计算机科学
COMPUTER SCIENCE

基于Geohash的增强型位置*k*-匿名隐私保护方案

李勇军, 祝跃飞, 白利芳

引用本文

李勇军, 祝跃飞, 白利芳. 基于Geohash的增强型位置*k*-匿名隐私保护方案[J]. 计算机科学, 2024, 51(9): 393-400.

LI Yongjun, ZHU Yuefei, BAI Lifang. Enhanced Location *K*-anonymity Privacy Protection Scheme Based on Geohash [J]. Computer Science, 2024, 51(9): 393-400.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230800183

基于 Geohash 的增强型位置 k-匿名隐私保护方案

李勇军[1,2]　祝跃飞[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　随着 LBS 的广泛应用,位置隐私保护势在必行。近年来,作为应用较为广泛的位置 k-匿名解决方案已成为研究热点,但 k-匿名方案易受到敌手背景知识攻击,虽有学者们不同程度地考虑了位置相关的信息,但都不全面,并且当前形成匿名区的方案大多较为耗时。基于此,为抵御敌手的语义攻击和查询及位置同质性攻击,提出了增强型位置 k-匿名方案,在匿名区构建时充分考虑与物理位置相关的语义信息、时间属性、查询概率及查询语义等信息;然后在进行位置选取时,保证所选位置相对分散;为降低匿名区构建时耗,采用 Geohash 进行位置编码;最后通过真实数据集上的实验表明,所提方案可提供较好的位置隐私保护。

关键词:Geohash;增强型位置 k-匿名;基于位置的服务;位置隐私;位置语义;查询概率;时间属性
中图分类号　TP309

Enhanced Location *K*-anonymity Privacy Protection Scheme Based on Geohash

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

Abstract　With the wide application of LBS,location privacy protection is imperative. In recent years,location k-anonymity solution has become a research hotpot which is widely used. However,k-anonymity schemes are vulnerable to background knowledge attacks. Although some scholars have considered location-related information to varying degrees,they are not comprehensive,and the current form in the anonymous scheme are relatively time-consuming. Based on this,in order to resist background knowledge attacks from adversaries,enhanced location k-anonymity scheme is proposed,which fully considers the semantic information,time attributes,query probability and query semantics related to physical location when constructing anonymous areas. When the location point is selected,it is necessary to ensure that the selected location is relatively scattered. In order to reduce the time consumption of anonymous area construction,Geohash is used to encode the location information. Finally,experiments on real data sets show that the proposed scheme can provide better location privacy protection.

Keywords　Geohash,Enhanced location k-anonymity,Location based services,Location privacy,Location semantic,Query probability,Time attribute

1　引言

随着无线通信技术和 GPS 的发展,基于位置的服务(Location Based Services,LBS)[1]被广泛应用到各个领域,并为用户提供大量的服务,如导航、推荐等。如同硬币的两面,用户在享受便利的同时,也面临着位置隐私泄露的风险。如何保护用户的位置隐私,成为当前 LBS 领域研究的热点问题之一。

在过去几年,很多学者提出了大量的解决方案[2-10]。根据保护机制,这些方案可分为假名法、假位置、空间转换、混淆机制等。在位置隐私保护中,这些机制对位置信息进行不同程度的修改,因此服务效用很难得到保证。近年来,位置 k-匿名技术因在可用性和安全性方面的优势而受到研究人员的青睐[11]。为抵御不同的攻击者,在进行 k-匿名时,研究者的关注点聚焦于位置的合理选择[12]及服务器不可信的问题[13]方面。

到稿日期:2023-08-29　返修日期:2024-01-08
基金项目:科技委基础加强项目(2020-JCJQ-ZD-021)
This work was supported by the Foundation Strengthening Project of Science and Technology Commission(2020-JCJQ-ZD-021).
通信作者:祝跃飞(yfzhu17@sina.com)

由于位置语义信息包含了丰富的信息，能更准确地反映用户的偏好和需求，而攻击者又易获取，因此在位置隐私保护过程中，研究者们开始逐步关注语义信息[14-19]。用户在享受LBS服务时，其位置信息、时间属性和查询操作，包括位置信息中包含的大量语义信息，通常都会暴露在外，若未对其进行任何保护或保护不当亦或保护不全，用户的隐私都会受到影响，并且在进行位置隐私保护时，当前大都是使用欧氏等平面距离方法进行位置的选取，匿名区的构建较为耗时。本文针对上述问题，提出一种基于Geohash算法并综合考虑位置信息、语义信息、查询操作及时间属性的LBS隐私保护方法。本文主要贡献如下：

1)为增强位置隐私，提出增强型的位置 k-匿名隐私保护方案。构建匿名集时，综合考虑用户发起查询的时间属性、查询语义、查询概率、位置语义等信息，保证匿名集中侯选位置的时间属性一致，查询概率接近，语义信息多样化；进行位置选取时，尽可能保证所选位置分散且距离最远。

2)为减少距离计算带来的时间消耗，采用Geohash算法。

3)基于真实数据集的评估显示，所提方法相比其他算法可达到更高的性能。

2 相关概念

2.1 LBS位置信息

LBS位置信息不仅包含物理位置，还应包含相应的语义信息和时间属性，以及发起查询的情况等附属信息。设 Ω 为地理空间数据集；位置信息 $p\in\Omega$，使用三元组来表示，$p=\langle u,l_u,\bar{l}_u\rangle$，$u$ 为用户标识，l_u 为位置的地理属性经纬度，即 $l_u=(lat,lon)$，$\bar{l}_u$ 为位置的附属属性，本文包括语义信息、时间属性和查询信息，具体表示为 $\bar{l}_u=(ls,t,q,pr)$，其中 ls 为 p 的语义信息；t 为 p 的时间属性，由周几 w 和时间区间 h 组成，记作 $t=(w,h)$；q 为用户 u 在 p 发起的查询内容；pr 为位置 p 被查询的概率。

