doi:10.11959/j.issn.1000-436x.2016176

位置大数据服务中基于差分隐私的数据发布技术

张琳[1,2,3]，刘彦[1]，王汝传[1,2,3]

（1. 南京邮电大学计算机学院，江苏 南京 210003；2. 江苏省无线传感网高技术研究重点实验室，江苏 南京 210003；
3. 南京邮电大学宽带无线通信与传感网技术教育部重点实验室，江苏 南京 210003）

摘 要：应对多组合复杂攻击及前景知识攻击，提出一种新的基于差分隐私保护机制的位置大数据发布模型，创新设计可用性评估反馈机制模块，引入时间变量动态地针对敏感属性以及身份识别等分析模型的服务质量，能在位置大数据与非位置大数据相结合、用户背景知识不确定等情况下保护用户的位置隐私。仿真实验基于多种空间索引技术验证了新发布模型能够在指定的私密性条件下，为位置查询服务发布具有更高准确率的匿名数据。

关键词：位置服务；大数据；差分隐私；数据发布；隐私保护

中图分类号：TP393 **文献标识码**：A

Location publishing technology based on differential privacy-preserving for big data services

ZHANG Lin[1,2,3], LIU Yan[1], WANG Ru-chuan[1,2,3]

(1.College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;
2.Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing 210003, China;
3. Key Lab of Ministry of Education Broadband Wireless Communication and Sensor Network Technology of Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Aiming at dealing with prospect knowledge and complex combinatorial attack, a new location big data publishing mechanism under differential privacy technology was given. And innovative usability evaluation feedback mechanism was designed. It gave corresponding solution details for the sensitive attributes and the identity recognition to analyze the quality of service, aimed at privacy protecting for location based big data under situations like combination of location information and non-location information and attacker's arbitrary background knowledge. Simulation results based on different spatial indexing technology proved that the new publishing model has a higher accuracy under specified privacy conditions for the location query service.

Key words: location service, big data, differential privacy, data publishing, privacy-preserving

1 引言

移动通信与传感设备等位置感知技术的快速发展使人们正式步入大数据信息时代，位置大数据服务产品不仅为用户提供了便捷的服务，如兴趣点（POI，points of interest）查找、社交网络位置分享等，针对收集到的海量位置数据也为企业提供了分析数据特点的查询业务，为企业挖掘有价值的商业信息提供便利。随着个人隐私保护意识的不断提高，对海量异构数据的安全发布与分析成为大多数企业亟待解决的问题。对位置大数据服务的恶意访问可能泄露个人大量的敏感信息[1,2]，如观察到用户出现在医院附近，可以推测出用户大致的健康状况；查询用户对兴趣地点的访问频率，可分析用户偏好

收稿日期：2015-10-26 **修回日期**：2016-08-11

基金项目：国家自然科学基金资助项目（No.61402241, No.61572260, No.61373017, No.61572261, No.61472192）；江苏省科技支撑计划基金资助项目（No.BE2015702）；江苏省属高校自然科学研究重大基金资助项目（No.12KJA520002, No.14KJA520002）

Foundation Items: The National Natural Science Foundation of China (No.61402241, No.61572260, No.61373017, No.61572261, No.61472192), Scientific and Technological Support Project of Jiangsu Province (No.BE2015702), Natural Science Key Fund for Colleges and Universities of Jiangsu Province (No.12KJA520002, No.14KJA520002)

和经济状况；此外，还可考虑用户轨迹开始和结束的地点，推测出用户的家庭住址等信息。特别是对大数据服务中的位置信息与社交网络中非位置信息相结合的攻击，将暴露用户更多个人信息。

针对以上问题，现有位置大数据的隐私保护技术可分为 3 类：基于启发式隐私度量、基于概率推测以及基于隐私信息检索的位置大数据隐私保护技术[3,6]。差分隐私[7,8]具有完全独立于攻击者背景知识的强隐私保护性质，已经有不少研究[9~11]将其应用到面向发布的数据模型中。文献[12]将差分技术与位置大数据服务相结合，针对发布数据聚集易受相似性攻击的问题，提出一种最大化差分隐私效果的匿名算法。文献[13]利用位置服务查询结果的相似性来辅助匿名服务器构造匿名区域，提出具有较高平衡性的 *k*-匿名位置隐私保护方法。为提高位置数据模式挖掘中差分技术的保护性能，文献[14]基于四叉树空间分解技术保持了位置数据的语义。社交网络中考虑推荐系统存在的隐私安全问题，文献[15]通过差分技术提出保护位置轨迹数据集隐私的解决方案。但它们的缺点在于忽略了将位置数据与非位置数据匹配的攻击模式，针对这个问题，本文提出差分保护下一种新的位置大数据服务的数据发布模型，在提供大数据查询服务中的同时保护位置数据和非位置数据的敏感信息，仿真实验证明算法能够在指定的私密性条件下，为查询服务发布具有更高准确率的位置数据。

2　问题描述

2.1　位置大数据服务中的隐私威胁

随着互联网大数据技术的不断发展和基于位置服务（LBS, location-based services）应用的不断增多，利用全球卫星导航定位系统和地面通信基站，结合云计算、海量数据处理技术，把数以亿计甚至数以千亿计的数据实时、安全地存储下来，再以可视化的方式展现给用户，这将产生巨大的社会价值和商业价值。通过手机、车载导航、内置传感器等移动设备中的 GPS、Wi-Fi 等定位设备，将 LBS 与大数据概念结合产生的产品越来越多。从国内看，2014 年，央视与百度合作在春运期间推出“百度迁徙”位置服务应用，启用百度地图定位可视化大数据播报国内春节人口迁徙情况，该项目利用百度 LBS 定位数据进行计算分析，展现春节前后人口大迁徙轨迹与特征位置，成为大数据位置服务进入大众生活的成功案例。

