计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20190360
57(4):847-858,2020

基于本地差分隐私的空间范围查询方法

张啸剑[1] 付 楠[1] 孟小峰[2]
[1](河南财经政法大学计算机与信息工程学院 郑州 450002)
[2](中国人民大学信息学院 北京 100872)
(xjzhang82@ruc.edu.cn)

Towards Spatial Range Queries Under Local Differential Privacy

Zhang Xiaojian[1], Fu Nan[1], and Meng Xiaofeng[2]
[1](*School of Computer & Information Engineering, Henan University of Economics and Law, Zhengzhou* 450002)
[2](*School of Information, Renmin University of China, Beijing* 100872)

Abstract User data collection and analysis with local differential privacy has attracted considerable attention in recent years. The trade-off among the domain size of user data, encoding method, and perturbation method directly constrains the accuracy of spatial range query. To remedy the deficiency caused by the current encoding and perturbating method, this paper employs grid and quadtree to propose an efficient solution, called GT-R, to answer spatial range query. GT-R uses a uniform grid to decompose the data domain, and generate unit sized regions. Based on these regions, an indexing quadtree is built. And then each user encodes his/her data with the quadtree shared from server, and runs the optimal randomized response on each node of the sampled level in the quadtree and reports the sampled level along with the perturbed value. The server accumulates reports from users to reconstruct a quadtree comprising sum of reports from all users. Besides, to boost the accuracy of range query, the server relies on post-processing skill for consistency on the frequency of each node. GT-R method is compared with existing methods on the large-scale real datasets. The experimental results show that GT-R outperforms its competitors, achieves the accurate results of spatial range query.

Key words local differential privacy; spatial range query; grid decomposition; randomized response; constrained inference

摘 要 基于本地差分隐私的用户数据收集与分析得到了研究者的广泛关注.用户数据的值域大小、编码机制以及扰动机制直接制约着空间范围查询的精度.针对现有编码机制与扰动机制难以有效响应空间范围查询的不足,提出了一种基于网格分割与四分树索引的空间范围查询响应方法 GT-R(grid-based quadtree range query),该方法利用网格对用户数据的值域进行均匀分割,产生大小均等的单元

收稿日期:2019-06-10;**修回日期**:2019-12-06
基金项目:国家自然科学基金项目(61502146,61572420,91646203,91746115);河南省自然科学基金项目(162300410006);河南省科技攻关项目(162102310411);河南省教育厅高等学校重点科研项目(16A520002);河南财经政法大学青年拔尖人才资助计划项目
This work was supported by the National Natural Science Foundation of China (61502146, 61572420, 91646203, 91746115), the Natural Science Foundation of Henan Province (162300410006), the Key Technologies Research and Development Program of Henan Province (162102310411), the Research Program of the Higher Education of Henan Educational Committee (16A520002), and the Young Talents Fund of Henan University of Economics and Law.

格区域.同时利用四分树结构对所有单元格区域进行索引.每个用户结合服务器共享的四分树副本,对所拥有的数据进行编码.借助于编码后的四分树进行层次随机采样,并利用优化随机应答机制对所采层次中的结点进行本地扰动处理.服务器利用每个用户的报告值重构四分树索引结构,并响应空间范围查询.GT-R 与现有的编码机制与扰动机制在真实的大规模空间数据集上实验结果表明,其分割精度以及响应范围查询效果优于同类算法.

关键词 本地差分隐私;空间范围查询;网格划分;随机应答;约束推理

中图法分类号 TP392

信息时代的飞速发展,用户空间数据(例如移动用户位置、GPS 位置、家庭住址等)的收集与分析能够改善 IT 企业的软件与服务质量,以及向用户提供更好的个性化体验.然而,不可信第三方对空间数据进行收集与分析时,个人的敏感信息有可能被泄露.例如不可信销售网站通过收集客户的位置信息,可以学习出客户的购物行为模式以及家庭住址.因此,在此情景下,用户通常无法掌控自己的空间隐私数据.本地差分隐私保护技术[1]的出现使得用户可以自己扰动自身数据之后再响应收集者的需求.目前基于本地差分隐私着眼于频率估计、均值估计等研究,而涉及空间范围查询的工作却很少.基于空间数据的范围查询是空间数据分析常用的技术之一.例如图 1 表示 100 万条纽约出租车位置数据(New York City data, NYC)散点图,查询框 Q_1 要求返回曼哈顿医院附近范围内的乘客数量.该查询对应的 SQL 语句可以表示为 Q_1:Select Count(*) from NYC where $-14\leqslant lat\leqslant 23$ and $-10\leqslant lon\leqslant 20$.

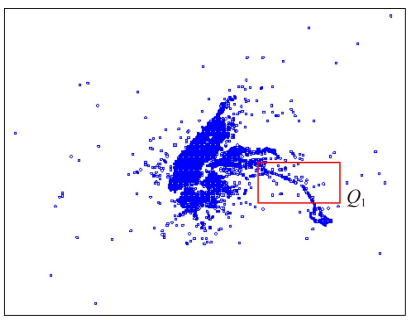


Fig. 1 Spatial range queries on NYC

图 1 基于 NYC 的空间范围查询

在 Q_1 查询中,用户位置所对应的经纬度为敏感信息.因此,用户在共享自身位置之前,需要本地化差分隐私保护处理.而在此过程中存在诸多挑战:1)收集者如何构建高效的空间索引结构收集用户的报告数据;2)由于空间数据的值域通常很大,用户采用什么样的编码机制与扰动机制处理自身的空间数据;3)如何设计高效的后置处理技术来提供空间范围查询精度.总而言之,目前还没有一个行之有效且满足本地差分隐私的空间范围查询方法能够同时克服上述 3 种挑战.为此,本文基于本地差分隐私技术提出了一种空间范围查询方法能够兼顾上述的查询需求.

本文主要贡献有 4 个方面:

1) 为了解决挑战 1,本文首先结合网格与四分树结构提出了 GT-R 方法.在该方法中,收集者首先利用网格结构均分空间数据值域,并形成大小均等的空单元格区域;然后基于所有空单元格区域构建四分树索引结构,并共享给每个用户.

2) 为了有效解决挑战 2,在 GT-R 方法中,每个用户结合收集者发来的四分树副本对自身位置数据进行编码,利用优化随机应答机制与随机采样技术本地扰动自身位置,并把所抽样的结点层次以及扰动值汇报给收集者.

3) 为了有效解决挑战 3,GT-R 方法结合四分树中父亲结点与其孩子结点所蕴含的逻辑关系,设计了一种有效的后置处理技术,该技术能够有效地提高空间范围查询精度.

4) 理论分析了 GT-R 方法满足 ε-本地差分隐私,以及响应范围查询的误差边界.通过真实数据实验分析,该方法具有较高可用性和空间范围查询准确性.

1 相关工作

基于中心化差分隐私的空间范围查询已存在多种方法.文献[1]利用均分网格自适应地划分 2 维空间数据,对单元格添加相应的拉普拉斯噪音,然后结合噪音单元格响应范围查询;文献[2]采用 kd-树对 2 维空间数据进行划分,利用指数机制选择分割中线以避免泄露实际空间点;文献[3]结合完全四分树划分空间数据,并通过结点计数的偏移值来减少

噪音.上述这些方法均是在假设数据收集者是可信的前提下才成立,不能直接应用于本地差分隐私环境.此外,这些方法在进行扰动时常依赖于拉普拉斯机制与全局敏感度的大小.而拉普拉斯机制通常导致误差很大的本地差分隐私统计结果.

