计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200800069

基于用户偏好和位置分布的假位置生成方法

王 辉 朱国宇 申自浩 刘 琨 刘沛骞
河南理工大学计算机科学与技术学院 河南 焦作 454003
(204932059@qq.com)

摘 要 传统的基于k-匿名机制的假位置生成算法生成的假位置的合理性较低,易被攻击者利用边信息进行攻击。针对此问题,提出了SPDGM算法。首先,定义语义加权有向图,描述语义的时间分布和语义转移关系;其次,为解决仅考虑位置历史概率产生的抵抗能力弱的问题,提出了位置可信度,统一考虑了位置历史概率和大众的评价信息;再次,为避免假位置分布过于密集,定义了离散度,以控制假位置的分布情况;最后,生成语义安全且分布稀疏的匿名集。实验证明,在语义攻击下SPDGM算法具有更低的被识别率,更高的隐私保护强度;在考虑语义攻击的算法中,SPDGM算法的运行时间更短。因此,SPDGM算法具有可行性与实用性。

关键词:假位置;语义转移;分布度量;用户偏好;位置隐私保护
中图法分类号 TP309

Dummy Location Generation Method Based on User Preference and Location Distribution

WANG Hui,ZHU Guo-yu,SHEN Zi-hao,LIU Kun and LIU Pei-qian
College of Computer Science and Technology,Henan Polytechnic University,Jiaozuo,Henan 454003,China

Abstract The traditional dummy location generation algorithm based on k-anonymity mechanism has low rationality and is vulnerable to attack by attackers using side information. Aiming at solving this problem,the SPDGM algorithm is proposed. Firstly,this algorithm defines the semantic weighted digraph to describe the time distribution and semantic transfer relationship of semantics. Secondly,for the sake of solving the problem of weak resistance caused by only considering the historical probability of location,this algorithm defines the location credibility,which considers the historical probability of location and the evaluation information of the public. Thirdly,in order to avoid the dense distribution of dummy location,the dispersion degree is defined to control the distribution of dummy location. Finally,this algorithm generates an anonymous set whose semantics safe and distribution sparsely. The experimental results show that the SPDGM algorithm has lower recognition rate and higher privacy protection strength under the semantic attack,and the running time of the algorithm considering semantic attack is lower. Therefore,SPDGM algorithm is feasibility and practicability.

Keywords Dummy location,Semantic transfer,Distribution metric,User preference,Location privacy-perserving

1 引言

近年来,随着智能终端的普及和物联网技术的发展,基于位置的服务(Location-Based Service,LBS)在日常生活中的应用愈加广泛[1],如百度地图服务、美团本地推荐服务等。用户下载并应用LBS应用程序后,可以通过向LBS服务器发送请求从而获取服务信息,如道路实时信息、附近的美食信息等。然而,位置信息反映了用户的习惯偏好,位置之间存在的关联关系是重要的个人隐私,同时这些信息面临着泄露风险。攻击者可以通过用户的位置隐私数据直接或者间接地推断出用户的个人敏感信息,如用户的社会关系、健康状况、家庭住址、经济情况等[2]。因此,在生成、处理、使用和存储这些隐私数据的过程中,如何保护其不被攻击者窃取成为亟待解决的问题。

近年来,国内外研究者提出了多种位置隐私保护方法,其中基于k-匿名模型的讨论相当活跃。位置k-匿名模型的基本思想[3-4]是:用一个包含目标用户在内的k个用户的匿名区域代替用户的真实位置,从而使攻击者无法区分出目标用户的真实位置,进而达到保护用户位置隐私的目的。

通过获取真实用户的数据或假位置数据均可以达到k-匿名的效果。对于前者,根据系统结构的不同,大致可以分为集中式结构[5]和分布式结构[6-8]。集中式结构的实现需要第三方的辅助,通过第三方获取周围$k-1$个用户的真实位置信息,将位置数据进行打包处理并发送给LBS服务器。该结构

到稿日期:2020-08-11 返修日期:2020-11-12 本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61300216)
This work was supported by the National Natural Science Foundation of China(61300216).
通信作者:申自浩(1124731892@qq.com)

中的第三方存在单点故障和瓶颈问题，但是其分担了客户端分析计算等任务开销，有效降低了隐私保护措施对客户端的要求。

相较于集中式结构，分布式结构一般不需要引入第三方，目标用户获取周围 $k-1$ 个用户的真实位置信息而不需要第三方提供帮助，一般通过点对点方式进行用户间的通信，可以利用广播或者 WIFI 接入点等技术实现。即使引入第三方，一般也仅用于加密等处理，第三方不直接获取或处理周围用户的真实位置信息。这种结构避免了单点故障和瓶颈问题，但是需要关注用户通信存在的延迟问题和串通攻击问题。

通过假位置数据实现 k-匿名的系统结构一般是客户端/服务器型结构，不需要第三方实体参与，只需要通过一定的算法生成满足需求的假位置数据，进而实现 k-匿名效果，达到保护位置隐私的目的。

日常生活中常见但不易察觉的场景（如晚上十点后，某人喜欢去酒吧；在搜索餐馆后有些人会搜索停车场，而有的人会搜索公交路线）反映出了重要的信息：人们的行为习惯存在规律性。即便处于新的环境，已有的偏好习惯在一定程度上也会得到保留，因此生活轨迹是可预测的。攻击者利用数据挖掘等技术可展开语义攻击，严重威胁了用户的位置隐私安全。针对这些问题，本文提出了基于用户语义偏好的假位置生成方法（Semantics Preference-based Dummy Generation Method，SPDGM），以保护用户的位置隐私。