然而，位置大数据应用服务在提供相应查询并发布相关用户信息时，将面临泄露客户个人隐私的危机。大数据时代，位置数据的来源极为广泛，据统计，每天使用百度定位服务产生的定位请求由 2013 年的 35 亿次上涨到 2015 年的 150 亿次，如果对其收集到的海量位置信息不加以控制地发布、不提供安全的挖掘分析机制，对于含大量敏感信息如医疗、金融行业以及情报收集等部门的隐私泄露将引发严重的社会恐慌。同时，将位置大数据中包含的移动对象不同时刻的位置信息与背景知识结合，可以让攻击者有效推测用户的行为模式，会泄露用户的健康状况、行为习惯、社会地位等敏感信息。攻击者可以根据在任意时刻 t 之前收集到历史位置数据 L，推测用户在 t 时刻处于某个敏感位置 l_n 的概率为 $P\left(\frac{l_n}{L}\right)$。为了量化位置大数据的隐私保护效果，对位置大数据的 θ 隐私定义如下。

定义 1　设任意用户 U 在任意时刻 t 处于位置 l_n 的推测概率为 $P(U_{l_n}^t)$，在 t 时刻之前收集到关于用户 U 的历史位置数据为 $L^t=\{l_1^t,l_2^t,\cdots,l_n^t\}$，则

$$P\left(\frac{U_{l_n}^t}{L^t}\right)-P(U_{l_n}^t)\leqslant\theta \tag{1}$$

其中，θ 是用户给定的隐私需求，也是攻击者能够获得的最大攻击效果。攻击者根据 t 时刻之前获取到的历史位置数据推测用户在 t 时刻处于位置 l_n 的后验概率 $P\left(\frac{U_{l_n}^t}{L^t}\right)$ 与其先验概率之差不能超过隐私需求 θ，从而量化了隐私保护效果，当 $\theta=0$ 时隐私保护效果达到最大，称为完美隐私。

2.2　位置服务基于差分隐私的数据发布模型

差分隐私是一种新的基于严格数学背景的隐私保护机制，提供了使隐私保护程度可量化、可评估、可证明的方法。差分隐私保护技术通过添加随机噪声扰动敏感数据，使某些数据在失真的同时保持其具有的统计性质，以便在进行数据挖掘等操作后得到高度近似的价值知识。将差分隐私技术运用到位置大数据发布模型中，能有效防止基于背景知识的恶意攻击。大多数位置数据发布系统结构基于“先收集、再匿名、后发布”的原则，即由一个数据收集服务器收集位置数据，并将海量原始数据存储到轨迹数据库中，然后结合云计算技术与一定的隐私保护机制进行隐私保护处理，最后为查询客户

提供可发布、可分析的安全轨迹数据。本文提出的差分隐私保护下位置大数据处理服务器主要有 3 个模块：数据预清洗模块、敏感信息隐藏模块和可用性评估模块，如图 1 所示。数据预清洗模块负责对收集到的轨迹数据进行等价类划分、轨迹同步等预处理操作；敏感信息隐藏模块结合差分隐私保护技术，负责对预处理后的隐私数据匿名操作；由可用性评估模块反馈隐私处理后的数据可用性，实现在高效挖掘信息蕴含知识与控制隐私泄露隐患之间找到最佳平衡点。最后将处理后的位置大数据发布给查询客户。

Dwork 等[7,8]在 2006 年首次提出差分隐私技术，根据这种强隐私保护模型，本文定义位置数据差分隐私如下。

定义 2 （位置差分隐私）对于至多相差一条位置记录的 2 个数据集 D 和 D'，即两者的线性相异距离 $\left|D-D'\right|\leqslant 1$，$M$ 为给定提供 $\varepsilon-$差分隐私保护的随机查询函数，$Range(M)$ 代表算法 M 的取值范围，若轨迹数据集 D 和 D' 在查询函数 M 下得到的任意位置 $L(L\subseteq Range(M))$ 满足式(2)，则 M 满足 $\varepsilon-$位置差分隐私。

$$\Pr\left[M(D)\in L\right]\leqslant \Pr\left[M(D')\in L\right]\mathrm{e}^{\varepsilon} \tag{2}$$

其中，概率 Pr[·] 表示隐私被泄露的风险，由算法 M 的随机性控制；参数 ε 为隐私保护预算，ε 越小隐私保护程度越高。

在查询函数的返回值中添加适量随机噪声是实现差分隐私的主要技术，常用的噪声机制有拉普拉斯机制[16]和指数机制[17]，给出定义 3 和定义 4。任意函数 $f:D\to\Re^d$，全局敏感度[7] $S(f)=\max\limits_{D,D'}\left\|f(D)-f(D')\right\|_1$ 是决定噪声适量添加的关键参数，其中，$\left\|f(D)-f(D')\right\|_1$ 是相邻数据集 D 和 D' 函数输出值的一阶范数距离。敏感度指添加或删除数据集中任一记录对函数输出值造成的最大改变。查询函数 f 的全局敏感度 $S(f)$ 由函数本身性质决定，与数据集无关。

定义 3 （拉普拉斯机制）任意函数 $f:D\to\Re^d$，若函数 f 的输出结果满足式(3)，则 f 满足 $\varepsilon-$差分隐私。

$$M(D)=f(D)+\left(\mathrm{Laplace}\left(\frac{S(f)}{\varepsilon}\right)\right)^d \tag{3}$$

