山东大学学报(理学版) 2018 年 9 月 第 53 卷 第 9 期
Journal of Shandong University (Natural Science), Vol.53, No.9, 2018
© 山东大学科技期刊社版权所有
E-mail: xblxb@sdu.edu.cn
http://lxbwk.njournal.sdu.edu.cn
Tel: +86-0531-88396917

文章编号: 1671-9352(2018) 09-0012-11 DOI: 10.6040/j.issn.1671-9352.0.2017.418

差分隐私密度自适应网格划分发布方法

晏燕[1,2],郝晓弘[1*]
(1.兰州理工大学电气工程与信息工程学院,甘肃 兰州 730050;
2.兰州理工大学计算机与通信学院,甘肃 兰州 730050)

摘要: 为了进一步均衡噪声误差和均匀假设误差对二维划分发布带来的影响,提出一种新的分层差分隐私位置信息划分发布算法。首先将位置空间聚类形成第一层密度自适应网格,然后对不同性质的密度区块采取不同的二次划分方法,在降低均匀假设误差的同时避免了大量空结点引入的噪声误差。在采用分层划分策略的同时,结合差分隐私模型的串行组合特性,对 2 个阶段的划分结果添加不同隐私预算的 Laplace 噪声,总体上实现对发布数据的 ε-差分隐私保护。实验证明,该算法在改善区域计数查询精度方面具有较好的效果,能够节省不必要的划分过程,有效提高了算法的运行效率。

关键词: 位置大数据; 差分隐私; 空间划分; 密度自适应网格
中图分类号: TP309; TP311 **文献标志码:** A
引用格式: 晏燕,郝晓弘.差分隐私密度自适应网格划分发布方法[J].山东大学学报(理学版),2018,53(9):12-22.

Differential privacy partitioning algorithm based on adaptive density grids

YAN Yan[1,2], HAO Xiao-hong[1*]
(1. School of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou 730050, Gansu, China;
2. School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, Gansu, China)

Abstract: In order to balance the influence of noise error and uniform hypothesis error for the two-dimensional partitioning publishing, a new hierarchical differential privacy partitioning algorithm DP-ADG is proposed. Firstly, the position space is clustered to form the density adaptive grids in the first layer. Then in the second layer, different partitioning methods are adopted for different density blocks. The noise error introduced by a large number of null nodes is avoided while reducing the uniform hypothesis error. While using the hierarchical partitioning strategy, different Laplace noise of different privacy budgets is added to the results of two phases according to the sequential composition of differential privacy, in order to realize the overall ε differential privacy protection for the publishing data. Experimental results show that the algorithm has good effect on improving the accuracy of range counting query, saving unnecessary spatial decomposition process, as well as improving the efficiency of the algorithm.
Key words: location big data; differential privacy; spatial decomposition; density adaptive grids

0 引言

移动互联网的迅速发展和智能终端的广泛普及,使得人们能够随时、随地、随身地获取信息和服务,同时促进了大数据时代的到来。鉴于大数据所蕴含的巨大商业价值,基于大数据的人口调查、公众健康研究、城市交通与道路规划、社会舆情分析与预测等应用得到了广泛的关注与研究[1]。作为互联网的“天然入口”,位置信息被众多大数据应用广泛采集和使用。车联网、智慧城市、基于位置的服务(location based services,

收稿日期: 2017-08-21; **网络出版时间:** 2018-04-27 09:39:43
网络出版地址: http://kns.cnki.net/kcms/detail/37.1389.N.20180426.1808.006.html
基金项目: 国家自然科学基金资助项目(61762059); 甘肃省青年科技基金计划项目(1310RJYA004)
第一作者简介: 晏燕(1980—),女,博士研究生,副教授,研究方向为隐私保护技术、多媒体信息安全. E-mail: yanyan@lut.cn
* **通信作者简介:** 郝晓弘(1960—),男,硕士,教授,博士生导师,研究方向为复杂系统理论、智能控制技术. E-mail: haoxh@lut.com

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

LBS)、移动社交网络等热门应用每天的数据量动辄超过千万。这些位置大数据及其统计信息在为政府、企业和公共服务部门提供数据支持的同时,也埋下了不小的安全隐患[2-5]。移动用户的物理位置和运动轨迹是能够反映出用户行为习惯的一个重要敏感因素,一旦暴露可能会导致用户生活习惯、兴趣爱好、健康状况、宗教信仰等个人隐私信息的泄露,甚至危害用户生命和财产安全[6-8]。虽然一些应用发布的是位置大数据的统计信息;但是攻击者通过结合背景知识仍然有较大的概率推断出个人的实时位置,因此,如何在发布位置大数据及其统计信息的同时防止用户隐私信息的泄露,是目前大数据和信息安全领域的热点问题,将直接关系到大数据的安全应用和进一步发展。

1　相关工作

划分发布是实现位置统计信息发布的一种有效形式,它依据一定的索引结构对位置信息集合进行划分,每一个索引区域采用其划分意义下的数据统计值进行标识,减小了用户真实位置信息的泄露风险。通过对索引区域的数据统计值添加差分隐私噪声,还可以进一步提高数据发布隐私保护的效果。针对二维位置信息的划分发布方法主要基于各种树结构或者网格结构[9-10],具体的划分过程既可以依据真实数据的分布情况,也可以设计独立于数据的索引结构。

Cormode 等[11]提出的 Quad-opt 划分方法采用与数据分布无关的完全四叉树对二维空间进行层次划分,并按照几何策略对不同层次分配差分隐私预算。该方法充分利用了差分隐私的串行和并行特性,通过对节点的加噪结果进行后置处理改善了计数查询的精度。吴英杰等[12]在四叉树划分方法的基础上,采用区域均匀性判断策略对四叉划分的结果自底向上进行调整合并,从而平衡噪声误差与均匀假设误差。虽然利用一致性约束对加噪和调整合并节点进行后置处理,在一定程度上提高了查询的精度,但是该方法仍然使用树的深度控制噪音的大小,存在自上而下分割数据空间时不能很好地确定树深度的问题。针对现有层次化分解方法对分割深度的依赖,Zhang 等[13]提出的 PrivTree 方法通过引入一个可控的偏差来决定是否进行分割,从而消除了对预先定义分割深度的限制。在空间数据的划分索引结构上,该方法沿用了完全四叉树结构,所以仍存在均匀假设误差比较高的问题。除了上述的树结构索引方式,Qardaji 等[14]提出的均匀网格划分方法(UG)采用更为简单的数据独立划分策略,将二维空间均匀地划分为 $m\times m$ 个相同尺寸的单元格,通过对每个单元格内的统计数据添加 Laplace 噪声实现差分隐私保护。Mir 等[15]针对大量蜂窝网络移动数据设计的 DP-Where 方法也采用均匀网格结构,通过添加受控噪声来修改 WHERE 模型实现差分隐私。整体而言,数据独立的空间划分方法通常无法兼顾噪音误差与均匀假设误差的均衡问题,在动辄几百万条数据的实际空间位置集上,无法获得理想的分割结果,范围查询响应的精度有限。