本文的创新工作如下：

（1）根据日常生活中的场景，分析位置语义信息以及人们行为习惯的规律性，引出语义转移关系，并提出了 SPDGM 算法。

（2）提出语义差异度，衡量多个语义在整体上的差异程度，挑选最小差异度的语义集合以隐藏真实语义。

（3）提出位置可信度，分析用户行为习惯，综合考虑历史查询数据和大众的评价信息，生成了更具合理性的假位置集合。

（4）提出离散度来衡量位置分布情况，量化假位置候选集合中各位置附近的密集程度，在生成匿名集时控制位置的分布。

2　相关工作

近年来，研究者们提出了众多利用假位置数据以达到匿名效果的算法，根据数据生成方式的不同，大致分为利用算法和利用历史信息两种类型。在前者的研究中，出现了匿名区域随机生成算法[9]、CirDummy 和 GridDummy[10]等，但是生成的假位置的合理性较差，如在河流和湖泊中等不寻常的位置，能被攻击者以较低的代价过滤掉。目前，涌现了一些新颖的算法，如 Song 等[11]考虑了不可达的障碍区域，提出了高效假位置生成算法。

在利用用户历史信息的研究中，研究者们利用历史查询频率来生成假位置集[12-15]。Niu 等[12]通过划分网格并计算各个网格的历史查询概率，提出了（Dummy Location Selection，DLS）算法和 Enhanced-DLS 算法，通过获取最大位置熵形成匿名集。Sun 等[15]提出了 ADLS 攻击算法，该算法是对 DLS 算法的进一步分析，并在此基础上提出了假位置隐私保护算法（Dummy Location Privacy-Preserving，DLP），防止攻击者通过概率攻击窃取用户真实信息。若只考虑历史查询概率，攻击者能够运用聚类、数据挖掘等技术，并通过分析用户偏好信息来攻击用户，降低了用户的隐私保护强度。

鉴于上述问题，研究者们开始关注位置语义和用户偏好信息[4,16-23]。但是这些属性丰富且驳杂，不同研究者的关注点不尽相同。

Zhu 等[18]考虑到用户移动类型和停留点偏好，根据数据挖掘的常用加权技术 TF-IDF 和 Web 结构挖掘技术 HITS 进行分析处理，对停留点的相似性和受欢迎程度进行量化。但是该处理方式存在局限性，即分析结果对数据的结构和属性依赖较大。

Wang 等[20]提出了一种基于位置语义和查询概率的最大最小假位置选择（Maximum and Minimum Dummy Selection，MMDS）算法。该算法量化了位置语义，并计算了位置间的语义差，在生成假位置时确保位置间的语义差达到阈值，从而避免同质性攻击。该研究提供了很好的思路，但是语义的度量较粗糙，所提出的语义树难以准确衡量语义间的关系，利用语义差衡量语义之间的差值的充分性不足。

Tu 等[22]提出了抵抗语义攻击和识别攻击的算法，通过分析用户频繁访问的位置，提取出用户的语义特征，并且满足了 k-匿名、l-多样性和 t-封闭性，保证了数据的可用性，但是其没有考虑到用户的生活习惯，即语义关联。

Hara[23]注意到了语义关联信息，从服务器的角度分析了用户偏好属性和用户保护度可调节问题，但是没有考虑生成的假位置可信度、假位置分散程度以及位置语义在时间上的分布，这给攻击者的攻击提供了可能。

现有方法对抵抗拥有边信息的攻击具有局限性，没有考虑真实用户的偏好习惯，攻击者能够利用这些信息低代价地过滤掉假位置，降低了用户的隐私保护强度。本文提出的 SPDGM 算法考虑了用户的偏好习惯，包括空间偏好及时间偏好，即考虑了从一种语义（位置）到另一种语义（位置）的可能性和语义在时间上的分布情况，可以个性定制防护措施，能有效抵抗位置攻击和推理攻击等，达到保护用户位置隐私的目的。

本文第 2 节介绍了此领域的发展情况；第 3 节详细描述了相关定义和模型结构；第 4 节给出了 SPDGM 算法的具体分析以及实现步骤；第 5 节给出了实验分析；最后总结全文并展望未来。

3　相关定义和模型结构

3.1　相关定义

位置语义是多组经纬度坐标的文字符号表示的提炼，如北京市十中、北京市十九中、北京市 101 中学等所具有的语义为北京市中学。位置语义反映了用户在某时刻的偏好等敏感信息，如上午 10 点位于语义为“北京中学”的位置，说明该用户可能为中学生或者学校工作人员。本文定义的语义具有以下性质：同一语义可以拥有多个位置，一个位置也可以拥有多个语义。例如：某区域各种名字的酒吧都只有一个语义——

酒吧，而有的商家除了提供咖啡外还提供快餐服务。而语义设置的粒度要相对较小，如设置语义为北京市中学，而不是中学或者学校。

3.1.1 数据处理

为了分析语义出现的时间，以及语义转移发生的时间，本文定义了时间分布，用于评价语义集中时段，反映了用户偏好与时间的关系。

定义 1（时间分布 $T_{\mu,\sigma}$） 表示语义转移出现相对集中的时段。用 μ,σ 限定，其中 μ 为时间点，σ 为时间范围。(μ,σ) 表示时间区间 $(\mu-\sigma)$ 至 $(\mu+\sigma)$，如 (1 205,15) 表示时间段 11:50—12:20。搜索餐馆类的查询可能分布在一天的多个时段，则 $T_{\mu,\sigma}$ 含有多个 (μ,σ) 的数据，组成形式为 $((\mu_1,\sigma_1),(\mu_2,\sigma_2),\cdots,)$，用链表保存该数据项，在不致混淆的情况下用 T 表示。