其中，拉普拉斯分布的位置参数为 0，尺度参数为 $\dfrac{S(f)}{\varepsilon}$，噪声量大小与全局敏感度 $S(f)$ 成正比，与隐私预算 ε 成反比。拉普拉斯机制局限于返回值为实数类型的函数，对于非数值型的查询函数提出指数噪声机制。

定义 4 （指数机制）给定一个可用函数 $q:(D^n\times R)\to\Re$，r 是可用函数输出域 $Range$ 中的一个实体对象，若函数 q 的输出结果满足式(4)，则 q 满足 $\varepsilon-$差分隐私。

$$M(D,q)=\left\{r:\Pr\left[r\in Range\right]\propto\exp\left(\frac{\varepsilon q(D,r)}{2S(q)}\right)\right\} \tag{4}$$

其中，$S(q)$ 为可用函数的全局敏感度，对象被选中的概率正比于可用函数 q 的打分。指数机制以正比于 $\exp\left(\dfrac{\varepsilon q(D,r)}{2S(q)}\right)$ 的概率返回实体对象。

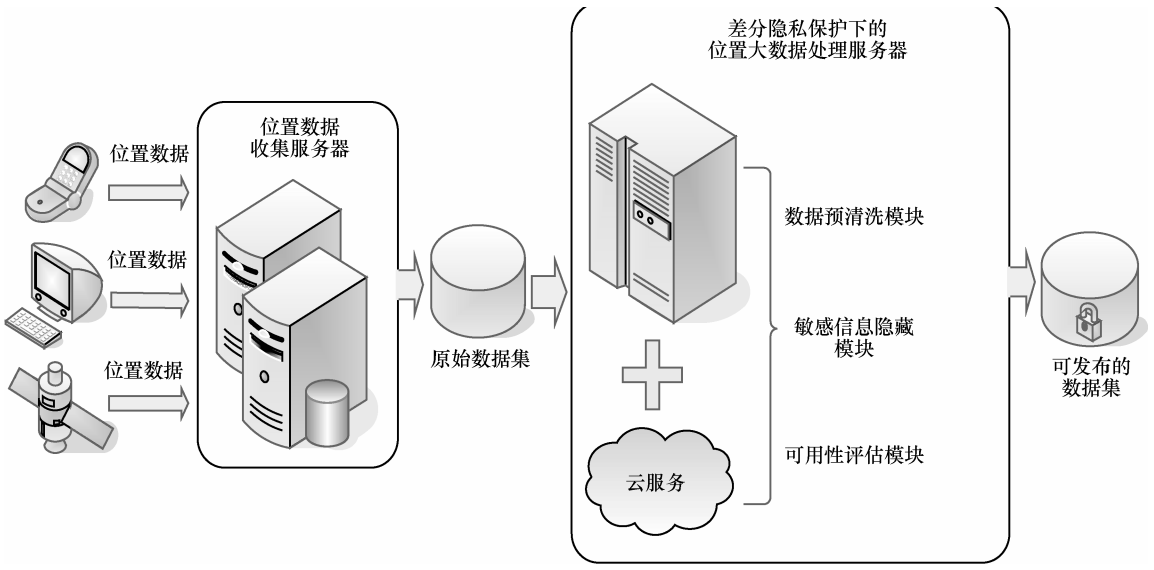

图 1 差分隐私保护下位置大数据发布模型

3　基于差分隐私的位置大数据发布方法

位置数据基于二维空间索引技术发布所处的地点信息，常见存储空间数据的数据结构包括四叉树（Quad-tree）、R 树及 R^+树系列（R-tree）、*k*-d tree（*k*-dimensional tree）等。这些数据结构大体可以分为 2 类[18]：数据无关的分解结构（data-independent decomposition）和数据相关的分解结构(data- dependent decomposition)。如当对同一位置信息数据库构造索引树时，四叉树按固定的递归算法分裂节点，树的构造结果与数据无关；而 R-tree，*k*-d tree 的节点分裂性质取决于位置覆盖面积大小以及分布方式的不同，其分裂形态将与数据库中的具体位置信息相关。因为地点信息的语义具有明显的垂直分层结构并且地点之间存在包含关系，所以用树结构存储的位置信息可以被划分保存，空间的分解也减少了访问查询的复杂度。由于空间数据的特殊性（海量、多维、空间拓扑特征、时间特征），空间分解技术采用分割原理，把查询空间划分为若干区域，同时将位置信息分层储存，形成可唯一标识空间要素。然后利用不同的数据结构对分割的区域进行组织，以达到快速访问数据项的目的。

本文提出的位置信息发布方法在基于空间分解技术的位置大数据添加差分隐私匿名泛化技术，使最终发布的数据在保持一定挖掘准确率的条件下保护个人隐私，达到安全访问的目的。图 2 为基于四叉树二维空间分解技术的位置兴趣点分布，其中，Q 和 Q'为对外提供的查询函数。按此数据结构实施差分隐私保护，为索引树添加噪声的过程如图 3 所示。

