一种满足差分隐私的轨迹数据发布方法

霍 峥[1),2)] 孟小峰[2)]

[1)](河北经贸大学信息技术学院 石家庄 050061)

[2)](中国人民大学信息学院 北京 100081)

摘 要 移动对象的轨迹数据包含丰富的时空信息,发布前需进行隐私保护处理以防止个人隐私信息的泄露.目前已有的隐私保护算法多以 k-匿名模型为基础,这类方法提供的隐私保护度不够,且隐私保护度强弱与背景知识高度相关.近年来出现的差分隐私技术是一种与背景知识无关的强隐私保护模型,针对发布数据进行统计查询的误差率可控.然而,针对统计信息的查询仍可能造成移动对象隐私的泄露,针对此问题,该文首先提出了两种攻击模型:稀疏位置攻击和最大运行速度攻击.然后,提出两种满足差分隐私的轨迹数据发布方法:在自由空间中,采用基于噪音四分树的轨迹数据发布方法,分别发布每个时刻的噪音数据,按噪音四分树的层次分割隐私预算,对每个区域中的移动对象计数值添加噪音;路网空间中采用基于噪音 R-树的轨迹数据发布方法,用 R-树索引路网中的路段,按层次分割隐私预算,对路段中的移动对象计数值添加噪音.在空间范围计数查询上,上述两种方法比 k-匿名模型的隐私保护度更高.差分隐私的基础是在原始数据中添加噪音,添加的独立噪音可能导致数据不一致问题.该文提出了一种基于移动对象最大运行速度的一致性处理算法.最后,该文在模拟数据集上对数据可用性和算法运行时间进行了实验,实验结果表明该文提出的算法具有良好的性能.

关键词 差分隐私;轨迹数据;数据发布;四分树;R-树

中图法分类号 TP309 **DOI 号** 10.11897/SP.J.1016.2018.00400

A Trajectory Data Publication Method under Differential Privacy

HUO Zheng[1),2)] MENG Xiao-Feng[2)]

[1)](*Information Technology School, Hebei University of Economics and Business, Shijiazhuang* 050061)

[2)](*Information School, Renmin University of China, Beijing* 100081)

Abstract Trajectory data have rich spatiotemporal information, which may cause users' personal privacy leakage. In order to avoid this, privacy-preserving techniques are required before the publication of trajectory data. Most of the existing trajectory privacy-preserving techniques are based on the k-anonymity model, which has two main drawbacks: one is low privacy guarantee; the other is heavily dependent on the background knowledge that the attackers know. In recent years, researchers proposed differential privacy, which is known as the strongest privacy-preserving model, and it is quite good at publication of statistical information. While, publication of statistical information may also cause leakage of users' location privacy. At first, we propose two attack models in publication of count values of location samples, called sparse location attack and maximum moving speed attack. Then we propose two trajectory data publication methods under differential privacy: in free space, we propose a differentially private trajectory data publication method based on noisy quad-tree. The privacy budget is divided according to the level of the quad-tree; Laplace noise is added into each level's count value of moving object. In road network

收稿日期:2016-06-13;在线出版日期:2017-07-20.本课题得到国家自然科学基金(91646203,61532010)、河北省自然科学基金(F2015207009,D2015207008)、河北省高等学校青年拔尖人才计划项目(BJ2016019,BJ2014021)资助.霍 峥,女,1982 年生,博士,讲师,中国计算机学会(CCF)会员,主要研究方向为隐私保护技术、移动对象数据管理等.E-mail:huozheng@heuet.edu.cn.孟小峰,男,1964 年生,博士,教授,博士生导师,中国计算机学会(CCF)会士,主要研究领域为大数据管理、移动对象数据库管理、隐私保护技术等.

space, we propose a differentially private trajectory data publication method base on noisy R-tree. While R-tree is used to index road segment, privacy budget is divided and Laplace noise is added to the count values of each road segment or the minimum bounding rectangles of road segment. Both method spublish noisy count values of location data of each timestamp, which form a sequence of count values. Differential privacy algorithms have higher privacy guarantee than traditional k-anonymity methods. Actually, published trajectory data is a location count value sequence of multiple consequent timestamps. Since the main idea of differential privacy is to add Laplace noise into original data, and the added Laplace noise are independent at each timestamp, it may cause data inconsistency, which may reduce the utility of the published data. The data inconsistency happens when publication of location data of two or more consequent timestamps. We propose a heuristic algorithm to solve this problem. This algorithm is an extreme value problem under certain constraints. The constraints we conclude are the maximum and minimum numbers of each area or road segment, according to the moving speed, area size or the length of a road segment. The extreme value function is the L2 distance between the noisy count value sequence and the consistency count value sequence, we may find a consistent count value sequence which is most close to the noisy data. At last, we conduct a set of experiments on two data sets, one is generated location data under Gaussian distribution, and the other is generated data under Oldenburg city road network constraints. We measure data utility and runtime of each algorithm before and after consistency procedure, the results show that the consistency algorithm may improve the data utility about 20% in average, which shows the effectiveness of the consistency algorithm. We also test runtime of each algorithm, with the incensement of the data set size, run time increases linearly, which means the algorithms are easy to extend.

Keywords　differential privacy; trajectory data; data publication; quad-tree; R-tree

1　引　言

近年来，个人智能设备及移动应用的快速发展使得大量的位置数据在个人不知情的情况下被第三方收集.收集到的数据被存储到移动对象数据库中，在移动对象数据库上的分析和挖掘可以支持多种与移动相关的应用，如检测道路拥堵路段，以合理规划公共交通车辆；查询人群密度，以提前进行预警和疏散等.然而，轨迹数据中通常包含移动对象的个人敏感信息，比如，敏感运行模式、访问位置信息等等.不恰当地使用轨迹数据易导致用户的隐私泄露.基于上述隐私泄露问题，研究者们在 k-匿名模型上对轨迹隐私进行了深入的研究.然而，笔者观察到：一些应用不需要移动对象在某时刻的具体位置，仅需发布某个区域在某时刻的移动对象个数即可.例如，检测交通拥堵区域，仅需要检查某个区域的移动对象个数是否超过某个拥堵数值即可；再如，某商场有几个候选地址可供开店，需根据客流量确定地址，仅需查询某个区域内移动对象的个数即可.受到该启发，本文主要关注轨迹数据发布中，针对空间范围计数查询(range count query)的隐私保护技术.计数查询也是多种数据分析方法的基础，针对计数查询的隐私保护方法也可扩展到其它数据分析方法中.

现有的轨迹数据隐私保护方法主要基于 k-匿名模型，这类方法存在以下缺点：(1) k-匿名模型的隐私保护效果受数据分布的影响较大，若数据分布过于稀疏，隐私保护后数据可用性会急剧下降；(2)以 k-匿名模型为基础的方法需事先知晓攻击者具备哪些背景知识，再设计相应的隐私保护算法.在背景知识未知的情况下不能保证隐私保护度.近年来出现的差分隐私技术是目前已知的最强的无条件隐私保护技术，即满足差分隐私的算法可在相应的统计查询结果上抵御任意背景知识、任意攻击模型的攻击[1-2].差分隐私对基于统计信息的隐私数据发布尤为擅长，计数值是一种简单的统计信息，因此，差分隐私可适用于针对位置数据的计数查询.

针对自由空间中的轨迹隐私保护问题，本文采

用文献[3]中的基于四分树的发布方法,利用差分隐私技术为四分树的每个结点的计数值添加拉普拉斯噪音;针对路网空间,本文提出一种基于噪音 R-树的发布方法,用 R-树索引路段信息,为每条路段上的移动对象计数值添加拉普拉斯噪音.由于添加的噪音是独立的,易造成数据不一致现象.本文利用移动对象的最大运行速度及路网限制等条件,提出了一种基于启发式规则的数据一致性处理方法,有效提高发布数据的可用性.本文重点关注轨迹数据发布中的隐私保护技术,即对两个相邻时刻(t_i和t_{i+1})的位置数据进行空间范围查询,依然满足差分隐私.即使攻击者知晓移动对象的最大运行速度,也不会导致任意时刻的移动对象位置隐私泄露.文献[3]主要关注静态噪音四分树的构建,并不能抵御本文定义的稀疏位置攻击和最大运行速度攻击.