定义 2（历史数据库 DB_h） 用于存储用户历史数据信息，$((x,y),label,t)$ 代表 DB_h 中的一条记录。其中 (x,y) 代表用户的地理坐标，即经纬度；$label$ 代表该位置的语义；t 代表产生请求数据的时刻。数据库的接口函数有：

(1) $\Gamma(label)$，统计数据库 DB_h 中语义为 $label$ 的记录个数。

(2) $\Psi(label_i,label_j)$，统计数据库 DB_h 中相邻的两个记录中的语义中，前一个语义为 $label_i$ 且后一个语义为 $label_j$ 的记录个数和语义转移发生的时段 T。当产生语义为 $label_j$ 的记录时，说明语义发生了转移，以语义为 $label_j$ 的时间分布代表语义转移发生的时段 T。

为了清晰地描述语义转移关系，本文定义了有向图来描述用户语义偏好习惯。

定义 3（语义加权有向图 G） 简称有向图 G，$G=(V,E,AV,AE,WV,WE)$（见图 1）。其中：

(1) V 代表有向图中所有语义节点的集合，$V=\{v_i \mid i=1,2,\cdots,n_v\}$，$v_i$ 为有向图节点，n_v 为有向图中所有节点的个数。

(2) E 代表有向图中有向边的集合，$E=\{e_i \mid i=1,2,\cdots,n_e\}$，$e_i$ 为有向图的边，n_e 为有向图中所有边的个数。

(3) AV 代表有向图中顶点集的属性集合，$\{\forall av_j \in AV \mid j=1,2,\cdots,n_v\}$ 对应 V 的属性集合，$av_j(av_j \in AV)$ 对应为节点 $v_j(v_j \in V)$ 的属性；AV 中的元素用二元组 $(label,num)$ 表示。$label$ 为语义类别，且每个节点均不相同；num 为该语义在历史数据库 DB_h 中出现的次数。

(4) AE 代表有向图中有向边集的属性集合，$\{\forall ae_j \in AE \mid j=1,2,\cdots,n_e\}$ 对应 E 的属性集合，$ae_j(ae_j \in AE)$ 对应为节点 $e_j(e_j \in E)$ 的属性，AE 的元素用三元组 $(\langle v_k,v_r\rangle,num,T)$ 表示，$\langle v_k,v_r\rangle$ 表示该边是由节点 $v_k(\in AV)$ 到节点 $v_r(\in AV)$；num 为该边出现的次数；T 为该边的时间分布。

(5) WV 为节点的权值集合，代表节点偏好支持度，$WV=\{wv_i \in WV \mid i=1,2,\cdots,n_v\}$，$wv_i$ 对应为节点 v_i 的语义支持度，其计算公式如下：

$$wv_i=\frac{av_i.num}{\sum_{av_j \in AV} av_j.num} \tag{1}$$

(6) WE 为边的权值集合，表示有向图节点转移概率或语义转移概率，$WE=\{we_i \in WE \mid i=1,2,\cdots,n_e\}$，$we_i$ 对应为边 e_i 的节点转移概率，其计算公式如下：

$$we_i=\frac{ae_i.num}{\sum_{ae_j \in AE} ae_j.num} \tag{2}$$

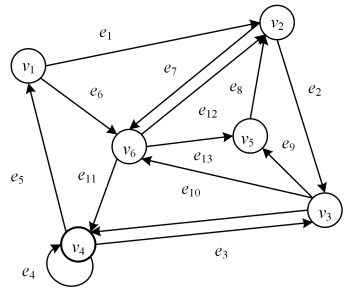

图 1 语义加权有向图

Fig. 1 Semantic weighted digraph

本文采用语义转移关系而不采用用户轨迹来描述用户偏好的主要原因是：1）分析轨迹数据时难以保证轨迹中每个语义的时间分布具有合理性；2）轨迹的起始点难以判定，轨迹的长度难以确定；3）各个轨迹出现的频率比较小，处理时误差较大；4）轨迹中各点暴露的可能性是不一样的，若轨迹中某个点暴露，则整条轨迹随之被过滤掉。

定义 4（地图数据库 DB_m） 用于存储处理后的地图数据，$((x,y),label,Lconf)$ 为数据库中 DB_m 的一条记录。其中 (x,y) 代表位置经纬度坐标；$label$ 代表该位置的语义；$Lconf$ 代表该位置的可信度，并满足下式：

$$Lconf=\alpha \cdot p_{(x,y)}+(1-\alpha)\cdot evaValue \tag{3}$$

其中，α 为用户可选择的权值；$evaValue$ 为从服务器获得的该位置的大众评价信息，用从服务器返回的该位置在服务器中被查询的总次数来表示，反映了该位置的受欢迎程度；$p_{(x,y)}$ 为位置 (x,y) 的查询次数在用户的历史查询总次数中的比例，满足：

$$p_{(x,y)}=\frac{\#\text{ records contains}(x,y)\text{ in }DB_h}{\#\text{ records in }DB_h} \tag{4}$$