目前本地差分隐私研究主要集中于频率估计[4-5]、Heavy hitter[6]、均值估计[7-10]、频繁模式挖掘[11]以及图数据统计[12]等研究.而涉及空间范围查询的工作却很少.文献[5]结合随机应答与1元编码提出了基本Rappor方法,该方法被嵌入谷歌Chrome平台收集用户的会话记录,并估计会话项的频率.然而,1元编码无法应对较大的值域d,通信代价为$\Theta(d)$.为此,文献[5]利用布隆过滤把值域d散列到相对较小的值域中,通信代价为$\Theta(k)$.不同于文献[5],文献[6]采用随机矩阵投影技术对值域d进行编码,通讯代价为$\Theta(\log m)$.此外,文献[6]利用2元本地散列技术对值域d进行编码,其通信代价同样为$\Theta(\log d)$.为了取得更好的扰动精度,文献[7]结合1元编码,提出了优化1元编码与优化本地散列方法.尽管2种方法在较大的值域上取得同样的精度,但本地散列的通信代价较小.

不同于上述方法中的频率估计,文献[8-10]集中研究均值估计.文献[8]结合随机应答机制估计连续区间$[-1,1]$中均值.而文献[9]结合文献[8]的高通信代价与计算代价,提出了高维数据上的均值估计方法.该方法取得较好的估计精度.然而文献[8-9]方法的输出是离散的,并且与输入区间$[-1,1]$差别非常大.为此,文献[10]结合文献[8]与拉普拉斯机制的各自优点,提出了一种输出为连续区间的谱方法,该方法能够支持高维数据上的均值估计与SVM(support vector machines)分类.

上述的频率与均值估计均无法直接应用于空间范围查询.近期文献[13]结合完全2叉树与Hadamard编码响应1维范围查询,然而由于索引结构的不同,该方法无法直接应用于空间范围查询.文献[14]提出了基于本地差分隐私的空间数据聚集方法,该方法结合用户个性化隐私需求与分类树来分析所有用户的位置分布.尽管该方法能够响应范围查询,但与本文的需求存在不同:文献[14]中的层次结构划分空间数据只是在语义层面,缺少实际的空间索引结构.因此,针对上述方法的不足,本文提出了一种基于网格与四分树结构的空间范围查询方法,该方法不但能够适应于大规模空间数据,还能够比较精确地响应不同粒度的范围查询.

2 定义与问题

2.1 本地差分隐私

不同于中心化差分隐私保护技术,本地差分隐私技术通常要求用户在本地保护自己的数据,把扰动之后的数据报告给不可信的收集者,从而实现隐私不被泄露.本地差分隐私的形式化定义为:

定义1. ε-本地差分隐私.给定一个随机算法A及其定义域$Dom(A)$和值域$Range(A)$,若算法A在任意2条不同空间位置l与l'($l,l'\in Dom(A)$)上得到相同输出结果O($O\in Range(A)$)的概率满足下列不等式,则A满足ε-本地差分隐私.

$$Pr[A(l)\in O]\leqslant \exp(\varepsilon)\times Pr[A(l')\in O], \quad (1)$$

其中ε为隐私预算,其值越小则算法A的隐私保护程度越高.

随机应答机制[15]与拉普拉斯机制[16]是实现本地差分隐私的常用技术.拉普拉斯机制通常需要计算出某操作的全局敏感性,利用拉普拉斯分布生成噪音对用户数值进行扰动.不同于拉普拉斯机制,随机应答机制在用户发送数据l_i之前,对其进行随机扰动.该机制的原始思想是用户在响应敏感的布尔问题时,以概率p真实应答,以$1-p$的概率给出相反的应答.为了使随机应答机制满足ε-本地差分隐私,通常设置$p=\exp(\varepsilon)/(1+\exp(\varepsilon))$或者更大的值(例如1/2),收集者获得所有应答后,即可对真实应答进行分析估计.

定理1[17]. 给定空间数据集D和n个随机算法$A_1,A_2,\cdots,A_n$,且A_i满足ε_i-本地差分隐私,则在D上的序列组合满足ε-本地差分隐私,且$\varepsilon=\sum_{i=1}^{n}\varepsilon_i$.

2.2 空间数据范围查询

空间数据通常包括空间位置信息、空间轨迹信息等,以2维散点图形式描述用户的空间位置.而空间范围查询是指在某一范围内所包含的用户位置个数.设$Dom(D)$为空间数据集D的值域,$l_i(x_i,y_i)$表示第i个用户的位置,其中x_i与y_i表示相应的经纬度.下面给出空间范围查询的形式化表示.

定义2. 空间范围查询.给定n个用户与一个空间范围查询框Q($Q\in D$)且$Q=[a,b]\times[c,d]$,则Q的查询结果可以表示为

$$Q_{[a,b][c,d]}=\sum_{i=1}^{n} I_{a\leqslant x_i\leqslant b,c\leqslant y_i\leqslant d}, \quad (2)$$

其中,I是标识函数,其值为1表示第i个用户的空间位置在Q内,其值为0表该用户位置不在Q内.

2.3 问题描述

给定多个用户与空间数据集 D,设$[a,b]\times[c,d]$为任意范围查询框,Q 为查询框$[a,b]\times[c,d]$中真实的响应结果,$\tilde{Q}$ 表示经过算法 A 处理后的结果.采用平方误差$(\tilde{Q}-Q)^2$度量$[a,b]\times[c,d]$的查询精度.本文要解决的问题是在设计满足本地差分隐私的空间范围查询方法的同时,要尽可能获得精度较高的查询结果.

3 基于本地差分隐私的空间范围查询方法

3.1 空间范围查询的原则

基于相关工作的分析,在设计新的基于本地差分隐私的空间范围查询方法时需要考虑 2 个原则:

1) 针对现有编码机制无法直接应对空间 2 维数据,所设计的方法尽可能利用空间几何结构对空间数据进行分割与索引;

2) 针对现有只对单个点的频率估计方法无法适应于 2 维空间范围查询,所设计的方法尽量能够保证较高的查询精度.

针对原则 1 与原则 2,本文利用网格与四分树对大规模空间数据进行分割与编码,在此基础上提出了一种有效的空间范围查询方法 GT-R,该方法能够满足本地差分隐私且输出较高精度的查询结果.

3.2 基于整个空间值域的范围查询方法

给定 n 个用户,每个用户拥有自己的空间位置.设 $c(l_i)$表示 $l_i(l_i\in Dom(D))$的用户计数.则式(2)可以重新表示为

$$Q_{[a,b][c,d]}=\sum_{x_i=a}^{b}\sum_{y_i=c}^{d}c(l_i(x_i,y_i)). \tag{3}$$

在响应空间范围查询 $Q=[a,b]\times[c,d]$时,最直接的方法是每个用户按照空间位置所在的值域 $Dom(D)$进行 2 进制编码,结合随机应答机制扰动 2 进制编码,收集者汇总 $Dom(D)$中所有的位置后再响应 Q 查询.假设 $error(A)$表示某种随机应答机制 A 扰动用户位置所产生的误差.则采用 A 直接响应范围查询Q 所产生的最坏误差为

$$Var(Q-\tilde{Q})=(b-a)\times(d-c)\times error(A), \tag{4}$$

其中,$(b-a)\times(d-c)$表示查询框 Q 的面积.

响应查询 Q 的误差随着 Q 的面积呈线性增加.例如利用拉普拉斯机制报告每个位置并响应查询 Q 所产生的误差为 $8(b-a)(d-c)n/\varepsilon^2$.由于直接方法是基于整个空间值域 $Dom(D)$对用户位置进行编码,过大的值域导致范围查询误差与查询框面积线性相关.因此,本文基于网格划分技术将空间值域分割成均匀单元格区域,将每个用户的位置信息压缩到一个单元格区域中.每个用户结合单元格区域所构建的四分树对自身位置进行编码.收集者通过重构四分树来响应空间范围查询.