具体来说,本文的创新工作如下:

(1)定义了两种在空间范围计数查询下,移动对象位置泄露的攻击模型:稀疏位置攻击和最大运行速度攻击;

(2)提出了一种满足差分隐私的移动对象轨迹数据发布方法,保证在任意时刻移动对象位置数据上的空间范围计数查询满足差分隐私;

(3)提出了一种路网空间中满足差分隐私的噪音 R-树构建方法.通过理论证明该发布方法满足差分隐私;

(4)利用移动对象的最大运行速度及路网限制等条件,提出一种基于启发式规则的一致性处理算法以提高发布数据的可用性;

(5)在两个模拟数据集上对数据可用性和算法的可扩展性进行了实验,验证了本文提出的算法具有良好的性能.

2 相关工作

近年来,研究者们对空间数据的差分隐私算法进行了研究.文献[4]提出了一种满足差分隐私的多维数据划分方法,分别设计了满足差分隐私的格划分和 KD-树划分及希尔伯特 R-树划分方法.文献[3]针对若干种空间索引结构设计了满足差分隐私的算法.作者将空间索引结构分为两大类:数据独立的空间索引结构和数据依赖的空间索引结构.四分树就是一种数据独立的索引结构,其结构及划分规则是预先定义的.唯一可能导致隐私泄露的数据是结点上存储的计数值,因此,在结点计数值上添加拉普拉斯噪音即可.对于数据依赖的索引结构,结构的划分与数据相关.例如,KD-树需要递归地按中值进行区域划分,直接发布树结构除了可能泄露结点上的计数值外,还可能造成中值的泄露.文献[5]提出了一种满足差分隐私的轨迹数据发布方法,该方法发布的数据面向频繁访问模式查询,采用前缀树方式存储轨迹数据,且在前缀树的每个结点上增加了噪音计数值.由于添加的噪音是独立的,可能造成数据不一致现象.文献[5]根据前缀树本身的特点(如父结点的噪音计数值必须大于子结点的噪音计数值等)对发布的噪音数据进行了一致性处理,以提高数据的可用性.文献[6]提出了一种满足差分隐私的轨迹数据处理技术,该方法以空间泛化为基础,利用指数机制,对同一时刻中,距离较近的采样位置进行合并.文献[7]提出了一种以"δ-位置集合"为基础的差分隐私技术,提出新的函数敏感度衡量方法,通过对轨迹上的位置抑制达到隐私保护目的.

文献[8-9]提出了一种路网上实时轨迹隐私保护方法,认为用户若处在路网上较为敏感位置时应进行匿名处理,将路网空间用无向图模拟,从敏感位置出发对无向图做广度优先遍历,直至找到 k 个语义位置以生成匿名框,该方法适用在路网空间中.文献[10]针对用户的连续查询提出了一种轨迹隐私保护的方法,该方法首先采用多跳的方式寻找用户,并进行 k-匿名,然后,生成假的连续查询发送给服务器以便于和真实用户的查询相混淆.上述三个研究工作均以 k-匿名模型为基础,与差分隐私的隐私保护度不同.文献[11]提出了一种基于用户与攻击者之间贝叶斯斯塔尔博格博弈方法进行轨迹隐私保护的方法,该方法可保护轨迹上过去、当前及将来的位置不被泄露.本文的研究工作与上述工作的不同之处在于:首先,本文关注的是在空间范围查询上满足差分隐私的轨迹数据发布,其隐私保护度明显与满足 k-匿名的轨迹隐私保护技术不同;其次,与仅考虑某个时刻的数据发布不同,本文考虑两个连续时刻采样位置数据之间的相关性,依据最大运行速度等参数对噪音数据进行了一致性处理.

3 问题定义

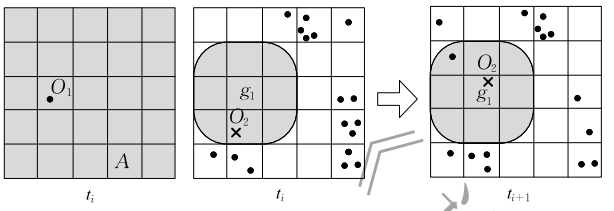
3.1 攻击模型

在定义攻击模型之前,首先介绍移动对象数据库及某个时刻的快照数据.

定义 1(移动对象数据库). n 个移动对象 $M=$

$\{O_1,O_2,\cdots,O_n\}$在 m 个离散时刻 $T=\{t_1,t_2,\cdots,t_m\}$ 的映射函数 $F:M\times T\rightarrow R^2$，即为移动对象数据库.

n 个移动对象在时刻 t_i 的位置数据存储到移动对象数据库中，将 t_i 时刻的快照数据记为 D_{t_i}. 按照时间顺序依次发布 $t_1,t_2,\cdots,t_m$ 时刻的快照数据. 在任一时刻的快照数据上可进行空间范围计数查询. 所谓空间范围计数查询指：给定某个时刻 t_i 的快照数据 D_{t_i} 及空间区域 A，查询结果返回 t_i 时刻区域 A 中移动对象的个数. 多次查询移动对象的空间范围计数值也存在隐私泄露风险，针对此问题，本文定义了稀疏位置攻击及最大运行速度攻击两种攻击模式.



(a) 稀疏位置攻击　　(b) 最大运行速度攻击

图 1　攻击模型

(1) 稀疏位置攻击是指：在移动对象分布稀疏的区域内，多次发起空间范围计数查询，逐渐缩小查询范围，导致移动对象位置泄露的情况，如图 1(a)所示. 假设在 t_i 时刻，移动对象 O_1 处于区域 A 中，区域 A 中没有其它移动对象. 若此时针对区域 A 发起空间范围计数查询，并逐渐缩小查询范围，经过多次查询之后，会造成移动对象 O_1 所处位置泄露.

(2) 最大运行速度攻击是指：在移动对象最大运行速度限制下，针对移动对象的最大运行范围的空间计数查询可能暴露移动对象运行轨迹的情况. 图 1(b)展示了 t_i 和 t_{i+1} 时刻的移动对象分布. 已知移动对象的最大运行速度，灰色的圆角矩形表示格 g_1 中移动对象的最大运行范围，也就是说，只有处于灰色圆角矩形中的移动对象才有可能在 t_{i+1} 时刻运行至格 g_1 中. 在 t_{i+1} 时刻，用“×”表示的移动对象肯定是 O_2. 针对圆角矩形区域进行空间范围查询时，移动对象从 t_i 至 t_{i+1} 的运行轨迹就泄露了.

3.2　差分隐私技术

差分隐私技术是 Dwork 等人在文献[12]中提出的，它是一种与背景知识无关的强隐私保护模型. 简单地说，如果隐私保护算法的输出对输入的细微变化不敏感，则该算法满足差分隐私，差分隐私如定义 2 所示.

定义 2(ε-差分隐私).　给定两个相邻数据集(两个数据集仅有一条记录不同) D_1 和 D_2 以及满足隐私保护条件的数据集 $\widetilde{D}\in Range(A)$，当 $\Pr[A(D_1)=\widetilde{D}]\leq e^{\varepsilon}\times\Pr[A(D_2)=\widetilde{D}]$ 成立时，隐私保护算法 A 满足 ε-差分隐私. 其中，ε 称为隐私参数，概率 $\Pr[\ ]$ 的值由隐私保护算法 A 的随机性决定. 研究者们提出拉普拉斯机制[12]和指数机制[13]使算法达到差分隐私. 在介绍两种机制之前，首先介绍函数敏感性的定义，它用来衡量当数据集中有一条记录发生变动时，函数 f 的输出结果可能发生的最大变化，如定义 3 所示.

定义 3(函数敏感性).　给定任意函数 $f:D\rightarrow\mathbb{R}^d$，对于任意两个相邻数据集 D_1 和 D_2，其敏感性 Δf 可定义为

$$\Delta f=\max_{D_1,D_2}\|f(D_1)-f(D_2)\|_1.$$

拉普拉斯机制是指对输出结果添加满足拉普拉斯分布的噪音数据，用以扰动原始输出，使得在两个相邻数据集 D_1 和 D_2 上产生的输出结果无法分辨. 噪音数据的概率密度函数为 $p(x|\lambda)=\frac{1}{2\lambda}e^{-|x|/\lambda}$，其中，$\lambda$ 是由函数 f 的敏感性 Δf 和隐私参数 ε 决定的，$\lambda=\frac{\Delta f}{\varepsilon}$. 假如拟发布关于函数 $f:D\rightarrow\mathbb{R}^d$ 的统计数据集，发布的数据集和 $f(D)$ 越接近，隐私保护效果越好. 可通过拉普拉斯机制设计隐私保护算法，如定理 1 所示.