3.1.2 语义分析

定义 5（语义差异度 H） 本文用语义差异度来度量语义转移概率之间的差异程度。对于图 1 中节点 v_6，其下一节点可能为 v_2，v_3 或者 v_4，根据定义 3 可知，其转移概率与边 e_5，e_7 和 e_8 的权值有关，本文用语义转移概率评估下一节点的选择可能性，用语义差异度 H 量化语义转移概率的差异程度。若边 e_5，e_7 和 e_8 的权值相等，则其语义差异度为 0，下一点选择为 v_2，v_3 或者 v_4 的可能性相等。

对于转移概率集合 $P=\{p_1,p_2,\cdots,p_n\}$，选出与语义转移概率 p 差异度最小的 m 个数据。集合 P 中的元素的差异性给攻击者的攻击提供了“支持”。为减少异常值的影响，对 m 个数据进行归一化处理，并对应表示为 $\{q_1,q_2,\cdots,q_m\}$，其均值为 $\frac{1}{m}$，q_i 与其均值的差越大，即 $\left|q_i-\frac{1}{m}\right|$ 越大，说明 q_i 对该组数据的差异性贡献越大，攻击者越容易过滤出真实语义。考虑到对称性，本文用 $\exp\left(\frac{(q_i-\frac{1}{m})^2}{2}\right)$ 表示 q_i 元素对于一组数据整体的差异程度。根据乘法原理，一组经过归一化处理的数据中的元素对整体差异性的贡献度为各自差异度之积。

一组数据的整体差异度的公式如下：

$$H=\prod_{i=1}^{m}\exp\left(\frac{\left(q_i-\frac{1}{m}\right)^2}{2}\right)$$
$$=\exp\left(\frac{\left(q_1-\frac{1}{m}\right)^2+\left(q_2-\frac{1}{m}\right)^2+\cdots+\left(q_m-\frac{1}{m}\right)^2}{2}\right)$$
$$=\exp\left(\frac{\sigma^2\cdot m}{2}\right) \tag{5}$$

$$\begin{cases} q_i=\dfrac{p_i}{\sum_{j=1}^{m}p_j} \\ \sum_{i=1}^{m}q_i=1 \\ \sigma=\sqrt{\dfrac{1}{m}\sum_{i=1}^{m}\left(q_i-\dfrac{1}{m}\right)^2}\ \text{i.e. } \sigma \text{ is Std. Deviation} \end{cases}$$

上式为式(5)的约束条件，m 为欲选取的语义个数。本文的目的是获得最小的差异度，差异度越小，攻击者越难识别真实用户查询的语义。m 个语义的查询概率相等，即标准差为 0 时，语义差异度最小。

定义 6(方差Var_C)　表示集合 C 的方差，用式(6)表示，其中 $\bar{p}$ 为集合 C 的均值。

$$Var_C=\frac{1}{\|C\|}\sum_{p_i\in C}(p_i-\bar{p})^2 \tag{6}$$

3.1.3　空间离散度

定义 7(离散度 K)　用以衡量集合内各点的离散度和密集分布区域。离散度量化值等于与位置 l 距离 num 个单位观察距离的点的个数。

$$K=Cnt(d_0,num,l) \tag{7}$$

其中，$d_0\triangleq$单位观察距离，$num\in\mathbb{N}^+$。位置分布示意图如图 2 所示。图 2 中空心小圆圈代表位置点。

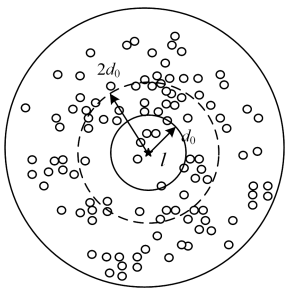

图 2　位置分布示意图

Fig. 2　Location distribution sketch map

定义 8(欧氏距离 $dist(l_i,l_j)$)　表示位置$l_i(x_i,y_i)$到位置$l_j(x_j,y_j)$间的距离，本文用欧氏距离表示。具体公式如下：

$$dist(l_i,l_j)=\sqrt{(x_i-x_j)^2+(y_i-y_j)^2} \tag{8}$$

3.2　客户端/服务器的模型

本文设计的模型主要分成两部分：客户端和服务器。该模型不需要第三方实体参与，有效避免了单点故障和瓶颈问题。

3.2.1　服务器端

服务器在模型中的角色是服务的提供者，拥有客户所需区域的地图信息，能够正确回答客户端的请求服务。

3.2.2　客户端

客户端通常是移动终端设备(如手机、智能平板)，在模型中的角色是服务的请求者，客户端需要具备的能力有：1)从服务器端获得某区域的地图信息，并处理相关信息；2)记录本地查询记录，分析处理历史记录；3)获取位置的经纬度。

4　SPDGM 算法分析

本文提出了基于语义偏好的假位置生成算法，综合分析了语义间关联关系和语义时间关联关系。当通过本文算法构建匿名集时，考虑了位置语义，目的是在构建匿名集时保证各位置的语义具有足够小的差异性。该算法在构造匿名集时保证了各位置点的语义选择概率差异尽量小、空间分布上足够分散、选择位置的可信度满足阈值。

