对 p 的先验为 $\pi_u(p)$，则质量损失如式(8)所示：

$$Q^{\text{loss}}(\pi_u,M)=\frac{\sum_{p,p'}\pi_u(p)Pr(p'\mid p'=M(p))d(p,p')}{S(p,p')\times T(p,p')} \tag{8}$$

5.3 实验结果

5.3.1 不确定性

5 种算法和理论值在不同 k 值下的信息熵值结果如图 4 所示。随着 k 值的增加，所有算法的信息熵值都呈上升趋势。DLS 算法从侯选集中随机生成 m 个侯选假位置集，从中选取熵值最大的作为最终的假位置集，熵值具有随机性，并受参数 m 的影响，熵值选取时依据的是不分时段的查询概率。enhanced-DLS 在 DLS 的基础上，考虑了假位置面积，这样在选取时可能会避开一些使熵值最大化的假位置，因此其信息熵值比 DLS 的低。对比 TSDLS，DLG_SI 以及本文提出的 MFDLS 的熵值与理论熵值，TSDLS 更接近理论值，DLG_SI 由于首先根据不分时段的查询概率进行筛选，再使用分时间查询概率进行筛选，会过滤掉使信息熵值最大的假位置，因此信息熵值略低于 TSDLS。本文在进行假位置选取时，考虑的因素较多，在使目标值最大的同时未必能够保证信息熵值最大，因此本文方法相比 TSDLS 和 DLG_SI 而言，熵值略低。从图中明显可以看出，DLS 和 enhanced-DLS 的信息熵值低于 TSDLS，DLG_SI，MFDLS 和理论值，这反映了使用分时段的查询概率相比不分时段的查询概率而言，信息熵值会更大。从本次实验结果来看，理论值的平均熵值为 3.21，DLS 的平均熵值为 2.70，enhanced-DLS 的平均熵值为 2.59，TSDLS 的平均熵值为 3.17，DLG_SI 的平均熵值为 3.13，本文所提方法 MFDLS 的平均熵值为 3.09，而 6 种方法的方差均为 0.95，说明本文算法足以抵御查询概率背景知识的攻击，能够有效地保护用户位置隐私。

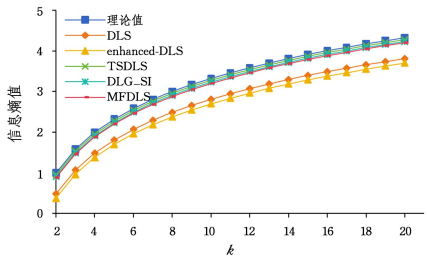

图 4 不同 k 值下的信息熵

Fig. 4 Information entropy with different k values

5.3.2 位置分散度

5 种算法在不同 k 值下假位置集形成的面积如图 5 所示。从图 5 中可看出，MFDLS 算法形成的区域面积稍小于 TSDLS，且随着 k 值的增加，在侯选位置数量较多的情况下，匿名集面积急剧增大。从本次实验运行数据来看，DLS 的平均分散度为 2.14，enhanced-DLS 的平均分散度为 5.39，TSDLS 的平均分散度为 8.77，DLG_SI 的平均分散度为 5.43，MFDLS 的平均分散度为 8.55。DLS 在进行假位置选取时，仅考虑了查询概率，因此所形成匿名区面积具有随机性；enhanced-DLS 在选取时虽考虑了假位置对面积的影响，但在选取时不是通过多边形面积，而是使用了距离乘积；TSDLS 在假位置生成算法中充分考虑了面积，并将其作为位置离散度指标，因此匿名区面积较大；DLG_SI 虽考虑了面积，但其使用假位置集中所有位置距离假位置集中心的欧氏距离来衡量位置离散度；MFDLS 在假位置选取算法中充分考虑了面积，但由于在位置选取时所受因素较多，匿名区面积并没有 TSDLS 的大。

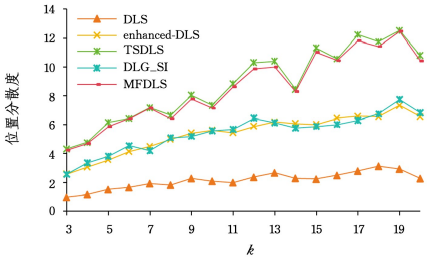

图 5 位置分散度

Fig. 5 Location dispersion

5.3.3 语义多样性

语义多样性通过语义差异性来衡量。5 种算法的语义差异性结果如图 6 所示。

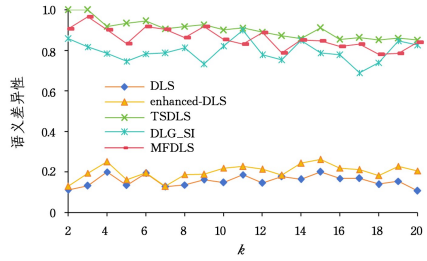

图 6 语义多样性

Fig. 6 Semantic diversity

从图 6 中可看出，DLS 和 enhanced-DLS 的语义差异性值低于其他算法，因为二者均未考虑语义差异性；后者比前者稍高，因为位置偏远时，语义类型一致的概率会变小。

