



文章编号:1001-9081(2018)01-0182-06

DOI:10.11772/j.issn.1001-9081.2017071676

基于语义位置保护的轨迹隐私保护的 k -CS 算法

霍 峥^{1*}, 崔洪雷², 贺 萍¹

(1. 河北经贸大学 信息技术学院, 石家庄 050061; 2. 浙江大学宁波理工学院 商学院, 浙江 宁波 315100)

(* 通信作者电子邮箱 huozheng@heuet.edu.cn)

摘要:针对轨迹数据隐私保护算法数据可用性低及易受语义位置攻击和最大运行速度攻击等问题,提出了一种在路网环境中基于语义轨迹的隐私保护算法—— k -CS 算法。首先,提出了两种路网环境中针对轨迹数据的攻击模型;然后,将路网环境中基于语义轨迹的隐私问题定义为 k -CS 匿名问题,并证明了该问题是 NP 难问题;最后,提出了一种基于图上顶点聚类的近似算法将图上的顶点进行匿名,将语义位置由相应的匿名区域取代。实验对所提算法和轨迹隐私保护经典算法(k, δ)-anonymity 进行了对比,实验结果表明: k -CS 算法在数据可用性、查询误差率、运行时间等方面优于(k, δ)-anonymity 算法;平均信息丢失率比(k, δ)-anonymity 算法降低了 20% 左右;算法运行时间比(k, δ)-anonymity 算法减少近 10%。

关键词:路网;隐私保护;轨迹数据;语义位置;聚类

中图分类号: TP311.13; TP393.08 **文献标志码:**A

***k*-CS algorithm: trajectory data privacy-preserving based on semantic location protection**

HUO Zheng^{1*}, CUI Honglei², HE Ping¹

(1. School of Information Technology, Hebei University of Economics and Business, Shijiazhuang Hebei 050061, China;

2. College of Business, Ningbo Institute of Technology, Zhejiang University, Ningbo Zhejiang 315100, China)

Abstract: Since the data utility would be sharply reduced after privacy-preserving process and several attack models could not be resisted by traditional algorithms, such as, semantic location attacks and maximum moving speed attacks, a trajectory privacy-preserving algorithm based on semantic location preservation under road network constraints, called k -CS (k -Connected Sub-graph) algorithm, was proposed. Firstly, two attack models in road network space were proposed. Secondly, the privacy problem of semantic trajectory was defined as the k -CS anonymity problem, which was then proven NP-hard. Finally, an approximation algorithm was proposed to cluster nodes in the road network to construct anonymity zones, and semantic locations were replaced with the corresponding anonymity zones. Experiments were implemented to compare the proposed algorithm with the classical algorithm, called (k, δ) -anonymity. The experimental results show that, the k -CS algorithm performs better than (k, δ) -anonymity algorithm in data utility, query error and runtime. Specifically, k -CS algorithm reduces about 20% in information loss than (k, δ) -anonymity, and k -CS algorithm deceases about 10% in runtime than (k, δ) -anonymity algorithm.

Key words: road network; privacy-preserving; trajectory data; semantic location; clustering

0 引言

近年来,随着智能移动手机和定位技术的发展,越来越多的位置数据被收集、存储、挖掘和分析。轨迹数据挖掘^[1]和语义轨迹抽取^[2]已经成为数据挖掘领域一个重要的研究方向。许多学者意识到对位置数据的分析和挖掘会导致移动对象的个人隐私泄露。近年来,学者们对轨迹数据的隐私保护技术进行了研究,这些研究主要可分为扰动法^[3]、抑制法^[4]和 k -匿名方法^[5-8]。近年来,出现了以差分隐私技术为基础的轨迹数据发布方法^[9],它能够保证无条件隐私,即对指定的统计信息进行分析,无法得到任何一个个体的信息。然而,差分隐私的主要问题在于其灵活性不足,仅能在有限的统计

信息上进行隐私保护。而传统的 k -匿名技术通过将 k 条轨迹上相应的采样位置匿名在同一区域中达到隐私保护的效果,实现简单,且应用环境灵活。在数据隐私保护技术中,隐私保护度和数据可用性是一对矛盾。隐私保护度越高,必然会造成数据可用性的下降;如果需要较高的数据可用性,必然以牺牲隐私保护度来换取。文献[7]提出了一种利用轨迹数据不确定性进行轨迹聚类的隐私保护算法,将移动对象轨迹上的各个采样位置都进行匿名处理,隐私处理后的数据可用性较低。文献[5]第一次在自由空间中提出:并不是轨迹上每一个采样位置都有必要进行隐私保护。例如:行进时所在的道路等位置信息不是移动对象真正访问的位置,并不会暴露用户的隐私,不必要的匿名会给轨迹数据造成严重的信息丢失,

