基于查询概率的假位置选择算法

吴忠忠 吕 鑫 李 鑫
(河海大学计算机与信息学院 南京 211100)

摘 要 位置服务(Location-based Service,LBS)已经成为日常生活的重要组成部分。用户在享受位置服务带来的巨大便利的同时,也面临着巨大的隐私泄露风险。针对传统的位置隐私保护中K-匿名机制没有考虑到攻击者具有背景知识或者边信息的问题,提出了一种改进的假位置选择算法来保护位置隐私。该方法首先对样本空间进行网格划分,并基于历史查询数据计算出每个位置单元的查询概率;再结合历史查询概率为用户寻找($K-1$)个假位置,使得这($K-1$)个假位置的历史查询概率与用户所在位置的历史查询概率尽量相同,并且使这K个位置尽量分散。实验结果证明了该算法在位置隐私保护方面的有效性。

关键词 位置服务,K匿名,查询概率,假位置,边信息
中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.05.024

Query Probability Based Dummy Location Selection Algorithm

WU Zhong-zhong LV Xin LI Xin
(College of Computer and Information,Hohai University,Nanjing 211100,China)

Abstract Location-based service(LBS) has become a vital part in daily life. While enjoying the convenience provided by LBS,users may have high risk of losing privacy. Traditional K-anonymity for location privacy does not consider that adversaries have some problems of background knowledge and side information. Therefore,this paper proposed an improved dummy locations selection algorithm to protect location privacy. Firstly,the space is divided into grids,and historical statistics are utilized to obtain the probability of queries of each grid of space. Then,the dummy locations are selected for users in order that the ($K-1$) grids has the same query probability as far as possible corresponding with the grid where the real user exists and the K locations are spread as far as possible. A series of experiments on synthetic datasets show that this algorithm is feasible and effective.

Keywords Location-based service,K-anonymity,Query probability,Dummy location,Side information

1 引言

随着定位技术和通信网络的迅猛发展,智能设备已经成为人们日常生活中必不可少的部分。目前,智能设备都基本配备了GPS(Global Positioning System)功能。在日常生活中,人们经常使用智能手机的各项服务,如地图查询、运动计步、个性推荐等,这些服务均需要用户实时提交位置信息。不可否认,这些服务使人们的生活变得更为便利,但另一方面,这些服务也带来了巨大的隐私泄露风险,因为用户提交的位置信息容易遭到泄露。攻击者针对位置数据的推理攻击不仅可以预测用户何时在何处,而且可以从这些位置数据中分析出一些目标用户的家庭地址、工作地址等敏感信息[1]。因此,位置隐私保护在大数据时代显得尤为重要,也受到了用户和研究者的广泛关注。目前,基于位置服务造成位置隐私泄露的场景主要有以下几种[2-4]:1)现有的很多隐私保护方法基于中心服务器架构,而中心服务器本身存在被攻击的风险,因此隐私泄露;2)攻击者获取用户所提交的位置信息、查询内容等,并基于此进行攻击,从而导致位置隐私的泄露;3)虽然目前大部分位置数据在发布前已被相关隐私保护算法处理过,但攻击者仍可通过一些数据挖掘的方法,如关联规则挖掘,来找到这些位置数据之间的内在联系,进而导致用户隐私的泄露。

位置服务中的隐私保护可以分为两类:位置隐私和查询隐私[5]。隐私保护主要解决以下几个问题[6]:1)保护轨迹上的敏感/频繁访问位置信息不被泄露;2)保护个体和轨迹之间的关联关系不被泄露,即保证个体无法与某条轨迹相匹配;3)防止移动对象的相关参数限制(最大速度、路网等)泄露移动对象的轨迹隐私问题。

为了解决位置隐私保护问题,学者们提出了很多方法[7-8],K-匿名技术就是其中之一。K-匿名技术最早由美国Carnegie Mellon大学的Latanya Sweeney提出,它指一条数据表示的个人信息和其他($K-1$)条数据不能区分,最初被使

到稿日期:2017-05-26 返修日期:2017-08-13 本文受国家重点研发计划(2016YFC0400910),国家重大专项(2017ZX07104-001)资助。
吴忠忠(1993-),男,硕士生,主要研究方向为信息安全;吕 鑫(1983-),男,博士,讲师,主要研究方向为密码学、网络信息安全,E-mail:lvxin.gs@163.com(通信作者);李 鑫(1992-),男,硕士生,主要研究方向为信息安全、密码学。