数据依赖的空间划分方法主要根据空间数据点的实际分布情况进行分割。Inan 等[16]提出的 Kd-tree 算法按照 Kd 树结构对二维空间进行层次划分,减少了均匀假设误差带来的问题,但是,每一层分割线的确定都依赖数据的真实分布情况,有可能在响应范围查询时泄露分割线上的真实数据信息,需要耗费一部分隐私预算对中值数加以保护。Cormode 等[11]提出的 Kd-hybrid 算法首先使用数据依赖的 Kd 树划分方法对原始数据进行 l 层划分,然后选用数据独立的四叉树方法对上述结果再次进行划分。实验结果表明综合的 Kd-hybrid算法在响应范围计数查询时可以获得更加精确的结果。Qardaji 等[14]提出的自适应网格划分方法(AG)首先以 $\alpha\cdot\varepsilon$($0<\alpha<1$)的隐私预算进行一次均匀网格划分,然后将每个粗粒度单元自适应地分割成 $m_2\times m_2$ 个细粒度单元,并添加$(1-\alpha)\cdot\varepsilon$ 的隐私预算。通过自适应划分策略,AG 方法避免了第二层单元过于密集或者过于稀疏的问题,但是第一层沿用了 UG 方法的均匀划分策略,没有考虑到数据本身的稀疏性,因此在减小均匀假设误差方面还有进一步改进的空间。黄泗勇等[17]提出的 Kd-PPDP 算法在 AG 算法每一层划分加噪后,使用平方和误差衡量当前网格的均匀性,对 2 个阶段的网格划分结果合并临近区域,以减少噪声误差的叠加。该方法的 2 次比较合并过程需要遍历所有网格的划分界限,计算复杂度较高。张琳等[18]指出通过将位置信息与其他渠道获得的与位置相关的非位置数据相结合,攻击者有可能获得更多关于用户的隐私信息,因此提出一种基于差分隐私保护的位置大数据发布框架,采取不同的噪声机制,对位置数据和非位置数据添加 Laplace 噪声和指数机制噪声实现差分隐私保护。关于位置信息的保护,该文献选用了常见的 Quad-tree、R-tree、Kd-tree 3 种空间索引技术,隐私预算根据索引树高度和叶子节点数量进行分配。所

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

以在发布位置信息的查询精度方面,该方法并没有突破上述3种索引方法的性能表现。文献[19]结合信号的相关性特点设计了一种相关性Laplace机制,将高斯白噪声通过特定冲激响应的滤波器,生成与待发布轨迹相关性一致的Laplace噪声,从而实现原始轨迹和发布轨迹序列的不可区分性。该方法能够在轨迹相关性背景知识公开的情况下,保证用户的隐私不被泄露。

2 差分隐私密度自适应网格划分发布算法

2.1 差分隐私保护模型

基于K匿名的传统隐私保护技术需要对攻击者的能力和背景知识进行假设估计,因此在实际应用中存在一定局限性。差分隐私保护模型[20-21]具备严格的数学特性,通过对待发布数据的随机扰动使得攻击者即便获得了除一条记录之外的所有敏感数据,依然无法根据查询输出结果判断这一条记录是否在原始数据表中。差分隐私模型与位置大数据的发布保护具有天然的匹配性,位置大数据的大规模性、动态多样性等使得在位置集中添加或者删除某个数据点对整体信息的影响非常小,这一特质与差分隐私定义的内涵相吻合。

定义1[21] 针对兄弟数据表T_1和T_2(T_1与T_2相比,存在且仅存在一条不同的记录)以及任意输出$S\subseteq \mathrm{Range}(K)$,如果隐私保护算法$K$得到的查询结果满足

$$\Pr[K(T_1)\in S]\leqslant e^{\varepsilon}\times\Pr[K(T_2)\in S], \tag{1}$$

则称算法K满足ε-差分隐私。

公式(1)表示以T_1和T_2作为隐私保护算法K的输入,得到的查询结果为S的概率非常接近,因此,通过观测算法的输出很难判断一条记录是来自T_1还是T_2,从而为该记录提供了隐私保护。差分隐私参数ε用来衡量隐私保护的强度,ε值越小,提供的隐私保护强度越高。

定义2[21] 对于非交互式的操作,可以通过向数据预先添加Laplace噪声实现差分隐私保护。对原始数据表T的查询请求可以被视为某个函数f作用于T上得到的值,查询结果可以表示为$f(T)$。为了实现ε-差分隐私保护,随机算法K的输出为$K(T)=f(T)+\eta$。其中η是满足Laplace分布的连续型随机变量,其概率密度函数为

$$p(\eta)=\frac{1}{2\lambda}e^{-\frac{|\eta|}{\lambda}}。 \tag{2}$$

定义3[20] 对原始数据表T的所有查询请求构成函数集合F,若$f(T)_{f\in F}\in R$,则函数集合F的敏感度定义为任意一对兄弟数据表查询结果的最大差异值:

$$S(F)=\max_{T_1,T_2}\left(\sum_{f\in F}|f(T_1)-f(T_2)|\right)。 \tag{3}$$

定理1[20] 对于敏感度为$S(F)$的函数集合F,K表示向F中每一个函数f的输出添加独立噪声的随机算法,如果该噪声为参数取值$\frac{S(F)}{\varepsilon}$的Laplace分布,则算法K满足ε-差分隐私。

定理1表明,实现差分隐私保护需要添加的Laplace噪声参数λ与敏感度$S(F)$和差分隐私参数ε密切相关。敏感度取值$S(F)$越大,隐私参数ε越小,隐私保护越强。

定理2[23] 假设有一组随机算法$\{F_1,F_2\cdots F_n\}$,其中$F_i(1\leqslant i\leqslant n)$满足数据表$T$上的$\varepsilon_i$-差分隐私,则由$\{F_1,F_2\cdots F_n\}$组合之后的算法能够实现数据表$T$上的$\sum_{i=1}^{n}\varepsilon_i$差分隐私。

定理2说明差分隐私保护算法具有串行组合特性,为了实现总体上隐私预算为ε的差分隐私保护,可以分不同阶段或者方法对数据表分别添加预算为ε_i的差分隐私噪声,满足$\varepsilon=\sum_{i=1}^{n}\varepsilon_i$。AG、Quad-opt、Kd-hybrid等分层差分隐私保护方法都应用了这一特性。

2.2 二维划分发布的主要噪声来源

位置信息的划分发布主要针对位置大数据的应用需求发布统计类型的信息。例如,查询某个范围内的用户数量,统计某条交通道路上的车流量等。其发布结果的误差主要来源于噪声误差和均匀假设误差2个方面。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(1) 噪声误差