定理 1.　给定函数 $f:D\rightarrow\mathbb{R}^d$，若隐私保护算法 A 满足 ε-差分隐私，当且仅当下述表达式成立：

$$A=f(D)+Laplace\left(\frac{\Delta f}{\varepsilon}\right).$$

指数机制(Exponential Mechanism)也是一种设计差分隐私保护算法的机制，它主要用于非数值输出的函数或添加数值噪音后无意义的函数. 指数机制的大致思想是：根据可用性函数 q 的分值，从输出域 R 中选择一个输出值 r，选择到该值的概率与分值成指数比. 函数 q 需对单个记录的变动不敏感，即函数敏感性低，其敏感性可表示为 $\Delta q=\max_{\forall r,D_1,D_2}|q(D_1,r)-q(D_2,r)|$. 可通过指数机制设计隐私保护算法，如定理 2 所示.

定理 2.　给定数据集 D 及可用性函数 $q:(D\times R)\rightarrow\mathbb{R}$，隐私保护机制 A 满足 ε-差分隐私，当且仅当下述表达式成立：

$$A(D,q)=\left\{return\ with\ \ probability\propto\exp\left(\frac{\varepsilon q(D,r)}{2\Delta q}\right)\right\}$$

差分隐私的分解性质(Composition Property)[12,14]是一种特殊的性质，满足差分隐私的算法进行多种

指定计算后,其隐私保护度可以根据该性质获取.

定理 3(顺序分解). 假定每个算法 A_i 满足 ε_i-差分隐私,则 A_i 的序列满足 $\left(\sum_i \varepsilon_i\right)$-差分隐私.

定理 4(平行分解). 假定 D_i 是关于原始数据集 D 的若干个不相交的分解,假定每个算法 A_i 在 D_i 上均满足 ε_i-差分隐私,则算法 A_i 的序列满足 ε_i-差分隐私.

上述性质将用到本文的隐私算法设计中.

3.3 数据可用性衡量

首先给出空间范围查询的定义.

定义 4(空间范围查询). 给定空间范围 R 及在时刻 t 的移动对象的快照位置数据集 D_t,在该数据集上的空间范围查询 Q 可表示为

$$Q(D_t)=\{x \mid x\in D_t, x\in range(R)\},$$

其中,$range(R)$表示在空间范围 R 中的移动对象的集合.

本文定义的空间范围查询与位置的时序性无关.本文采用文献[5,15]和[16]中使用的相对误差来衡量发布数据的可用性,相对误差的计算如式(1)所示.

$$Error=\frac{|Q(D_t')-Q(D_t)|}{\max(Q(D_t,s))} \tag{1}$$

式(1)计算了在可发布数据集 D_t' 上执行查询 Q 与在原始数据集 D_t 上执行查询 Q 相比,产生的误差.其中,参数 s 是为了防止查询 Q 的选择性太强而设置的阈值,以避免分母为 0.所谓查询的选择性是指:所有满足查询条件的记录数量占总数的百分比.

4 满足差分隐私的轨迹数据发布

本节介绍满足差分隐私的轨迹数据发布方法及一致性处理算法.

4.1 噪音四分树构建

四分树是自由空间中常用的空间索引方法[17].其每个结点对应一个空间区域,当某个空间区域内移动对象的数目大于上限阈值 l 时,则将空间划分为大小相等的四个子空间,直至每个子空间中的移动对象数目都不大于 l 为止.在四分树中,移动对象的位置记录存储在叶子结点上,中间结点仅存储其覆盖区域及相应的统计数值.

满足差分隐私的移动对象数据库的发布过程如算法 1 所示.输入原始数据集 D_t,隐私预算 ε,首先为生成一棵噪音四分树 QT;然后,采用一致性处理算法对 QT 进行处理,以提高其可用性,最终发布数据集 D_t'.

算法 1. 自由空间中满足差分隐私的数据发布.

输入:某时刻位置数据集 D_t;隐私预算 ε;最大运行速度 $v_{\max}$;子空间移动对象个数上限 l

输出:可发布数据集 D_t'

1. 创建以 QT 为根结点的四分树;
2. $\bar{\varepsilon}=\frac{\varepsilon}{h}$; //按层次平分隐私预算
3. $j\leftarrow 1$;
4. WHILE $j<h$ DO
5. FOR 第 j 层上每个结点 u_i DO
6. $\tilde{c}(u_i)\leftarrow NoisyCount(u_i,\bar{\varepsilon})$;
7. $j++$;
8. $D_t'\leftarrow$一致性处理$(QT, v_{\max})$;
9. END;

算法 1 的处理对象是一个时刻的位置数据集 D_t.首先,根据差分隐私顺序分解的性质,将隐私预算 ε 平均分成若干份,每份的大小为 $\bar{\varepsilon}=\frac{\varepsilon}{h}$,用于对噪音四分树 QT 的一层添加噪音.当四分树 QT 的深度 j 小于 h 时,算法用同样的隐私预算 $\bar{\varepsilon}$ 构造四分树的下一层.四分树的中间结点存储了其对应区域的移动对象计数值.算法 1 使用隐私预算 $\bar{\varepsilon}$ 为四分树结点的计数值添加拉普拉斯噪音,得到每个结点的噪音计数值 $\tilde{c}(u_i)$.稍后介绍一致性处理算法,使得最终生成的可发布数据可用性最大化.

4.2 噪音 R-树构建

一般情况下,移动对象在道路网络空间中运行,在此环境下,四分树对位置数据的划分通常不均匀,易造成较大的查询误差.此外,四分树是内存索引结构,当移动对象的位置频繁更新时,维护树结构的代价较大.本文采用噪音 R-树发布路网空间中的位置数据.R-树是 Guttman 于 1984 年提出一种空间索引结构[18],也是目前应用最广泛的一种空间索引结构.R-树用空间对象的最小边界矩形(Minimum Bounding Rectangle,MBR)近似表达空间对象.本文将路网模拟为无向图 $G(E,V)$,E 和 V 分别表示路段和路段交叉点的集合,利用 R-树对路网信息及位置数据进行索引.每条路段用其 MBR 表示,对应于 R-树中的叶子结点.在 R-树的每个结点上添加一个计数值属性,表示该结点对应 MBR 中移动对象的数量.首先,利用隐私预算 ε 构建一棵噪音 R-树,稍后介绍对噪音数据进行一致性处理.具体过程如算法 2 所示.

算法 2. 路网空间中满足差分隐私的数据发布.

输入：某时刻位置数据集 D_t；隐私预算 ε；最大运行速度 $v_{\max}$；路网数据 $G(V,E)$

输出：可发布数据集 D_t'

1. $\bar{\varepsilon}=\frac{\varepsilon}{h}$；
2. 按位置数据所处路段对 D_t 进行划分；
3. c_{rs_i}←路段 rs_i 上的移动对象个数；
4. 以 RT 为根结点创建 R-树；
5. c_{mbr_i}←mbr_i 中移动对象个数；
6. RT←将 c_{rs_i} 和 mbr_i 加入噪音 R-树中；
7. FOR RT 中的每个叶子结点 DO
8. $\tilde{c}_{rs_i}$←$NoisyCount(\tilde{c}_{rs_i},\bar{\varepsilon})$；
9. FOR RT 中的每个中间结点 DO
10. $\tilde{c}_{mbr_i}$←$NoisyCount(\tilde{c}_{mbr_i},\bar{\varepsilon})$；
11. D_t'←一致性处理$(RT,v_{\max},G)$；
12. RETURN D_t'；

R-树针对相对稳定的路网结构进行索引，不同时刻的 R-树索引，仅有结点上移动对象计数值发生改变，噪音 R-树的维护代价较低. 根据差分隐私顺序分解的性质，算法 2 将隐私预算 ε 分成大小相等的几部分，每部分的大小为 $\bar{\varepsilon}=\frac{\varepsilon}{h}$，其中，$h$ 表示 R-树的深度. 随后，将快照位置数据按照移动对象所处的路段进行分组，计算每条路段上的移动对象计数值. 算法 2 基于路段的 MBR 构建根结点为 RT 的 R-树，计算每个中间结点的 MBR 的移动对象计数值，将路段及 MBR 的移动对象计数值添加到 R-树的结点上. 分别对叶子结点和中间结点的计数值添加噪音，R-树中每层结点的隐私预算为 $\bar{\varepsilon}$. 噪音 R-树每个结点都存储了其对应的 MBR 中的移动对象噪音计数值.