3.3 基于网格分割的空间范围查询方法

网格分割是空间划分常用技术之一,其主要特点是在不涉及实际数据分布的情况下将空间分割成大小相等或不等的单元格区域.单元格区域是空间范围查询的最小响应单元.

3.3.1 基于均匀网格的空间范围查询方法

本节首先基于均匀网格提出 GT-R 算法,该算法包括基于网格与四分树结构的空间划分和索引、收集者重构四分树、用户利用四分树扰动自身位置以及响应查询 4 种操作.该算法具体细节详见算法 1:

算法 1. GT-R 算法.

输入:数据集 D、用户的位置 $l_i(1\leqslant i\leqslant n)$、查询框 Q、隐私预算 ε;

输出:满足本地差分隐私的 $\tilde{Q}$.

① $\boldsymbol{G}_{m\times m}\leftarrow UG(Dom(D))$; /* 利用 UG 方法均分 $Dom(D)$的空间值域 */

② 收集者构建一个包含 m^2 个叶子结点的空四分树 T;

③ 收集者与用户共享四分树 T;

④ for each user u_i do

⑤ u_i 发送 $(h_i^*,\boldsymbol{Z}_i)\leftarrow LRR(l_i,\varepsilon)$; /* 用户扰动 l_i */

⑥ end for

⑦ for each level h of T in $[1,1+2\log_4 m]$ do /* 收集者利用 n 个用户的$(h_i^*,\boldsymbol{Z}_i)$重构 T */

⑧ for each $(h_i^*,\boldsymbol{Z}_i)$ do

⑨ if $h_i^*=h$ then

⑩ for each node v_j in h level do

⑪ $c'(v_j)\leftarrow c'(v_j)+z_j$; /* h 层聚集每个报告值,z_j 是向量 $\boldsymbol{Z}_i$ 中与结点 v_j 对应值 */

⑫ end for

⑬ end if

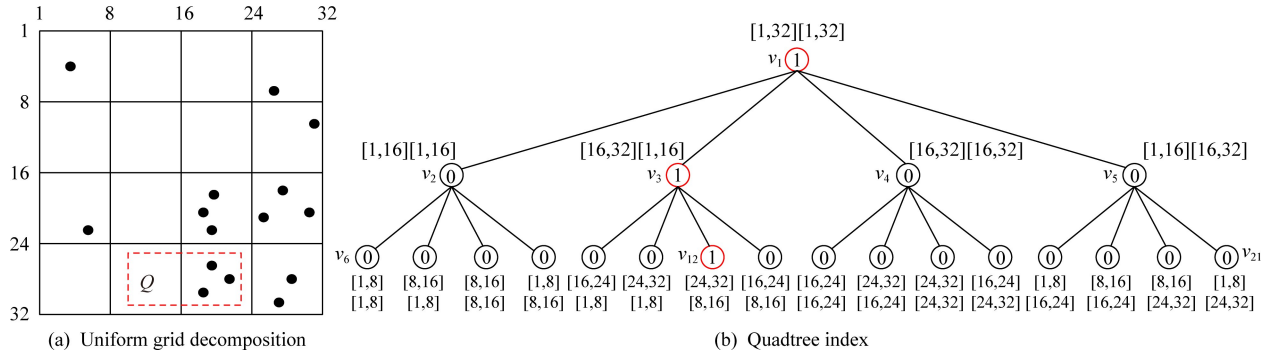
⑭ end for

⑮ end for

⑯ return $\tilde{Q}\leftarrow TDT(T,Q)$.

GT-R 算法基于文献[2-3]中的 UG 方法均匀分割值域 $Dom(D)$,即 $m=\sqrt{|Dom(D)|\varepsilon/10}$

(行①).例如图 2 所示,给定空间数据集 D 的域值范围为 $Dom(D)=32\times32$,利用均匀网格划分方法将其划分为 $4\times4(m=4)$的单元格.收集者创建空树 T 并共享给每个用户(行②③),如图 2 所示.每个用户通过 LRR(local randomized response)方法本地扰动自己位置数据并报告给收集者(行④~⑥).收集者聚集所有用户的报告值后,重构四分树 T(行⑦~⑮).遍历 T 获得最终的范围查询结果(行⑯).



(a) Uniform grid decomposition (b) Quadtree index

Fig. 2 Grid decomposition and quadtree index

图 2 网格划分与四分树索引

用户结合已被赋值的四分树 T,向收集者报告自己的空间位置.LRR 算法详见算法 2:

算法 2. LRR 算法.

输入:用户 u_i 的位置 $l_i(1\leqslant i\leqslant n)$、$T$、隐私预算 ε;

输出:扰动之后的值$(h_i^*,\boldsymbol{Z}_i)$.

① $\boldsymbol{Z}_i\leftarrow\varnothing$; /* 向量 $\boldsymbol{Z}_i$ 用来存放的扰动值 */

② 用户 u_i 遍历 T 来判断 l_i 所属的路径 p;

③ for each node v_i in T do /* 编码过程 */

④ if v_i in p then

⑤ $w_i(v_i)\leftarrow1$; /* 权重为 1 */

⑥ else

⑦ $w_i(v_i)\leftarrow0$;

⑧ end if

⑨ end for

⑩ 用户 u_i 随机采样 $h_i^*\leftarrow[1,1+2\log_4 m]$;

⑪ 生成一个向量 $\boldsymbol{V}_i\in\{0,1\}^{4^{2\log_4 m+1-h_i^*}}$;

⑫ for each $w_i(v_j)$ in V_i do /* 扰动过程 */

⑬ if $w_i(v_j)=1$ then

⑭ $Pr[z_i=1|w_i(v_j)]=1/2$;

⑮ else

⑯ $Pr[z_i=0|w_i(v_j)]=1/(1+\exp(\varepsilon))$;

⑰ 以概率 $Pr[z_i|w_i(v_j)]$采样一个伯努利变量 z_i;

⑱ $\boldsymbol{Z}_i\leftarrow\boldsymbol{Z}_i\cup\{z_i\}$;

⑲ end if

⑳ end for

㉑ return$(h_i^*,\boldsymbol{Z}_i)$.

收集者把空的四分树 T 共享给用户后,每个用户均拥有一棵 T 的副本(算法 1 行③).结合 LRR 算法,用户首先遍历四分树 T,寻找到包含自身位置的路径,并判断自己的空间位置属于 T 中哪个叶子结点(算法 2 行②),找到所属叶子结点之后,该叶子结点至根结点路径的权重均被赋值为 1,其他路径的权重为 0(算法 2 行③~⑨).例如给定用户 u_i 的空间位置 $l_i(x_i=28,y_i=12)$.u_i 结合图 2 中的四分树 T,判断 $l_i(x_i=28,y_i=12)$属于结点 v_{12}([24,32]×[8,16]),则路径 $v_{12}-v_3-v_1$ 上的权重均被赋值为 1,如图 2 所示.接下来每个用户在 T 中随机选择一层,并产生由 0/1 构成的向量(算法 2 行⑩⑪).例如,随机选择图 2 中四分树的第 2 层,则所生成的向量为 $\boldsymbol{V}_i=(0,1,0,0)$.最后利用优化随机应答机制[7]生成报告结果(算法 2 行⑫~⑳).由于每个用户在 T 中随机选择一层报告,则 LRR 算法最坏的通信代价为 $O(2\log_4^m+1)$.

定理 2. LRR 算法满足 ε-本地差分隐私.