为了防止攻击者通过大量查询结果推测出合法用户的位置信息,划分发布方法采用 Laplace 机制或指数机制[20,23]对统计数据添加噪声,使之满足差分隐私保护的需求。添加的噪声对查询结果造成的影响就称之为噪声误差,定义如公式(4):

$$\mathrm{NoiseErr}(P) = |C(P) - C(P')|, \tag{4}$$

其中,P 代表划分的区块,$C(P)$ 和 $C(P')$ 分别表示该划分区块内的原始统计值和加噪统计值。如果划分过程中产生大量的空结点(即真实计数值为 0 的区块),不但会消耗差分隐私预算,还会引入较大的噪声误差,从而影响发布结果的查询精度,降低了发布数据的可用性。

(2) 均匀假设误差

由于无法将位置数据空间精确划分到单个点,二维数据划分发布方法往往需要假设某个区域内的数据分布是均匀的。此时,范围计数查询算法根据查询区域 Q 与假设区域相交的面积比例计算查询结果。均匀假设误差就是由于采用均匀假设估计而造成的范围计数查询误差,其定义如公式(5)所示:

$$\mathrm{UniErr}(Q) = \left| \sum_{i=1}^{m} r_i \cdot C(P_i) - C(Q) \right|, \tag{5}$$

其中,$P_i(i=1,2\cdots,m)$ 表示与查询区域 Q 相交的划分区块,满足 $\underset{1\leqslant i,j\leqslant m\wedge i\neq j}{P_i\cap P_j} = \varnothing$。$r_i(i=1,2,\cdots,m)$ 代表查询区域 Q 在该区块中所占的面积比例。均匀假设误差的大小与发布数据的真实分布情况和区块划分的数量紧密相关,如果数据的真实分布情况较为均匀,利用均匀假设估计能够减小查询结果中的噪声误差,此时区块划分的越多,噪声误差被分摊到越多的区块当中,有助于减小噪声误差;而当真实数据分布不均匀时,区块划分虽然有助于降低噪声误差,但却会造成较大的均匀假设误差。

2.3　差分隐私密度自适应网格算法

为了进一步均衡噪声误差和均匀假设误差对二维划分发布结果带来的影响,本文提出一种新的分层差分隐私位置信息划分发布算法(DP-ADG)。该方法首先结合位置信息的真实分布情况,通过判定横坐标和纵坐标的分布密度,将位置空间聚类成第一层密度自适应网格划分。第二层划分过程对不同性质的密度区块采取不同的处理方法:对于用户分布最为稀疏的区块(X 轴与 Y 轴同时位于稀疏区域),停止划分以避免产生大量空结点;剩余密度区块则借鉴自适应网格划分或四叉树划分方法,将粗粒度区块进一步划分为细粒度单元,用于分担区域内的噪声误差。在采用上述分层策略的同时,结合差分隐私保护算法的串行组合特性,对 2 个阶段的划分结果添加不同隐私预算的 Laplace 噪声,总体上实现对发布数据的 ε-差分隐私保护。

算法　DP-ADG 算法

输入　待发布的位置信息集合 L,差分隐私预算 ε,隐私预算分配比例 α;

输出　添加 Laplace 噪声的发布结果。

步骤

Step 1: 以待发布的位置信息 L 为初始矩阵,创建 2 个集合:一个只包含所有横坐标位置(记为 L_X),另一个只包含所有纵坐标位置(记为 L_Y)。

Step 2: 定义横坐标和纵坐标方向上的密度阈值 $\mathrm{Dens}_X = k_x \cdot \dfrac{|L_X|}{\Delta x}$, $\mathrm{Dens}_Y = k_y \cdot \dfrac{|L_Y|}{\Delta y}$,分别对集合 L_X 和 L_Y 进行密度判定,形成横坐标和纵坐标的稠密区间和稀疏区间。其中 $|L_X|$ 和 $|L_Y|$ 代表集合中的元素个数,Δx 和 Δy 是在不同坐标方向上划分的区间长度,k_x 和 k_y 是分配的权重系数。

Step 3: 综合上述 2 个方向的区间划分结果,对位置信息集合 L 进行密度自适应网格划分,得到第一层粗粒度区块;

Step 4: 对于横坐标和纵坐标同时位于稀疏区域的第一层区块,直接对其原始计数结果添加隐私预算为 ε 的 Laplace 噪声;

Step 5: 其余区块按照密度自适应网格划分结果,首先添加隐私预算为 $\alpha \cdot \varepsilon$ 的 Laplace 噪声,然后转向下一层划分;

Step 6: 根据第一层密度区块的噪音计数结果 N',进行第二层自适应网格划分(或者 Quad-tree 划分),并添加隐私预算为 $(1-\alpha) \cdot \varepsilon$ 的 Laplace 噪声;

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

Step 7: 对两层加噪后的结果进行一致性约束调整,得到最终发布结果 N_p。

3 实验结果与分析

本节从范围查询的精度和算法运行的效率两个方面,将本文提出的密度自适应网格划分发布方法与均匀网格划分方法[14](UG)、自适应网格划分方法[14](AG)、四叉树划分方法[11](Quad-opt)、Kd-tree 与四叉树结合的方法[11](Kd-hybrid)进行了比较分析。

3.1 实验环境与数据

本文实验在 Intel(R) Core(TM) i7-6700/3.4GHz/8GB 硬件平台和 Windows 7/MATLAB R2013a 软件环境下进行。实验使用的位置数据集分别选自基于位置的社交网站 Gowalla 上的用户 Checkin 信息①、infochimps 提供的美国 48 个州的地标位置信息集 Landmark② 和存储设施位置信息集 Storage③,各数据集的分布状态如图 1 所示。按照参考文献[14]的方法,设置 6 种大小不同的查询区域,针对每种类型的查询区域随机生成 500 个查询,计算相对误差的平均值。相对误差的定义如式(6)所示:

$$\mathrm{RE}(Q)=\frac{|C_M(Q)-C(Q)|}{\max\{C(Q),\rho\}}, \tag{6}$$

其中,$C(Q)$ 为查询原始数据集的结果,$C_M(Q)$ 是查询发布数据集得到的结果,$\rho=0.001\times|D|$,$|D|$ 代表数据集的大小。实验数据集与查询区域的性质如表 1 所示。