4.3　数据一致性处理

添加到计数值中的拉普拉斯噪音是独立同分布的，易造成数据不一致的现象，而发布的数据不一致会严重影响数据的可用性，造成较大的查询误差.

4.3.1　自由空间中的数据一致性问题

本文根据移动对象的运行速度研究数据一致性问题. 假定已知移动对象的最大运行速度 $v_{\max}$，则可计算每个区域中的移动对象在下一个时刻的最大运行边界(Maximum Moving Boundary，MMB)[19]，即按照半径 $v_{\max}\times(t_{i+1}-t_i)$ 拓展该区域. 在四分树索引结构中，假定已知移动对象可能的最大运行速度 $v_{\max}$，则可计算每个结点对应区域的 MMB，即按照半径 $v_{\max}\times(t_{i+1}-t_i)$ 拓展每个子区域，如图 2 所示. 其中，t_i 和 t_{i+1} 表示两个连续时刻，节点 u_i 的最大运行边界用 $MMB(u_i)$ 表示.

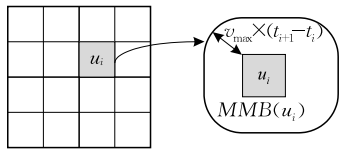


图 2　四分树结点的最大运行边界

对比自由空间中两个时刻发布的数据，会发现这种数据不一致问题. 图 3 展示了两个连续时刻 t_i 和 t_{i+1} 的某个区域的位置分布情况，上面两图为原始数据，下面量图为添加了拉普拉斯噪音的数据. 在 t_i 时刻，区域 u_i 中不包含任何移动对象，u_i 的 MMB 中有两个移动对象，也就是说，在 t_{i+1} 时刻，最多有两个移动对象运行至 u_i 中. 而在 t_{i+1} 时刻，u_i 中有 6 个移动对象.

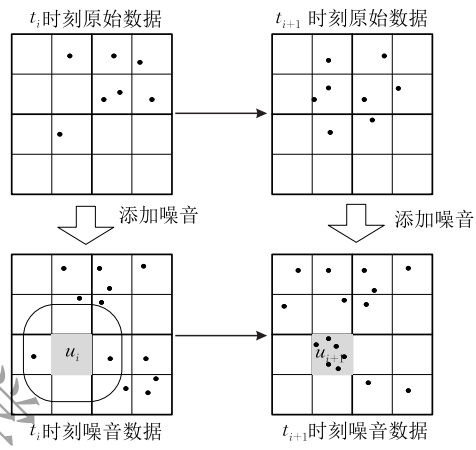


图 3　噪音四分树中的数据不一致性

定义 5(自由空间中的一致性限制条件).　给定移动对象最大运行速度 $v_{\max}$ 及噪音四分树 QT，一致性限制条件如下：

(1) 对任意两个连续时刻 t_i 和 t_{i+1}，结点 u_i 的计数值满足 $0\leq\tilde{c}_{t_{i+1}}(u_i)\leq\tilde{c}_{t_i}(MMB(u_i))$；

(2) 对任意时刻 t_i，结点 u_i 的计数值满足 $\tilde{c}_{t_i}(u_i)\leq l$.

将上面定义的限制条件表示为 γ_c，该限制条件由两部分组成：对任意两个连续时刻 t_i 和 t_{i+1} 的噪音计数值，t_{i+1} 时刻 u_i 的噪音计数值大于等于 0，且小于 t_i 时刻 u_i 的 MMB 中的移动对象数量；对于任意时刻 t_i 的任意结点 u_i，其噪音计数值小于等于结点最大容量阈值 l.

一致性处理的目标是：使发布的数据在满足限制条件 γ_c 的情况下，与原始噪音数据最接近，从而提高发布数据的可用性.

4.3.2　路网空间中的数据一致性问题

路网空间中噪音数据的不一致问题也是添加独

立同分布的拉普拉斯噪音导致的.如果知晓移动对象的最大运行速度 $v_{\max}$ 和路网状况,与前述区域最大运行边界一样,每条路段 rs_i 有一个最大运行边界 $MMD(rs_i)$,它是以 $v_{\max}\times(t_{i+1}-t_i)$ 为半径对路段 rs_i 进行拓展生成的圆角矩形.其中,t_i 和 t_{i+1} 是两个连续时刻.其中,$v_{\max}$ 表示移动对象的最大运行速度,t_i 和 t_{i+1} 是两个连续时刻.图 4 展示了路段 rs_i 的最大运行边界.

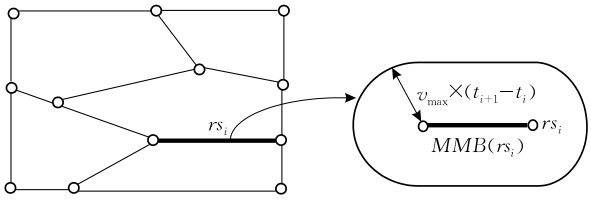


图 4 路段的最大运行边界

值得注意的是:移动对象在路网空间中不能随意地运行至 MMB 中的任意位置,仅能运行到被最大运行边界覆盖的,且与路段 rs_i 连通的路段上.此处的数据不一致产生原因与四分树中的情况类似,在这里就不做过多赘述.下面给出了路网空间中的一致性处理限制条件.

定义 6(路网空间中的一致性限制条件). 给定移动对象最大运行速度 $v_{\max}$、噪音 R-树 RT 以及路网状况,数据一致性需满足下述条件:

(1)对任意两个连续时刻 t_i 和 t_{i+1},路段 rs_i 的计数值满足:$0\leq\tilde{c}_{t_{i+1}}(rs_i)\leq\tilde{c}_{t_i}(MMB(rs_i))$;

(2)对任意时刻 t_i,路段 rs_i 的噪音计数值满足:$\tilde{c}_{t_i}(rs_i)\leq r_{\max}$.

其中,$r_{\max}$ 表示一条路段上能容纳的移动对象的最大数量,通常是与路段长度有关的常数.定义 6 中的限制条件包含两部分含义:(1)对于两个连续时刻 t_i 和 t_{i+1},路段 rs_i 在 t_{i+1} 时刻的噪音计数值需大于等于 0,小于等于 t_i 时刻 $MMD(rs_i)$ 中的移动对象计数值.如果与 rs_i 连通的路段仅有部分被 $MMD(rs_i)$ 覆盖,按均匀性假设计算计数值;对于任意时刻 t_i,路段 rs_i 的噪音计数值必须小于等于该路段上可容纳的移动对象最大值 $r_{\max}$.

4.3.3 一致性处理算法

一致性处理的规则指:给定噪音计数值序列 $\tilde{C}$ 和限制条件 γ_c,找到满足限制条件 γ_c,且与原始噪音计数序列 $\tilde{C}$ 距离最近的计数值序列,记为 $\bar{C}$."距离最近"是用 L_2 距离衡量的.本文根据文献[20]中提出的一致性处理规则设计了一致性处理算法.

定义 5 和定义 6 中的一致性限制条件可用同一种一致性处理方法解决.给定原始噪音计数值序列 $\tilde{C}=\{\tilde{c}_1,\tilde{c}_2,\cdots,\tilde{c}_n\}$,假定一致性处理后的噪音计数值序列 $\bar{C}=\{\bar{c}_1,\bar{c}_2,\cdots,\bar{c}_n\}$ 满足限制条件 γ_c,$\tilde{C}$ 和 $\bar{C}$ 的 L_2 距离由 $f(\bar{C})$ 表示,形式化表述如下:

$$f(\bar{C})=\sqrt{(\tilde{c}_1-\bar{c}_1)^2+(\tilde{c}_2-\bar{c}_2)^2+\cdots+(\tilde{c}_n-\bar{c}_n)^2} \quad (2)$$

$$\gamma_c=\begin{cases}\bar{c}_1,\bar{c}_2,\cdots,\bar{c}_n>0\\ \bar{c}_1,\bar{c}_2,\cdots,\bar{c}_n<con_1\\ \bar{c}_1,\bar{c}_2,\cdots,\bar{c}_n<con_2\end{cases} \quad (3)$$

其中,con_1 和 con_2 是两个常数,可对应为定义 5 或定义 6 中的上限限制条件.在定义 5 和 6 中,l 和 $r_{\max}$ 均为常数值,给定 t_{i+1} 时刻的移动对象噪音计数值,$\tilde{c}_{t_i}(MMB(u_i))$ 和 $\tilde{c}_i(MMB(rs_i))$ 也是常数.式(2)和(3)构成了二次规划问题,在满足 γ_c 的条件下,使得目标函数 $f(\bar{C})$ 的值最小.本文提出一种基于启发式规则的算法,以获取近似解.