收稿日期:2017-07-06;修回日期:2017-08-22。 基金项目:国家自然科学基金资助项目(61502279);河北省自然科学基金资助项目(F2015207009);河北省高等学校科学技术研究项目(BJ2016019, QN2016179);宁波市软科学项目(2016A10066)。

作者简介:霍峥(1982—),女,河北邯郸人,讲师,博士,CCF 会员,主要研究方向:隐私保护、移动对象数据库; 崔洪雷(1976—),女,辽宁沈阳人,讲师,博士,主要研究方向:大数据环境下的金融数据应用; 贺萍(1982—),女,山东莱阳人,讲师,博士,主要研究方向:无线传感器网络、图优化算法。



导致数据不可用。例如,在利用轨迹数据进行交通流量信息分析时,道路上的数据如果不会暴露用户隐私而不加以匿名处理,其数据可用性更高,分析结果更加精确。而用户运行轨迹中真正能够暴露隐私的是用户访问过的地图上的语义位置,例如酒吧、医院、宾馆等。也就是说,攻击者更容易通过语义位置获取用户隐私。若仅对语义位置进行隐私保护,则能极大地提高数据的可用性。

笔者注意到,语义轨迹上某些重要的停留位置需要进行隐私保护处理,而其余位置并不需要隐私保护处理。采用上述思路能够减少需匿名的采样位置数量,从而提高轨迹数据的可用性。针对上述问题,本文提出一种路网空间中基于语义轨迹的轨迹隐私保护技术。具体来说,本文的主要贡献如下:

- 1) 本文提出一种基于语义轨迹的轨迹隐私保护方法,在隐私保护过程中,仅仅对轨迹上的语义位置进行匿名处理,对一般的位置不作隐私保护,能够提高数据可用性。
- 2) 本文将路网环境中基于语义位置的轨迹隐私保护问题定义为一个 k -CS(k -Connected Sub-graph)匿名问题,且证明了该问题是一个 NP(Non-deterministic Polynomial) 难问题。
- 3) 提出了一种基于图上顶点聚类的近似算法,得到地图上语义位置的 k -CS 匿名区域,并通过算法将轨迹上的停留位置进行匿名处理,保护停留位置的隐私。
- 4) 在真实数据集上对 k -CS 匿名算法的数据可用性、查询误差率和运行时间进行了实验,实验结果表明本文提出的方法比传统 k -匿名方法的数据可用性高 20% 左右。

1 预备知识

下面介绍本文算法的预备知识。

1.1 系统结构

在轨迹数据隐私保护技术中,使用图 1 所示的系统架构。该架构包括客户端、轨迹数据库、隐私保护服务器三个组件。客户端将自己的位置数据发送给数据采集方,隐私保护服务器负责将收集到的轨迹数据进行语义位置提取、地图匿名区域生成及轨迹匿名处理。匿名后的数据形成可发布轨迹数据,可供其他应用程序进行挖掘或统计。



图 1 轨迹数据隐私保护系统结构

Fig. 1 System architecture of trajectory data privacy-preserving

1.2 相关定义

定义 1 语义轨迹。语义轨迹 T 是一系列带有注释的停留位置和移动位置的序列,通常用如下元组表示:(轨迹标示符,移动对象标示符,注释,轨迹上重要位置,语义停留位置,轨迹片段)。

语义轨迹和原始轨迹数据不同,它将原始轨迹上的采样位置语义化为移动对象访问的位置及访问时间等重要信息。原始轨迹是指从位置采集设备上收集到的位置及采样时间,位置信息是用经纬度表示的。在语义轨迹的信息中,本文的算法主要关注轨迹上重要位置和语义停留位置,本文将其统称

为语义位置。

定义 2 路网。路网 $G = (V, E, W)$ 是一个无向图,其中 V 是所有兴趣位置(Points of Interests, POI)的集合,集合中的元素表示为 v_i ; E 是边的集合,当且仅当两个顶点 v_i 和 v_j 之间有一条不包含任何顶点的路段时,两者之间有一条边 (v_i, v_j) ; W 表示边权的集合,元素 w_{ij} 表示顶点 v_i 到顶点 v_j 的距离,即边 (v_i, v_j) 的长度。

1.3 攻击模式

针对语义轨迹主要有以下两种攻击模式。

第 1 种 语义位置攻击。在路网环境中,并不是轨迹上的任意采样位置都是攻击者感兴趣的,仅有那些用户真正访问过的语义位置才是攻击者重点攻击的对象,例如,如果某个移动对象访问了医院,攻击者可能推导出该用户患了某种疾病,然而,如果用户仅仅是从医院门口经过,并不能得出上述结论,此为语义位置攻击。