证明. 假设 $\boldsymbol{V}_i$ 是用户 u_i 基于四分树 T 随机生成的向量且结点个数 $|\boldsymbol{V}_i|=4^{2\log_4 m+1-h_i^*}$,其中 h_i^* 为该用户基于 T 随机选择的树层.$\boldsymbol{Z}_i$ 为 LRR 算法响应 $|\boldsymbol{V}_i|$ 次的输出结果.结合定理 1 则等式成立:

$$Pr[\boldsymbol{Z}_i|\boldsymbol{V}_i]=\prod_{j=1}^{|\boldsymbol{V}_i|}Pr[z_j=1|w_j(v_j)],$$

同理,给定 $\boldsymbol{V}_i'$,等式成立:

$$Pr[\boldsymbol{Z}_i|\boldsymbol{V}_i']=\prod_{j=1}^{|\boldsymbol{V}_i'|}Pr[z_j=1|w_j'(v_j)].$$

则:

$$\frac{Pr[\boldsymbol{Z}_i|\boldsymbol{V}_i]}{Pr[\boldsymbol{Z}_i|\boldsymbol{V}'_i]}=\frac{\prod_{j=1}^{|\boldsymbol{V}_i|}Pr[z_j=1|w_j(v_j)]}{\prod_{j=1}^{|\boldsymbol{V}'_i|}Pr[z_j=1|w'_j(v_j)]}\leqslant$$

$$\frac{Pr[z_j=1|w_j(v_j)=1]Pr[z_j=0|w_j(v_j)=0]}{Pr[z_j=1|w'_j(v_j)=0]Pr[z_j=0|w'_j(v_j)=1]}=$$

$$\frac{1/2}{1/(1+\exp(\varepsilon))}\times\frac{\exp(\varepsilon)/(1+\exp(\varepsilon))}{1/2}=\exp(\varepsilon).$$

根据定义 1 可知,LRR 算法满足 ε-本地差分隐私. 证毕.

尽管通过 LRR 算法可以重构四分树 T,但我们期望 T 中每个结点的噪音计数满足无偏性.

定理 3. 假设 v_i 是收集者重构四分树 T 中的任意结点,$c(v_i)$与 $c'(v_i)$分别是结点 v_i 中真实的用户位置数与估计数,则无偏估计 $E[c'(v_i)]=c(v_i)$ 成立.

证明. 设 l 为结点 v_i 所在的层次,n_l 为该层次中的用户数,也是结点 v_i 中的用户报告数目.根据算法 1 的行⑪$c'(v_i)=c'(v_i)+z_i$ 可知,n_l 个用户把自己的位置信息汇总到 l 层每个结点中去.设 $I(v_i)$表示结点 v_i 中 1 的个数.为了证明方便,结合算法 2 设 $Pr[z_i=1|w_i(v_j)]=p$,以概率 p 生成 $z_i=1$,否则以概率 $Pr[z_i=0|w_i(v_j)]=q$ 生成 $z_i=0$.则$I(v_i)=p\times c'(v_i)+q\times(n_l-c'(v_i))$成立.进而可以获得随机变量 $c'(v_i)$:

$$c'(v_i)=\frac{I(v_i)-q\times n_l}{p-q},$$

则:

$$E[c'(v_i)]=\frac{E[I(v_i)]-q\times n_l}{p-q}$$

成立.

根据 $I(v_i)=p\times c'(v_i)+q\times(n_l-c'(v_i))$可知:

$$E[c'(v_i)]=$$

$$\frac{p\times E[c'(v_i)]+q\times(n_l-E[c'(v_i)])-q\times n_l}{p-q}.$$

我们期望 $E[c'(v_i)]=c(v_i)$,则:

$$E[c'(v_i)]=$$

$$\frac{p\times c(v_i)+q\times(n_l-c(v_i))-q\times n_l}{p-q}=$$

$$\frac{c(v_i)\times(p-q)}{p-q}=c(v_i)$$

成立. 证毕.

由于每个用户本地扰动自身的空间位置,T 中每个结点的计数不可避免地产生误差.定理 4 给出了每个结点所产生的方差.

定理 4. 假设 v_i 是收集者重构四分树 T 中的任意结点,n_l 为 l 层次中的用户数,$p=1/2$,$q=1/(1+\exp(\varepsilon))$,$c(v_i)$与 $c'(v_i)$分别是结点 v_i 中真实的用户位置数与估计数,则:

$$Var[c'(v_i)]=\frac{n_l\times 4\times\exp(\varepsilon)}{(\exp(\varepsilon)-1)^2}+c(v_i).$$

证明. 根据定理 3 可知,$c'(v_i)=\frac{I(v_i)-q\times n_l}{p-q}$,则:

$$Var[c'(v_i))]=Var\left[\frac{I(v_i)-q\times n_l}{p-q}\right]=$$

$$\frac{Var[I(v_i)]}{(p-q)^2}=\frac{n_l\times q\times(1-q)}{(p-q)^2}+$$

$$\frac{c(v_i)\times(1-(p+q))}{p-q}=\frac{n_l\times 4\times\exp(\varepsilon)}{(\exp(\varepsilon)-1)^2}+c(v_i).$$

证毕.

由文献[12]可知,T 中很多结点真实位置计数非常小,因此,$Var[c'(v_i)]\approx n_l\times 4\times\exp(\varepsilon)/(\exp(\varepsilon)-1)^2$.进而可知每个结点的方差仅与四分树每层所分配的用户个数 n_l 线性相关.

尽管通过定理 4 可以估算出每个结点产生的方差,而如何度量 $c(v_i)$与 $c'(v_i)$之间的最大偏差是个挑战性问题.

定理 5. 假设 v_i 是收集者重构四分树 T 中的任意结点,设 l_j 是第 j 个用户的位置.$c(v_i)$与 $c'(v_i)$分别是结点 v_i 中真实的用户位置数与估计数,则等式至少以概率 $1-\beta$ 成立:

$$|c'(v_i)-c(v_i)|=O(\sqrt{n\times\mathrm{lb}(1/\beta)}/\varepsilon\sqrt{h}),$$

其中,n 为用户个数,$h=1+2\log_4 m$ 为四分树高度.

证明. 设 $c'(v_i)-c(v_i)$为一个随机变量,根据定理 3 可知其均值为 0.根据算法 1 与算法 2 可知:

$$|c'(v_i)-c(v_i)|=$$

$$\left|\sum_{j=1}^{n_l}\frac{z_j-q}{p-q}-\sum_{j=1}^{n_l}\frac{w_j(v_i)-q}{p-q}\right|=$$

$$\sum_{j=1}^{n_l}\left|\frac{z_j-w_j(v_i)}{p-q}\right|. \tag{5}$$

其中,$w_j(v_i)$与 z_j 分别表示第 j 个用户的真实值与估计值.

同理可知 $z_j-w_j(v_i)$也是一个随机变量.由于 $z_j\in\{0,1\}$,$w_j(v_i)\in\{0,1\}$,则 $z_j-w_j(v_i)\in\{-1,0,1\}$.