如图 2 所示，假设某用户在一段时间内形成移动轨迹，攻击方联合收集用户在该区域出现的标记点数量，基于统计背景知识对该区域包含的部分区域请求位置大数据服务，发出连续的查询函数 Q、Q'等，攻击方对比背景知识与连续查询函数的结果，可以分析推测用户在某个时间点出现的具体位置。所以，基于这种位置数据发布可能出现隐私泄露的问题，本文提出的算法在位置大数据服务中结合差分隐私技术，可实现隐私四叉树的安全发布。如图 3 所示，标记位置 $L1$~$L7$ 为用户的兴趣位置查询点，给此用户标记的位置信息添加拉普拉斯或指数噪声，方框中的数据是对真实数据匿名化的结果，其中，P 定义为某个位置兴趣点被查询的概率，a~e 定义为位置分割节点。使位置服务发布的匿名化数据能在实现隐私保护的同时保证给查询函数提供一定可用性数据是本文算法的研究重点。

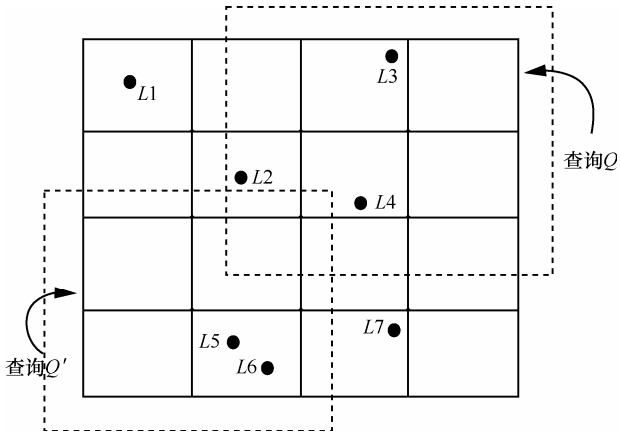

图 2　基于四叉树空间分解的位置分布

3.1　隐私位置数据发布流程

现有的位置大数据服务中，攻击者可以从多种渠道获得用户和位置数据相关的其他类型数据，并结合位置数据共同推测用户的隐私信息。如许多社交软件的签到服务、分享定位等操作，通过用户的个性设置或者行为模式对其位置数据与非位置数据进行匹配，会将用户大量敏感的个人信息发布给基

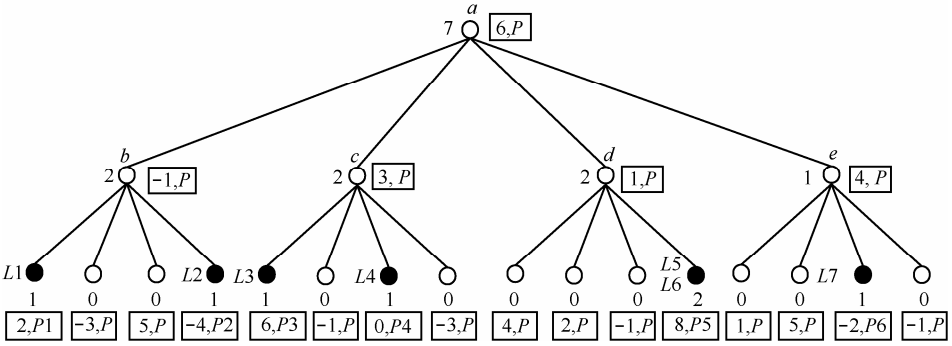

图 3　添加噪声的索引树数据

于位置服务的攻击者，带来隐私泄露的潜在危机。本文对基于位置大数据服务器的恶意攻击途径主要为获取对系统的先验知识，攻击者对不同知识掌握的准备度也直接影响着攻击者的攻击能力。为解决上述问题，构建基于含有安全客户信任中间件的 LBS 系统，实现差分隐私保护下位置大数据服务器对隐私位置数据的安全发布，本文提出差分位置发布（Diff_PriLocation）算法步骤如下。

步骤 1 位置数据预处理。根据空间分解技术（SD）将二维位置分解为基本数据集，依据空间位置自然的包含关系划层次存储每个位置元素，保留位置元素的语义信息得到相应完整的树形数据结构，为以后目标点搜索提高了执行效率。

步骤 2 合理分配隐私预算 ε。根据数据预处理过后的树高度 h 和叶子节点的数量 N_{leaf} 合理分配迭代过程中的隐私预算。

步骤 3 数据匿名化处理。基于不同的空间分解方法遍历树的所有节点，获取每个标记点的位置数据 L 与非位置数据 D。用差分隐私技术分别对位置和非位置数据添加噪声，对不同属性类别的敏感信息采取不同匿名化处理机制。

步骤 4 反馈机制。为了保证发布数据具有较高的数据可用性，建立查询函数 Q 的反馈机制。根据查询用户的反馈信息分析数据的泛化程度，调节发布数据的隐私预算 ε。

步骤 5 安全发布位置大数据。对数据库中位置数据与非位置数据的分类处理，是为了保证基于位置大数据的服务质量。重组匿名处理后的位置数据，最后对查询函数发布满足差分隐私的数据。

在算法的整个流程中，创新提出综合考虑位置数据与非位置数据来保护用户隐私，设计新的应对多组合复杂攻击及前景知识攻击的差分隐私保护数据发布方案，针对敏感属性以及身份识别等提出相应的数据细节处理方案，使发布框架具有良好的开放性。