2.2 位置 k-匿名

位置 k-匿名要求每个匿名区(即一组在某些“识别”属性方面彼此不可区分的位置记录)至少包含 k 条位置记录。

设 A 为某一匿名区中所有位置构成的集合，$A\subseteq\Omega$，$A_{\min}$ 为 A 对应匿名区的最小面积，$A_{\max}$ 为 A 对应匿名区最大面积，对于任一物理位置 $p\in\Omega$，设 f 为匿名映射函数，将 p 映射到 A 中的某个元素，则 $f:p\to A$，那么函数 f 应满足以下性质：

1) $\exists\, p_i=p$，其中 $p_i\in A$；

2) $\forall\, p_i,p_j\in A,p_i\neq p_j$；

3) $pr=Pr(l(u)=l_u)\in[1/k,1]$，表示敌手识别出 u 位置的概率，其中函数 $l(u)$ 表示 u 当前位置；

4) $A_{\min}\leqslant Area(A)\leqslant A_{\max}$。

2.3 增强型位置 k-匿名

增强型位置 k-匿名是位置 k-匿名的扩展，设 f' 为匿名映射函数，将 p 映射到 A' 中的某个元素，则 $f:p\to A'$，那么函数 f' 除满足 f 的性质外，还需满足如下性质：

1) $Pr(l(u,t)=l_u)\leqslant pr$，表示在 f' 下，敌手借助于 t 识别出 u 位置的概率相比 f 并不会增加；

2) $Pr(l(u,t|ls)=l_u)\leqslant pr$，表示在 f' 下，敌手借助于 t 和 ls 识别出 u 位置的概率相比 f 并不会增加；

3) $Pr(l(u,q)=l_u)\leqslant pr$，表示在 f' 下，敌手借助于 q 识别出 u 位置的概率相比 f 并不会增加；

4) $Pr(l(u,pr)=l_u)\leqslant pr$，表示在 f' 下，敌手借助于 pr 识别出 u 位置的概率相比 f 并不会增加；

5)空间分布相对均匀。以用户真实位置所在矩形区域的中心点为坐标原点，绘制直角坐标系，设 n 为任一象限中的位置数，则 $n\in[\lfloor k/4\rfloor,\lfloor k/4\rfloor+1]$。

2.4 Geohash编码

Geohash算法[20]将二维经纬度坐标转换为一维字符串的地理位置编码，主要思想是将地球看成一个二维矩形平面，利用类二分法递归划分经纬度范围，使用Base32进行一维编码。如对于经纬度坐标(39.975269,116.342241)，使用Geohash进行编码时，若要求精度为7则编码为wx4ermm，精度为8时编码为wx4ermmb，精度为9时编码为wx4ermmbf。具体编码过程可参考文献[21]。

2.5 语义类别相似度

设 c_p 为 p 的语义信息类别编码，$\forall\, p_1,p_2\in\Omega$，$p_1,p_2$ 的语义类别相似度如式(1)所示：

$$Ls(p_1,p_2)=\frac{L_1(c_{p_1},c_{p_2})+L_2(c_{p_1},c_{p_2})+L_3(c_{p_1},c_{p_2})}{|c_{p_1}|} \tag{1}$$

其中，$L_1(c_{p_1},c_{p_2})$，$L_2(c_{p_1},c_{p_2})$ 和 $L_3(c_{p_1},c_{p_2})$ 表示 p_1,p_2 的语义类别大类、中类、小类是否相同。若相同，则值为2，否则为0。

2.6 时间属性相似度

$\forall\, p_1,p_2\in\Omega$，$t_1$ 和 t_2 为 p_1 和 p_2 的时间属性，则 p_1 和 p_2 的时间属性相似度如式(2)所示：

$$T_s(p_1\parallel t_1,p_2\parallel t_2)=\frac{\sum_{i=1}^{7}t_1\cdot w[i]\times t_2\cdot w[i]}{\sqrt{\sum_{i=1}^{7}(t_1\cdot w[i])^2}\times\sqrt{\sum_{i=1}^{7}(t_2\cdot w[i])^2}}\times\frac{|t_1\cdot h\cap t_2\cdot h|}{|t_1\cdot h\cup t_2\cdot h|} \tag{2}$$

时间相似度由两部分乘积得出：第一部分是 w 的相似度，采用one-hot编码方式，使用七维空间的余弦相似度进行计算；第二部分是 h 的相似度，使用区间数相似度计算。

2.7 查询语义相似度

设 q_1 和 q_2 为从 p_1 和 p_2 发起的查询，t_1 和 t_2 为 p_1 和 p_2 的时间属性，则 q_1 和 q_2 的查询语义相似度如式(3)所示：

$$Qs(p_1\parallel q_1\parallel t_1,p_2\parallel q_2\parallel t_2)=\begin{cases}\dfrac{q_1\cdot q_2}{\parallel q_1\parallel\parallel q_2\parallel}, & 若\ T_s(p_1\parallel t_1,p_2\parallel t_2)=1\\ 0, & 若\ T_s(p_1\parallel t_1,p_2\parallel t_2)\neq 1\end{cases} \tag{3}$$