本文采用可行方向方法,即由迭代中已经得到的点 c_k,以及某个方向 z_k,由 $\tilde{C}$ 中某点出发沿着方向 $\bar{Z}$ 求一个步长 λ_k,使得目标函数值最小.例如求 λ_k,使得 $\min\limits_{\lambda\geq0}f(c_k+\lambda z_k)=f(c_k+\lambda_k z_k)$,得到一个新的点 $c_{k+1}=c_k+\lambda_k z_k$.再由 c_{k+1} 出发沿某一方向 z_{k+1},求目标函数的最优解.其中,$f(\bar{C}+\lambda\bar{Z})=f(x_1+\lambda z_1,x_2+\lambda z_2,\cdots,x_n+\lambda z_n)$ 是关于单变量 λ 的函数,求 $\min\limits_{\lambda}f(\bar{C}+\lambda\bar{Z})$ 是一个单变量极值问题.本文采用"成功-失败"方法,设计了一种启发式规则以解决此问题,如算法 3 所示.

算法 3. 一致性处理算法.

输入:步长 g;误差 δ;函数 $F(\lambda)$ 及限制条件 γ_c;噪音计数值序列 $\tilde{C}$

输出:一致的噪音技术值序列 $\bar{C}$

```
1. FOR C̃ 中的每个数据项 c̃_i DO
2.     λ=λ_0;
3.     g 初始化为一个常数;
4.     F_0=F(λ);
5.     F_1=F(λ+g);
6.     WHILE |g|>δ DO
7.         WHILE F_1>F_0 DO
8.             F_0=F_1;
9.             λ=λ+g;
10.            g=2g;
11.            F_1=F(λ+g);
12.        END WHILE
13.        g=-g/4;
14.    END WHILE
15. c̄_i≈λ;
16. END FOR
```

基于启发式规则，算法 3 提出了一个计算函数极值的方法. 算法的主要思想是：取初始点 λ_0 以及初始步长 g，采用不同的步长值计算函数值 $F(\lambda_0)$ 和 $F(\lambda_0+g)$，根据上述两个函数值的对比结果调整步长值，直至在误差范围 δ 内找到 $F(\lambda)$ 的最小值. 两个函数值的对比有成功和失败两种可能：

搜索成功，即 $F(\lambda_0+g)<F(\lambda_0)$：下一步由点 λ_0+g 沿先前步长的两倍，继续向前搜索，即 $F(\lambda_0+g)$ 和 $F((\lambda_0+g)+2g)$；

搜索失败，即 $F(\lambda_0+g)\geq F(\lambda_0)$：下一步由点 λ_0 沿先前步长的 1/4，向相反的方向搜索，即比较 $F(\lambda_0)$ 和 $F(\lambda_0-g/4)$.

算法 3 的输入是步长值 g，误差范围 δ，函数 $F(\lambda)$ 和限制条件 γ_c. 算法的输出是满足限制条件的计数值序列 $\bar{C}$. 首先，将步长变量 λ 和步长增量 g 初始化为两个常数，并分别计算 F_0 和 F_1 的值，F_1 是步长增加 g 后的值. 当步长增量 g 大于误差范围 δ 时，对比 F_0 和 F_1 的值. 如果 F_1 的值较大，则说明当前的搜索方向是正确的，需要继续计算 F_1，把步长增加为 $\lambda+2g$；如果 F_1 的值较小，则说明当前的搜索方向是错误的，这时要检查外部循环条件. 如果步长增加量 g 大于误差范围 δ，则把步长增加量设为 $-h/4$，再次计算 F_0 和 F_1 的值. 如果步长增长量 g 小于误差范围 δ，则证明已经找到 $F(\lambda)$ 的近似最小值，当前的 λ 值即为 $\bar{c}_i$ 的值. 算法 3 根据上述方法依次处理 $\tilde{C}$ 中的每条数据项，最终得出 $\bar{C}$.

4.4　查询处理

噪音四分树的查询处理. 噪音四分树的每个结点存储了位置数据的噪音计数值. 空间范围查询可表示为查询矩形，四分树的每个结点也对应一个空间划分的矩形. 在噪音四分树上执行空间范围查询时，按下述步骤处理：

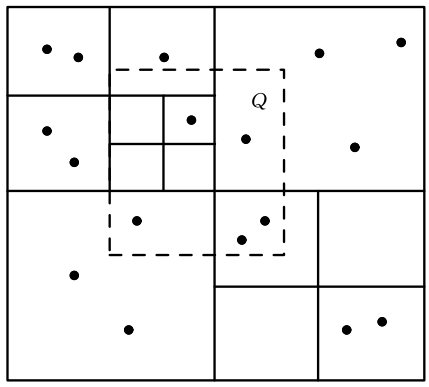
(1) 从根结点开始，依次向下层遍历，访问所有与查询矩形相交的或被查询矩形覆盖的结点；

(2) 若结点 u_i 被查询矩形完全包含，则将该结点的计数值 $\bar{c}(u_i)$ 加入查询结果中；

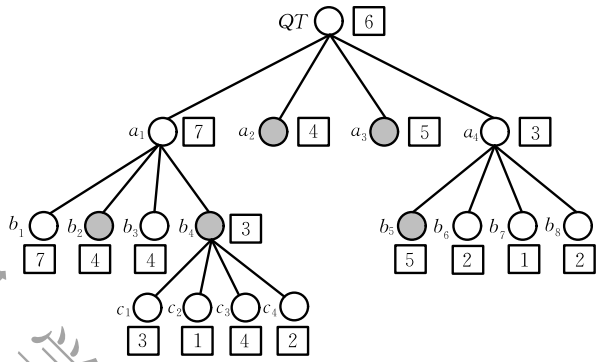
(3) 否则，递归地查找结点 u_i 的子结点，直到找到被查询矩形完全包含的结点或叶子结点为止，被完全包含的结点按第(2)步中的方式处理；

(4) 如果直到叶子结点也没有完全被查询矩形覆盖，仅和查询矩形相交，则按均匀分布的假设，根据节点对应的区域和查询矩形重叠的面积比例，计算被查询矩形覆盖的噪音计数值，加入查询结果中.

噪音四分树及其对应空间划分的例子如图 5 所示. 图 5(a)是四分树对应的空间划分，矩形 Q 表示某个空间范围查询. 图 5(b)为通过处理后生成的噪音四分树 QT，圆形结点表示四分树上的结点，矩形表示结点对应的噪音计数值.



(a) 噪音四分树对应的空间划分



(b) 噪音四分树

图 5　噪音四分树示例

(1) 当查询 Q 到来时，首先遍历 QT 的第一层结点，结点 a_1，a_2，a_3，a_4 均和查询 Q 相交；

(2) 依次遍历结点 a_1，a_2，a_3，a_4 的子结点：

① 结点 a_1 的子结点 b_2 和 b_4 与 Q 相交，结点 b_2 已经是叶子结点，直接按照均匀分布假设，将相应的噪音计数值记入查询结果中；结点 b_4 完全被查询 Q 覆盖，直接将结点 b_4 的噪音计数值记入查询结果中；

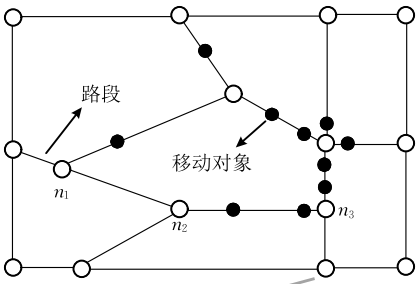
② 结点 a_2 和结点 a_3 是叶子结点，按照均匀分布假设，将相应的噪音计数值记入查询结果中；

③ 结点 a_4 是非叶子结点且与查询 Q 部分相交. 遍历 a_4 的子结点，叶子结点 b_5 与查询 Q 部分相交，按照均匀性假设将相应的噪音计数值记入查询结果中.