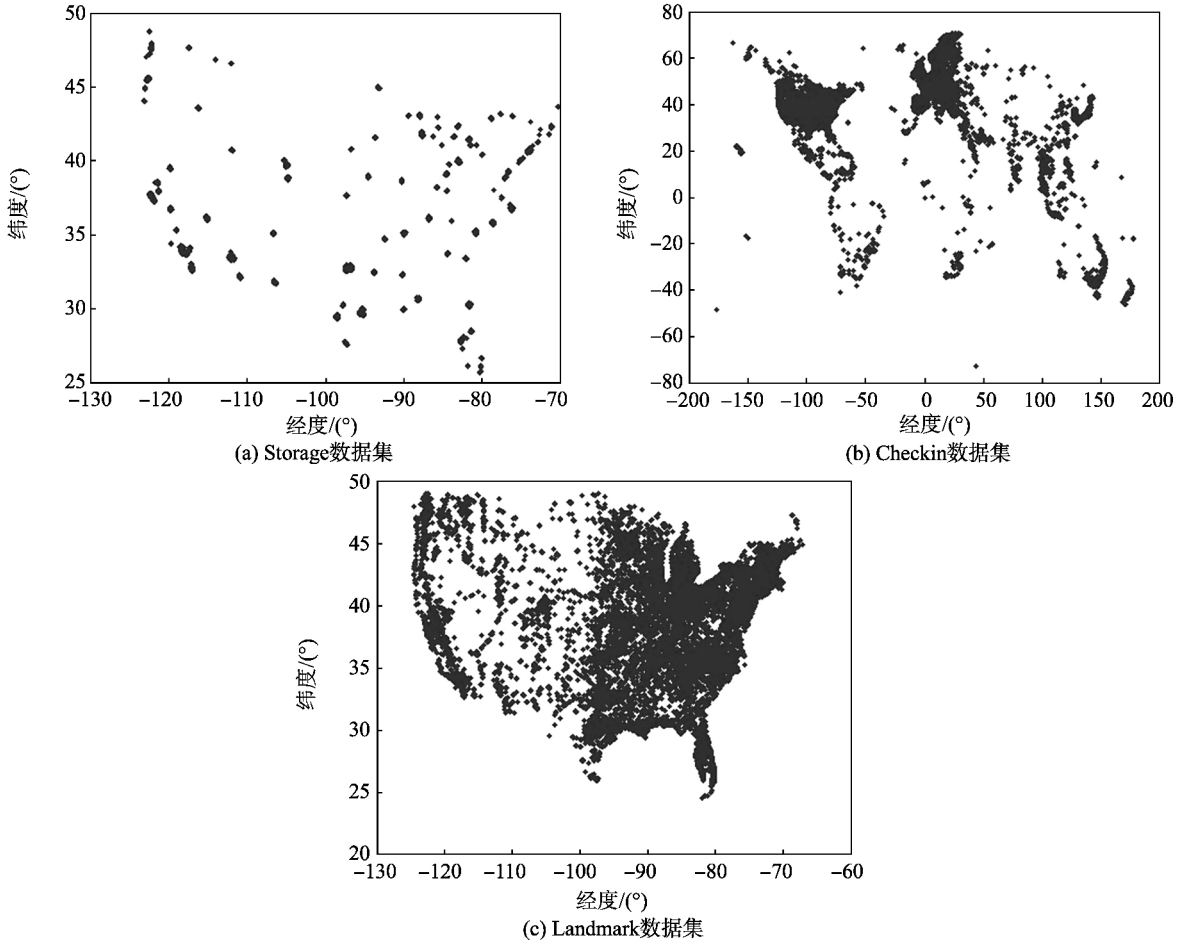


图 1 实验数据集的分布状态

Fig.1 Distribution status of testing database

① http://snap.stanford.edu/data/loc-gowalla.html。

② http://www.infochimps.com/datasets/storage-facilities-by-landmarks。

③ http://www.infochimps.com/datasets/storage-facilities-by-neighborhood—2。

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 1　实验测试数据集信息

Table 1　Information of testing database

参数	数据集		
	Storage	Checkin	Landmark
总数据个数	8 938	1 000 000	870 052
覆盖范围(经度范围×维度范围)	60×40	360×150	60×40
q_1 尺寸(经度范围×维度范围)	1.25×0.625	6×3	1.25×0.625
q_2 尺寸(经度范围×维度范围)	2.5×1.25	12×6	2.5×1.25
q_3 尺寸(经度范围×维度范围)	5×2.5	24×12	5×2.5
q_4 尺寸(经度范围×维度范围)	10×5	48×24	10×5
q_5 尺寸(经度范围×维度范围)	20×10	96×48	20×10
q_6 尺寸(经度范围×维度范围)	40×20	192×96	40×20

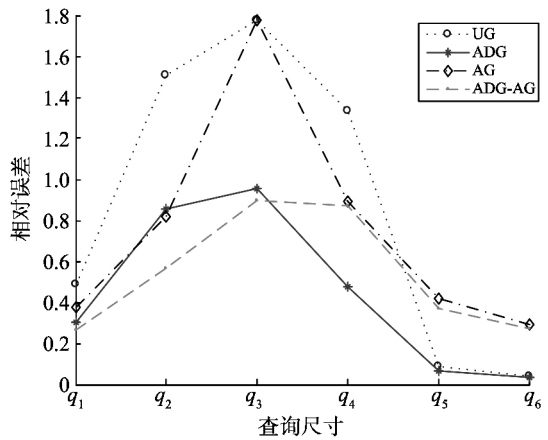
3.2　范围查询精度

实验分为 2 组进行。第 1 组实验将本文提出的密度自适应网格划分方法与基于网格的划分发布算法进行比较。为了与均匀网格划分方法 UG 进行对比,采用单层的密度自适应网格划分,对划分结果添加预算为 ε 的 Laplace 噪声实现差分隐私保护(记为 ADG 算法)。为了与自适应网格划分方法 AG 进行对比,采用 2 层划分策略(记为 ADG-AG),第 1 层使用密度自适应网格划分,隐私预算分配比例 $\alpha=0.5$; 第 2 层使用与 AG 算法相同的自适应网格划分策略,划分粒度 $m_2=\left\lceil\sqrt{\frac{N'(1-\alpha)\varepsilon}{c_2}}\right\rceil$, $c_2=\frac{c}{2}$; 最后按照式(7)和式(8)的一致性约束条件[20]对 2 层加噪的结果进行调整,其中 $N^{*\prime}$ 代表调整之后 2 层划分单元格的噪声计数值。

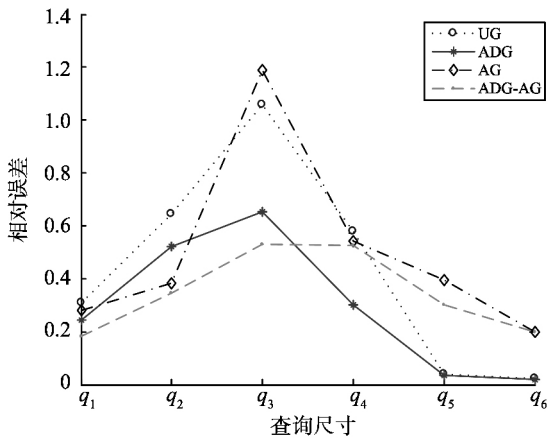
$$N_P=\frac{\alpha^2 m_2^2}{(1-\alpha)^2+\alpha^2 m_2^2}N'+\frac{(1-\alpha)^2}{(1-\alpha)^2+\alpha^2 m_2^2}\sum N^*\ , \tag{7}$$