2.8 位置相似度

$\forall\, p_1,p_2\in\Omega$，设 q_1 和 q_2 为从 p_1 和 p_2 发起的查询内容，$Pr(p)$ 为位置 p 被查询的概率，则位置 p_1 和 p_2 的位置相似度如式(4)所示：

$$Sim(p_1,p_2,q_1,q_2,t_1,t_2,Pr(p_1),Pr(p_2))=(\alpha\times Ls(p_1,p_2)+\beta\times Qs(p_1\parallel q_1\parallel t_1,p_2\parallel q_2\parallel t_2)+\gamma\times(1-|Pr(p_1)-Pr(p_2)|))/3 \quad (4)$$

其中，α，β 和 γ 分别表示位置语义类别相似度、查询语义相似度和位置被查询相近度的权重，且 $\alpha+\beta+\gamma=1$。

2.9 增强位置熵

设 $p\in A$ 为 u 的真实位置，q 为 t 时刻从 p 发起的请求，$Pr(p)$ 为 p 被查询的概率，$p_i\in A(i\in[0,k))$ 为候选位置，q_i 为 t_i 时刻从 p_i 发起的请求，$Pr(p_i)$ 为 p_i 被查询的概率，那么 p_i 成为真实位置的概率为 $Pr(i)=Sim(p,p_i,q,q_i,t,t_i,Pr(p),Pr(p_i))/\sum_{i=1}^{k-1}Sim(p,p_i,q,q_i,t,t_i,Pr(p),Pr(p_i))$，增强位置熵为 $ELE=-\sum_{i=1}^{k}Pr(i)\log(Pr(i))$。候选位置的增强位置熵越大，隐私度就越高。

ELE 用于刻画匿名区域内位置的平均隐私信息量，也就是匿名区域的隐私不确定程度。ELE 越大，隐私泄露的可能性就越小，因此它也可用于衡量隐私的保护程度。在没有外部条件影响时，该值是一个确定的值。易证明，这种隐私信息熵满足 *Shannon* 信息熵的基本性质[22]，即具有非负性、可加性、扩展性等。

3 相关工作

3.1 位置 k-匿名方法

Zhang 等[23]提出了一种基于地理语义的 k-匿名的位置隐私保护方法。该方法利用最大最小距离多中心聚类算法构建候选集，基于语义相似度生成虚拟位置。Kuang 等[14]在路网情况下，使用双向 k 扰动算法进行用户位置及查询目标的保护，当 k 值增加时，所需要的路段数会增多。Yang 等[11]使用 k-匿名进行 LBS 隐私保护，通过差分隐私方案构建匿名区，建立了 Stacklberg 博弈模型来寻找一个最优的虚拟位置集，以抵御背景知识攻击，但对匿名候选位置不能形成满足用户需求的 k 值未进行处理。Xing 等[24]提出了一种改进的基于双 k-匿名的隐私保护方案，该方案隐藏了用户的位置和请求信息。Yang 等[25]为解决 k-匿名中的欺骗行为和服务摇摆问题，引入一种去摆动声誉评价方法(DREM)，构建一个可信的隐形区域来保护请求者的位置隐私。但随着 k 值的增大，其计算时间几乎呈线性增加；当用户处于隐私保护要求较高的敏感位置时，该方案的延迟时间较高。

3.2 基于语义的方法

Kuang 等[14]综合考虑用户位置和查询位置的语义信息，并根据用户对不同位置语义敏感度的不同，产生敏感权重文档，使用 k-匿名技术完成隐私的保护，可抵抗攻击者对用户位置和查询位置的语义攻击。但当用户位置或查询位置的语义信息较少时，该方案效果并不佳，语义信息采用随机生成方案完成。Xu 等[15]使用兴趣点(Point of Interest，POI)作为位置语义信息进行下一个位置的预测。Yang 等[11]考虑到敌手具有位置语义信息对安全性的影响，提出基于 k-匿名的 USP-PM，位置语义信息的获取借助于高德地图完成。Shi 等[16]针对位置语义攻击，提出基于位置语义量化的虚拟位置产生算法，文中的位置语义通过用户在不同时间段访问的数量来进行量化，此种方式下不能抵御位置同质性攻击。Jiao 等[17]针对当前背景知识中的查询概率随时间段的变化而变化，提出基于时间段的假位置选择算法，在选择假位置时，考虑到语义距离，语义信息由百度地图的逆地理编码 API 获取。Xing 等[18]针对敌手的背景知识攻击，考虑到匿名区中用户的兴趣点及行为爱好，兴趣点使用洛阳市洛龙区的 POI 数据，用户社会行为数据使用微博中的点赞、评论、转发和相互关注数据等。Tu 等[19]使用贪心法完成假位置的生成，再使用枚举方案完成假位置的优化，将地名作为语义信息。

3.3 基于 Geohash 的方法

Zhou 等[26]为避免将用户精确位置发送至 LBS 服务器，使用 Geohash 算法将用户位置坐标转化为字符串编码序列，将序列发送至 LBS 服务器。Feng 等[27]针对物联网频谱分享过程中用户对隐私泄露的顾虑，提出 Geohash 编码前缀和二进制后缀相结合的 k 匿名区域位置编码方式。二进制编码对区域进行细粒度划分，Geohash 编码快速查询该区域的用户情况。Yin 等[21]为保证在庞大历史记录查询中为服务提供即时位置服务，使用 Geohash 编码和反向检索方式完成匿名位置集的构建。Li 等[28]为提高查询效率，利用 Geohash 编码时空数据并按历史数据分布分区。Liu 等[29]为抵御中心服务器攻击，使用 Geohash 编码和伪随机序列产生技术来完成匿名区域的构建。