第 2 种 最大运行速度攻击。最大运行速度攻击最早是在文献[10]中提出的。在路网环境中,最大运行速度攻击的效果更具威胁性。假设图 2(a) 中灰色区域为移动对象在 t_i 时刻的匿名区域,图 2(a) 中的圆角矩形表示在假如移动对象在 t_i 到下一个时刻 t_{i+1} 之间可能运行到的最大运行区域 (Maximum Moving Boundary, MMB)。道路一般有限速(环路一般不超过 80 km/h,普通道路一般不超过 60 km/h),移动对象的最大运行速度用 v_{\max} 来表示,则 t_i 时刻的最大运行边界可由式 $v_{\max} \times (t_{i+1} - t_i)$ 计算得出,那么攻击者可以推导出在 t_{i+1} 时刻,移动对象肯定处于被 Z_i 的最大运行边界覆盖的区域中,导致 t_{i+1} 时刻的匿名区域中的大部分区域变为不可达的,隐私保护度降低。

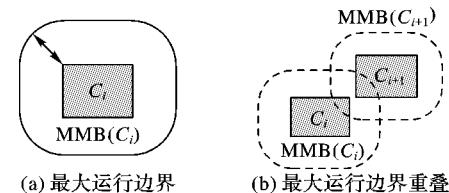


图 2 最大运行速度攻击

Fig. 2 Maximum moving speed attack

2 基于语义轨迹的隐私保护算法

算法主要由 3 个步骤构成:

第 1 步 从原始轨迹数据上抽取出重要位置和停留位置,即原始轨迹语义化;

第 2 步 根据路网数据和语义轨迹数据,将地图上的 POI 进行 k -CS 匿名,即将 k 个地图上的 POI 组成一个匿名区域;

第 3 步 将语义轨迹进行匿名,即,将语义位置由相应的地图匿名区域取代,而经过位置可不作处理。

2.1 语义位置抽取

原始轨迹中的采样位置只是经纬度,并不包含语义信息。作者认为,假如不对原始轨迹中的采样位置加以区分,进行同样的匿名处理,势必会造成数据可用性的严重下降。本节主要讲述如何从原始数据中抽取重要位置或停留位置。所谓停留位置,是指移动对象真正访问过的位置。与移动对象仅经过未访问的位置不同,停留位置包含了更多的语义信息。若



移动对象访问了某个专科医院,可以推断该用户患了何种类型的疾病;如果移动对象的轨迹上有对应于该专科医院的采样位置,但用户并未在此处停留,则无法推导出用户访问过该位置的结论,即前述语义位置攻击。

轨迹上停留位置抽取方法有两种:一种是基于停留时间的抽取方法;一种基于采样位置密度的抽取方法。第1种方法是为了识别轨迹上速度为0的停留,这时需设置一个时间阈值 th_t ,但凡两个连续采样位置的时间间隔大于 th_t ,则认为该采样位置发生了停留。第2种基于采样密度的方法主要是为了识别速度不为0的游览型访问,比如,在公园中游览等,此时,移动对象的速度并不为0,但是游览速度较慢,因此,位置的采样密度较大。

经过语义位置抽取过程,原始轨迹 T_i 可以转换为语义位置和停留时间的序列,即, $T_i = \{(L_1, td_1), (L_2, td_2), \dots, (L_n, td_n)\}$ 。其中: L_i 表示语义位置, td_i 表示在相应位置的停留时间。

2.2 地图匿名区域生成

抽出的轨迹停留位置是地图上的POI是一个经纬度,可由反向地址解析器解释得到确切的地址信息,此处不再赘述。

经由停留位置抽取之后,路网数据 $G = (V, E)$ 中的顶点和边都已生成。其中, V 中包含的顶点除了路网数据中的POI之外,还有从轨迹停留位置中抽取出来的POI,并将抽取出的POI放置在距离其最近的一条路段上,作为一个顶点。

文献[5]采用了语义和自由空间欧氏距离的混合距离对空间中存在的POI进行聚类,生成包含 k 个顶点的匿名区域。而在路网空间中,不能简单地用距离对POI进行聚类,这样会产生某些匿名区域中POI不可达的问题,导致隐私保护度下降,具体如图3所示。

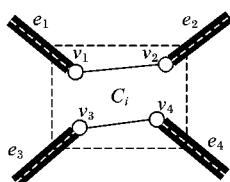


图3 不连通的4-匿名区域

Fig. 3 4-anonymity area of unconnected sub-graph

图3展示了一个4-匿名区域 C_i ,其中包含了 v_1, v_2, v_3, v_4 四个顶点,另有 e_1, e_2, e_3, e_4 四条边与该匿名区域连通。按照匿名模型的定义,一旦达到4-匿名,意味着隐私泄露的概率为 $1/4$,即,攻击者无法识别移动对象处于 v_1, v_2, v_3, v_4 中哪个位置。然而,图3所示的例子有路网信息作为背景知识,4-匿名无法达到应有的隐私保护度。若有轨迹 T_i 从道路 e_1 进入匿名区域中的某个语义位置并停留一段时间,则将该停留位置用匿名区域 C_i 取代进行匿名。然而,攻击者可能发现移动对象只能处在 v_1 或 v_2 位置上,不可能处于 v_3 或 v_4 ,因为路径 e_1 没有路径连通到 v_1 和 v_2 。同理,从路径 e_2, e_3, e_4 进入匿名区域 C_i 的轨迹,均无法达到4-匿名的效果,只能保证隐私泄露的概率不大于 $1/2$ 。