$$N^{*\prime}=N^*+\left(N_P-\sum N^*\right)。 \tag{8}$$

图 2 显示了上述基于网格划分策略的差分隐私保护算法在不同数据集上的范围查询相对误差。位置信息分布较为均匀的 Landmark 集具有较高的查询精度,Checkin 数据集次之,位置信息分布稀疏的 Storage 集查询误差较大。其原因在于稀疏网格引入的噪声误差和不均匀网格产生的均匀假设误差。在相同的数据集下,各种算法的相对误差都随着隐私预算 ε 的增大而逐渐减小。因为在敏感度 S 不变的情况下,隐私预算 ε 的增大导致添加的 Laplace 噪声减小,所以发布结果与真实数据的偏差减小。比较各种算法的结果不难看出,UG 算法存在较大的均匀假设误差和噪声误差,在各种数据集和隐私预算下的查询误差都较大; AG 算法虽然在第 2 层采用了自适应的网格划分策略,但很难抵消第 1 层均匀划分引入的误差; ADG 算法通过对位置信息分布密度的聚类降低了稠密区域的均匀假设误差,减小了稀疏区域的噪声误差,当查询区域较小时精度甚至优于 AG 算法; ADG-AG 算法进一步结合了密度网格和自适应网格划分的优势,在各种数据集和隐私预算下的查询误差都有较好的表现。

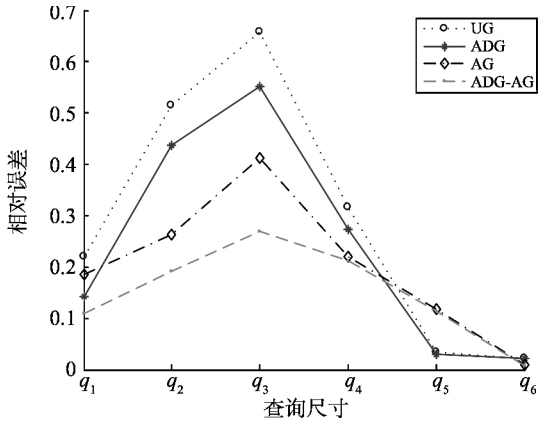


(a) Storage数据集，ε=0.1

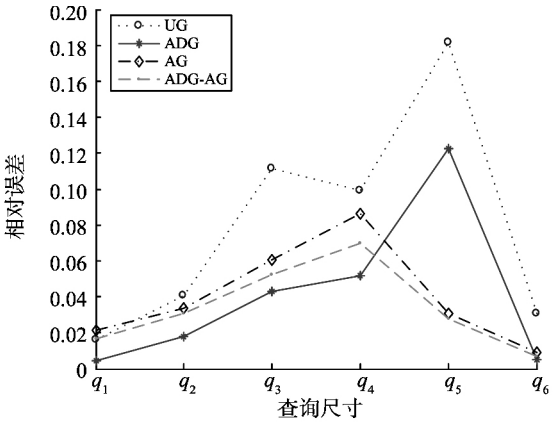


(b) Storage数据集，ε=0.5

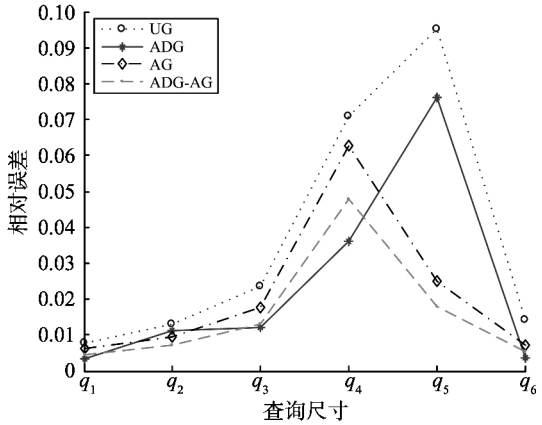
(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net