3.4 基于查询概率的方法

Fei 等[30]使用大众点评中的 POI 总评价数作为 LBS 用户的查询概率；Yang 等[11]使用模拟的方式生成查询概率；Yin 等[21]在使用 k-匿名进行位置隐私保护时，使用与文献[31]一样的方式，将数据集中的签到数据作为历史查询概率。Tu 等[19]先使用贪心法进行假位置生成，再使用枚举方案完成假位置的优化，文中考虑了历史查询概率，此值为每个网格访问时间的时长与整个区域访问时长的比值。

3.5 基于时间的方法

Yang 等[11]指出位置语义信息中除包括地图语义特征，还包括时间语义特征，时间语义特征使用停留时长来表示。Jiao 等[17]考虑到不同时间段查询概率不尽相同，对位置查询概率进行时间分片，对时间段进行细粒度划分。Tu 等[19]针对连续查询场景，在选取假位置时，考虑了时间可达性。

4 增强型位置 k-匿名隐私保护方案

4.1 出发点

当前基于 Geohash 算法的匿名方案效率较高，这主要是基于编码区，也就是网格。假设地图栅栏化(20 * 20)，用户位置的分布情况如图 1 所示。

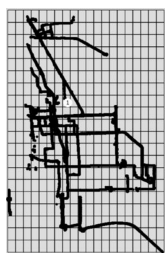


图 1 位置分布

Fig. 1 Distribution of location

从图 1 中可看出，网格中用户分布并不是均匀的，大部分都是集中在网格的某一位置，如左下角或左上角等。如①所在网格中点做匿名化时，若本网格内匿名失败，就自然向外扩一圈。假设区域内所有位置均满足约束条件，则在进行位置选取时，会选四周熵值大或距离最远的位置，这一方面保证了多样性，另一方面保证了所选位置比较分散。

4.2 增强型的位置 k-匿名保护方案

为避免匿名区构建时位置计算导致的时间代价问题，采取 Geohash 算法对位置进行编码；为抵御同质性攻击和语义攻击，需要匿名区所选位置时间一致，查询概率接近，查询内容语义和位置语义尽可能多样化。因此本方案首先读取用户位置，随后进行位置信息的 Geohash 编码[20]，使用第三方接口获取每个位置的语义信息、时间属性和查询概率，接着进行匿名区的构建，最后根据用户需求进行位置的选取。具体过程如图 2 所示。

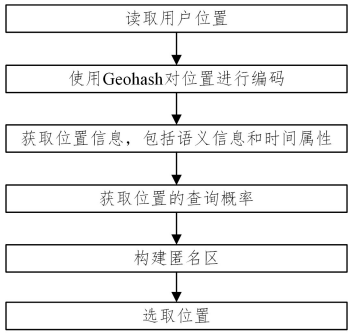


图 2 k-匿名保护方案流程

Fig. 2 Process of k-anonymity protection scheme

为防止语义隐私漏洞和背景知识攻击，本方案重点是保证匿名区中位置语义及查询语义的多样性，并且语义类别数据尽可能相等。因此在构建匿名区时，一方面要保护匿名位置之间的语义尽可能不同，另一方面要保证匿名区中的语义又尽可能多样化，同时需要考虑用户密集时匿名区域面积过小或用户分布稀疏时匿名区域过大等问题。匿名集的构建具体如算法 1 所示。

算法 1 构建匿名集

输入：用户位置信息 p，所有位置信息 plist，隐私等级参数 k，EPSION，最小区域 A_{min}，最大区域 A_{max}

输出：匿名集中位置列表 AP

1. maxLen←根据 A_{min} 确定 Geohash 编码的最长位数；
2. minLen←根据 A_{max} 确定 Geohash 编码的最短位数；
3. lc←对当前用户位置进行长度为 maxLen 的 Geohash 编码；
4. if len(lc) ≥ minLen：# 当前位置的 Geohash 编码长度在 minLen 和 maxLen 之间
   4.1. n←统计当前位置所在编码区内位置数；
   4.2. 若 n≥4k：
      tn←筛选出编码区中位置时间属性与 p 的时间一致的位置；
      若 tn<2k，转 4.3
      pn←筛选出编码区中位置查询概率与用户位置查询概率 p 相差 EPSION 的位置；
      若 pn<2k，转 4.3
      qn←筛选出编码区中位置查询语义与用户位置查询语义相同的位置；
      若 qn<2k，转 4.3
      sn←统计每个语义类别的位置数；
      min←求 sn 中最小值；# 计算语义类别中的最小值
      cn←min * 语义类别数；
      若 cn<2k，转 4.3
      AP←求编码区中满足 ELE 最大的 2k 个位置。
      return AP；
   4.3. 否则 maxLen←maxLen-1，转 3；
5. else：
   5.1. an←计算当前编码区内位置个数；
   5.2. AP←AP. append(an)；
   5.3. r←随机生成(4k−an)个虚假位置，要求时间属性同 p 的时间一致，查询概率与 p 的查询概率误差在 EPSION 内，位置熵值最大；
   5.4. AP←AP. append(r)；
   5.5. return AP。