SPDGM 算法的流程如图 3 所示，图中的“否”分支仅说明未达到阈值时可以返回前述步骤，修改参数可继续执行，或者匿名失败。

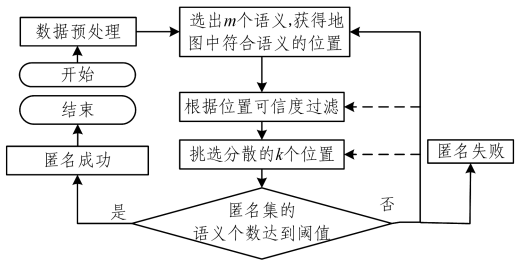

图 3　SPDGM 算法流程

Fig. 3　Flow chart of SPDGM

4.1　数据处理

在历史数据库DB_h中，为避免同一请求的多次查询，对于其中相邻的两条记录$((x_i,y_i),label_i,t_i)$与$((x_j,y_j),label_j,t_j)$，满足$|t_i-t_j|\leqslant threshold$，$threshold$ 为用户设定的阈值，用户每完成一次查询请求，就将请求数据保存在 DB_h中。

语义转移概率在用户偏好习惯方面的价值和两次查询的时间间隔相关。时间间隔越长，两次连续查询的语义之间的联系越低，当时间足够长时，语义转移概率的可用性急剧下降，而用户语义支持度对用户偏好的影响更大，用户偏好评价值如下：

$$Prefer=\frac{\lambda p_{i,j}}{\Delta t}+(1-\frac{\lambda}{\Delta t})p_{label_j} \tag{9}$$

其中，$p_{i,j}$ 代表语义$label_i$到语义$label_j$的语义转移概率，由式(2)计算得到；p_{label_j}为语义 $label_j$的语义支持度，由式(1)计算得到；Δt 为本次查询与上一次查询的时间间隔，单位为 min；λ 为用户设置的常数。

根据式(9)和有向图 G 的权值信息，获得语义$label_i$到语义 $label_j$的语义转移数据$(\langle label_i,label_j\rangle,Prefer,T)$，其中 T 为语义 $label_j$的时间分布。称$label_i$为前驱语义，$label_j$为后继语义。上一次查询语义集合为 S_p，只需要处理有向图 G 中前驱语义属于集合 S_p 的语义转移数据，并记为 C_{tran}。

当用户在某个区域内首次进行请求位置服务时，需要从服务器获得该区域的地图信息，可以从 LBS 服务器等获得以

进行离线使用。信息包括该区域的位置点的经纬度、名称和大众的评价信息等简要信息。结合有向图 G、本地历史数据库 DB_h 和定义 4，构造地图数据库 DB_m。

4.2 假位置候选集合构造算法

假位置候选集合构造算法的流程如下。

(1)假设目标用户欲查询内容的语义为 S，上一次查询语义集合 S_p，欲选取的语义个数为 m，本次查询的真实偏好评价值为 $Prefer_0$，位置可信度阈值为 $Lconf_0$，存放候选语义集的集合用 C_S 表示。

(2)将 C_{tran} 中的位置语义的时间分布包含查询时间的所有元素放入链表 $list_{tran}$ 中，根据式(9)计算其评价值。

(3)将各元素的偏好评价值 $Prefer$ 按升序排列。计算 $list_{tran}$ 中元素的偏好评价值等于 $Prefer_0$ 的个数，记为 M_S，若 $M_S \geqslant m$，则将这 M_S 个元素全部放入集合 C_S 中，转到步骤(6)；若 $M_S < m$，将此 M_S 个元素全部放入链表 $list$ 中，并从元素 S 两边各取 $(m-M_S)$ 个元素放入 $list$ 中，按升序处理，执行步骤(4)。

(4)将此 M_S 个元素全部放入集合 C_S 中，用 $P_{\max}$ 和 $P_{\min}$ 分别代表集合 C_S 中元素的最大值与最小值；用 $P_{\max-\min}$ 表示链表 $list$ 中非 C_S 集合的元素中大于 $P_{\max}$ 的元素的最小值；用 $P_{\min-\max}$ 表示链表 $list$ 中非 C_S 集合的元素中小于 $P_{\min}$ 的元素的最大值。将之分别与集合 C_S 中的元素结合，根据式(3)分别计算其方差，分别用 $Var_{\max-\min}$ 和 $Var_{\min-\max}$ 表示，将方差较大者的元素添加到集合 C_S 中。

(5)重复步骤(2)直到集合 C_S 的模为 m。

(6)从处理后的地图数据中获取符合集合 C_S 中语义的所有位置，并放入假位置候选集 C_d 中。对 C_d 中的每一个元素根据式(3)进行处理，然后进行判断，若 C_d 中的元素的可信度小于可信度阈值 $Lconf_0$，则将其从 C_d 中删除。

4.3 空间离散度分析

空间离散度是对一系列点的离散程度或混乱程度的度量。因为实际位置的分布多种多样，为了能够度量不同位置分布下的混乱程度，本文提出了一种度量方法。该方法逐点计算了距离该点的 num 个单位观察距离的点的个数，$num \geqslant 1$ 且为整数，即利用离散度度量混乱程度。对于位置点的离散度分析，仅分析 num 为 1 的情况，在不致混淆的情况下用 $Cnt(l)$ 或 Cnt 表示。从 4.2 节获得的集合 C_d 中得到最终的匿名集 C 的步骤如下。

(1)将用户欲查询的位置 l_{real} 放进匿名集 C，然后删除假位置候选集合 C_d 中距离 l_{real} 一个单位观察距离以内的点。

(2)逐点计算集合 C_d 的 $Cnt(l)$，并将结果升序放置在链表 $list_{cnt}$ 中，从 Cnt 最大的点开始的 c 个点内，选出首个点的语义不存在于匿名集 C 中的点进行分析，若未找到，则取最大值点进行分析，并记为 $l_{\max}$。将 $l_{\max}$ 放进匿名集 C，然后删除链表 $list_{cnt}$ 中距离 $l_{\max}$ 一个单位观察距离以内的点，更新链表 $list_{cnt}$。