从本次实验运行结果来看，DLS 的平均语义差异性为 0.16，enhanced-DLS 的平均语义差异性为 0.20，TSDLS 的平均语义差异性为 0.90，DLG_SI 的平均语义差异性为 0.79，MFDLS 的平均语义差异性为 0.86。TSDLS 在假位置生成时，所选位置与假位置集中已有位置的语义差异性最大，因此语义差异性值较大；DLG_SI 在假位置选择时，假位置语义满足语义阈值，并不是极值，因此语义差异性值比 TSDLS 低；MFDLS 在假位置选取时，选取思路同 TSDLS，由于假位置选取时的约束条件较多，语义差异性比 TSDLS 略低。

5.3.4 时间一致性

时间一致性使用式(5)来计算衡量。5 种算法的时间一致性结果如图 7 所示。

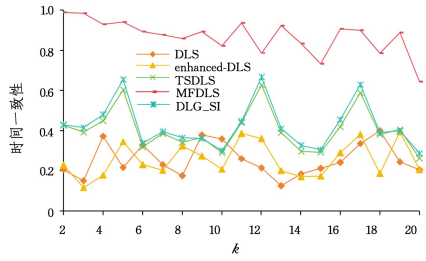

图 7 时间一致性

Fig. 7 Time consistency

从图 7 中可以看出，DLS 和 enhanced-DLS 的时间一致性低于其他算法，因为二者均未考虑时间属性。二者的时间特征并无规律，enhanced-DLS 相比 DLS 虽考虑了距离，但距离并不能严格意义上表明位置的时间属性关系。

从本次实验结果来看，DLS 的时间一致性值为 0.26，enhanced-DLS 的时间一致性值为 0.26，TSDLS 的时间一致性值为 0.40，DLG_SI 的时间一致性值为 0.42，MFDLS 的时间一致性值为 0.88。TSDLS 和 DLG_SI 虽未考虑到时间属性，但由于位置语义和时间属性有一定的关系，如相同语义位置的时间属性具有一定的相似性(即使虽未必完全相同)，相对而言，二者的时间一致性相对更高，而 DLG_SI 的时间一致性整体上比 TSDLS 的更好。本文所提方法 MFDLS 在时间一致性上效果最好，因为在选取假位置时，考虑到了所选位置与假位置集中已有位置的时间最为相似，因此时间一致性值较大。

5.3.5 敏感语义多样性

敏感语义多样性使用敏感差异性来衡量。5 种算法的敏感语义差异性结果如图 8 所示。

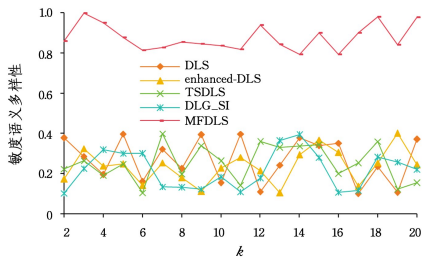

图 8 敏感语义多样性

Fig. 8 Sensitive semantic diversity

从图 8 中可看出，DLS，enhanced-DLS，TSDLS 和 DLG_

SI 的敏感语义差异性值较低，无明显变化规律；TSDLS 和 DLG_SI 虽考虑了语义差异性，但却忽略了用户敏感语义，选取的假位置集语义虽丰富，但可能与用户的敏感语义相违背，因此敏感语义差异性值并不比未考虑语义的 DLS 和 enhanced-DLS 的高。MFDLS 在进行假位置选择时，需保证所选位置语义与用户敏感位置集中的语义差异性较大，因此敏感差异性值较大，从而保证了敏感语义的多样性。

5.3.6 敌手错误

敌手错误使用式(7)来计算。5 种算法的敌手错误如图 9 所示。可以看出，DLS 和 enhanced-DLS 算法的敌手错误较低，TSDLS 和 DLG_SI 的敌手错误相对较高，而本文所提方法 MFDLS 的敌手错误更大。