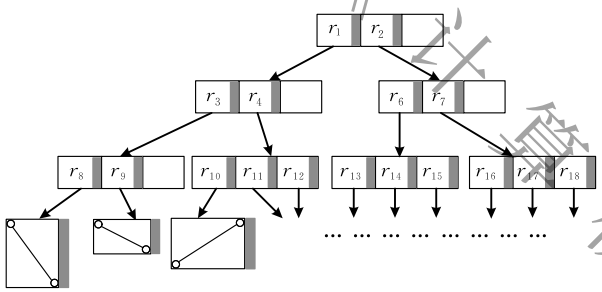
通过上述查询过程可知，最终的查询结果包含结点 a_2，a_3，b_2，b_4，b_5（即图 5(b)中灰色结点）的部分或全部噪音计数值.

噪音 R-树的查询处理. 由于噪音 R-树是对路网进行索引，因此支持基于路段空间范围查询. 给定

一条路径 $n_1 \rightarrow n_2 \rightarrow n_3$，查询结果返回每个路段上的移动对象数目，如图 6 所示. 查询从根结点开始，访问所有与查询路径相交的结点. 递归地查找这些结点的子结点，直到叶子结点位置. 如果叶子结点中包含的路段即为查询中的路段，则将噪音计数值记入查询结果中. 如果该叶子结点只包含部分路段，则按照均匀分布假设将相应的噪音计数值记入查询结果中.



(a) 路网空间与移动对象数据



(b) 噪音R-树

图 6 噪音 R-树示例

4.5 算法隐私保护度与可用性分析

4.5.1 隐私保护度分析

差分隐私算法的隐私保护度与噪音数据的产生机制有关. 在轨迹数据发布中，除了快照位置数据中的移动对象计数值，并未发布有关原始数据的其它统计信息. 定理 5 证明了本文提出的算法满足 ε-差分隐私.

定理 5. 给定隐私预算 ε，算法 1 和算法 2 分别满足 ε-差分隐私.

证明. 算法 1 和算法 2 分别包含两个主要步骤：噪音树构建和一致性处理.

首先证明算法 1 和算法 2 的噪音树构建过程满足 ε-差分隐私. 算法 1 采用了噪音四分树索引结构，发布自由空间中的移动对象计数值；算法 2 采用噪音 R-树索引结构发布路网空间中的移动对象计数值. 无论是四分树或 R-树结构，其每层各个结点所包含的移动对象都是互不相交的. 根据差分隐私的平行分解和顺序分解性质，从根结点到叶子结点的隐私预算为 $\sum_{i=1}^{h}\frac{\varepsilon_i}{h}=\varepsilon$，算法 1 和算法 2 的第一步均满足 ε-差分隐私.

下面证明算法 1 和算法 2 的一致性处理步骤满足 ε-差分隐私. 用 Q 表示在限制条件 γ_c 下的查询序列. 给定噪音查询序列 $\tilde{q}=\tilde{Q}(D)$，用向量 $\bar{\boldsymbol{q}}$ 表示 $\tilde{q}$ 的 L_2 距离最小解，$\bar{\boldsymbol{q}}$ 满足限制条件 γ_c 且满足 $\|\tilde{q}-\bar{\boldsymbol{q}}\|_2$ 的值最小. 算法 1 和算法 2 的一致性处理过程正是采用 L_2 最小距离求解的. 对于任意查询输出 $\bar{\boldsymbol{q}}$，用 $N(\bar{\boldsymbol{q}})$ 表示带噪音的查询输出，则 $\tilde{q}\in N(\bar{\boldsymbol{q}})$. 给定数据集 D 及 D 的邻居数据集 D'，则

$$\begin{aligned}\Pr[\bar{Q}(D)=\bar{\boldsymbol{q}}]&=\Pr[\tilde{Q}(D)\in N(\bar{\boldsymbol{q}})]\\&\le\Pr[\tilde{Q}(D')\in N(\bar{\boldsymbol{q}})]\cdot\exp(\varepsilon)\\&=\Pr[\bar{Q}(D')=\bar{\boldsymbol{q}}]\cdot\exp(\varepsilon).\end{aligned}$$

算法 1 和算法 2 的噪音树构建及一致性处理步骤均满足 ε-差分隐私，则算法 1 和算法 2 满足 ε-差分隐私.

4.5.2 数据可用性分析

在空间范围计数查询上，算法 1 和算法 2 满足差分隐私. 定理 6 证明了这两个算法生成的数据集同样可达到 (ε,δ)-可用性.

定理 6. 算法 1 和算法 2 生成的数据在空间范围查询上可达到 (ε,δ)-可用性.

证明. 假定空间范围查询 Q 覆盖了输出域中的 n 个项，Q 在数据集 D 上的精确查询结果为 $Q(D)=\sum_{i=1}^{n}Q(S_i)$，其中 S_i 表示查询 Q 覆盖的项；在噪音数据集 $\bar{D}$ 上的查询结果表示为 $Q(\bar{D})=\sum_{i=1}^{n}(Q(S_i+N_i))$，其中 N_i 表示添加的噪音. 根据 (ε,δ)-可用性的定义可知，需证明 $\Pr[|Q(D)-Q(\bar{D})|\le\varepsilon]>1-\delta$，其中，$|Q(D)-Q(\bar{D})|$ 可以表示为式(4)的形式：

$$\begin{aligned}|Q(D)-Q(\bar{D})|&=\left|\sum_{i=1}^{n}Q(S_i+N_i)-\sum_{i=1}^{n}Q(S_i)\right|\\&=\left|\sum_{i=1}^{n}N_i\right|\le\sum_{i=1}^{n}|N_i|\le\varepsilon\end{aligned}\qquad(4)$$

根据差分隐私的定义，N_i 是一个满足拉普拉斯分布的独立同分布变量. 如果对于每条 N_i 来说，都有 $|N_i|\le\varepsilon_i$，则 $\sum_{i=1}^{n}|N_i|\le\varepsilon$，其中，$\varepsilon_i=\varepsilon/n$. 如果 $|N_i|>\varepsilon_i$，则称为一次失败，用 FAILURE 表示.

由文献[11]可知，$\Pr[\text{FAILURE}]=\exp(-\lambda\varepsilon_i)$. 因此，可得出 $\Pr[|N_i|\le\varepsilon]>(1-\exp(-\lambda\varepsilon_i))^n$. 文献[4]证明了 $(1-\exp(-\lambda\varepsilon_i))^n\ge1-n\cdot\exp(-\lambda\varepsilon_i)$，根据上述结论，可以得出：$\Pr[|Q(\bar{D})-Q(D)|\le\varepsilon]\ge1-n\cdot\exp\left(-\frac{\varepsilon\lambda}{n}\right)$，这样就证明了算法 1 和算法 2 生成的噪音数据满足 (ε,δ)-可用性.

5　实验分析

本节主要从数据可用性和算法效率两个方面进行实验分析，以验证本文提出算法的性能. 本文与 k-匿名模型为基础的轨迹隐私保护算法在隐私保护度上不同，而其他满足差分隐私的轨迹保护算法，均考虑单个时刻发布的数据满足差分隐私，并未考虑两个连续采样时刻位置数据的关系. 本实验主要验证本文提出算法的数据可用性与算法可扩展性.

5.1　实验设置

实验采用了两个模拟数据集：在自由空间中，采用模拟的高斯分布数据集，表示为 Gaussian. 该数据集包含了 10000 个移动对象在 3 个连续采样时刻的位置数据. 该数据集中的移动对象运行在区域大小为 5000 m×5000 m 的空间中. 由于算法主要考量两个相邻时刻 t_i 和 t_{i+1} 上的空间范围计数查询，再进行一致性处理. 在模拟数据集上取 3 个采样时刻可以实现上述目的. 在路网空间中，采用德国城市 Oldenburg 的路网数据，整个城市区域的面积约为 23.57 km×26.92 km，它包含了 6105 个结点和 7035 条边. 该数据集表示为 OLDEN. 另外，使用 Brinkhoff Generator[21] 生成器生成了 100 000 个移动对象在 20 个时刻的位置数据.