3.2 平衡隐私预算

隐私预算 ε 的不合理分配会破坏算法的隐私保护性能，算法本身也就失去了意义。本算法依据位置数据提供者对隐私保护的需求 B 以及空间分解后输出的索引树高度 h，合理分配迭代过程中的隐私预算 ε。对于数据不相关的空间分解技术，如基于四叉树索引的空间分解算法，由于其输出的数据结构与数据性质无关，仅与索引树的高度以及叶子节点包含的标记点数 n 有关，所以对每次匿名迭代分配的隐私预算为 $\varepsilon=\dfrac{B}{2n(h-1)}$。而对于数据相关的空间分解方法数据的性质会直接影响隐私预算的分配，所以平衡隐私预算的方法要具体考虑如树根节点到某叶子节点路径上的中位数、节点数等参数。索引树高度 h（$1\leqslant i\leqslant h$），取路径上每个节点消耗隐私的预算最大和为 ε，且令根到叶子的每条路径上隐私消耗相等，定义为

$$\varepsilon=\sum_{i=1}^{h}\varepsilon_i^m+\sum_{i=1}^{h}\varepsilon_i^c \tag{5}$$

其中，ε_i^m 为路径内节点根据中位数算法所消耗的隐私预算，ε_i^c 为统计路径节点数算法所消耗的隐私预算，平衡二者的隐私分配可提高发布数据的可用性。具体关于数据相关空间分解索引的隐私平衡算法可参考文献[14]。

3.3 匿名处理

给定隐私预算 ε，对遍历收集到的位置数据 L 和非位置数据 D 添加噪声，使之满足差分隐私的要求。对于非位置数据 D 的属性集 $A=\{A_1,\cdots,A_k\}$，其属性类别包括连续值属性 $C_j^A=\{A_j\in C\bigcap(A=A_j)\}$ 和非连续值属性 $E_j^A=\{A_j\in E\bigcap(A=A_j)\}$，针对数据属性的不同分别采取不同的差分噪声机制。对连续值添加拉普拉斯噪声 $\text{LapNoise}_\varepsilon$ 扰动真实值，指数机制以正比于 $\exp\left(\dfrac{\varepsilon}{2}|Q(E)-Q(E')|\right)$ 的概率输出离散值，其中，Q 为查询函数。差分隐私保护下的数据匿名处理（Diff_Anonmity）算法描述如算法 1 所示。

算法 1 差分隐私保护下的数据匿名处理算法

输入 ε—隐私保护预算，L—位置数据集，D—非位置数据集，$A=\{A_1,\cdots,A_k\}$—数据属性集合，C_j^A—连续属性数据集合，E_j^A—离散属性数据集合

输出 满足差分隐私保护要求的匿名化数据集 *Diff_AnonymData*

1) begin Procedure Diff_Anonmity(ε, L, D, A, C_j^A, E_j^A)

2) $\varepsilon=\dfrac{B}{(2+n)d+2}$，其中，$d$ 是空间树的高度，如果接收到反馈机制的请求，则对 ε 做出相应调节。

3) if 数据属于位置数据 L

4) for L 集合中任一元素 L_i

5)　$N_{L_i}=\mathrm{LapNoise}_{\varepsilon}\left[Q(L_i)\right]$，其中，$Q$ 为用户的查询函数，具有全局敏感度 $S(Q)$

6)　end　for

7) else　if　数据属于非位置数据 D

8)　for　D 集合中任一元素 D_i，并且元素满足 $D_i\in\left\{D\cap A_j\in\left\{A_1,\cdots,A_k\right\}\right\}$

9)　if　A_j 为连续值属性，则 $D_i\in C_j^A$

10)　$N_{D_i}=\mathrm{LapNoise}_{\varepsilon}\left[Q(D_i)\right]$ 拉普拉斯机制

11)　else　if A_j 为离散值属性，则 $D_i\in E_j^A$

12)　$N_{D_i}=\mathrm{ExpMech}_{\varepsilon}\left[Q(D_i)\right]$ 指数机制

13)　end　if

14)　end　for

15) end　if

16) return　*Diff_AnonymData*(N_{L_i}、N_{D_i})

17) end　Procedure

3.4　性能分析及反馈机制

3.4.1　反馈机制

本文提出位置大数据处理服务器的可用性评估模块包含反馈机制以及性能分析模块，反馈机制针对敏感属性以及身份识别等提出相应的数据细节处理方案。在位置大数据动态发布的背景下，考虑准标识符和敏感属性随着时间发生变化会影响构建模型的服务质量，可引入时间变量 t 标记数据集的改动，并在一定时间内分析动态集改动操作的敏感度，若在统计过程中敏感度超过设定的阈值 Γ，则需重新迭代位置索引树获取更新的数据信息，并结合用户反馈信息发布新的数据模型。由于动态发布的记录数据集中项目的增加、删除、修改等操作都会影响执行效率并导致查询结果的不同，确立适当的反馈机制实时监管变化的数据，实施可控地隐私保护，可防止恶意攻击方联合位置数据与非位置数据对用户身份识别，同时将对整体发布数据的压缩和匿名化处理的影响降到最低，使整体设计的发布框架具有良好的开放性。

3.4.2　数据可用性分析

分析算法的性能以及收集用户对服务质量的反馈信息，提供交互接口可对位置数据的发布过程进行优化。根据位置大数据服务对象对查询精确度的要求，可以适当调整发布数据的隐私保护程度。对于相同的查询函数 Q，数据在匿名前和匿名后输出查询结果的近似度体现了隐私保护算法对数据的可用性的影响。设 $G(Q)$ 为原始数据的查询结果，$G^{\wedge}(Q)$ 为匿名后数据的查询结果，则近似度 S_Q 可定义为两输出结果的一阶范数距离 $S_Q=\left\|G(Q)-G^{\wedge}(Q)\right\|_1$。对连续查询 $Q_i\in\left\{Q_1,\cdots,Q_k\right\}$，位置服务发布数据的可用性定义为

$$E=\frac{1}{|Q_i|}\sum_{Q_i}\left(\frac{S_{Q_i}}{\mathrm{e}^{\frac{\varepsilon}{2}}}\right) \tag{6}$$