随机变量 $z_j-w_j(v_i)$的方差可以表示为

$$Var[z_j-w_j(v_i)]=E[(z_j-w_j(v_i))^2]-$$

$$(E[z_j-w_j(v_i)])^2=E[(z_j-w_j(v_i))^2]=$$

$$\frac{\exp(\varepsilon)+3}{4(1+\exp(\varepsilon))}-\left(\frac{\exp(\varepsilon)-1}{4\times(1+\exp(\varepsilon))}\right)^2=\frac{3\times\exp(\varepsilon)+2}{4\ (1+\exp(\varepsilon))^2}+\frac{3}{16}=O(1/\varepsilon^2).$$

此外,由式(5)可知:

$$|z_j-w_j(v_i)|\leqslant 2(\exp(\varepsilon)+1)/(\exp(\varepsilon)-1)$$

成立.根据 Bernstein 不等式可知:

$$Pr\left[|c(v_i)-c'(v_i)|\geqslant\lambda\right]=Pr\left[\sum_{j=1}^{n_l}\left|\frac{z_j-w_j(v_i)}{p-q}\right|\geqslant\lambda\right]\leqslant 2\times\exp\left(-\frac{\frac{1}{2}\lambda^2}{\sum_{j=1}^{n_l}Var\left[\frac{z_j-w_j(v_i)}{p-q}\right]+\frac{\lambda}{3}\times\frac{2(\exp(\varepsilon)+1)}{\exp(\varepsilon)-1}}\right)=2\times\exp\left(-\frac{\frac{1}{2}\lambda^2}{n_lO(1/\varepsilon^2)+\lambda O(1/\varepsilon)}\right).$$

结合上述不等式可知 $\lambda=O(\sqrt{n\times\mathrm{lb}(1/\beta)}/\varepsilon\sqrt{\hbar})$,此时 $n_l=n/\hbar$.则可知 $|c'(v_i)-c(v_i)|<\lambda$ 至少以概率 $1-\beta$ 成立. 证毕.

结合收集者重构的四分树 T 响应空间范围查询 Q,具体细节如算法 3 描述.

算法 3. TDT 算法

输入:四分树 T、空间范围查询 Q;

输出:响应结果 $\tilde{Q}$.

① $T\leftarrow post\_processing(T)$;
② 标记根结点 v_1,$unvisited(v_1)=1$;
③ while $unvisited(v_i)=1$ do /* 自顶向下遍历 T */
④ 标记结点 v_i,$unvisited(v_i)=0$;
⑤ if v_i 与 Q 不相交 then
⑥ 忽略 v_i;
⑦ else if Q 完全包含 v_i then
⑧ $\tilde{Q}\leftarrow\tilde{Q}\cup v_i$;
⑨ else if v_i 不是叶子节点且与 Q 部分相交 then
⑩ 标记 v_i 的孩子结点 v,$unvisited(v)=1$;
⑪ $\tilde{Q}\leftarrow TDT(v,Q)$;
⑫ else if v_i 是叶子节点且与 Q 部分相交 then
⑬ 计算 Q 和 v_i 的重叠部分;
⑭ $\tilde{Q}\leftarrow\tilde{Q}\cup overlap(Q,v_i)$;
⑮ end if
⑯ end while
⑰ return $\tilde{Q}$.

收集者利用 GT-R 算法重构四分树 T 之后,首先进行后置处理(行①),再利用 TDT 算法自动向下遍历 T 即可响应空间范围查询 Q.如果某结点与 Q 无交叉,则忽略该结点(行⑤⑥).若 Q 完全包含该结点,即可把该结点中的空间点数添加响应结果中(行⑦⑧).若 Q 部分包含该结点且该结点不是叶子结点,则重新遍历该结点的孩子结点(行⑨~⑪).否则计算 Q 与该结点的重合部分,并把重合部分中的空间点数添加响应结果中(行⑫~⑭).

定理 6. 结合算法 1 与算法 3,任意空间范围查询 Q 的最大误差 $error(Q)$ 满足:

$$error(Q)\leqslant 8\times\frac{n}{\eta}\times\frac{4\times\exp(\varepsilon)}{(\exp(\varepsilon)-1)^2},$$

其中,η 为某一常数.

证明. 四分树中每个结点产生的误差只与该结点所在层次中用户个数线性相关.因此,只要自顶向下估算出每一层最多有多少个结点响应 Q,即可估算出 Q 的查询误差上界.任意给定 T 中一层 l($1\leqslant l\leqslant\hbar$),$c_l$ 表示该层中响应查询 Q 的最多结点个数.设 $C(Q)$ 表示 T 中能够响应查询 Q 的结点总数,则 $C(Q)=\sum_{l=1}^{\hbar}c_l$ 成立.T 为四分树,则 $c_l\leqslant 8\times 2^{\hbar-l}$.其原因是 Q 查询框与 $l-1$ 层中至多有 8 个结点相交.因此,每一层最多产生的误差为 $8\times 2^{\hbar-l}\times Var[c'(v_i)]$.

结合 $Var[c'(v_i)]\approx n_l\times 4\times\exp(\varepsilon)/(\exp(\varepsilon)-1)^2$,可知不等式成立:

$$error(Q)\leqslant\sum_{l=1}^{\hbar}8\times 2^{\hbar-l}\times Var[c'(v_i)]=\frac{\sum_{l=1}^{\hbar}8\times 2^{\hbar-l}\times 4n_l\times\exp(\varepsilon)}{(\exp(\varepsilon)-1)^2}=8\times\frac{4\times\exp(\varepsilon)}{(\exp(\varepsilon)-1)^2}\times\sum_{l=1}^{\hbar}2^{\hbar-l}\times n_l. \quad (6)$$

为了估算 $error(Q)$,设 $f=\sum_{l=1}^{\hbar}2^{\hbar-l}\times n_l$.为了使不等式(6)成立,可以利用优化技术解决该问题.

$$\begin{cases}\min(f)=\min\left(\sum_{l=1}^{\hbar}2^{\hbar-l}\times n_l\right)\\ \text{subject to }\sum_{l=1}^{\hbar}n_l=n.\end{cases}$$

基于目标函数 $\min(f)$ 与柯西-施瓦茨不等式可知不等式成立：

$$\left(\sum_{l=1}^{h} n_l\right)\left(\sum_{l=1}^{h} 2^{h-l}\times n_l\right)\geqslant\left(\sum_{l=1}^{h}\sqrt{n_l\times 2^{h-l}\times n_l}\right)^2. \tag{7}$$

当 $n_l=\eta\times 2^{h-l}\times n_l$ 时，则不等式(7)中的等号成立.可知 $2^{h-l}=1/\eta$.该等式说明四分树中响应查询 Q 的结点个数是定值，则 $\min(f)=n/\eta$.进而可知：

$$error(Q)\leqslant 8\times\frac{4\times\exp(\varepsilon)}{(\exp(\varepsilon)-1)^2}\times\frac{n}{\eta}.$$ 证毕.

结合定理 6 与式(4)可知，GT-R 算法与最直接方法均可以采用方差度量误差.若在最直接方法中每个用户也利用优化随机应答机制扰动自身位置，可知 $error(A)=4n\times\exp(\varepsilon)/(\exp(\varepsilon)-1)^2$.因此，只要范围查询 Q 的查询面积大于 $8/\eta$，即可知 GT-R 算法优于最直接方法.

3.3.2 基于重构四分树的求精处理方法

四分树本身蕴含着内在的计数特征，即任何一个非叶子结点中的计数等于其孩子结点中计数的和.随机给定一个非叶子结点 v_i，$c(v_i)$ 为其计数，u_j 是 v_i 的孩子结点，则 $c(v_i)=\sum_{u_j\in child(v_i)} c(u_j)$ 成立.此外，结点 v_i 中的计数与其孩子结点计数之和满足无偏估计，即 $E[c(v_i)]=E[\sum_{u_j\in child(v_i)} c(u_j)]$ 成立.然而，由于算法 2 中的本地扰动噪音与用户基于四分树的随机采样，打乱了四分树本身内在的约束机制.因此，本文利用文献[18]中的一致性约束处理方法，对四分树结点中的计数进行求精处理(算法 3 行①).设 $c'(v_i)$ 和 $\bar{c}(v_i)$ 分别表示结点 v_i 中的用户位置的估计数与求精处理之后的计数.求精处理包含 2 种操作：

1) 自底向上的平均化处理

$$c''(v_i)=\begin{cases}c'(v_i), & v_i \text{ 是叶子结点},\\ \dfrac{4^i-4^{i-1}}{4^i-1}c'(v_i)+\dfrac{4^{i-1}-1}{4^i-1}\displaystyle\sum_{u_j\in child(v_i)} c''(u_j), & \\ \quad \text{其他}, & \end{cases}$$

其中，u_j 表示结点 v_i 的孩子结点.