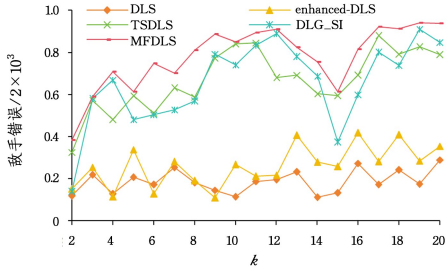

图 9 敌手错误

Fig. 9 Adversary error

从实验结果数据来看，DLS 的平均敌手错误为 373.11，enhanced-DLS 的平均敌手错误为 520.56，TSDLS 的平均敌手错误为 1 336.38，DLG_SI 的平均敌手错误为 1 310.56，本文所提方法 MFDLS 的平均敌手错误为 1 556.51。

DLS 仅考虑了查询概率；enhanced-DLS 在查询概率的基础上，进一步考虑了位置分布；TSDLS 和 DLG_SI 在此基础上，加入了语义考量；MFDLS 除上述因素外，额外考虑了用户敏感位置、时间属性等因素。因此在基于背景知识攻击的情况下，MFDLS 抵御攻击的能力更强。

5.3.7 质量损失

质量损失使用式(8)来计算。5 种算法的质量损失如图 10 所示。

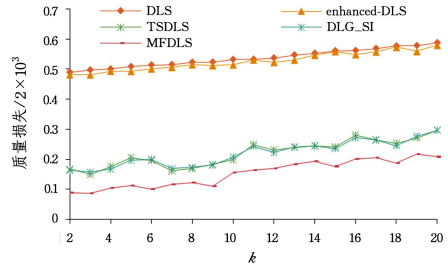

图 10 质量损失

Fig. 10 Quality loss

从图 10 中可看出，质量损失并不如敌手错误明显，DLS 和 enhanced-DLS 算法的质量损失较高，TSDLS 和 DLG_SI 的质量损失相对较低，而本文方法 MFDLS 的质量损失更低，且随着匿名区域的增大，质量损失呈上升趋势(虽增幅不是特别明显)。从本次实验结果数据来看，DLS 的平均质量损失为 1 075.76，enhanced-DLS 的平均质量损失为 1 053.60，TSDLS 的平均质量损失为 440.44，DLG_SI 的平均质量损失为 438.99，MFDLS 的平均质量损失为 307.50。

从结果中可看出，不同算法选取假位置时考虑的因素不同。在选取假位置时，考虑因素越多，选取位置越符合真实位置特征，与真实位置偏离越小。

结束语 针对当前假位置隐私保护算法中未考虑到位置的时间属性、用户敏感偏好等因素的问题，本文首先全面分析了位置隐私中存在的攻击类型，随后设计了一种综合多因素的假位置选取算法。算法综合考虑了分时间段的查询概率、地理位置分布、位置语义、位置时间属性和用户敏感偏好等因素，确保所选假位置集在相应时间段内的查询概率与真实位置相近，地理位置上相对分散，具有多样性特征，在特定时间的营业情况与真实位置相符且与敏感地点在语义上相似但不完全相同。实验从不同角度进行了验证，包括不确定性、位置分散度、语义多样性、时间一致性、敏感语义多样性等。实验结果表明，本文方法能够有效抵御基于背景知识的攻击，尤其是对于基于时间的位置链接攻击和敏感同质攻击具有显著效果。此外，从敌手错误和质量损失两个方面来看，本文方法能够满足用户的隐私需求，具有较好的性能。本文方法更适用于用户密集区域，在较为稀疏区域如何更好地保护用户隐私，是之后要解决的问题。

参 考 文 献

[1] LIAN H, QIU W, YAN D, et al. Privacy-Preserving Spatial Query Protocol Based on the Moore Curve for Location-Based Service [J]. Computers & Security, 2020, 96(3): 101-125.

[2] ZENG S, MU Y, HE M, et al. New approach for privacy-aware location-based service communications [J]. Wireless Personal Communications, 2018, 101: 1057-1073.

[3] ZHANG L, SONG G, ZHU D, et al. Location privacy preservation through kernel transformation [J]. Concurrency and Computation: Practice and Experience, 2022, 34(16): e6014.

[4] ZHANGX J, YANG H Y, LI Z, et al. Differentially Private Location Privacy-preserving Scheme with Semantic Location [J]. Computer Science, 2021, 48(8): 300-308.

[5] LIU Z P, MIAO D W, LIU Q N, et al. Location privacy protection through local differential privacy under k-anonymity [J]. Application Research of Computers, 2022, 39(8): 2469-2473.