用在关系数据库的数据发布隐私保护中。最早将 K-匿名技术应用到位置隐私保护中的是 Marco Gruteser，他提出了位置 K-匿名：当一个用户的位置和其他($K-1$)个用户的位置不能相互区别时，则称该用户的位置满足 K-匿名。学者们在此基础上对 K-匿名技术做了深入的研究，并且提出了很多方案，如 HC(Hilbert Cloak)，NNC(Nearest Neighborhood Cloaking)，New Casper 等。这些技术的提出对运用 K-匿名技术来保护隐私起到了积极的作用，但是它们都存在着如何选择匿名区域的问题。传统的 K-匿名方法[9-11]在选择匿名区域时都假设攻击者没有一些边信息，如用户的历史位置数据、用户的背景知识等[12-13]，在生成匿名区域时主要采取随机生成策略[10]、矩形或圆形匿名区域生成策略[14]等。当攻击者掌握了一定的边信息时，这些方法的效果则较差。例如，根据一些匿名区域策略生成的匿名区域内包含湖泊、河流等，攻击者如果掌握了一定的地图知识，就会很容易地把这些匿名区域给过滤掉，因此 K-匿名的有效性很难得到保证；又如，使用添加假位置的传统方法往往随机选择假位置，但是每个位置的查询概率都可能不一样，如果选择的假位置的历史查询概率比较小(如湖泊、森林等)，则攻击者很容易结合地图知识把这些假位置过滤掉，这给用户的隐私安全造成了巨大威胁。

例如，张某在家发起了一次匿名度为 3 的查询服务，那么需要找到包含张某家在内的 3 个位置提交给 LBS 服务器，这样攻击者识别出张某家的真实位置的概率就为 1/3。传统的假位置隐私保护算法由于没有考虑到边信息，因此可能选出几乎无人发送查询服务的两个假位置，如在某个沼泽地或者某条河内，把选出的这两个假位置连同张某家的位置一起发送给 LBS 服务器进行查询。在这种情况下，攻击者很容易结合地图学的知识将除张某家所在位置之外的两个假位置过滤掉，因为在这两个假位置发出查询的可能性几乎为零，张某的位置也就随之被暴露。

对上述场景进行抽象分析可以得出，在选择假位置时除要考虑到所选假位置的查询概率与真实用户所在位置的查询概率相同或相近外，还要考虑到位置之间的分散性。如上述场景中，应选择与张某家所在位置查询概率相同或相近的位置作为候选假位置；并且基于位置之间的分散性选取李某家、张某家附近的小商店作为假位置，即可有效防止攻击者通过结合边信息将一些假位置过滤掉，从而使张某的位置隐私得到有效保护。针对上述场景中攻击者通过结合查询概率易将假位置过滤掉和位置之间的分散性这两个问题，本文对传统的假位置隐私保护算法进行改进。改进算法由于充分考虑了边信息和位置之间的分散性，因此能更好地适用于上述位置隐私保护的场景。

本文的主要工作是提出一种基于历史查询概率的位置隐私保护算法 IDLAS。该算法从两个方面来改善用户位置隐私保护的效果。首先，基于用户所在位置的历史查询概率来寻找另外($K-1$)个与真实用户位置的历史查询概率相同或者尽量相同的位置单元作为假位置，并将这 K 个位置一起发送给 LBS 服务器；其次，在寻找($K-1$)个位置的过程中还要使这 K 个位置单元尽量分散。本文算法通过改进这两个方面，使得攻击者很难结合边信息将一些假位置过滤掉，有效弥补了传统的利用假位置的位置隐私保护算法中攻击者在具有相应边信息的情况下可以容易地排除一部分假位置的缺陷；同时，本文算法所提出的位置最分散思想使得攻击者将真实用户锁定在一个很小的区域中的可能性大大降低，进一步改善了位置隐私保护的效果。从位置熵、所选位置的分散度和平均匿名时间 3 个方面将所提 IDLAS 算法与已有的算法进行对比，实验结果验证了 IDLAS 算法的有效性。

2 预备知识

2.1 样本空间位置单元划分

如图 1 所示，将样本空间按照网格进行划分，每个网格即一个位置单元，则可以得到每个位置单元的中心坐标。算法开始前，用一个数据训练集来训练每个位置单元的查询概率。根据公式 $p_i=\frac{\text{位置单元 } i \text{ 中的历史查询次数}}{\text{整个样本空间的历史查询次数}}$($i=1,2,\cdots,n^2$)，定义每个位置单元的历史查询概率。样本空间中所有位置单元的历史查询概率之和为 1，即 $\sum_{i=1}^{n^2}=1$。