服务对象可定义允许的误差范围 δ，则可容忍的信息丢失率为 $1-\delta$，误差范围 δ 与数据可用性 E 成反比。其中，基于空间距离的查询函数敏感度为 $\sigma(f)=1$，而针对海量位置数据集中一组连续分区查询策略 $F_v=\left\{f_p:D\to(R^+)^d\right\}$，定义其全局敏感度为其中任意查询函数敏感度的最大值，表示为 $\sigma(F_v)=\max_{i\in v}\sigma(f_i)$。当匿名发布数据可用性低于用户的查询要求时，可以对发布模型反馈误差信息，在差分保护范围内要求适当降低隐私保护干扰，从而降低差分隐私对数据集的泛化程度，发布可用性更高的数据。数据可用性和算法复杂度是评判发布算法性能的重要标志，能在提高添加噪声执行效率的同时减少噪声带来的误差，兼顾发布数据安全性与可用性。

3.4.3　算法复杂度分析

最优化差分技术为数据原项添加噪声的算法复杂度，会直接影响匿名发布模型的执行效率。首先 Diff_Anonmity 算法采用贪心方法自顶而下递归索引树的时间复杂度为 $O(\mathrm{lb}n)$，然后分类处理每个节点包含的数据信息。针对位置数据集添加噪声消耗时间复杂度为 $O(|L|)$，对于非位置数据集中每个连续值属性中出现的值要逐一排序并统计重复值出现的次数，处理连续值属性找出最佳断点并加入拉普拉斯噪声的时间复杂度为 $O(|D_1|\mathrm{lb}|D_1|)$，完成对连续值属性的泛化处理后，Diff_Anonmity 算法最后再对离散值属性处理平均消耗的时间复杂度 $O(|D_2|\mathrm{lb}|D_2|)$，整个算法过程消耗的总时间复杂度还受到空间分解生成树高度的影响。因为大数据服务的海量信息结构复杂多样，所以收集服务器在前期对异构集合进行必要的数据清洗处理，将大幅降低匿名算法迭代处理的复杂程度。

4　仿真实验

为了研究本文提出算法的可行性，采用的系统硬件配置为主频 2.4 GHz 的 Intel(R)Core(TM)i3 兼

容 PC 机，2 GB 内存，200 GB 以上的可用磁盘空间；软件配置平台为 WIN 2008 Server 操作系统以及 Microsoft SQL Server 数据库系统，C/S 结构的运行模式。实验基于 3 个数据集 GeoLife、Amazon Access Samples 以及 Diversification Dataset Div400。数据集的来源数据库分别是：GeoLife GPS Trajectories，UCI Machine Learning Repository 和 UMass Trace Repository。实验的数据集都包括用户的位置信息和非位置信息，其数据集大小以及属性特征如表 1 所示。

表 1　　实验数据集

数据集	属性集	
	位置数据	非位置数据
GeoLife (38 494)	纬度	交通方式
	经度	生活方式
	国家、城市、街道	社交信息搜索行为
Diversification Dataset Div400 (43 418)	纬度	共享信息
	经度	文字图片
	国家、城市、街道	个人数据文件夹
Amazon Access Samples (60 000)	用户定位信息	用户账号
		用户交易信息
		用户职业
		用户工作单位
		家庭基本信息

常用的空间索引技术可以提高空间信息数据库的操作效率，Quad-tree 将位置空间递归划分为不同层次的树结构，当空间数据对象分布比较均匀时，具有比较高的空间数据插入和查询效率；R-tree 广泛应用于原型研究和商业应用中，为了使 R-tree 能在海量的空间数据库中发挥重要作用，将优化插入路径选择减少 R-tree 插入节点后目录矩形重叠面积，以及保持在当前层实现节点分裂的独立性；*k*-d tree 索引用于多维检索的结构形式，对数据点在 *k* 维空间中划分，并在每一层都根据该层的分辨器对相应对象做出分枝决策，对于精确的点能匹配查找具有和二叉树一样良好的性能（平均查找长度为 1+4lb*n*）；本文提出的发布框架基于以上 3 种空间索引技术 Quad-tree、R-tree、*k*-d tree，对实验的 3 种数据集进行了 2 类实验分别测试差分隐私保护下的发布算法与传统位置数据发布算法的数据可用性以及安全性能，并对实验结果进行分析，验证本文算法 Diff_PriLocation 和 Diff_Anonmity 的有效性。

仿真实验在隐私预算 ε 逐步递增的条件下，比较经过本文发布框架输出的匿名化数据的准确程度，满足隐私要求在同等测试环境下分析 3 个数据集的实验结果，分析本文算法在不同隐私保护预算 ε 下的数据可用性，其中，差分隐私的可用函数取查询函数的信息熵增益值 $Gain(Q)$，考虑算法中拉普拉斯变换对海量数据的噪声敏感度问题，选取不同的变换尺度参数 $\frac{S(f)}{\varepsilon}$，噪声量大小与全局敏感度 $S(f)$ 成正比，与隐私预算 ε 成反比。算法首先根据给定的数据集及其 n 个属性集产生原始的空间划分，得到所有区域单位格为 $cell_i$，则其中的位置点频数 $count(cell_i)$。当采用树划分算法为所有频数加入 Laplace 噪声时，在每一次的递归迭代中，首先计算当前分区 P_i 中频数分布的紧密程度 $L(P_i)=\sum_{cell_i\in D_r}\left|count(cell_i)-a_i\right|$，其中，$a_i$ 是 P_i 中所有数据格频数的平均值，若 $L(P_i)$ 大于预定义的阈值，则将该分区划分为 2 个子分区。利用树的层次特征分而治之地把海量数据集分割处理，并根据树的高度 $1+\log_m n$ 调整分区方案的敏感性，能精确地响应较长范围的计数查询。