[6] JIAO Z X, ZHANG L, LIU Q P. Fine-grained spurious location selection algorithm for distinguishing time periods [J]. Journal of Nanjing University of Posts and Telecommunications: Natural Science Edition, 2022, 42(6): 106-114.

[7] ZHANG X J, YANG Y X, LI J L, et al. A false location generating algorithm against background information inference attacks [J]. Computer Science, 2023, 50(S2): 879-887.

[8] WANG S, LI F H, NIU B, et al. Research Progress on Location-preserving Technique [J]. Journal of Communication, 2016, 37(12): 124-141.

[9] WU L, PAN X, PIAO C H, et al. Based on personalized privacy protection against sensitive homogeneity attacks in location services [J]. Computer applications, 2014, 34(8): 2356-2360.

[10] KIDO H, YANAGISAWA Y, SATOH T. An anonymous com-

munication technique using dummies for location-based services [C]// ICPS'05. IEEE,2005:88-97.

[11] LU H,JENSEN C S,YIU M L. PAD:privacy-area aware,dummy-based location privacy in mobile services[C]// ACM International Workshop on Data Engineering for Wireless & Mobile Access. ACM,2008:16-23.

[12] NIU B,LI Q,ZHU X,et al. Achieving k-anonymity in privacy-aware location-based services [C]// IEEE INFOCOM 2014-IEEE Conference on Computer Communications. NJ: IEEE, 2014:754-762.

[13] ZHANG A,LI X H,LI B. Location privacy desensitization algorithm based on false location selection [J]. Application Research of Computers,2022,39(5):1551-1556.

[14] ZHANG Y B,ZHANG Q Y,LI Z Y,et al. Pseudo-location K-anonymous location privacy protection method based on approximate matching [J]. Control and Decision,2019,35(1):65-73.

[15] WANG H,ZHU G Y,SHEN Z H,et al. False location generation method based on user preference and location distribution [J]. Computer Science,2021,48(7):164-171.

[16] SONG C,ZHANG Y D,PENG W P,et al. Research on K-anonymous privacy protection Scheme based on Bilinear pairs [J]. Applied Research of Computers,2019,36(5):1529-1532.

[17] HARA T. Dummy-based location anonymization for controlling observable user preferences[C]// 2019 IEEE Global Communications Conference(GLOBECOM). IEEE,2019:1-7.

[18] WANG J,WANG C,MA J F,et al. Pseudoposition selection algorithm based on position semantics and query probabilities. [J]. Journal of Communications,2020,41(3):53-61.

[19] TU S P,ZHANG L,LIU X P. Double false location selection algorithm based on Behavior Association [J]. Computer Science, 2023,50(5):348-354.

[20] XIA X Y,BAI Z H,LI J,et al. Location anonymity algorithm based on false position and Stackelberg game [J]. Journal of Computer Science,2019,42(10):2216-2232.

[21] LI C,ZHANG X,YAN F,et al. False position generation Scheme based on User Preference [J]. Computer Engineering and Design,2019,40(4):914-919,1195.

[22] CHOWC Y,MOKBEL M F. Enabling private continuous queries for revealed user locations[C]// International Symposium on Spatial and Temporal Databases. Springer,2007:258-275.

[23] XIAO P,ZHEN X. Survey of location privacy-preserving [J]. Journal of Computer Science and Frontiers, 2007, 1(3): 268-281.

[24] MACHANAVAJJHALA A,KIFER D,GEHRKE J,et al. L-diversity:Privacy Beyond k-anonymity [J]. ACM Transaction on Knowledge Discovery from data,2007,1(1):3-5.

[25] WANG S,LI F H,NIU B,et al. Research Progress on Location-preserving Technique [J]. Journal of Communication, 2016, 37(12):124-141.

[26] ZHENG Y,XIE X,MA W Y. GeoLife:A Collaborative Social Networking Service among User,location and trajectory [J]. IEEE Data Engineering Bulletin,2021,33(2):32-40.

[27] TAKAGI S,CAO Y,ASANO Y,et al. Geo-graph-indistinguishability:Protecting location privacy for LBS over road networks [C]// 33rd Annual IFIP WG 11.3 Conference on Data and Applications Security and Privacy (DBSec). Springer, 2019: 143-163.

LI Yongjun, born in 1983, Ph.D. Her main research interest is privacy protection.

ZHU Yuefei, born in 1962, professor, Ph.D supervisor. His main research interests include network security and cryptography.

(责任编辑:何杨)