2) 自顶向下的均值一致性处理

$$\bar{c}(v_i)=\begin{cases}c''(v_i), & v_i \text{ 是叶子结点},\\ c''(v_i)+\dfrac{1}{4}\left[\bar{c}(p(v_i))-\displaystyle\sum_{u_j\in child(v_i)} c''(u_j)\right], & \\ \quad \text{其他} & \end{cases}$$

其中，$p(v_i)$ 表示结点 v_i 的父亲结点.

结合图 2，以例子说明求精处理过程.给定只有 2 层的四分树 T，v_i 为其根结点.假设 $c'(v_i)=5$，其孩子结点的计数分别为 $c'(u_1)=2$，$c'(u_2)=1$，$c'(u_3)=1$，$c'(u_4)=2$.由于 $c'(v_i)\neq c'(u_1)+c'(u_2)+c'(u_3)+c'(u_4)$，则需要求精处理.处理之后 $c'(v_i)=3.9$，$c'(u_1)=2$，$c'(u_2)=1$，$c'(u_3)=1$，$c'(u_4)=2$.

4 实验结果与分析

实验平台是 4 核 Intel i7-4790 CPU(4 GHz)、8 GB 内存、Win7 系统.所有算法均采用 Python 实现.实验采用 2 个数据集 Checkin 与 Landmark，其中 Checkin 数据集从基于地理位置的社交网站 Gowalla 获取，该数据集记录了在 2009-02—2010-10 期间，Gowalla 用户签到的时间和位置信息，包含 100 万条记录；Landmark 数据集从 infochimps 平台获得，该数据集记录了 2010 年人口普查时用户到过的美国 48 个州的地标位置，总共包含 87 万条数据.2 种数据集具体细节与可视化结果分别如表 1 与图 3 所示.

结合上述数据集，采用相对误差(relative error，RE)度量 KRR[19]，RAPPOR[5]，PSDA[14]，OUE[7]，GT-R，N-PSDA，N-GT-R，QT-RAPPOR，QT-KRR 方法的范围查询精度.其中 N-PSDA 与 N-GT-R 表示无空间结构索引下的方法；QT-RAPPOR 与 QT-KRR 表示四分树索引下的 RAPPOR 与 KRR 方法.RE 的表示为

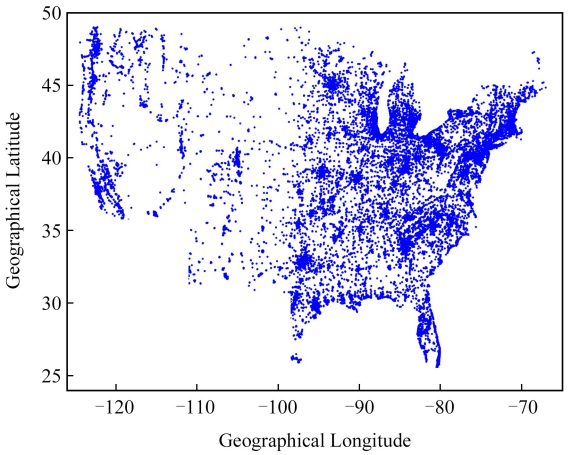
$$RE(\tilde{Q}(D))=\frac{|\tilde{Q}(D)-Q(D)|}{Q(D)}, \tag{8}$$

其中，$Q(D)$ 表示 D 上真实的范围查询结果，$\tilde{Q}(D)$ 表示 D 上范围查询的噪音结果.

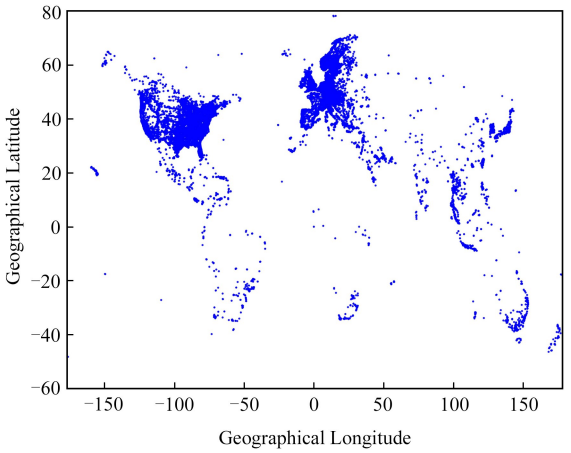
Table 1 Characteristics of Datasets

表 1 数据集的属性

Dataset	Coordinate Range	Real Size	Sample Size
Landmark	$[-124.4,-67.0]\times[24.6,49.0]$	8.70051×10^5	5×10^5
Checkin	$[-176.3,177.46]\times[-48.2,90.0]$	1×10^6	5×10^5



(a) Visualization of Landmark



(b) Visualization of Checkin

Fig. 3 Visualization of two datasets

图 3 2 种数据集可视化结果

本文设置的隐私预算参数 ε 的取值为 0.1,0.3,0.5,0.7,0.9.实验中范围查询 Q 的查询范围分别覆盖 Landmark 与 Checkin 这 2 种数据集的[10%,50%],[15%,55%],[20%,60%],在每种查询范围内随机生成 500 次查询.

1) 基于 Landmark 数据集的多种方法 *RE* 值比较

图 4(a)~(f)描述了 KRR,RAPPOR,OUE,N-PSDA 以及 N-GT-R 算法的 *RE* 值比较结果.由图 4(a)~(c)的实验结果可以发现,当范围查询 Q 固定时,ε 从 0.1 变化到 0.9,5 种方法的 *RE* 值均减少.然而,N-GT-R 的范围查询精度优于其他 4 种方法.在 $\varepsilon=0.7$,Q 为[15%,55%]时,N-GT-R 所取得的查询精度远优于 KRR 方法和 RAPPOR 方法,从实验结果可以看出,其精度是 KRR 方法的近 3 倍,是 RAPPOR 方法的 2 倍多,是 N-PSDA 与 OUE 方法的近 1 倍.从图 4(d)~(f)的实验结果可以发现,当 ε 固定,Q 的范围从[5%,45%]变化到[20%,60%]时,5 种方法的 *RE* 值均增大,其原因在于范围查询的查询范围越大,包含的查询单个点数越多,累计误差越大,导致精度随着查询范围的增大而降低.从图 4(d)~(f)中还可以发现 N-GT-R 方法的精度优于其他 4 种方法,其原因在于 N-GT-R 算法采用均匀网格分割整个值域,缩小了值域空间.

图 4(g)~(l)描述了 QT-KRR,QT-RAPPOR,PSDA,GT-R 这 4 种方法的 *RE* 值比较结果.由图 4(g)~(i)可以发现,当范围查询 Q 固定时,ε 从 0.1 变化到 0.9,4 种方法的 *RE* 值均减少,且 GT-R 方法明显优于其他 3 种方法.当 $\varepsilon=0.5$ 且 Q 为[20%,60%]时,GT-R 所取得的精度是 QT-RAPPOR 的近 4 倍,是 QT-KRR 和 PSDA 的近 3 倍.其原因是 GT-R 方法采用均匀网格与四分树对空间值域进行重新编码与索引,而其他 3 种方法均是基于整个值域响应范围查询.此外,GT-R 通过后置处理使结果不仅达到了无偏还保证了结果的一致性.由图 4(j)~(l)还发现,当固定 ε,查询范围 Q 从[5%,45%]变化到[15%,55%]时,4 种方法的查询精度随着查询范围的增大而降低.但查询范围 Q 从[15%,55%]变化到[20%,60%]时,4 种方法的查询精度出现随着查询范围的增大而提高的现象,其原因在于使用四分树索引后,Q 的查询精度与其所含盖的树中结点个数成反比,随着查询面积的增大,查询范围在四分树中含盖的索引结点个数可能变少,反而使精度更加精准.