对于依赖树划分的海量空间数据，添加的拉普拉斯噪声量将与空间分区方案密切相关。使用基于区域密度的分割停止条件来构造差分隐私保护技术，使分区中数据格的频度彼此接近，从而在大型数据集中可仅通过添加少量的噪声，就能达到高级别的隐私保护。同时，为了验证 Diff_PriLocation 算法对不同空间分解技术的开放性，实验分别基于 Quad-tree、R-tree 以及 *k*-d tree 这 3 种常用的空间索引技术，设计 20 组实验，分别在不同数量级和不同隐私预算要求下测试算法的性能，实验结果取平均值。对 3 个实验数据集匿名化处理后，得到满足差分隐私保护要求的可发布数据 *Diff_AnonymData*，分析 *Diff_AnonymData* 的准确率可体现本文位置大数据处理服务器在满足查询要求的条件下处理发布的数据集可用性。各实验数据集的发布数据准确性如图 4~图 6 所示。

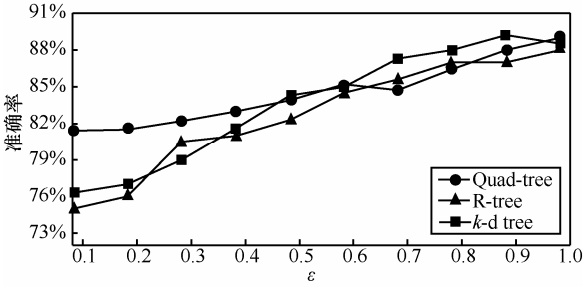

图 4　GeoLife 数据集可用性

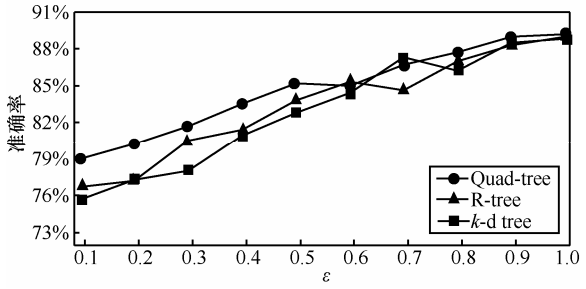

图 5　Amazon Access Samples 数据集可用性

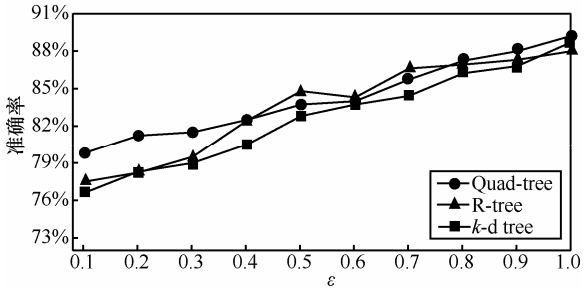

图 6　Div400 数据集可用性

隐私预算 ε 与隐私保护程度成反比，当隐私预算越小、数据泛化程度越大且当隐私保护程度 ε=0 时，即达到完美隐私。从实验结果可以看出，随着隐私保护预算的递增，差分算法对分布数据的泛化程度减小，所以数据的可用性都有所上升。从仿真实验结果中也可以看出，因为基于四叉树索引生成的位置空间树结构与实验数据集的性质无关，所以基于四叉树索引的差分隐私发布数据会具有相对较高的数据可用性，同时保持一定的算法稳定性。从表 1 可以看出 3 个实验数据集所含的数据项呈上升趋势，随着实验数据集的增加，与数据相关的 R-tree、*k*-d tree 空间索引技术受到数据统计性质的影响，在隐私保护程度要求较高的情况下会具有较差的数据可用性，而且 Diff_Anonmity 算法的执行效率也趋向下降。同时，较一般非差分隐私保护的安全发布算法[19]，本文算法也维持了近似的信息损失度和数据可用性。

为了研究算法的隐私保护性能，在数据可用性与隐私保护程度之间找到最佳平衡，考虑海量数据集的噪声敏感度问题，实验在不同拉普拉斯变换尺度参数 $b=\frac{S(f)}{\varepsilon}$ 下分析比较各个数据集的平均隐私保护程度。在基于数据相关与数据无关 2 类不同的空间分解技术下，实验数据集被差分隐私匿名保护的结果如图 7 和图 8 所示。从实验结果可以看出，Diff_Anonmity 算法的隐私保护程度基本能达到 80%以上，随着匿名要求的提高不会大幅降低算法的执行效率，在隐私保护程度要求比较高的情况下，差分隐私的性能具有较好的稳定性。