匿名区构建完成后，侯选位置数量是 $4k$，如何从 $4k$ 中选取 k 个位置，需要考虑到所选位置的分布情况，以及位置的语义多样化等。具体如算法 2 所示。

算法 2 位置选取

输入：用户位置信息 cp，AP，隐私等级参数 k。

输出：匿名位置列表 ap。

1. if AP≠NULL：
   1.1. for i=0 to 3：
      n_QUAD[i]←0 # 初始化各象限的位置数为 0；
      p_QUAD[i]←NULL # 初始化各象限的位置信息为 NULL；
      sp←NULL # 记录 cp 与匿名位置的增强位置熵；
   1.2. for p in AP：
      计算 cp 与 p 之间的相似度；
      ele=计算 cp 与 p 的增强位置熵；
      angle←计算 cp 与 p 之间的夹角；
      n_QUAD[angle/90]←n_QUAD[angle/90]+1 # 增加相应象限的位置数；
      p_QUAD[angle/90]←p_QUAD[angle/90]. append(cp) # 增加相应象限的位置信息；
      sp. append(ele)；
   1.3. sort(sp) # 按照增强位置熵进行降序排列；
   1.4. min←求 4 个象限中 n_QUAD[i]的最小值；
   1.5. 若 4 * min≥k−1：# 选取的位置在 4 个象限中可以均匀分布
      ap←ap. append(4 个象限中前 min 个位置)；
   1.6. 否则 other←k−1−4 * min；
      ap←ap. append(4 个象限中前 min 个位置)；
      op←从 4 个象限中余下的位置中选取距离最远的 other 个位置；
      ap←ap. append(op)；
   1.7. ap. append(cp)；# 加入当前位置
   1.8. return ap；
2. else：
   return NULL。

4.3 算法分析

4.3.1 安全性分析

1)位置安全性分析

算法 1 中保证 A 所形成区域的面积在 A_{min} 和 A_{max} 之间，其中有 $4k$ 个位置，算法 2 从其中选取 $k-1$ 个位置，保证能形成位置 k-匿名区，因此方案满足 2.2 节中位置 k-

匿名的性质 1)－4)。

设 f 作用下，nt 为与 p 有相同时间属性的位置数，ns 为匿名区中语义的种类数。

2)时间属性安全性分析

算法 1 中步骤 4.2 要求所有侯选位置与真实位置的时间属性一致，即与 p 相同时间属性的位置数为 k，显然 $k \geqslant nt$，所以敌手根据时间属性推测出用户位置的概率并不会增加，因此方案满足 2.3 节中增强型位置 k-匿名的性质 1)。

3)语义安全性分析

算法 1 中步骤 4.2 要求计算所有语义类别，显然此时 $k \geqslant ns$，因此方案满足 2.3 节中增强型位置 k-匿名的性质 2)。

4)查询语义安全性分析

算法 1 中要求侯选位置的查询语义与 p 的查询语义相同，因此敌手通过查询语义识别出用户真实位置的概率并不会增加，因此方案满足 2.3 节中增强型位置 k-匿名的性质 3)。

5)查询概率安全性分析

算法 1 中要求侯选位置的查询概率与 p 的查询概率相差不到 $EPISION$，敌手通过查询概率识别出用户真实位置的概率基本上等同于随机猜测，成功率接近 $1/k$，相比 f 来言，本方案的概率值较小，满足 2.3 节中增强型位置 k-匿名的性质 4)。

6)位置离散性分析

算法 2 要求匿名区内所选位置在 4 个象限中均匀分布，并要求尽可能选择远的位置，这样匿名区面积相对较大，使攻击者无法精准确定用户位置，因此方案满足 2.3 节中增强型位置 k-匿名的性质 5)。

4.3.2 时间复杂度分析

算法 1 的时间复杂度主要在第 4 步正常构建匿名区，第 4.1 步统计当前编码区中位置的个数，时间复杂度为 $O(N)$，N 为所有位置个数。第 4.2 步如果构建匿名区成功，则时间消耗集中在时间属性的筛选、查询概率筛选、语义类别统计、增强位置熵最大的 $4k$ 个位置选取方面。假设位置保持不变，为最大 n，则时间属性和查询概率筛选的复杂度均为 $O(n)$；假设语义类别个数为 m，则语义类别统计的时间复杂度为 $O(n*m)$；增强位置熵最大的 $4k$ 个位置选取，实质上就是 Top-$4k$ 问题，时间复杂度为 $O(n+4k\log^n)$。因此，总的时间复杂度为 $O((m+3)*n+4k\log^n) \leqslant O(n)$。由于 $n \leqslant N$，所以算法 1 的时间复杂度为 $O(N)$。

算法 2 的时间复杂度主要在第 1 步：步骤 1.1 完成初始化，时间为线性；步骤 1.2 是对匿名区中的位置进行遍历，由于位置个数是 $2k$，因此时间复杂度为 $O(2k)$；步骤 1.3 是排序问题，假如使用基于比较的排序，则时间复杂度为 $O(2k\log^{2k})$；步骤 1.4 遍历 4 个象限求最小值，时间消耗为常量值，因此时间复杂度为 $O(1)$；步骤 1.5 取每个象限前 min 个位置，时间消耗为 $O(4*min)$；步骤 1.6 是在步骤 1.5 的基础上，加上求余下 $(k-4min)$ 个位置的代价，时间消耗为 $O(k)$；步骤 1.7 的时间消耗为线性。因此，算法 2 的时间复杂度为 $O(2k\log 2k) \approx O(n\log n)$。