为了避免上述问题的发生,需要重新设计地图匿名算法,避免仅用距离衡量哪些POI应该处于同一个匿名区域中。匿名算法中需要满足下述3个要求:1) 每个匿名区域中包含至少 k 个语义位置;2) 匿名区域中的 k 个匿名位置须构成一个

连通子图;3) 由于数据可用性需求,匿名区域面积越小越好。本文将满足上述3个条件的问题定义为 k -CS匿名问题。该问题的形式化定义如下。

定义3 k -CS匿名问题。给定路网数据 $G = (V, E, W)$, k -CS匿名问题需要找到满足下述条件的匿名区域:

1) 图 $G = (V, E, W)$ 最多可划分为 $m = \lfloor n/k \rfloor$ 个子图(n 为图 G 中顶点个数): $G_1 = (V_1, E_1, W_1), G_2 = (V_2, E_2, W_2), \dots, G_i = (V_i, E_i, W_i)$;每个子图中至少包含 k 个顶点,每个子图即为一个匿名区域;

2) $V_1 \cap V_2 \cap \dots \cap V_m = \emptyset$;

3) $V_1 \cup V_2 \cup \dots \cup V_m \subseteq V$;

4) 连通子图 V_i 的区域面积最小。

然而,在图中计算不规则多边形的面积并非一个简单工作,本文采用路网距离之和来取代匿名区域面积。所谓路网距离是指:如果顶点 v_i 和顶点 v_j 之间有一条通路,则顶点 v_i 到顶点 v_j 的路网距离就是两点之间的最短距离。如果 v_i 和 v_j 不连通,则其路径长度为 $+\infty$ 。

文献[11]中,已经证明了图上的 k -way划分是NP-hard问题,下面证明本文提出的 k -CS匿名问题也是NP-hard问题。

定理1 k -CS匿名问题是NP-hard问题。

证明 k -way划分是指:给定含有 n 个顶点的图 $G = (V, E, W)$,将 G 划分为若干个顶点之间没有交集的非空子图 V_1, V_2, \dots, V_k ,使得 $V_1 \cup V_2 \cup \dots \cup V_k = G$,且 $Ecost = \sum_{i=1}^m w_i$ 最小。其中, w_i 是连接两个划分到不同子图中的边的权值。下面将 k -CS划分归结为 k -way问题,给定图 G ,其边的权值之和是固定的, $Ecost$ 最小意味着 $ICost$ 最大,所谓 $ICost$ 是指 V_i 的内部权值之和。显然,存在一个常数 N 和一组数字 u_i ,边权可表示

为 $w_i = N - u_i$,将 $Ecost = \sum_{i=1}^m w_i$ 最小化,也就是将 $\sum_{i=1}^m u_i$ 最大化。如果将 u_i 看作图的边权的倒数,则正好是 k -CS匿名问题。

证毕。

本文提出了一种近似算法求解 k -CS匿名问题,算法以聚类算法 k -medoids为基础,区别在于:一般的 k -medoids聚类是在自由空间中进行的,以欧氏距离为基础;而在路网空间中,距离的衡量公式都会发生改变。如图4所示,如果按照欧氏距离的定义进行聚类,会产生图4(a)的聚类效果,显然, v_2, v_3, v_4, v_5 构成的匿名区域是一个不连通的子图,会产生前述的隐私泄露问题。尽管图4(b)中的匿名区域 C_2 面积大于图4(a)中的匿名区域 C_1 ,但是由 v_2, v_4, v_5, v_6 构成的子图是一个连通子图,不会造成前述隐私泄露问题,因此,本文采用路网距离进行聚类。

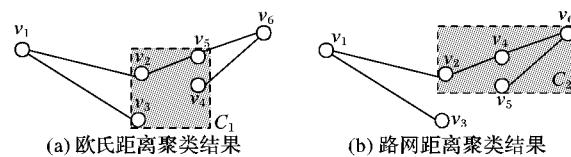


图4 不同距离的聚类结果

Fig. 4 Clustering results based on different distance measure

算法1 k -CS匿名区域生成($G(V, E, W), k$)。

输入:路网数据 G ,隐私保护参数 k 。

输出:匿名区域 C_1, C_2, \dots, C_m 。



```

FOR 图 G 中的每个连通分量  $G_i$ 
    任选  $G_i$  中的一个顶点  $v_i$  作为聚类中心;
    按照距离相远的原则选择  $\lfloor n_i/k \rfloor$  个聚类中心;
    FOR 每个聚类中心  $v_i$ 
        WHILE  $|C_j| < k$  DO
             $C_j \leftarrow$  距离质心路网距离最近的顶点;
             $|C_j|++$ ;
        END WHILE
        FOR 每个簇  $C_j$ 
            计算  $P_{score} \leftarrow$  各个顶点到质心的距离之和;
            按顺序每个顶点  $v_m$  作为聚类中心;
            重新计算  $P_{score}$ ;
            选择  $P_{score}$  最小的那个簇的顶点为聚类中心;
            重复上述步骤,直到簇不变化为止。
        END FOR
    END FOR
    RETURN  $C_1, C_2, \dots, C_p$ 