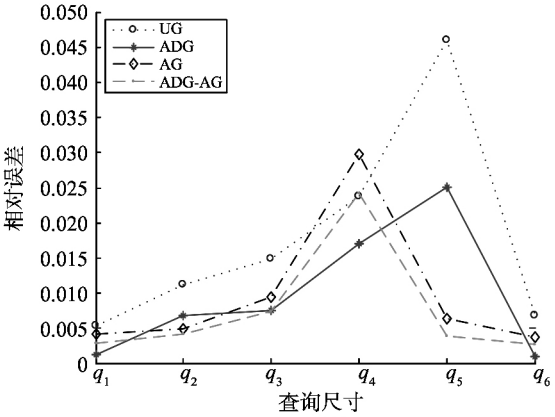
(c) Storage数据集，ε=1.0



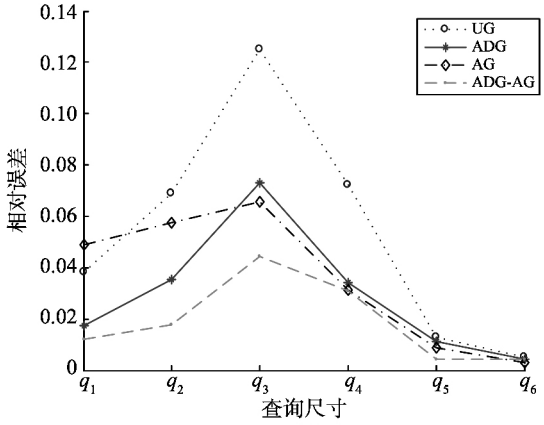
(d) Checkin数据集，ε=0.1



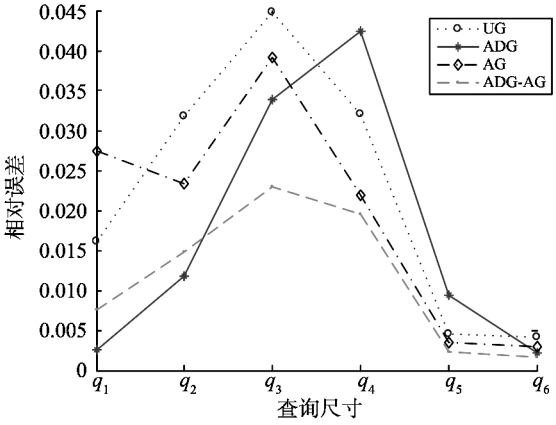
(e) Checkin数据集，ε=0.5



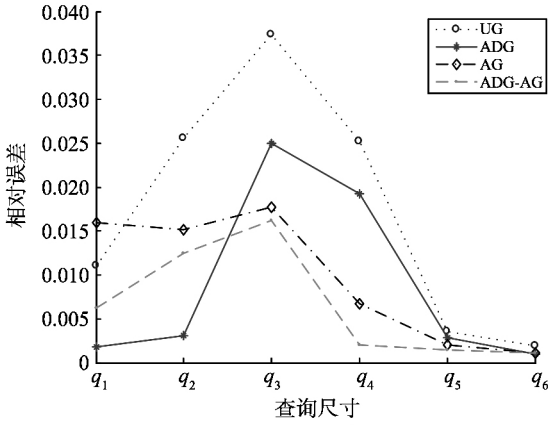
(f) Checkin数据集，ε=1.0



(g) Landmark数据集，ε=0.1



(h) Landmark数据集，ε=0.5



(i) Landmark数据集，ε=1.0

图 2 网格结构划分算法的查询精度比较

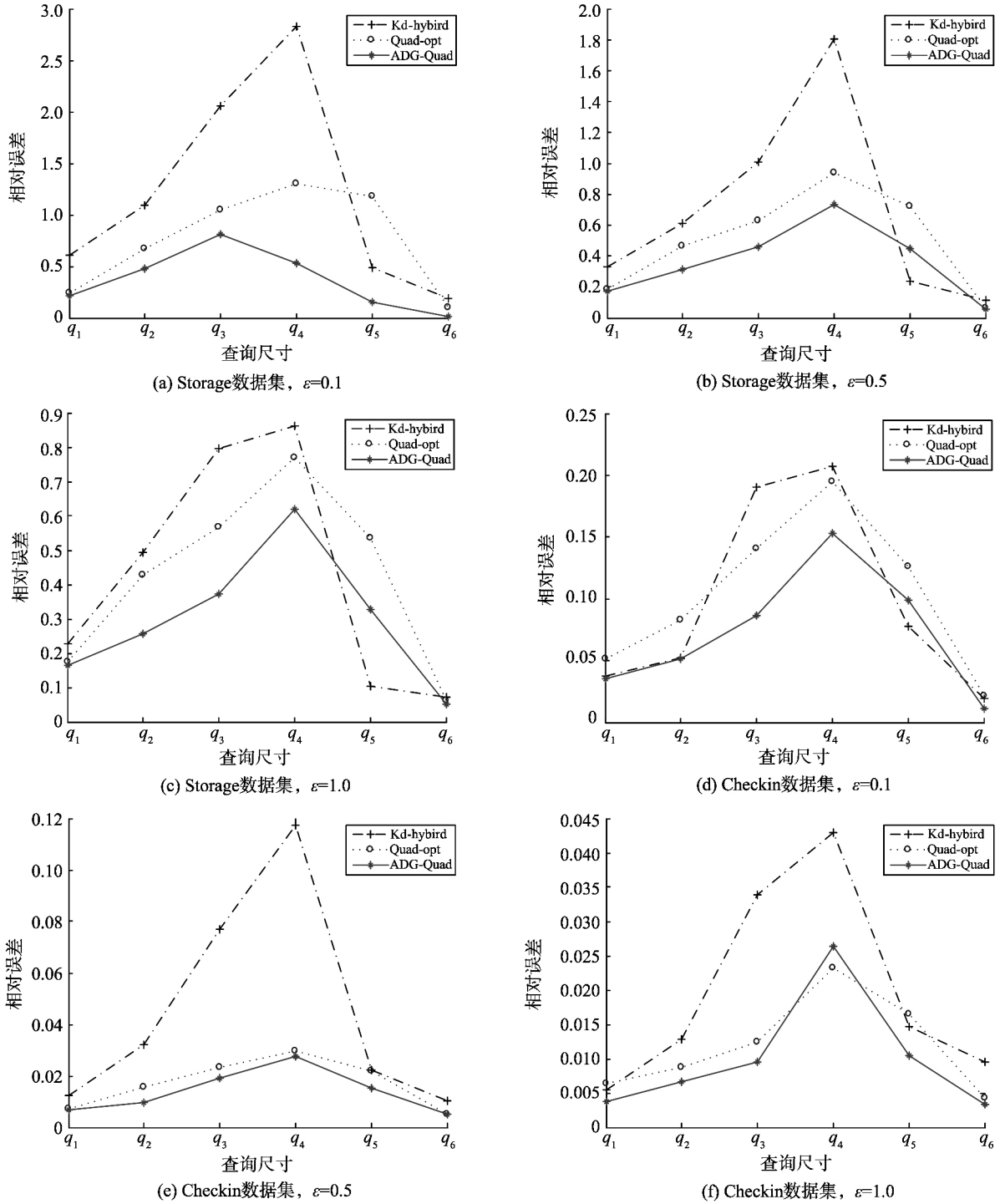
Fig.2 Comparison of query accuracy of partitioning algorithms based on grids

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第 2 组实验比较本文提出的密度自适应网格划分算法与基于树结构的划分发布算法在响应范围查询时的精度。其中,Quad-opt 算法采用的是与数据分布无关的完全四叉树划分,划分层次 $h=8$,按照公式(9)所示的几何策略[11]对不同层次分配差分隐私预算并添加噪声

$$\varepsilon_i = 2^{\frac{h-i}{3}} \cdot \varepsilon \cdot \frac{\sqrt[3]{2}-1}{2^{\frac{h+1}{3}}-1}。\tag{9}$$

密度自适应网格划分算法采用分层策略(记为 ADG-Quad),第 1 层使用密度自适应网格划分,隐私预算分配比例 $\alpha=0.5$; 第 2 层采用数据独立的 Quad-tree 方法将第一层密度单元逐个进行完全四叉树划分,划分层次 $h=4$; 最后按照 Quad-opt 算法的后置处理方法对 2 层划分加噪的结果进行调整。Kd-hybrid 算法也采用分层策略,首先使用数据依赖的 Kd-tree 划分方法对原始数据进行 h_1 层划分; 然后采用数据独立的 Quad-tree 划分策略对上述结果进行 h_2 层划分,$h_1=h_2=4$。