4.3.3 匿名成功率分析

k 匿名机制的目标是形成至少由 k 个用户形成的匿名区，最低限度是要达成 k 个用户形成匿名区，满足用户数量至少为 k，以便被保护对象与其他至少 $k-1$ 个用户是无法区分的。本文在形成匿名区时，要求侯选位置有 $4k$ 个满足时间属性相同、查询概率接近的位置，若匿名区内位置数大于 $4k$，则能够正常形成匿名区，本方案的匿名成功率与文献[21]是一致的；但若形成匿名区时，原有满足条件的位置不足，则使用随机生成的方法完成匿名区的构建，当 $A_{\min}$ 和 $A_{\max}$ 之间的区域较小，或者 k 值较大时，随机生成的 $(4k-an)$ 个位置可能会存在重复的情况，即若生成的 $4k$ 个位置中存在重复位置而导致最终有效位置数小于 k 时，本方案将匿名失败。当然若匿名成功，本方案抵御敌手攻击的能力会比文献[21]的强。

5 实验及分析

5.1 实验环境及数据准备

实验从两个方面对所提方案进行验证：一是处理时间，二是匿名效果。

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。

数据集：GeoLife[32]，选取 2008.10.15-19 具有出行方式的 10 个用户代表：010，062，068，084，085，126，128，153，167，179，共 80198 个位置信息，位置语义信息和时间属性(POI 营业时间)由高德地图[33]获悉，位置的查询概率使用大众点评[34]获取 POI 总评价数计算得出。

对所选用户位置进行分析，所选用户所在区域为 39.839367，40.0927916，116.483975，116.266164，南北最大跨度为 18.643223944960727 km，东西跨度为 28.137742967356 1km。用户的整体位置分布如图 3 所示，每周不同天的位置分布情况如图 4 所示，每天不同时刻的位置分布情况如图 5 所示，不同用户的位置数如图 6 所示。位置语义采取 500 m 最近 POI 方案获取，其中 1 个位置 500 m 内无 POI 信息，扩展到 1 000 m，整体 POI 按大类统计信息如图 7 所示。

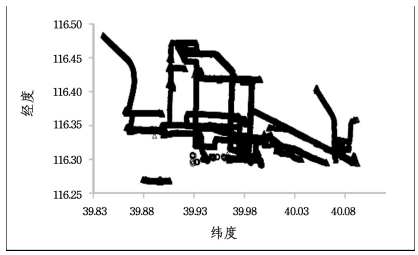


图 3 用户位置分布

Fig. 3 Distribution of user' location

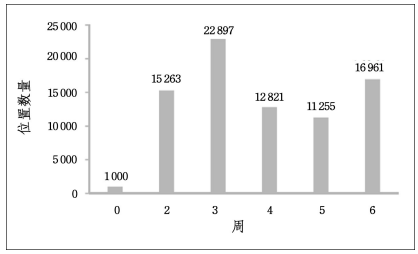


图 4 DoW 分布情况

Fig. 4 Distribution of day of week

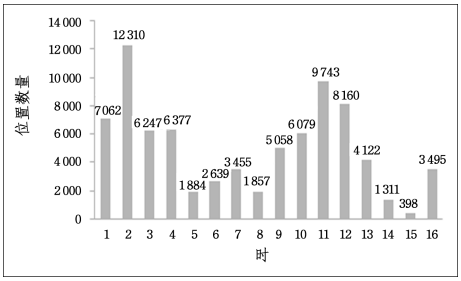


图 5 不同时刻的位置数

Fig. 5 Number of locations at different times

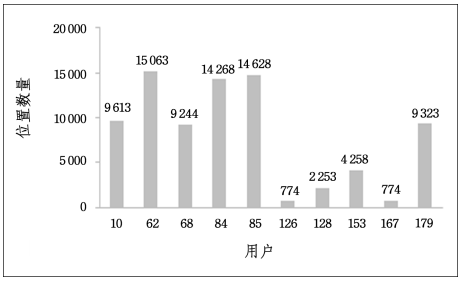


图 6 不同用户的位置个数

Fig. 6 Number of locations of different users

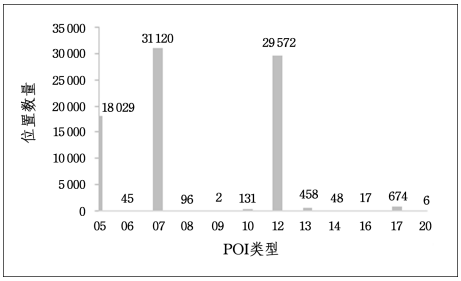


图 7 不同 POI 类型的位置数

Fig. 7 Number of different POI type

从图 4 中可以看出，周三位置数最多，为 22 897；周一无位置数据；周日最少，为 1 000。从图 5 中可以看出，并不是所有时刻都有位置数，而有位置的时刻中，2 时人位置数最多，为 12 310；而 15 时的位置数最少，为 398。图 6 展示了不同用户的位置数量，如用户 62 的位置数最多为 15 063；而最少的是 126 和 167 两个用户，位置数相同，均为 774。从图 7 可以看出，POI 大类编码为 07 的位置数最多，为 31 120，对应的类型为生活服务；位置数相对较多的 POI 大类编码为 05 和 12，对应的 POI 类型为餐饮服务和商务住宅；而 POI 大类编码为 09 的位置数最少，值为 2，对应的 POI 类型为医疗保健。

本文在进行位置相似度计算时，设 $\alpha=\beta=\gamma=1/3$，语义信息使用城市功能的 POI 代替，语义类别编码即是 POI 编码。对于位置 p，$|c_p|=6$，两个位置的 POI 语义类别要么不同，要么是大类相同，或大类、中类同时相同，亦或是大类、中类和小类均相同，也就是式(1)的分子只能为 0，2，4，6，这样两个位置的语义类别相似度值只能是 0，1/3，2/3 或 1。Geohash 算法编码后的字符串长度为 9，其精度可保持 2m 左右，能够满足实际需求。由于进行的是快照位置 k-匿名，因此出行方式在本实验中并未使用到。