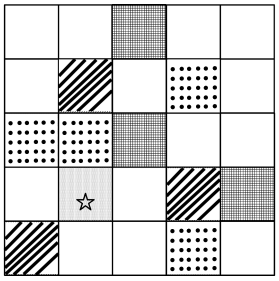

图 1 样本空间划分图

Fig. 1 Division of sample space

2.2 系统框架

本文采取“用户-位置匿名服务器-位置访问服务器”的结构[15]，如图 2 所示。由于本文主要研究提高匿名精度的问题，因此假设攻击者能够成功攻击位置服务器，并假设攻击者所掌握的边信息包括用户的实时查询数据、历史数据、匿名机制等，攻击者可以基于此推断出用户的一些敏感信息。而传统的研究大多未考虑到攻击者结合用户的边信息进行攻击的情况[16-17]。

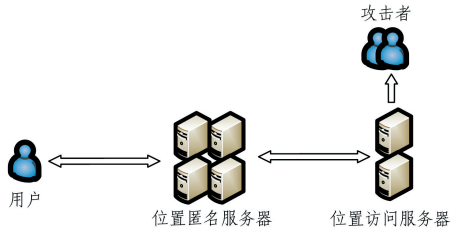

图 2 系统框架图

Fig. 2 System framework

2.3 位置单元间距离的计算

位置单元间的距离采用欧氏距离，即将样本空间均分之后，可得到每个正方形位置单元的中心坐标，根据欧氏距离公式 $d=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$ 来计算两个位置单元之间的距离，其中 x，y 分别表示位置单元的经纬度。

2.4　假位置分散思想

图 3 给出了选择假位置的两种情况。假设 $K=3$，图3(a)选择的假位置单元是两个距离用户真实位置单元较近且与用户真实位置单元的历史查询概率相同的位置单元；图 3(b)选择的假位置单元是两个距离用户真实位置单元较远且与用户真实位置单元的历史查询概率相同的位置单元。虽然图 3(a)中的假位置数量满足了匿名度要求，但是由于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，用户隐私很容易被泄露；而图 3(b)中添加的假位置单元更加分散，用户被成功攻击的可能性就会大大降低。

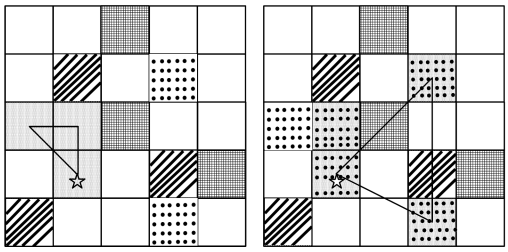

图 3　匿名区域分散思想

Fig. 3　Dummies with bigger cloaking region

2.5　利用海伦公式寻找最分散位置

如图 4 所示，四边形 $ABCD$ 的 4 个顶点 A,B,C,D 对应 4 个位置单元的中心(设 A 为真实用户所在位置单元的中心)，那么由这 4 个位置单元的中心连接起来的多边形区域的面积可以通过下述方法计算：由 n 边形总能划分出$(n-2)$个三角形的定理，将这个四边形划分成 2 个三角形，每个三角形的面积记为 S_1 和 S_2，求出 AB,AC,AD,BC,BD 的距离，再利用海伦公式分别计算这 2 个三角形的面积，所得面积之和即为该区域的面积。

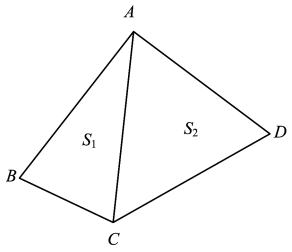

图 4　四边形 $ABCD$

Fig. 4　Quadrilateral $ABCD$

海伦公式为：$s=\sqrt{p(p-a)(p-b)(p-c)}$，其中 a,b,c 分别为三角形的边长，p 为三角形的半周长，s 为三角形的面积。

计算多边形区域的面积时也是先对其进行分割，将其划分成三角形之后求出每个三角形的面积，再通过对所有三角形面积求和得到整个多边形区域的面积。

2.6　用户请求

用户发送一次位置服务请求的格式为$\{ID,(x,y),K\}$，其中，每个用户的 ID 都是唯一的，在同一时刻不允许两个相同的 ID 发送位置服务请求；x 表示用户所在位置的经度；y 表示用户所在位置的纬度；K 表示用户设置的隐私保护等级，针对不同用户设置的隐私保护等级可能不同。