2) 基于 Checkin 数据集的多种方法 *RE* 值比较

图 5(a)~(f)描述了 KRR,RAPPOR,OUE,N-PSDA,N-GT-R 这 5 种方法的 *RE* 值比较结果.由图 5(a)~(f)可以发现,当范围查询 Q 固定,ε 从 0.1 变化到 0.9 时,5 种方法的 *RE* 值均减少,然而 N-GT-R 的查询误差的稳定性与查询精度优于其他 5 种方法.当 $\varepsilon=0.3$ 且查询范围为[15%,55%]时,N-GT-R 方法所取得的查询精度是 KRR 方法的近 4 倍,是 RAPPOR 方法的近 3 倍,是 OUE 方法和 N-PSDA 的近 0.5 倍.当 ε 固定,随着查询范围扩大,5 种方法查询精度均呈现下降趋势,但 N-GT-R 方法的查询精度与误差稳定性同样优于其他 4 种方法.其主要原因是 N-GT-R 方法减小了值域空间,在做相同的范围查询时比其他方法所累积的误差少.

图 5(g)~(l)描述了 QT-KRR,QT-RAPPOR,PSDA,GT-R 这 4 种方法的 *RE* 值比较结果.由图 5

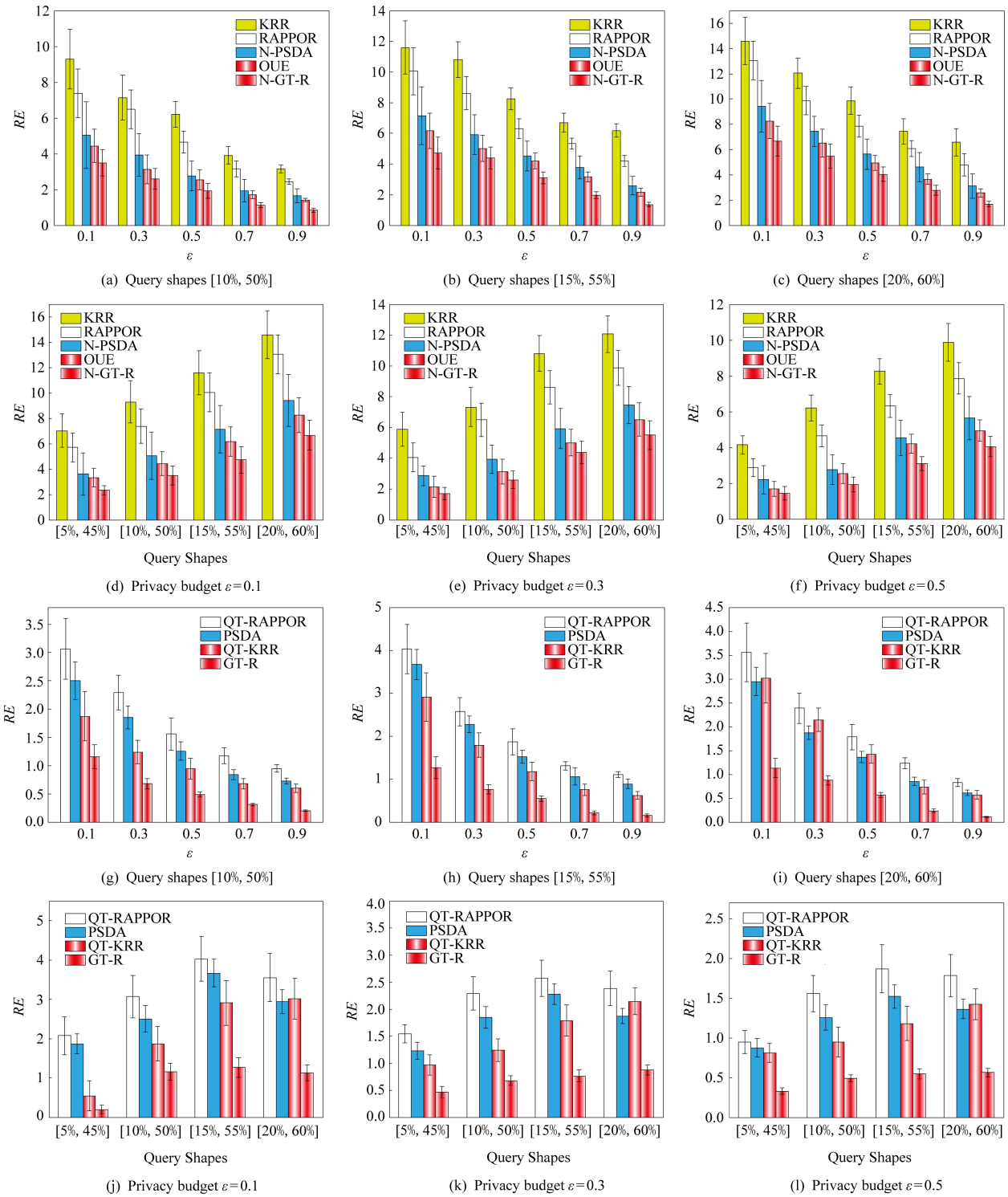


Fig. 4 Results of range queries on Landmark

图 4 Landmark 数据集范围查询结果

(g)～(i)可知,当范围查询 Q 固定且 ε 增大时,4 种方法的误差均减小.特别是查询范围为[10%,50%]且 $\varepsilon=0.9$ 时,GT-R 方法的精度是 PSDA 方法和 QT-RAPPOR 方法的近 7 倍,是 QT-KRR 的近 6 倍.此外,由图 5(j)～(l)显示,当 ε 固定,查询范围 Q 从[5%,45%]变化到[20%,60%]时,GT-R 方法在各个范围的查询精度虽然变化相对不大,但优于其他 3 种方法.在 $\varepsilon=0.5$ 时,GT-R 方法取得的精度是其他 3 种方法的 1 倍以上.其主要原因是 GT-R 方法利用网格与四分树对整个空间值域进行压缩与重新