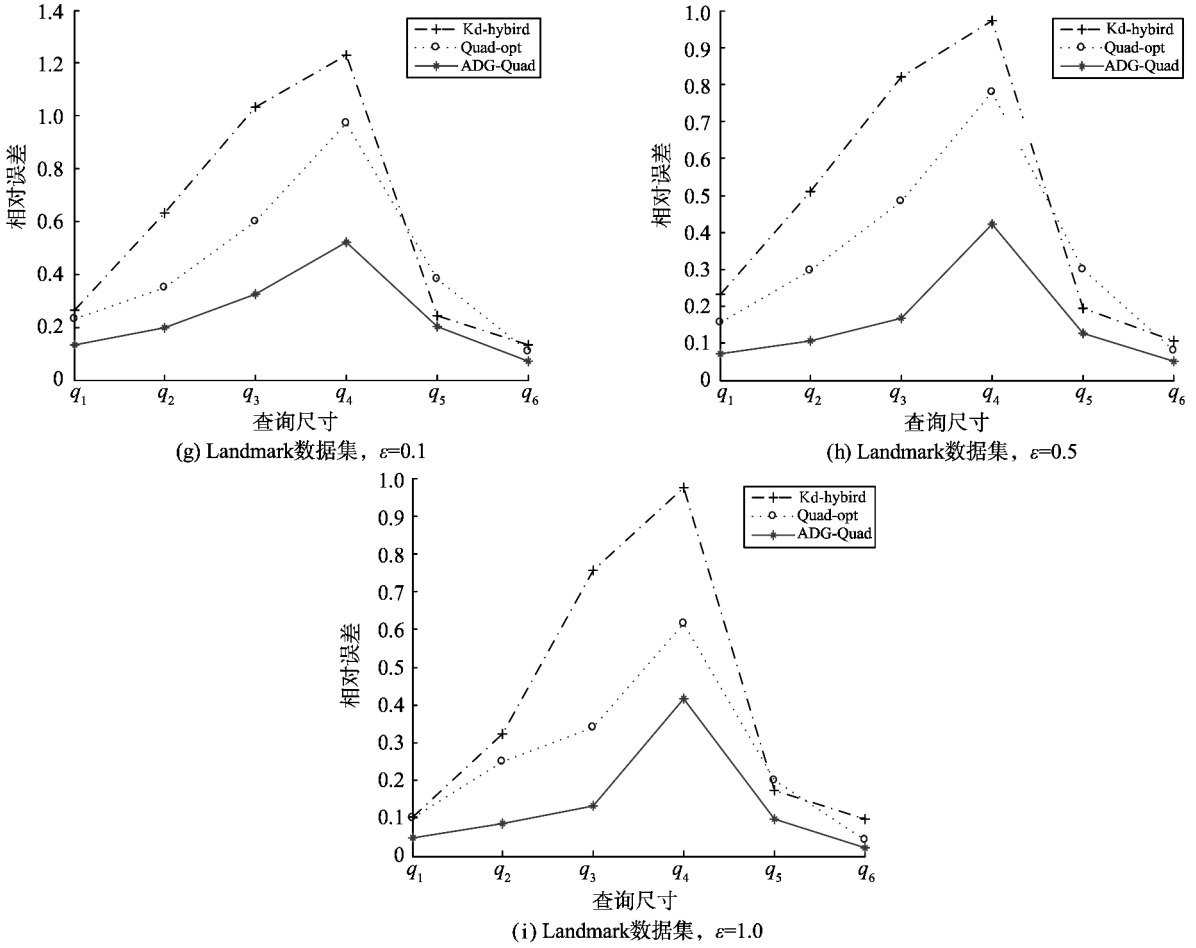


(a) Storage数据集，ε=0.1　　(b) Storage数据集，ε=0.5

(c) Storage数据集，ε=1.0　　(d) Checkin数据集，ε=0.1

(e) Checkin数据集，ε=0.5　　(f) Checkin数据集，ε=1.0

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net



(g) Landmark数据集，ε=0.1　(h) Landmark数据集，ε=0.5

(i) Landmark数据集，ε=1.0

图 3　树结构划分算法的查询精度比较

Fig.3　Comparison of query accuracy of partitioning algorithms based on trees

图 3 显示了上述算法在不同数据集上的范围查询相对误差。位置信息分布较为集中的 Checkin 集具有较小的查询误差，相对分散的 Landmark 数据集次之，位置信息分布最为稀疏的 Storage 集查询误差较大。在相同的数据集下，各种算法的相对误差都随着隐私预算 ε 的增大而逐渐减小，原因同上。比较各种算法的结果不难看出，Kd-hybrid 算法在查询区域大小适中时存在较大的相对误差，因为该算法虽然在第 1 层采用了数据依赖的 Kd 树划分，但其划分依据是数据的平均个数而不是分布密度，所以仍然存在较大的均匀假设误差；当查询区域较大时，Kd-hybrid 算法表现出较低的相对误差，体现出基于树结构的划分发布方法在稀疏数据大范围查询上的优势。Quad-opt 算法通过采用隐私预算几何分配策略和后置处理改进了范围查询的精度。ADG-Quad 算法结合了数据依赖划分与数据独立的树结构划分方法的优势，在各种查询区域下的查询误差都有较好的表现。

3.3　算法运行效率

二维划分发布算法的运行效率主要比较不同算法用于构造索引结构并添加差分隐私噪声形成发布数据的整体用时。一般而言，不依赖于数据特征的独立划分结构要比数据依赖的划分结构节省时间；混合了数据独立划分与数据依赖划分的算法用时介于两者之间。表 1 比较了本文算法与上述基于网格结构和树结构的差分隐私保护算法在不同数据集上的构造时间。划分层次和后置处理方法与前两组实验相同，隐私预算分别取 $\varepsilon=0.1$、$\varepsilon=0.5$、$\varepsilon=1.0$，各种算法运行 500 次求平均时间。

表 2 的结果反映出：同一种算法在不同数据集上的构造时间差异很大，主要原因是实验数据集的样本点数量和分布特性存在较大差异。整体上随着样本点数量的增加，算法的构造时间也增长。相比于网格结构的划分方法 UG 和 AG，本文提出的密度自适应网格划分方法通过对所有数据点进行密度判定形成了“密度聚类”，能够节省不必要的网格划分过程，从而有效提高算法的运行效率。对比基于树结构的划分方法，在划分层次相同的情况下，数据独立的四叉树划分方法 Quad-opt 在各种测试集上表现基本一致。混合了数据

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

依赖划分与数据独立划分的 Kd-hybrid 算法,用时稍高于 Quad-opt 算法。本文提出的 ADG-Quad 方法结合了密度划分与四叉树划分的优势,在样本数据量增大并且分布具有明显的稠密和稀疏区间时具有明显的优势。

表 2　各种算法的平均构造时间

Table 2　Average construction time for different algorithms　　ms

算法类型	算法名称	Storage 集	Landmark 集	Checkin 集
网格结构	UG	60.52	679.67	787.79
	ADG	15.13	442.92	518.89
	AG	66.71	786.36	842.43
	ADG-AG	29.74	458.09	534.18
树结构	Kd-hybrid	969.75	1 174.32	1 227.68
	Quad-opt	878.54	995.38	1 104.96
	ADG-Quad	19.55	748.53	925.42

综合上述分析,本文提出的密度自适应网格划分方法在第 1 层采用的是数据依赖的划分策略,算法的构造时间主要取决于待发布位置信息的具体分布。对于位置信息分布较为分散的情况,密度划分过程需要构建较多的网格,算法耗时较长; 对于样本数据具有明显稠密和稀疏区间的情况,密度划分反而能够节省不必要的网格划分,算法运行效率较高。另外,对于差分隐私保护的位置信息发布而言,隐私空间的分解过程不应当作为算法的瓶颈,因为空间索引结构的构造过程是“一次性代价”,所以在高性能设备和云计算环境下,上述算法构造时间的差异可以忽略不计。

4　总结

本文针对位置大数据的划分发布应用需求,对比研究了现有数据依赖和数据独立划分方法的特点,分析了导致发布误差的主要来源。在此基础上,提出了分层的差分隐私密度自适应网格划分算法。该方法能够结合待发布位置信息的真实分布情况进行有效的密度区块划分,对不同的划分区域采用不同的策略进一步细致划分,在降低均匀假设误差的同时避免了大量空结点引入的噪声误差。结合差分隐私的串行组合特性,对 2 个阶段的划分结果分别添加 Laplace 噪声,实现总体上的 ε-差分隐私保护。通过与传统划分发布算法的实验对比,证明了该算法在改善区域计数查询精度方面的优势; 在样本数据量增大并且分布具有明显的偏态性时,具有更好的运行效率。