(3)重复步骤(2)，若 $\|C\| = k$ 或 $list_{cnt}$ 中没有数值大于 0 的元素，则终止运算，执行步骤(4)；否则，统计 $list_{cnt}$ 中为 0 的元素个数，并记为 k_0。如果 $\|C\| + k_0 \geqslant k$，则从 k_0 中随机选择 $(k-k_0)$ 个元素放进匿名集 C。

(4)若匿名集 C 中位置的语义总数大于或等于 S $(S \leqslant m)$，则匿名成功；否则，匿名失败。

空间离散度分析时，查询区域取点示意图如图 4 所示。其中实心五角星为 l_{real}，其单位观察距离内的点已丢弃；实心圆点为 Cnt 最大点；r 为查询范围半径。

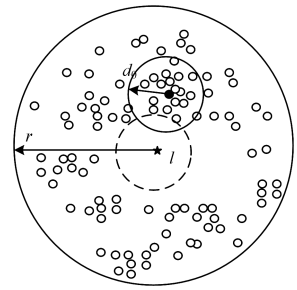

图 4 查询区域取点示意图

Fig. 4 Sketch map of taking points in query area

4.4 数据规模分析

本文构建的两个数据库和有向图不需要在每次查询时重新生成，数据的少量变化对用户偏好分析的影响很小，且只需在查询前完成即可；构造 C_{tran} 时，只处理有向图 G 中前驱语义属于集合 S_p 的语义转移数据，不需要对整个有向图数据进行处理；空间离散度分析时，通过 4.2 节获得的数据规模很小，逐点计算的复杂度很小。

4.5 安全性分析

本文算法能够抵抗利用边信息的攻击。边信息包括但不限于用户历史查询概率、用户习惯偏好等用户敏感信息。本文分析用户的位置语义转移概率，通过 4.2 节的处理，产生候选集合 C_d，其元素的语义在概率上差异性最小且满足时间分布。这使得攻击者无法根据时间信息以及概率信息分析出用户真实位置，从而使算法具备抵抗利用边信息攻击的能力。

本文算法能够避免匿名位置过于密集的问题，能有效抵抗位置同质性攻击。对位置点进行空间分布处理时，其考虑了位置可信度和匿名集离散度，保证了生成的假位置点之间的距离大于单位观察距离，匿名集中的假位置在地理位置上具有分散性。

5 实验与性能分析

为了评估 SPDGM 算法的性能，本文采用 GeoLife 数据集[24]进行了分析，该数据集包括 182 名真实用户的 17 621 条真实轨迹，该数据集分布于 2007 年 4 月到 2012 年 8 月间，用户的生活轨迹记录时长从几周到数年不等，采样时间为 1～5 s。

为了方便分析讨论，本文观察时间窗口为一天，即 0:00—23:59。通过 $\Gamma(label)$ 获取历史数据库中语义为 $label$ 的查询发生的个数，根据聚类分析手段，如基于密度的聚类算法 DBSCAN[25]，利用 $\Psi(label_i, label_j)$ 获取语义 $label_i$ 到语义 $label_j$ 的语义转移发生的时间分布 T。截取的 GeoLife 数据如表 1 所列。

表 1　GeoLife 数据表

Table 1　Table of data in GeoLife

纬度	经度	保留	高度	时间 1	日期	时间 2
39.984683	116.31845	0	492	39744.120254629 6	2008-10-23	02:53:10
39.984686	116.318417	0	492	39744.1203125	2008-10-23	02:53:15
39.984688	116.318385	0	492	39744.120370370 4	2008-10-23	02:53:20
…	…	…	…	…	…	…
39.995224	116.326677	0	509	39748.497673611 1	2008-10-27	11:56:39
39.995287	116.326372	0	511	39748.497731481 5	2008-10-27	11:56:44
39.995361	116.326374	0	512	39748.497789351 9	2008-10-27	11:56:49

数据预处理时，时间数据使用表 1 中的“时间 2”一列，并随机间隔 5～15 min 提取该行数据，不考虑“日期”一列。大众的评价信息采用随机值。

实验环境是 Inter(R)Core(TM)i5-7200U 2.5 GHz 处理器，8 GB 内存，Windows10 操作系统。在 PyCharm 2019.3.1 环境下，通过 Python 编程实现算法。实验参数如表 2 所列。

表 2　实验参数

Table 2　Experimental parameters

参数	默认值	取值范围
r	5 km	—
λ	5	—
m	10	[2,12]
k	8	[5,20]
d_0	1.5 km	[0.5,5]，步长 0.5
α	0.2	[0.1-0.8]，步长 0.1