3　算法设计

划分完样本空间后，根据查询概率可以得出：有很多位置单元的查询概率与用户所在位置单元一样(一般来说与用户查询概率一样的位置单元数会多于 K，因为样本空间一般都会很大且样本空间的划分粒度较小)，也有很多位置单元的查询概率与用户所在位置单元的不一样。接着，需要选取$(K-1)$个假位置。本文生成一个位置单元数为 $2K$ 的候选集 H 来进行选择，目的是使得所选出的假位置尽量分散，避免因集中在很小的区域而容易被攻击者攻击。

算法 1　改进的假位置匿名算法(Improved Dummy Location Anonymous Selection Algorithm，IDLAS)

Step1　划分样本空间，确定真实用户所在位置单元的位置 C_{real}。

Step2　确定每个位置单元的坐标，并根据历史查询数据统计每个位置单元的历史查询概率(以下简称概率)。

Step3　用队列 G 来存放所有位置单元。在 G 中将概率与用户所在位置单元概率相同的放在用户左侧，将概率不同的以概率降序的方式放在用户右侧。

Step4　在 G 中，从用户左侧随机选取 K 个位置单元，并从用户右侧顺序选取 K 个位置单元，将这 2K 个位置单元放入队列 H 中。

Step5　从队列 H 中选择一个距用户最远的位置单元作为第 1 个假位置 C_1。

Step6　计算 H 中除 C_1 外的其余位置单元到 C_1 的距离，利用海伦公式计算并选取能与 C_{real} 和 C_1 围成最大面积的位置单元作为第 2 个假位置 C_2。

Step7　计算 H 中除 C_1 和 C_2 外的剩余位置单元到 C_2 的距离，然后按照计算面积的方法选出与 C_{real}，C_1，C_2 围成最大面积的位置单元作为第 3 个假位置 C_3。

Step8　重复上述步骤，直到找到(K－1)个假位置。

Step9　把选出的(K－1)个假位置和用户的真实位置一起发送给 LBS 服务器。

算法的流程图如图 5 所示。

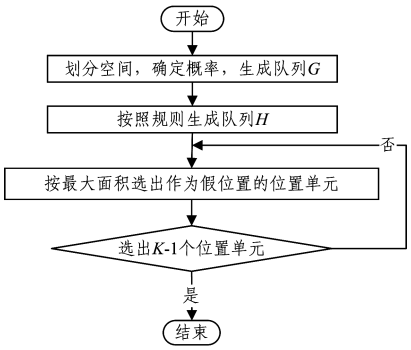

图 5　所提算法的流程图

Fig. 5　Flow chart of proposed algorithm

4　实验结果与分析

本文实验采取的数据集是 OLDEN，它由 Brinkhoff 基于网络的移动对象数据生成器生成。该数据集也是位置隐私实验中最常用的数据集，以城市 Oldenburg 的交通路网作为输入，生成随机分布的移动对象。

本算法采用 Java 语言实现，运行在 Window 7 操作系统上，硬件配置为：Intel Core i7-6700HQ 四核处理器，8GB 内存。

实验方案：从位置熵、所选假位置的分散度和平均匿名时间这 3 个方面来验证所提算法的有效性。将 enhanced-DLS[18] 作为对比算法，该算法和本文所提算法都是针对单个用户在快照查询过程中产生的位置隐私进行保护，且本文算法是对 enhanced-DLS 算法在匿名精度上的改进。

将所选取的样本空间均匀地划分为 1000000 个单元格，共 2565797 个样本轨迹点作为历史数据。实验过程中选取 620494 个数据点作为不同的用户，每个采样点模拟一次向 LBS 系统发起查询。为了便于实验，设置初始匿名度 K 为 2，然后以 2 为匿名度增量进行实验，实验分为 5 组，分别设置匿名度 K 为 2,4,6,8,10。

4.1 位置熵

本实验采用位置熵作为指标来度量算法的隐私保护效果，位置熵也被广泛应用于隐私保护中。位置熵主要用来衡量真实位置的不确定性，熵值越大表明匿名化程度越高，反之表示匿名化程度越低。位置熵的计算如下：

$$H=-\sum_{i=1}^{k} q_i \log_2 q_i$$

其中，q_i 表示得到用户位置服务请求之后，将每个位置单元的历史查询概率归一化，即 $q_i=\frac{p_i}{\sum_{i=1}^{k} p_i}$。

两种算法在不同 k 值下的位置熵如图 6 所示。

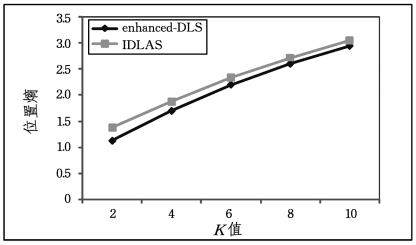

图 6 位置熵

Fig. 6 Location entropy

从图 6 中可以看出，在相同的 K 值下，IDLAS 算法计算得到的位置熵大于 enhanced-DLS 算法计算得到的位置熵；随着 K 值的增大，IDLAS 算法的位置熵也逐渐增大，用户真实位置的不确定性也会随之增加，从而使用户的隐私得到有效保护。