5.2　数据可用性衡量

发布的数据主要用来做空间范围计数查询，数据可用性由式(1)定义的相对误差来衡量. 本实验主要验证：查询 Q 在噪音数据集 $\tilde{D}$ 和在一致性处理后的数据集 $\bar{D}$ 上的准确率的对比. 根据文献[12]的介绍，选择性参数 s 设置为数据集大小的 1%. 对比算法由如下方式表示：

(1) 噪音四分树算法表示为 Noisy-QT；

(2) 加上一致性处理的噪音四分树算法表示为 Cons-QT；

(3) 噪音 R-树算法表示为 Noisy-RT；

(4) 加上一致性处理的噪音 R-树算法表示为 Cons-RT.

5.2.1　噪音四分树的数据可用性实验

实验在隐私预算 ε 取不同值的情况下，测试算法的相对误差. 通过变更查询范围的大小，随机生成了 10000 个空间范围查询，查询区域面积 Size(Q) 设为数据集总面积的 5%到 50%不等. 根据 Size(Q) 的大小将查询分成不同的组. 图 7 是实验结果，即 Noisy-QT 算法在 ε 取值从 0.5 到 1.5 的查询相对误差，每个值都是执行 10 次查询的平均值.

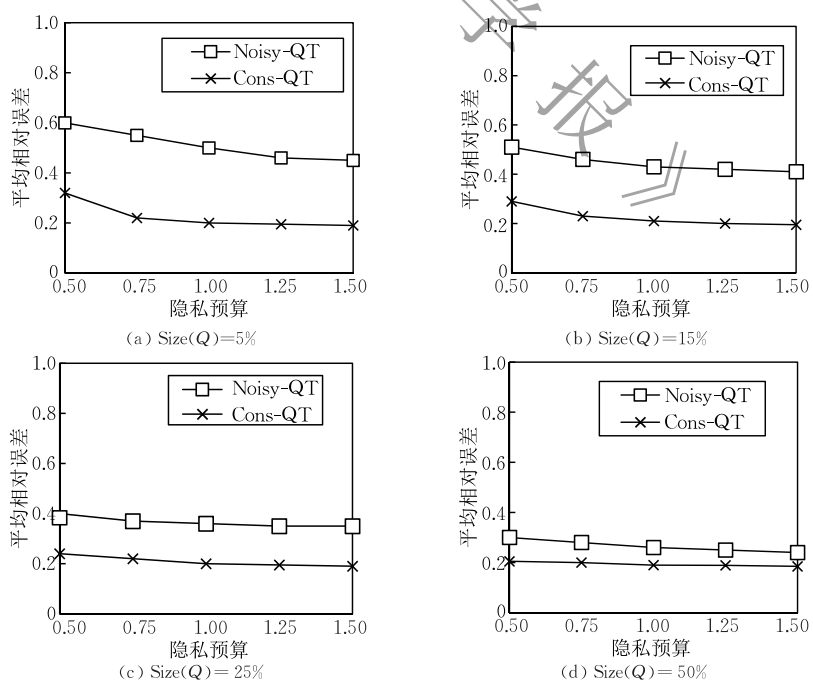


图 7　噪音四分树算法的平均相对误差

图 7 中，x-轴表示不同的隐私预算，y-轴表示相对误差. 查询面积 Size(Q) 分别取为 5%，15%，25%及 50%的数据集空间的总面积. 从图中可以看出，平均相对误差随着隐私预算 ε 的增加而减小，这是由于随着 ε 的增大，构建噪音四分树时添加的噪音数量逐渐减小，使得查询结果更加精确. 从图中可

以看出一致性处理算法的效果，Cons-QT 算法的表现明显优于 Noisy-QT 算法. 随着查询面积的增大，一致性处理算法降低的相对误差越来越小. 随着 Size(Q)的增加，相对误差逐渐减少. 这是由于查询面积越大，其包含的噪音数据的比例就越小. 从实验结果中可以看出，本文提出的方法在不同的查询面积上是稳定的. 按照图 7 的数据趋势来看，随着 Size(Q)的增加，当其大于 50%时，相对误差值会更小.

5.2.2 噪音 R-树的数据可用性实验

本实验在隐私预算 ε 取不同值的情况下测试算法的相对误差. 通过改变查询路段的长度随机生成了 10000 个查询，查询路段的长度 Length(L)设为 5 到 20 个结点不等. 根据 Length(L)的大小将查询分成 4 个不同的组. 每个相对误差值都是执行 10 次查询的平均值. 隐私预算 ε 取值为 0.5 到 1.5 时. 实验结果如图 8 所示.

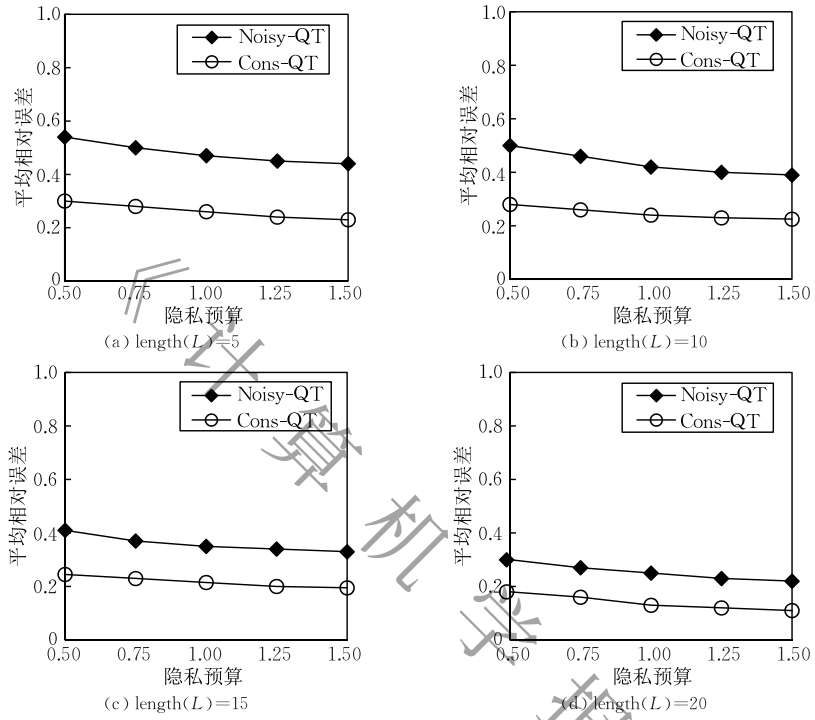


图 8 噪音 R-树算法的平均相对误差

图 8 中，x-轴表示不同的隐私预算值，y-轴表示平均相对误差值. 查询长度 length(Q)分别设为 5，10，15，20 个路网结点. 平均相对误差随着隐私预算 ε 的增大而减小，与前述实验相同：根据差分隐私的性质，隐私预算 ε 越大，构建噪音 R-树时添加的噪音量越小，查询结果越精确. 从图 6 中可以看出数据一致性处理算法的作用，其明显提高了数据的可用性. 随着 length(Q)的增加，相对误差也在减小，这是由于查询范围越大，包含的噪音比例相对较小. 从实验结果中可以看出，本文提出的方法在数据可用性上是稳定的.

5.3 算法运行效率

本实验测试了 4 个算法的运行效率. 算法的运行时间长短与数据集大小 $|D|$ 相关，因此实验测试在不同数据集大小时的运行时间. 图 9 中，x-轴表示数据集大小，y-轴表示算法的运行时间. 本实验的结果是在隐私预算 $\varepsilon=1.0$ 时的运行时间. 图 9(a)的数据集是在 Gaussian 数据集中随机截取大小不同的子集作为数据集进行查询，以测试算法的运行时间，查询区域 Size(Q)的大小取值为总数据集的 15%；图 9(b)的数据集是从 OLDEN 中随机截取相应大小的子集，查询长度 length(Q)=10. 图 9 中的结果是多次运行算法得到的平均运行时间.

从图 9 的实验结果可以看出，算法的运行时间随着数据集的增加近似成线性增长. 经过一致性处理的算法比简单的噪音树算法运行时间长，但运行时间增加的幅度有限. 这也说明使用经过一致性处理的算法有良好的效率和数据可用性. 实验结果表明，运行时间与数据集大小近似成正比. 因此，两个算法可以运行在较大规模的数据集上，算法有较好的可扩展性.