5.1　隐私保护强度分析

为了对比考虑语义信息对匿名保护的优势，将本文算法与未考虑语义信息的 DLP 算法[13]进行对比。实验采用语义攻击[22]，即根据语义判断位置点的合理性。被攻击者识别的位置越少，说明用户的隐私保护强度越强。利用式(10)评估隐私保护强度，η 为被识别概率，其值越小，隐私保护强度越强。本文就隐私保护等级 k 将本文算法与 DLP 算法[13]、MMDS 算法[20]和 HARA 算法[23](因为该算法在文献[23]没有命名，本文以其作者名字表示)进行对比实验，其他参数按照表 2 默认设置。

$$\eta=\frac{\#\,identified\ locations}{\#\,total\ locations} \tag{10}$$

图 5 给出了在相同的攻击下、不同算法下，用户真实位置被识别的概率。

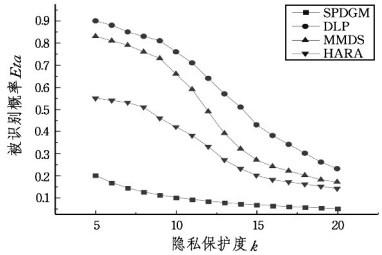

图 5　不同算法的被识别概率

Fig. 5　Recognition probability of different algorithms

实验结果表明：随着隐私保护度 k 的增大，不同算法的被识别概率均下降；由于 DLP 算法仅考虑了用户的历史查询概率，没有考虑位置语义信息，所以隐私保护强度较低，被识别概率高于其他算法；尽管 SPDGM 算法、MMDS 算法和 HARA 算法均考虑了位置语义信息，但是 MMDS 算法仅考虑了位置间的语义差，对语义转移信息未做分析，而 HARA 算法虽然考虑了语义转移信息，但是不能保证每一位置的语义均符合历史时间分布，所以 SPDGM 算法的被识别概率低于其他算法。

5.2　时间开销分析

本文采用客户端/服务器体系结构，客户端需要具有一定的计算能力，因此假位置生成时间是反映假位置生成算法性能的重要指标。将本文算法与 DLP 算法、MMDS 算法和 HARA 算法进行对比分析，图 6 的实验结果表明：随着隐私保护度 k 增加，运行时间均会增加；由于 DLP 算法没有进行语义信息分析处理，其运行时间在相同条件下最少；SPDGM 算法的运行时间低于 MMDS 算法和 HARA 算法，且 SPDGM 算法的运行时间随隐私保护度的增加而增长，但增长趋势较平缓。

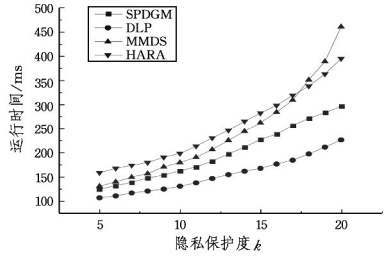

图 6　不同算法的运行时间对比

Fig. 6　Runtime of different algorithms

5.3　匿名成功率分析

在实验性能评估中，因为 m，d_0，α 均是本文中独有的参数，本文仅对 m，d_0，α 分别进行控制变量实验，除控制变量外，其他参数均按照表 2 设置，不进行对比实验，分别分析不同参数的匿名成功率的变化情况。

图 7 给出了用户隐私保护要求的语义个数与匿名成功率之间的关系。当语义需求越多时，产生的候选匿名集合的元素越多，越容易挑选出匿名集，所以匿名成功率随着语义数的增加而增加。

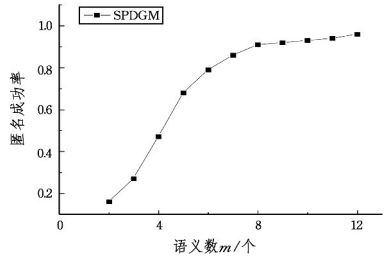

图 7　语义数和匿名成功率的关系

Fig. 7　Relationship between semantic number and anonymity success rate

图 8 给出了单位观测距离与匿名成功率的关系。在同一区域中，对于符合约束条件的位置集合，位置越分散，匿名效果越好。当查询半径确定时，随着单位观测距离的增大，匿名成功率会降低。在d_0在 3 km 以内时，变化趋势不明显；在d_0大于 3km 时，匿名成功率急剧降低。空间离散度分析处理时，将位置点周围单位观察距离内的位置点全部丢弃，使得当单位观察距离大于 3km 时难以找到其他位置点，从而使匿名成功率降低。

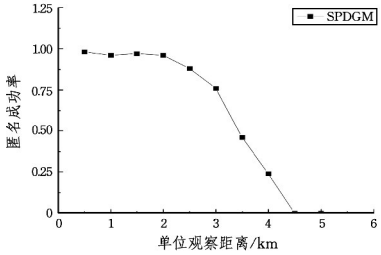

图 8 单位观察距离和匿名成功率关系

Fig. 8 Relationship between unit observation distance and anonymous success rate

图 9 给出了权值α与匿名成功率的关系。经常访问区域和陌生区域的实验结果分别用 α-FV 和 α-SA 表示。从图 9 可以看出：随着权值逐渐增大，匿名成功率逐渐降低；在陌生区域时，位置的历史查询概率为 0，相同条件下 α-SA 的匿名成功率低于α-FV。

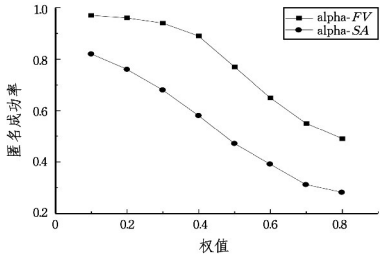

图 9 权值与匿名成功率关系

Fig. 9 Relationship between weight and anonymous success rate

结束语 本文通过分析常见的生活场景，引出了位置语义和用户偏好习惯的联系，提出了 SPDGM 算法来抵抗拥有边信息的攻击者的攻击。本文充分分析了位置语义间的关联性，通过求取具有最小差异度的语义集合，选择出 m 个语义，再通过综合历史访问概率以及大众评价信息来评估用户选择该位置的可能性，然后选择地理分布上分散的 $k-1$ 个假位置来满足用户的隐私保护需求。直观上讲，位置语义之间是存在差异性的，如湘菜馆与小吃店之间的语义差值比湘菜馆和银行之间的语义差值小，若选出的位置之间的语义差异性小，则在一定程度上会泄露用户隐私，如何量化语义所蕴含的信息，进而计算语义差值，将是后续研究的关注点。

参 考 文 献

[1] PAULET R, KAOSAR M G, YI X, et al. Privacy-Preserving and Content-Protecting Location Based Queries[C]// IEEE 28th International Conference on Data Engineering. 2012: 44-53.