```

算法1展示了生成k-CS匿名区域的过程。为了减小计算代价,算法将连通子图匿名区域面积最小用路网距离之和最小取代。算法对图G中的每个连通分量作处理,每个连通分量各自选取 $\lfloor n_i/k \rfloor$ 个聚类中心,其中 n_i 表示第*i*个连通分量中顶点的个数。当每个簇中的顶点个数小于k时,将距离簇质心最近的顶点加入到簇 C_j 中,然后重新计算簇的质心,并使簇中的顶点数增加1,质心的选择是聚类效果的关键。当簇中只有一个顶点时,该顶点就是该簇的质心;当簇中顶点数多于2个时,选择距离簇中其他各个顶点距离之和最小的作为簇的质心。然后再加入其他的顶点到当前簇中,直到簇中包含k个顶点为止。对于每个簇 C_j ,按照顺序依次选择一个顶点作为质心,计算该簇的得分 P_{score} ,即,每个顶点到簇中心的路网距离之和。哪个顶点作为质心得到的 P_{score} 最小,就用哪个顶点作质心,重新调整簇。重复这些步骤,直到簇不再变化为止。输出由顶点构成的簇 C_1, C_2, \dots, C_p ,每个簇为一个匿名区域。下面证明用算法1得到的匿名区域中,每个子图都是一个包含的k个顶点的连通子图。

定理2 用算法1得到的每个匿名区域都是包含k个顶点的连通子图。

证明 算法1将距离质心路网距离最小的顶点加入到簇中。假定簇 C_j 中质心为 v_i ,那么距离 v_i 路网距离最近的顶点 v_j 一定被加入到 C_j 中。根据Dijkstra算法,距离 v_i 的路网距离最近的顶点分为两种情况:

第1种 v_j 有一条直接路径到达 v_i 路网距离最近;

第2种 v_j 经过另外一个中间节点 v_m 到达 v_i 的距离最近。

第1种情况 v_j 和 v_i 连通的,因为两个顶点之间有一条直接路径。第2种情况下, v_m 到 v_i 的距离比 v_j 到 v_i 的距离更近,因此, v_m 必定先于 v_j 加入到以 v_i 为质心的簇中, v_j 通过 v_m 和 v_i 连通。依此类推, C_j 中有k个顶点时,子图也是连通的。证毕。

2.3 轨迹匿名处理

得到匿名区域 C_1, C_2, \dots, C_p 之后,将采集到的轨迹数据进行匿名处理。匿名处理的过程就是将收集到的轨迹数据中的采样位置分为停留位置和经过位置。如前所述,停留位置携带的敏感信息更多,能够真正反映移动对象的隐私信息,因此,需要对轨迹上的停留位置进行匿名处理。此外,轨迹匿名处理后的数据还应抵御最大运行速度攻击。具体如算法2所

示。

算法2 轨迹匿名处理。

输入:匿名区域 C_1, C_2, \dots, C_p ,原始轨迹数据库 D 。

输出:可发布轨迹数据 T_i^* 。

FOR D 中的每条轨迹 T_i

 转换为语义位置序列 $T_i = \{(L_1, td_1), (L_2, td_2), \dots, (L_n, td_n)\}$

 将停留位置 T_i 由地图上的匿名区域 C_i 替换;

 计算 $MMB(C_i)$ 和 $MMB(C_{i+1})$;

 IF $MMB(C_i)$ 不能够完全覆盖 C_{i+1} ,延长 td_i ,使得 $MMB(C_i)$ 完全覆盖 C_{i+1}

 END IF

END FOR

RETURN T_i^*

算法2将原始轨迹数据转换为匿名后的数据。首先,通过语义位置提取方法将轨迹转换为语义位置序列,然后,用语义位置对应的匿名区域将语义位置替换。替换完成后,计算两个相邻时刻区域 C_i 和 C_{i+1} 的最大运行边界 $MMB(C_i)$ 和 $MMB(C_{i+1})$,检查 $MMB(C_i)$ 是否能够完全覆盖 C_{i+1} ;如果能,则这种匿名方式可以抵御最大运行速度攻击;若不能,则需延长 td_i ,即停留时间的长度,以增大 $MMB(C_i)$,使其能够完全覆盖 C_{i+1} 。算法最后返回匿名后的轨迹数据 T_i^* 。