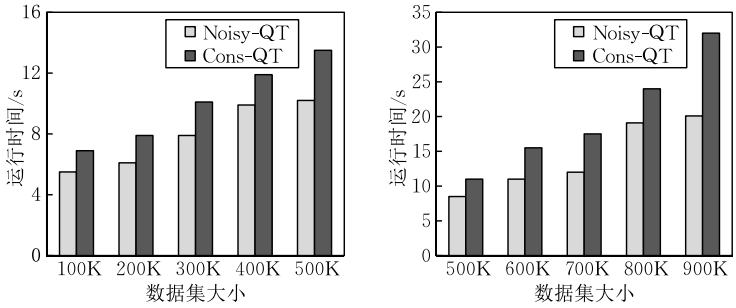


图 9　算法运行时间测试

6　总结与展望

本文提出了一种满足差分隐私的轨迹数据发布方法.针对自由空间移动对象的运行特点,提出了支持空间范围计数查询的噪音四分树构建方法,及基于最大运行速度限制的一致性处理算法;在路网空间提出了基于路段计数查询的噪音 R-树构建方法,及在路网限制条件下的一致性处理算法.本文通过理论证明了上述两个算法满足 ε-差分隐私,且生成的数据在空间范围查询上可达到(ε,δ)-可用性.最后,本文在模拟数据集上进行了数据可用性和算法可扩展性实验,实验结果证明本文提出的方法具有较低的相对查询误差率及良好的可扩展性,算法可用在较大规模数据集上.

本文的研究工作是基于数据收集者是"可信"的假设之上的.然而,越来越多的实际案例表明:所谓的"可信"第三方数据收集者并不可靠,许多数据收集者有意或无意地泄露用户的原始数据,造成大量用户的个人隐私数据泄露.在今后的研究中,可以考虑采用本地化差分隐私技术(Local differential privacy)进行位置或其它个人敏感数据的收集和发布[22],使得第三方收集者也无法获取原始数据,从源头上遏制由个人数据泄露造成的用户隐私泄露.此外,在大数据环境中,个人位置数据和其它个人数据关联起来造成的个人隐私泄露风险量化和攻击模型建模也是值得研究和关注的问题.

参　考　文　献

[1] Dwork C. Differential privacy//Proceedings of the 33rd International Colloquium on Automata, Languages and Programming(ICALP'06). Venice, Italy, 2006: 1-12

[2] Zhang Xiao-Jian, Meng Xiao-Feng. Differential privacy in data publication and analysis. Chinese Journal of Computers, 2014, 37(4): 927-949(in Chinese)
(张啸剑,孟小峰. 面向数据发布和分析的差分隐私保护. 计算机学报, 2014, 37(4): 927-949)

[3] Cormode G, Procopiuc C, Srivastava D, et al. Differentially private spatial decompositions//Proceedings of the IEEE 28th International Conference on Data Engineering (ICDE'12). Washington, USA, 2012: 20-31

[4] Xiao Y, Xiong L, Yuan C. Differentially private data release through multidimensional partitioning//Proceedings of the 7th VLDB Workshop on Secure Data Management. Singapore, 2010: 150-168

[5] Chen R, Fung B C M, Desai B C, Sossou N M. Differentially private transit data publication: A case study on the Montreal transportation system//Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining(KDD'12). Beijing, China, 2012: 213-221

[6] Hua J, Gao Y, Zhong S. Differentially private publication of general time-serial trajectory data//Proceedings of the 2015 IEEE Conference on Computer Communications(INFOCOM). Hong Kong, China, 2015: 549-557

[7] Xiao Y, Xiong L. Protecting locations with differential privacy under temporal correlations//Proceedings of the ACM SIGSAC Conference. Denver, USA, 2015: 1298-1309

[8] Yigitoglu E, Damiani L M, Abul O, Silvestri C. Privacy-preserving sharing of sensitive semantic locations under road-network constraints//Proceedings of the 13th IEEE International Conference on Mobile Data Management (MDM'12). Bengaluru, India, 2012: 186-195

[9] Silvestri C, Yigitoglu E, Damiani L M, Abul O. SAWLnet: Sensitivity aware location cloaking on road-networks//Proceedings of the 13th IEEE International Conference on Mobile Data Management (MDM'12). Bengaluru, India, 2012: 336-339

[10] Peng T, Liu Q, Meng D, Wang G. Collaborative trajectory privacy preserving scheme in location-based services. Information Science, 2017, 387: 165-179

[11] Theodorakopoulos G, Shokri R, Troncoso C, et al. Prolonging the hide-and-seek game: Optimal trajectory privacy for location-based services//Proceedings of the 13th Workshop on Privacy in the Electronic Society (WPES 2014). Scottsdale, USA, 2014: 73-82

[12] Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis//Proceedings of

the 3rd Theory of Cryptography Conference(TCC'06). New York, USA, 2006: 265-284

[13] McSherry F, Talwar K. Mechanism design via differential privacy//Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07). Providence, Rhode Island, 2007: 94-103

[14] McSherry F. Privacy integrated queries: An extensible platform for privacy-preserving data analysis//Proceedings of the 2009 ACM SIGMOD/PODS Conference(SIGMOD'09). Providence, Rhode Island, 2009: 19-30

[15] Chen R, Mohammed N, Fung B C M, et al. Publishing set-valued data via differential privacy. Proceedings of the VLDB Endowment, 2011, 4(11): 1087-1098

[16] Xiao X, Wang G, Gehrke J. Differential privacy via wavelet transforms//Proceedings of the 26th IEEE International Conference on Data Engineering (ICDE'10). Long Beach, USA, 2010: 225-236

[17] Finkel A R, Bentley L J. Quad trees: A data structure for retrieval on composite keys. Acta Informatica, 1974, 4: 1-9

[18] Guttman A. R-Trees: A dynamic index structure for spatial searching//Proceedings of the 1984 ACM SIGMOD Conference (SIGMOD'84). Boston, USA, 1984: 47-57

[19] Pan X, Xu J, Meng X. Protecting location privacy against location-dependent attacks in mobile services. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(8): 1506-1519

[20] Hay M, Rastogi V, Miklau G, Suciu D. Boosting the accuracy of differentially private histograms through consistency. Proceedings of the VLDB Endowment, 2010, 3(1): 1021-1032

[21] Brinkhoff T. A framework for generating network-based moving objects. GeoInformatica, 2002, 6(2): 153-180

[22] Chen R, Li H, Qin A K, et al. Private spatial data aggregation in the local setting//Proceedings of the IEEE International Conference on Data Engineering(ICDE'16). Helsinki, Finland, 2016: 289-300

HUO Zheng, born in 1982, Ph. D., lecturer. Her research interests include privacy-preserving techniques, mobile data management, etc.

MENG Xiao-Feng, born in 1964, Ph. D., professor, Ph. D. supervisor. His research interests include big data management, mobile data management, privacy-preserving techniques, etc.

Background

This research is partially supported by the Grants from the Natural Science Foundation of China (Nos. 91646203, 61532010), the Natural Science Foundation of Hebei Province (Nos. F2015207009, D2015207008) and the Young Prominent Talent Project of Hebei Province Higher School (Nos. BJ2016019, BJ2014021).

Recent years, with the development of location-aware devices, more and more locations and traces of moving objects are collected and then published for mobile-related applications. Analyzing trajectories of passengers in an area may help people making commercial decisions, such as where to build a restaurant. It also can be seen in traffic control systems, analyzing trajectories of vehicles in a city may help government to optimize traffic control strategy. Although publishing trajectories is beneficial for mobility-related decision making processes, it may represent serious threats to individual privacy, since trajectories contain rich spatio-temporal information, which may reveal individual's habits, health condition, social customs, etc. In order to protect trajectory privacy, more and more attentions have been given to this area. In the past years, researchers proposed several techniques based on k-anonymity, however, they could not provide sufficient privacy guarantee since they are highly depend on the background knowledge, the attackers hold. This paper proposes a method under differential privacy, which is a strong privacy model, as we know.

WAMDM lab have studied location privacy-preserving since 2006, several key problems are solved, Related research findings are published in CIKM, ACM SIGSPATIAL GIS and IEEE TKDE, we have studied trajectory privacy-preserving techniques since 2010, related research findings are published in DASFAA, Chinese Journal of Computers, etc. We have studied differential privacy since 2012, publications include SDM, DASFAA, Frontiers of Computers, etc.