位置大数据不但数据量庞大,而且随发布时间快速变化,现有的静态空间分割索引方法无法直接应用于位置大数据的动态发布环境。另一方面,来源于实际生活的位置大数据通常在分布特性上呈现偏斜,导致传统树结构与网格结构划分结果出现大量零节点,直接影响加噪发布结果的查询精度。而本文采取的密度自适应网格划分方法,可以通过对位置信息的密度聚类避免偏态分布带来的问题,进一步结合动态密度聚类方法,将有望解决位置大数据的动态空间划分与发布问题。

参考文献:

[1] 工业和信息化部. 工业和信息化部关于印发大数据产业发展规划(2016—2020 年)的通知,工信部规[2016]412 号[R].北京: 工业和信息化部,2016.

[2] 冯登国, 张敏, 李昊. 大数据安全与隐私保护[J]. 计算机学报, 2014, 37(1): 246-258.
FENG Dengguo, ZHANG Min, LI Hao. Big data security and privacy protection[J]. Chinese Journal of Computers, 2014, 37(1): 246-258.

[3] 刘雅辉, 张铁赢, 靳小龙, 等. 大数据时代的个人隐私保护[J], 计算机研究与发展, 2015, 52(1): 229-247.
LIU Yahui, ZHANG Tieying, JIN Xiaolong, et al. Personal privacy protection in the era of big data[J]. Journal of Computer Research and Development, 2015, 52(1): 229-247.

[4] 孟小峰, 张啸剑. 大数据隐私管理[J], 计算机研究与发展, 2015, 52(2): 265-281.
MENG Xiaofeng, ZHANG Xiaojian. Big data privacy management[J]. Journal of Computer Research and Development, 2015, 52(2): 265-281.

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

[5] 方滨兴，贾焰，李爱平，等. 大数据隐私保护技术综述[J].大数据，2016,2(1):1-18.
FANG Binxing, JIA Yan, LI Aiping, et al. Privacy preservation in big data: a survey [J]. Big Data Research, 2016, 2(1): 1-18.

[6] EMC Education Services. Data science and big data analytics: discovering, analyzing, visualizing and presenting data [M]. New York: Wiley, 2015.

[7] WERNKE M, SKVORTSOV P, DUR F, et al. A classification of location privacy attacks and approaches [J]. Personal and Ubiquitous Computing, 2014, 18: 163-175.

[8] TERROVITIS M, POULIS G, MAMOULIS N, et al. Local suppression and splitting techniques for privacy preserving publication of trajectories [J]. IEEE Transactions on Knowledge & Data Engineering, 2017, 29(7): 1466-1479.

[9] 张啸剑，孟小峰. 面向数据发布和分析的差分隐私保护[J]. 计算机学报，2014，37(4):927-949.
ZHANG Xiaojian, MENG Xiaofeng. Differential privacy in data publication and analysis [J]. Chinese Journal of Computers, 2014, 37(4): 927-949.

[10] 彭慧丽，张啸剑. 基于差分隐私的空间分割研究综述[J]. 燕山大学学报，2016，40(3):263-269.
PENG Huili, ZHANG Xiaojian. Survey on spatial data decomposition with differential privacy [J]. Journal of Yanshan Univisity, 2016, 40(3): 263-269.

[11] CORMODE G, PROCOPIUC C, SRIVASTAVA D, et al. Differentially private spatial decompositions [C] // Proceedings of the 28th International Conference on Data Engineering (ICDE). Washington: IEEE, 2012: 20-31.

[12] 吴英杰，卢清，蔡剑平，等. 基于四分树的差分隐私二维数据划分发布算法[J].华中科技大学学报(自然科学版)，2016，44(3):99-104.
WU Yingjie, LU Qing, CAI Jianping, et al. Differential privacy two-dimensional data partitioning publication algorithm based on quad-tree [J]. Journal of Huazhong University of Science & Technology (Natural Science Edition), 2016, 44(3): 99-104.

[13] ZHANG Jun, XIAO Xiaokui, XIE Xing. PrivTree: a differentially private algorithm for hierarchical decompositions [C] // Proceedings of the 36th ACM International Conference on Management of Data. CA: San Francisco, 2016: 155-170.

[14] QARDAJI W, YANG W N, LI N H. Differentially private grids for geospatial data [C] // Proceedings of the 29th International Conference on Data Engineering (ICDE). Brisbane: IEEE, 2013: 757-768.

[15] MIR D J, ISAACMAN S, CACERES R, et al. DP-Where: differentially private modeling of human mobility [C] // Proceedings of 2013 IEEE International Conference on Big Data, 2013: 580-588.

[16] INAN A, KANTARCIOGLU M, GHINITA G, et al. Private record matching using differential privacy [C] // Proceedings of the 13th International Conference on Extending Database Technology. Lausanne: ACM, 2010: 123-134.

[17] 黄泗勇，陈婷婷，卢清，等. 基于Kd-树的差分隐私二维空间数据划分发布方法[J].山东大学学报(工学版)，2015，45(1):24-29.
HUANG Siyong, CHEN Tingting, LU Qing, et al. Differential privacy two-dimensional dataset partitioning publication algorithm based on Kd-tree [J]. Journal of Shandong University (Engineering Science), 2015, 45(1): 24-29.

[18] 张琳，刘彦，王汝传. 位置大数据服务中基于差分隐私的数据发布技术[J]. 通信学报，2016，37(9):46-54.
ZHANG Lin, LIU Yan, WANG Ruchuan. Location publishing technology based on differential privacy-preserving for big data services [J]. Journal on Communications, 2016, 37(9): 46-54.

[19] 王豪，徐正全，熊礼治，等. CLM:面向轨迹发布的差分隐私保护方法[J]. 通信学报，2017，38(6):85-96.
WANG Hao, XU Zhengquan, XIONG Lizhi, et al. CLM: differential privacy protection method for trajectory publishing [J]. Journal on Communications, 2017, 38(6): 85-96.

[20] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis [C] // Proceedings of the 3th Theory of Cryptography Conference (TCC). Berlin: Springer-Verlag, 2006: 265-284.

[21] DWORK C. Differential privacy [C] // Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP). Berlin: Springer-Verlag, 2006: 1-12.

[22] MCSHERRY F. Privacy integrated queries: an extensible platform for privacy-preserving data analysis [C] // Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD). New York: ACM, 2009: 19-30.

[23] MCSHERRY F, TALWAR K. Mechanism design via differential privacy [C] // Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (Focs'07). Washington: IEEE Computer Society, 2007: 94-103.

(编辑:许力琴)

(C)1994-2022 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net