2.4 算法分析

算法1和算法2能够保证语义轨迹上的停留位置被匿名在一个包含有k个POI的匿名区域中,因此,移动对象在停留位置隐私泄露的概率至多为 $1/k$ 。轨迹数据的匿名处理对数据造成了一定的可用性丢失,信息丢失主要是由两方面造成的:一方面是由于停留位置匿名导致的,从一个采样位置泛化为一个面积,导致精度下降;另一方面是由停留时间 td_i 的延长导致的,致使某个时刻移动对象所处位置不精确。

针对第一方面的信息丢失(Information Loss, IL),通常采用文献[12]提出的标准进行衡量,即:将一个点泛化为一个区域后,移动对象被识别概率降低了多少,如式(1)所示:

$$IL = \left[\sum_{i=1}^n \sum_{j=1}^k (1 - 1/\text{area}(C_i)) \right] / (n \times k) \quad (1)$$

其中: n 表示移动对象数据库中轨迹的条数, k 表示一条轨迹上的语义位置数目。只有语义位置才需要被匿名区域取代,因此,停留位置会造成信息的丢失。

此外,空间范围查询的误差率也是衡量信息丢失的重要标准,它能够衡量上述两方面的信息丢失。所谓空间范围计数查询是指:查询某个时间段内某个空间区域中的移动对象数目。在对语义轨迹进行匿名之后,空间范围计数查询必然产生一定的误差,该误差用error表示,可由式(2)计算:

$$\text{error} = \frac{\min(Q(D), Q(D^*))}{\max(Q(D), Q(D^*))} \quad (2)$$

其中, $Q(D)$ 表示在原始轨迹数据上进行空间范围计数查询时得到的值; $Q(D^*)$ 表示在隐私保护处理后的数据上,进行空间范围计数查询时得到的值。本文主要衡量两种空间范围计数查询:一种是PSI(Possibly Sometimes Inside)查询;一种是DAI(Definitely Always Inside)查询。

3 实验分析

实验采用北京市路网数据以及Geolife的真实数据进行。北京市路网数据中包括了17万个路网顶点及43万余条边。



轨迹数据 Geolife 采集了 155 个志愿者在北京市的 8 000 多条轨迹,该数据集中大约包含 230 万条采样位置信息,采样位置主要分布在北京五环区域内。数据分布如图 5 所示。

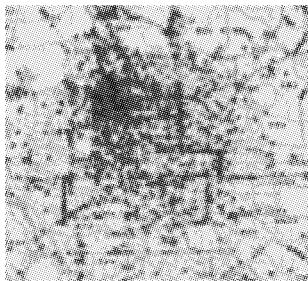


图 5 数据分布

Fig. 5 Data distribution

本实验的实验环境为 Intel i5 3.30 GHz 处理器、8 GB 内存、Windows 7 操作系统。根据前述的停留位置抽取算法,将时间阈值设为 20 min,共抽取近 8 万个停留位置。实验对算法的信息丢失率、范围查询相对误差及算法的运行时间进行了测试。对比算法是轨迹隐私保护的经典算法 (k, δ) -anonymity 算法,它是在自由空间中通过轨迹聚类进行隐私保护的一种算法,其中, δ 是给定的匿名区域半径,在本实验中, δ 的取值与文献[7]中的取值相同,即从 1 000 到 4 000,每次增长 1 000,实验中展示的结果是在不同 δ 值上的平均值。

3.1 信息丢失率

本节主要展示 k -CS 算法和 (k, δ) -anonymity 算法在数据可用性上的对比结果,其中,信息丢失率由式(1)计算得出。实验证了在不同隐私参数 k 的取值下,信息丢失率的变化情况,如图 6 所示。两个算法的信息丢失率随着 k 的增加逐渐增大,当隐私参数 k 取值到 12 时, k -CS 算法的信息丢失率为 40% 左右,而 (k, δ) -anonymity 算法的信息丢失率接近 80%,这是由于 (k, δ) -anonymity 算法将轨迹上的各个采样位置都进行了泛化,造成了较大的信息丢失。而 k -CS 算法只对轨迹上的语义位置进行泛化,因此,信息丢失率较小。

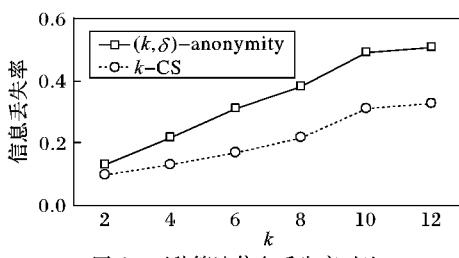


图 6 两种算法信息丢失率对比

Fig. 6 Comparison of information loss rate for two algorithms

3.2 查询误差率

查询误差率实验主要评估在轨迹数据集 D^* 与原始数据集 D 上执行空间范围查询的误差率,该标准是隐私保护算法常用的衡量标准之一。查询误差率的计算由式(2)计算得出。实验结果如图 7 所示。由于 DAI 查询比 PSI 查询的选择性更强,所以在图 7(b)中,两个算法得查询误差率均高于图 7(a)中的算法查询误差率。 (k, δ) -anonymity 算法的查询误差率明显高于 k -CS 算法,而 k -CS 算法的查询误差率在两种查询上均低于 20%,显示出良好的数据可用性。