5.2 匿名处理时间

将本文方案与文献[11，21]的方案进行比较，三者的匿名处理时间如图 8 所示。文献[11]生成匿名集时，需要进行查询概率的排序、位置的选择和侯选位置的语义多样性熵(Semantic Diversity Entropy，SDE)计算等操作，相对来讲，匿名集生成时间就会较长；而文献[21]的 Geohash 编码方案生成匿名集时，仅是直接将二维的空间数据转换为一维的字符串，并未考虑位置的语义等其他信息；本文方法在文献[21]的基础上，进行了位置的筛选，并预先进行了不同编码区位置的统计工作，所以匿名时间比文献[21]方案要长但相差不大，如本次实验本文方法运行的平均时长比文献[21]长 0.208 s，最大时差 0.54 s，最小时差 0 s。

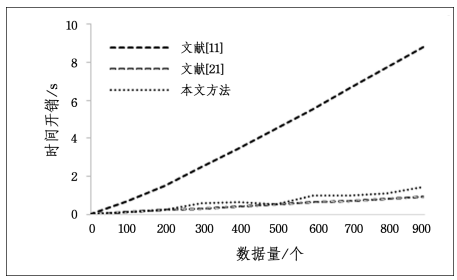
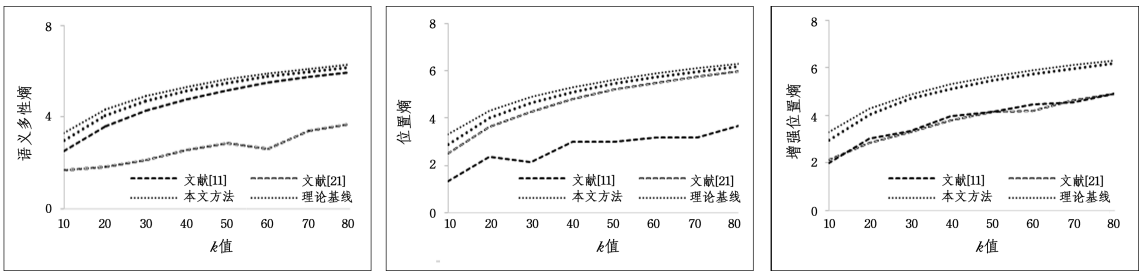


图 8 不同方法的匿名处理时间

Fig. 8 Anonymization times of different methods

5.3 匿名效果

将本文方案与文献[11，21]的方案进行比较，不同熵与 k 值的关系如图 9 所示。



(a)基于语义多样性熵 (b)基于位置熵 (c)基于增强位置熵

图 9 不同熵与 k 值的关系

Fig. 9 Relationship between different entropy and k value

文献[11]使用语义多样性熵来计算，与语义信息相关，关注的是不同语义类别位置的数量。文献[21]使用位置熵进行度量，位置熵的计算仅与查询概率相关。本文方法不仅考虑位置语义类别相似度、查询语义相似度和查询概率，而且考虑到时间一致，增强位置熵的计算使用位置相似度，其相比语义多样性熵，更偏重位置的实际语义信息，且考虑了查询概率、

查询语义及时间属性；与位置熵相比，除考虑了查询概率，还考虑了位置语义信息、时间属性和查询语义等，可更好地抵御由背景知识导致的隐私泄露。

从图 9 中可以看出，随着 k 值的增大，3 种保护方案都逐渐接近理论保护效果，而本文所提方案无论是基于哪种熵的度量方式，其保护效果均比另外两种好。如果使用熵的方式来度量，由于熵的计算方式不同，结果不尽相同。如图 9(a)中，当使用语义多样性熵进行效果度量时，文献[11]的效果比文献[21]的效果更好，平均值相差约 2.09，最大值相差约 2.27，最小值相差约 0.81。而如果使用位置熵进行度量，结果相反，如图 9(b)所示，平均值相差约 1.95，最大值相差约 2.28，最小值相差约 1.16。当使用本文所提方案进行度量时，二者的效果不分上下，如图 9(c)所示。从图中可以看出，使用增强位置熵来度量时，本文所提方案的熵值与理论基线趋势一致，且随着 k 值的增大，越来越接近理论 k 值。本次实验中，文献[21]的平均值比文献[11]的平均值约大 0.05，而文献[11]的最大值和最小值比文献[21]的约大 0.002 和 0.12。

结束语 针对当前位置 k-匿名解决方案中未充分考虑敌手具有的各种背景知识信息以及匿名区构建及位置选取时距离方式计算的耗时问题，本文提出增强型位置 k-匿名方案。在匿名区构建时，充分考虑与物理位置相关的背景知识，如时间属性、位置语义、查询概率等，并在位置选择时保证位置的分散性，这样既可以抵御语义导致的位置隐私泄露，同时还可以保证时间不一致或查询概率偏差过大导致的隐私泄露。对于空间位置，采取 Geohash 算法进行编码，可以使用反向检索快速完成匿名区的扩大，缩短匿名处理时间，同时可抵御位置分布攻击。但本文主要针对的是快照位置隐私保护，如何将此方案应用到轨迹隐私保护，需再拓展哪些语义信息，是下一步要研究的工作。

参 考 文 献

[1] HE Y, CHEN J. User location privacy protection mechanism for location-based services[J]. Digital Communications and Networks, 2021, 7(2): 264-276.