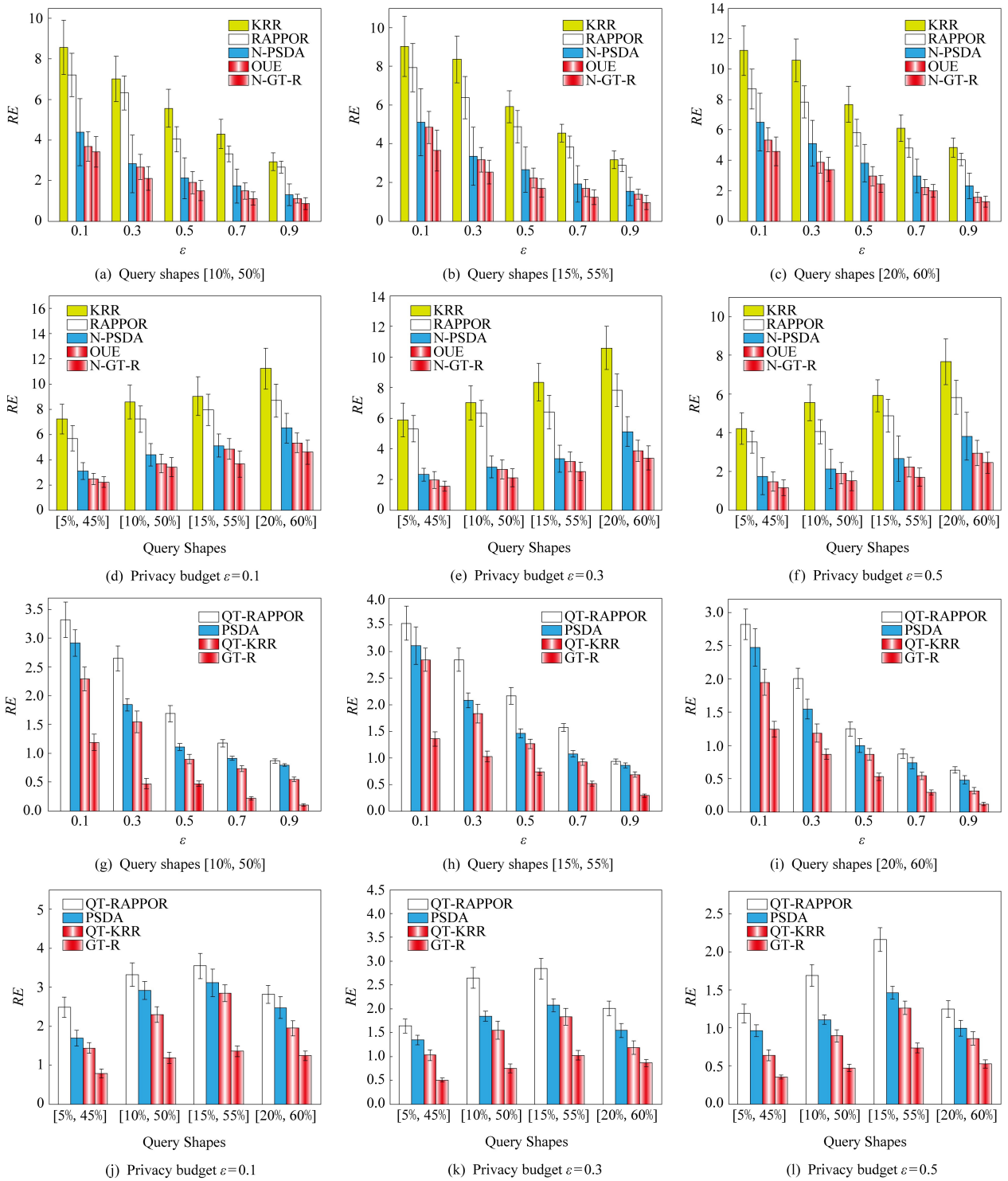


Fig. 5 Results of range queries on Checkin

图 5 Checkin 数据集范围查询结果

编码,既缩减了值域空间又减少了统计时的累计误差点数,因此精度优于其他方法.

5 结束语

针对本地差分隐私保护下收集用户空间位置存在的问题,本文结合现有的用户数据收集方法存在的不足,提出了基于网格与四分树索引的空间位置收集方法 GT-R.该方法通过均匀网格缩小空间值域,通过四分树对用户位置进行重新编码.从本地差分隐私定义角度分析 GT-R 满足 ε-本地差分隐私.最后通过 2 种真实的大规模数据集验证了 GT-R 方法

的范围查询精度.实验结果表明:GT-R 明显优于现有的同类方法.未来工作考虑动态环境下的隐私空间数据收集与分析问题.

参 考 文 献

[1] Qardaji W H, Yang Weining, Li Ninghui. Differentially private grids for geospatial data [C] //Proc of the 29th IEEE Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2013: 32-33

[2] Cormode G, Procopiuc C M, Srivastava D, et al. Differentially private spatial decompositions [C] //Proc of the 28th IEEE Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2012: 20-31

[3] Zhang Jun, Xiao Xiaokui, Xie Xing. PrivTree: A differentially private algorithm for hierarchical decompositions [C] //Proc of the 2016 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2016: 155-170

[4] Erlingsson Ú, Pihur V, Korolova A. Rappor randomized aggregatable privacy-preserving ordinal response [C] //Proc of the 2014 ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2014: 1054-1067

[5] Bassily R, Smith A. Local, private, efficient protocols for succinct histograms [C] //Proc of the 47th Annual ACM on Symp on Theory of Computing. New York: ACM, 2015: 127-135

[6] Wang Tianhao, Bloci J, Li Ninghui, et al. Locally differentially private protocols for frequency estimation [C] // Proc of the 26th USENIX Security Symp. Berkeley, CA: USENIX Association, 2017: 729-745

[7] Duchi J C, Jordan M I. Local privacy and statistical minimax rates [C] // Proc of the 54th Annual IEEE Symp on Foundations of Computer Science. Piscataway, NJ: IEEE, 2013: 429-438

[8] Duchi J C, Jordan M I, Wainwright M J. Minimax optimal procedures for locally estimation [J]. Journal of the American Statistical Association, 2018, 113(521): 182-201

[9] Nguyên T T, Xiao Xiaokui, Yang Yin, et al. Collecting and analyzing data from smart device users with local differential privacy[OL]. (2016-06-05) [2019-05-03]. https://arxiv.org/abs/1606.05053

[10] Wang Ning, Xiao Xiaokui, Yang Yin, et al. Collecting and analyzing multidimensional data with local differential privacy [C] // Proc of the 35th IEEE Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2019: 638-649

[11] Wang Tianhao, Li Ninghui, Jha S. Locally differentially private frequent itemset mining [C] //Proc of IEEE Symp on Security and Privacy. Piscataway, NJ: IEEE, 2018: 127-143

[12] Qin Zhan, Yu Ting, Yang Yin, et al. Generating synthetic decentralized social graphs with local differential privacy [C] //Proc of ACM Conf on Computer and Communications Security. New York: ACM, 2017: 425-438

[13] Kulkarni T, Cormode G, Srivastava D. Answering range queries under local differential privacy [C] //Proc of the 2019 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2019: 1832-1834

[14] Chen Rui, Li Haoran, Qin Kai, et al. Private spatial data aggregation in the local setting [C] //Proc of the 32nd IEEE Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2016: 289-300

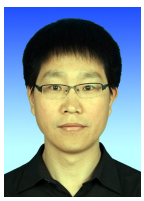
[15] Warner S L. Randomized response: A survey technique for eliminating evasive answer bias [J]. The American Statistical Association, 1965, 60(309): 63-69

[16] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis [C] //Proc of the 3rd Theory of Cryptography Conf. Berlin: Springer, 2006: 363-385

[17] Dwork C, Lei Jing. Differential privacy and robust statistics [C] //Proc of the 41st Annual ACM Symp on Theory of Computing. New York: ACM, 2009: 371-380

[18] Michael H, Vibhor R, Gerome M. Boosting the accuracy of differentially private histograms through consistency [J]. Proceedings of the VLDB Endowmen, 2010, 3(1): 1021-1032

[19] Kairouz P, Oh S, Viswanath P. Extremal mechanisms for local differential privacy [J]. Journal of Machine Learning Research, 2016, 17(17): 43-51



Zhang Xiaojian, born in 1980. PhD, associate professor in the School of Computer and Information Engineering, Henan University of Economics and Law. His main research interests include differential privacy, data mining, and graph data management.

Fu Nan, born in 1988. Master candidate in the School of Computer and Information Engineering, Henan University of Economics and Law. His main research interests include local differential privacy, data mining.

Meng Xiaofeng, born in 1964. Professor and PhD supervisor at Renmin University of China. Executive director of CCF. His main research interests include cloud data management, Web data management, native XML databases, and flash-based databases, privacy-preserving, and etc.