3.3 运行时间

本实验验证了算法的运行时间。 k -CS 算法的运行时间

仅计算了聚类所需。两个算法的运行时间都随着隐私保护参数 k 的取值增大而减少。可以看出 k -CS 算法的运行时间优于 (k, δ) -anonymity 算法,但是两种算法的运行时间都在百秒级别,并不适合在实时环境中使用。由于本文所提出的算法主要用在离线轨迹数据处理之上,算法运行时间可以满足要求。

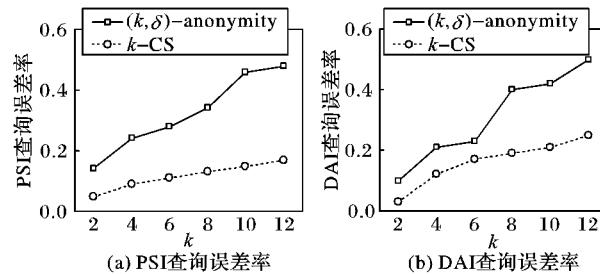


图 7 两种算法查询误差率对比

Fig. 7 Comparison of query error rate for two algorithms

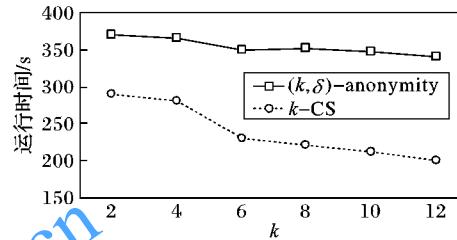


图 8 两种算法的运行时间对比

Fig. 8 Runtime comparison for two algorithms

4 结语

本文提出一种路网环境中基于语义位置的隐私保护方法。该方法的最大贡献在于,仅对轨迹上停留的语义位置进行匿名,信息丢失率较低。本文首先提出路网环境中针对轨迹数据的两种攻击模型,将隐私保护问题定义为 k -CS 匿名问题,并证明了该问题是 NP-hard 问题。随后提出一种基于图中顶点聚类的近似算法,将图中的顶点进行聚类。最后实验证了该算法的数据可用性和算法效率。后续工作包括继续优化图上的聚类算法,提高其精确度及运行效率。

参考文献 (References)

- [1] ZHENG Y. Trajectory data mining: an overview [J]. ACM Transactions on Intelligent Systems and Technology, 2015, 6(3): 1–41.
- [2] PARENT C, SPACCAPICTRA S, RENSO C, et al. Semantic trajectories modeling and analysis [J]. ACM Computing Surveys, 2013, 45(4): 42.
- [3] DAI J, HUA L. A method for the trajectory privacy protection based on the segmented fake trajectory under road networks [C]// Proceedings of the 2nd International Conference on Information Science and Control Engineering. Piscataway, NJ: IEEE, 2015: 13–17.
- [4] 赵婧, 张渊, 李兴华, 等. 基于轨迹频率抑制的轨迹隐私保护方法 [J]. 计算机学报, 2014, 37(10): 2096–2106. (ZHAO J, ZHANG Y, LI X H, et al. A trajectory privacy protection approach via trajectory frequency suppression [J]. Chinese Journal of Computers, 2014, 37(10): 2096–2106.)
- [5] HUO Z, MENG X, HU H, et al. You can walk alone: trajectory privacy-preserving through significant stays protection [C]// Proceedings of the 17th International Conference on Database Systems