4.2 位置分散度

本文在选取假位置时考虑到了所选出的各个假位置（包括假位置和真实位置）之间最分散的问题。所选假位置和真实位置所围面积越大，所选出的假位置（包括假位置和真实位置）就越分散，这样攻击者就不能把用户位置缩小在一个很小的范围内，从而提高用户的隐私保护效果。定义参数 α 来表示位置分散度，$\alpha=\frac{1}{s}$，s 表示各个假位置和真实位置所围成区域的面积。α 越大，表示各个位置之间越紧密，反之则表示各个位置之间越分散。本文使用两种算法的位置分散度比 β 来衡量所选位置的分散情况。$\beta=\frac{\alpha_{\text{enhanced-DLS}}}{\alpha_{\text{IDLAS}}}=\frac{S_{\text{IDLAS}}}{S_{\text{enhanced-DLS}}}$，$\beta$ 值越大，表示本文算法所选位置比 enhanced-DLS 算法所选位置更分散。

位置分散度的比值结果如图 7 所示。

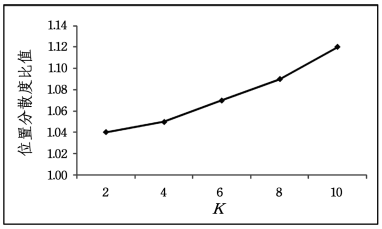

图 7 位置分散度比值

Fig. 7 Ratio of region dispersion

从图 7 中可以看出，当 K 值较小时，本文所提算法生成的区域面积相比对比算法略有提高；随着 K 值逐渐增大，本文算法生成的区域面积与 enhanced-DLS 算法的比值在逐渐增大，即所选的假位置（包括假位置和真实位置）之间更加分散；当 $K=10$ 时，本文所提算法相比 enhanced-DLS 算法的位置分散度提高了 14%左右，说明本文算法能够有效地使得所选的假位置更加分散。

4.3 平均匿名时间

平均匿名时间主要是指用户发送一个位置请求服务之后，算法寻找（$K-1$）个假位置所消耗的时间。本实验对比了 IDLAS 和 enhanced-DLS 算法的平均匿名时间，如图 8 所示。

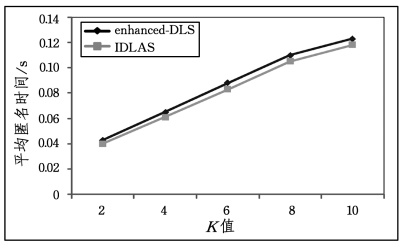

图 8 平均匿名时间

Fig. 8 Average anonymous time

从图 8 看出，IDLAS 算法和 enhanced-DLS 算法在不同 K 值下的平均匿名时间都相差不大，原因在于这两个算法都需要先通过计算位置中心之间的距离来寻找分散的假位置，enhanced-DLS 算法是基于距离的概率来寻找假位置，而本文所提算法是通过直接比较面积来寻找假位置，寻找出的假位置更加分散。两种算法的时间复杂度是相同的，且随着 K 值逐渐增大，平均匿名时间也都会增加，因为要寻找的假位置增多，产生的时间开销也会增大。根据上文位置分散度实验得出，在与对比算法的平均匿名时间相近的情况下，本文所提算法有效地增加了假位置的分散程度，改善了隐私保护的效果。

首先，传统的利用假位置的隐私保护算法没有考虑到边信息这一因素，而 IDLAS 算法通过考虑边信息，使得攻击者很难结合边信息过滤一些假位置；其次，IDLAS 算法提出了一种新的位置最分散的计算方法，并给予了理论说明。实验结果表明，IDLAS 算法能够有效提高位置熵的值，即 IDLAS 算法能够有效地应用于位置隐私保护场景中。

结束语 随着物联网技术的发展和移动终端设备的普及，由位置服务产生的隐私保护问题逐渐受到人们的重视。在选择假位置时不仅要考虑匿名的强度，还要考虑到攻击者是否具备相应的背景知识或者边信息，本文所提算法对此做

（下转第 162 页）