[2] ZHAO K, TU Z, XU F L, et al. Walking Without Friends: Publishing Anonymized Trajectory Dataset Without Leaking Social Relationships[J]. IEEE Transactions on Network and Service Management, 2019, 16(3): 1212-1225.

[3] ZHOU B, PEI J. The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks[J]. Knowl. Inf. Syst., 2011, 28(1): 47-77.

[4] ZHANG S B, LI X, TAN Z Y, et al. A caching and spatial K-anonymity driven privacy enhancement scheme in continuous location-based services[J]. Future Generation Computer Systems, 2019, 94: 40-50.

[5] PAN X, XU J L, MENG X. Protecting Location Privacy against Location-Dependent Attack in Mobile Services[J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 24: 1506-1519.

[6] CHOW C Y, MOKBEL M F, LIU X. Spatial cloaking for anonymous location-based services in mobile peer-to-peer environments[J]. Geoinformatica, 2011, 15(2): 351-380.

[7] WU D D, LYU X. Location Anonymous Algorithm Based on User Collaboration under Distributed Structure[J]. Computer Science, 2019, 46(4): 158-163.

[8] JIANG J, FU C Y. Location Privacy Protection Method Based on Query Fragment and User Collaboration[J]. Journal of Chinese Computer Systems, 2019, 40(5): 935-940.

[9] KIDO H, YANAGISAWA Y, SATOH T. An anonymous communication technique using dummies for location-based services [C]// ICPS. IEEE, 2005: 88-97.

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

[11] SONG D, SONG M, SHAKHOV V, et al. Efficient dummy generation for considering obstacles and protecting user location [J]. Concurrency and Computation: Practice and Experience, 2019, 33(2): e5146.

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

[13] SUN G, CHANG V, RAMACHANDRAN M, et al. Efficient location privacy algorithm for Internet of Things(IoT) services and applications[J]. Journal of Network and Computer Applications, 2017, 89: 3-13.

[14] DU Y W, CAI G, ZHANG X J, et al. An Efficient Dummy-Based Location Privacy-Preserving Scheme for Internet of Things Services[J]. Information, 2019, 10(9): 278.

[15] SUN Y M, CHEN M, HU L, et al. ASA: Against statistical attacks for privacy-aware users in Location Based Service[J]. Future Generation Computer Systems, 2017, 70: 48-58.

[16] ZHU L, XU C Q, GUAN J F, et al. SEM-PPA: A semantical pattern and preference-aware service mining method for personalized point of interest recommendation[J]. Journal of Network and Computer Applications, 2017, 82: 35-46.

[17] TAO L Q, CAO J L, LIU F. Dynamic feature weighting based on user preference sensitivity for recommender systems [J]. Knowledge Based Systems, 2018, 149: 61-75.

[18] ZHU L，XU C Q，GUAN J F，et al. A preference aware trajectory privacy-preserving scheme in location-based social networks [C]// IEEE INFOCOM. IEEE Conference on Computer Communications Workshops. 2017.

[19] NI W W，GU M Z，CHEN X. Location privacy-preserving k nearest neighbor query under user's preference[J]. Knowledge-Based Systems，2016，103：19-27.

[20] WANG J，WANG C R，MA J F，et al. Dummy location selection algorithm based on location semantics and query probability[J]. Journal on Communications，2020(3)：53-61.

[21] ZHANG Y B，ZHANG Q Y，LI Z Y，et al. A k-anonymous location privacy protection method of dummy based on approximate matching[J]. Control and Decision，2020，35(1)：65-73.

[22] TU Z，ZHAO K，XU F L，et al. Protecting Trajectory from Semantic Attack Considering k-Anonymity，l-diversity and t-closeness[J]. IEEE Transactions on Network and Service Management，2019，16(1)：264-278.

[23] HARA T. Dummy-based location anonymization for controlling observable user preferences[C]// IEEE Global Communications Conference. 2019.

[24] ZHENG Y，XIE X，MA W Y. GeoLife：A Collaborative Social Networking Service among User，location and trajectory[J]. Invited paper，in IEEE Data Engineering Bulletin，2010，33：32-40.

[25] FENG Z H，QIAN X Z，ZHAO N N. Greedy DBSCAN：an improved DBSCAN algorithm on multi-density clustering[J]. Application Research of Computers，2016，33(9)：2693-2696，2700.

WANG Hui，born in 1975，Ph.D，professor，Ph. D supervisor，is a member of China Computer Federation. His main research interests include mobile Internet privacy protection，network information security，information system development and simulation and so on.

SHEN Zi-hao，born in 1980，Ph. D，lecturer，is a member of China Computer Federation. His main research interests include network and information security，information simulation，intelligent information processing and so on.