- for Advanced Applications. Berlin: Springer, 2012: 351 – 366.
- [6] CAI Z F, YANG H X, SHUANH W, et al. A clustering-based privacy-preserving method for uncertain trajectory data [C] // Proceedings of the 2014 International Conference on Trust, Security and Privacy in Computing and Communications. Piscataway, NJ: IEEE, 2014: 1 – 8.
- [7] ABUL O, BONCHI F, NANNI M. Anonymization of moving objects databases by clustering and perturbation [J]. Information Systems, 2010, 35(8): 884 – 910.
- [8] 霍峰, 孟小峰, 黄毅. PrivateCheckIn: 一种移动社交网络中的轨迹隐私保护方法 [J]. 计算机学报, 2013; 36(4): 716 – 726. (HUO Z, MENG X F, HUANG Y. PrivateCheckIn: trajectory privacy-preserving for check-in services in MSNS [J]. Chinese Journal of Computers, 2013, 36(4): 716 – 726.)
- [9] HUA J, GAO Y, ZHONG S. Differentially private publication of general time-serial trajectory data [C] // Proceedings of the 2015 IEEE Conference on Computer Communications. Piscataway, NJ: IEEE, 2015: 549 – 557.
- [10] PAN X, XU J, MENG X. Protecting location privacy against location-dependent attacks in mobile services [J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(8): 1506 – 1519.
- [11] CHOI T Y. A linear-time heuristic algorithm for k -way network partitioning [J]. Journal of the Korea Safety Management and Science, 2004, 7(8): 1183 – 1194.
- [12] YAROVY R, BONCHI F, LAKSHMANAN L, et al. Anonymizing moving objects: how to hide a MOB in a crowd? [C] // Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology. New York: ACM, 2009: 72 – 83.
- [13] CHEN R, LI H, QIN K A, et al. Private spatial data aggregation in local setting [C] // Proceedings of the 32nd IEEE International Conference on Data Engineering. Piscataway, NJ: IEEE, 2016: 289 – 300.
- [14] SU S, TANG P, CHENG X, et al. Differentially private multi-party high-dimensional data publishing [C] // Proceedings of the 2016 International Conference on Data Engineering. Piscataway, NJ: IEEE, 2016: 205 – 216.
- [15] QIN Z, YANG Y, YU T, et al. Heavy hitter estimation over set-valued data with local differential privacy [C] // Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2016: 192 – 203.
- [16] 孟小峰, 张啸剑. 大数据隐私管理 [J]. 计算机研究与发展, 2016, 52(2): 265 – 281. (MENG X F, ZHANG X J. Big data privacy management [J]. Journal of Computer Research and Development, 2016, 52(2): 265 – 281.)

This work is partially supported by the National Natural Science Foundation of China (61502279), the Natural Science Foundation of Hebei Province (F2015207009), the Scientific Research Projects in Colleges and Universities in Hebei Province (BJ2016019, QN2016179), the Soft Science Project of Ningbo City (2016A10066).

HUO Zheng, born in 1982, Ph. D., lecturer. Her research interests include privacy-preserving, mobile object database.

CUI Honglei, born in 1976, Ph. D., lecturer. Her research interests include economic data applications under big data environment.

HE Ping, born in 1982, Ph. D., lecturer. Her research interests include wireless sensor network, graph optimization algorithm.

(上接第 158 页)

- [14] ZHENG W B, TANG H, QIAN Y T. Collaborative work with linear classifier and extreme learning machine for fast text categorization [J]. Journal of World Wide Web, 2013, 18(2): 1 – 18.
- [15] ZENG H J, HE Q C, CHEN Z. Learning to cluster Web search results [C] // SIGIR '04: Proceedings of the 27th Annual International ACM SIGIR Conference On Research and Development in Information Retrieval. New York: ACM, 2004: 210 – 217.
- [16] GRBOVIC M, DJURIC N, RADOSAVLJEVIC V. QueryCategorizr: a large-scale semi-supervised system for categorization of Web search queries [C] // WWW 2015: Proceedings of the 24th International Conference on World Wide Web Companion. New York: ACM, 2015: 199 – 202.
- [17] TWEEDIE L, SPENCE R, WILLIAMS D, et al. The attribute explorer [C] // CHI '94: Proceedings of the 1994 Conference Companion on Human Factors in Computing Systems. New York: ACM, 1994: 435 – 436.
- [18] WANG C, CAO L B, WANG M C. Coupled nominal similarity in unsupervised learning [C] // CIKM '11: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. New York: ACM, 2011: 973 – 978.
- [19] 王秀红, 鞠时光. 用于文本相似度计算的新核函数 [J]. 通信学报, 2012, 33(12): 43 – 48. (WANG X H, JU S G. Novel kernel function for computing the similarity of text [J]. Journal of Communications, 2012, 33(12): 43 – 48.)
- [20] BORIAH S, CHANDOLA V, KUMAR V. Similarity measures for categorical data: a comparative evaluation [C] // Proceedings of the 2008 SIAM International Conference on Data Mining. Atlanta, Georgia: [s. n.], 2008: 243 – 254.
- [21] GUTTMAN A. R-trees: a dynamic index structure for spatial searching [C] // Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data. New York: ACM, 1984: 47 – 57.
- This work is partially supported by the National Natural Science Foundation of China (61401185), the General Project of Liaoning Province Education Department (LJYL018), the Natural Science Foundation of Liaoning Province (201705 40418).
- BI Chongchun**, born in 1992, M. S. candidate. His research interests include spatial data analysis and query.
- MENG Xiangfu**, born in 1981, Ph. D., associate professor. His research interests include Web database query, spatial data analysis.
- ZHANG Xiaoyan**, born in 1983, Ph. D. candidate, engineer. Her research interests include spatial data query, city calculation.
- TANG Yanhuan**, born in 1992, M. S. candidate. His research interests include spatial data mining, recommender system.
- TANG Xiaoliang**, born in 1980, Ph. D., lecturer. His research interest include machine learning.
- LIANG Haibo**, born in 1995. His research interest include data mining, database query.