[2] WANG Y, ZUO K, LIU R, et al. Dynamic pseudonym semantic-location privacy protection based on continuous query for road network[J]. Int. J. Netw. Secur, 2021, 23: 642-649.

[3] MUSHTAQ M, ULLAH A, ASHRAF H, et al. Anonymity Assurance using efficient pseudonym consumption in Internet of vehicles[J]. Sensors, 2023, 23: 5217-5234.

[4] ZHOU J Q, LI Y J. Personalized dummy generation method based on spatiotemporal correlations and location semantics[J]. Ruan Jian Xue Bao/Journal of Software, 2019, 30(S1): 18-26.

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

[7] LIANH, 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.

[8] ZHANG L, SONG G, ZHU, D, et al. Location privacy preservation through kernel transformation[J]. Concurrency and Computation: Practice and Experience, 2020, 34(16): e6014.

[9] ZHANG X 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.

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

[11] YANG X, GAO L, WANG H, et al. A User-related semantic location privacy protection method in location-based service [C]// Proceedings of the IEEE 27th International Conference on Parallel and Distributed Systems (ICPADS). 2021: 14-16.

[12] YANG X, GAO L, ZHENG J, et al. Location privacy preservation mechanism for location-based service with incomplete location Data[J]. IEEE Access, 2020(8): 95843-95854.

[13] ZHANG L, LIU D, CHEN M, et al. A user collaboration privacy protection scheme with threshold scheme and smart contract [J]. Information Sciences, 2021, 560: 183-201.

[14] KUANG L, WANG Y, ZHENG X, et al. Using location semantics to realize personalized road network location privacy protection[J]. J Wireless Com Network, 2020, 1(2020): 1-16.

[15] XU M J, HAN J. Next location recommendation based on semantic-behavior prediction[C]// Proceedings of the 2020 5th International Conference on Big Data and Computing. 2020: 65-73.

[16] SHI X, ZHANG J, GONG Y. A dummy location generation algorithm based on the semantic quantification of location[C]// Proceedings of 2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA). 2021: 172-176.

[17] JIAO Z X, ZHANG L, LIU X P. A dummy location selection algorithm based on differentiated time-segment and fine-grained [J]. Journal of Nanjing University of Posts and Telecommunications (Natual Science Edition), 2022, 42(6): 106-114.

[18] XING L, ZHANG D, WU H, et al. Distributed K-Anonymous Location Privacy Protection Algorithm Based on Interest Points and User Social Behavior[J]. Electronics, 2023, 12: 2446.

[19] TU S P, ZHANG L, LIU X P. Double Dummy Location Selection Algorithm Based on Behavior Correlation[J]. Computer Science, 2023, 50(5): 348-354.

[20] WikiPedian. Geohash [OL]. [2019-11-20]. http://http://en.wikipedia.org/wiki/Geohash.

[21] YIN F M, CHEN H. K Anonymous Location Privacy Preservation Scheme Based on Geohash Coding[J]. Journal of Wuhan University (Natural Science Edition), 2022, 68(1): 73-82.

[22] CHEN Y. Information Theory and Coding (2nd ed)[M]. Beijing: Publishing House of Electronics Industry, 2012.

[23] ZHANG Y B, ZHANG Q Y, LI Z Y, et al. A k-anonymous Location Privacy Protection Method of Dummy Based on Geogra-

phical Semantics[J]. Int. J. Netw. Secur, 2019, 21: 937-946.
[24] XING L, JIA X, GAO J, et al. A location privacy protection algorithm based on double K-anonymity in the social Internet of vehicles[J]. IEEE Communications Letters, 2021, 25: 3199-3203.
[25] YANG M X, YE B P, CHEN Y L, et al. A trusted de-swinging k-anonymity scheme for location privacy protection [J]. Journal of Cloud Computing, 2022, 11(1): 1-10.
[26] ZHOU Y H, LI G H, YANG Y G, et al. Location Privacy Preserving Nearest Neighbor Querying Based on GeoHash[J]. Computer Science, 2019, 46(8): 212-216.
[27] FENG J Y, YANG J W, ZHANG R T, et al. A Spectrum Sharing Incentive Scheme Against Location Privacy Leakage in IoT Networks[J]. Journal of Computer Research and Development, 2020, 57(10): 2209-2220.
[28] LI Z Y, ZHAO Z F. Index and Query Method Based on Spatial-Temporal Distribution of Trajectory Big Data[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2022, 54(3): 528-536.
[29] LIU K, HAN Y L, WANG J J, et al. Location Privacy Protection Method Based on Geohash Coding and Pseudo-Random Sequence[C]// Proceedings of the 2022 3rd Information Communication Technologies Conference(ICTC). 2022: 178-183.
[30] FEI F, LI S, DAI H, et al. A K-Anonymity Based Schema for Location Privacy Preservation[J]. IEEE Transactions on Sustainable Computing, 2019, 4(2): 156-167.
[31] SHAHID A R, PISSINOU N, IYENGAR S S, et al. Delay-aware privacy-preserving location-based services under spatiotemporal constraints[J]. International Journal of Communication Systems, 2021, 34: e4656.
[32] ZHENG Y, XIE X, MA W Y. GeoLife: A Collaborative So cial Networking Service among User, location and trajectory[J]. IEEE Data Engineering Bulletin, 2021, 33(2): 32-40.
[33] Gaode Map[OL]. https://www.amap.com/.
[34] Dianping[OL]. http://www.dianping.com/.

LI Yongjun, born in 1983, doctoral student. Her main research interests include privacy protection and so on.

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

（责任编辑：柯颖）