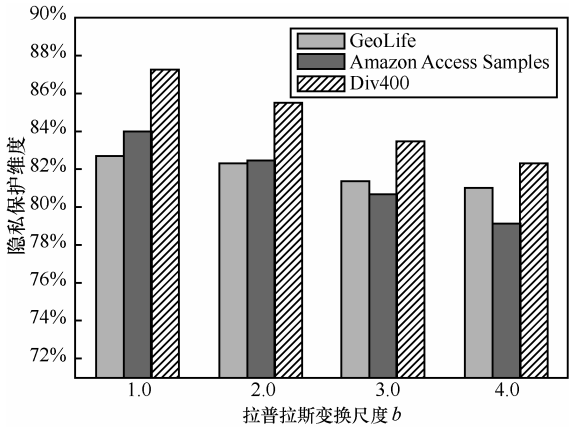

图 7　基于数据无关分解结构

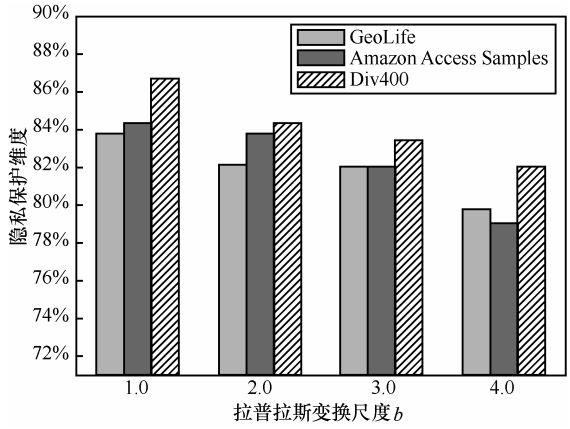

图 8　基于数据相关分解结构

面向单一攻击的安全数据发布框架已经无法适应多组合复杂攻击的大数据环境。一方面，大部分数据发布框架设计忽略了真实环境数据的动态变化，单一考虑位置数据而忽略了非位置数据的隐私保护，无法估计数据多次发布所带来的潜在风险；另一方面，现有数据发布模型无法兼顾数据安全性

与可用性，及时对用户的反馈做出调整，未能高效地添加适当噪声减少匿名泛化过程中带来的误差，以提高数据的可用性。本文算法在实验过程中表现的稳定性有效解决了以上问题，实现了在未来大数据位置服务中应用差分隐私保护技术的可能性。

5 结束语

针对现有海量位置信息服务中的不足，本文提出了一种基于差分隐私保护机制的位置大数据发布算法，将位置数据与非位置数据区别分析，同时引入随机噪声干扰敏感数据，有效地降低数据挖掘过程中隐私泄露的风险。仿真实验基于不同数据集、不同的空间索引方法验证了本文提出算法的有效性。从实验结果可看出发布数据匿名算法 Diff_Anonmity 在实现隐私保护的同时具有更良好的数据可用性。研究如何保护位置大数据隐私的技术将具有广阔的研究前景和实用价值，现将差分隐私技术运用在大数据服务上的有关研究还比较少，本文后续工作将继续深入对该方向的研究，提出满足更高要求的大数据安全访问方法。

参考文献：

[1] 王璐，孟小峰．位置大数据隐私保护研究综述[J].软件学报，2014,25(4):693-712.
WANG L，MENG X F. Location privacy preservation in big data era: a survey[J]. Journal of Software, 2014,25(4):693-712.

[2] 霍峥，孟小峰.轨迹隐私保护技术研究[J]. 计算机学报, 2011, 34(10): 1820-1830.
HUO Z, MENG X F. A survey of trajectory privacy-preserving techniques[J]. Chinese Journal of Computers, 2011, 34(10): 1820-1830.

[3] KIDO H, YANAGISAWA Y, SATOH T. Protection of location privacy using dummies for location-based services[C]//The 21st International Conference on Data Engineering Workshops, ICDEW IEEE Computer Society. Washington, DC, 2005: 1248-1254.

[4] MOKBEL M F, CHOW C Y, AREF W G. The new casper: query processing for location services without compromising privacy[C]//The 32nd International Conference on Very Large Data Bases, VLDB Endowment. 2006: 763-774.

[5] CHOW C Y, MOKBEL M F. Trajectory privacy in location-based services and data publication[J]. ACM SIGKDD Explorations Newsletter, 2011,13(1):19-29.

[6] BAMBA B, LIU L, PESTI P, et al. Supporting anonymous location queries in mobile environments with privacygrid[C]//WWW, ACM, 2008: 237-246.

[7] DWORK C. Differential privacy[C]//The 33rd International Colloquium on Automata, Languages and Programming (ICALP). Venice, Italy, 2006:1-12.

[8] DWORK C. Differential privacy: a survey of results[C]//The 5th International Conference on Theory and Applications of Models of Computation (TAMC). Xi'an, China, 2008:1-19.

[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 X J, MENG X F. Differential privacy in data publication and analysis[J]. Chinese Journal of Computers, 2014, 37(4):927-949.

[10] 丁丽萍，卢国庆. 面向频繁模式挖掘的差分隐私保护研究综述[J]. 通信学报，2014,35(10):200-209.
DING L P, LU G Q. Survey of differential privacy in frequent pattern mining[J]. Journal on Communications, 2014, 35(10):200-209.

[11] CHEN R, ACS G, CASTELLUCCIA C. Differentially private sequential data publication via variable-length *n*-grams[C]//The 2012 ACM Conferenceon Computer and Communications Security, CCS, ACM, NewYork, 2012: 638-649.

[12] DEWRI R. Local differential perturbations: location privacy under approximate knowledge attackers[J]. IEEE Trans on Mobile Computing, 2013,12(12): 2360-2372.

[13] 叶阿勇，李亚成，马建峰，等. 基于服务相似性的 *k*-匿名位置隐私保护方法[J]. 通信学报, 2014,35(11):162-169.
YE A Y, LI Y C, MA J F, et al. Location privacy-preserving method of k-anonymous based on service similarity[J]. Journal on Communications, 2014, 35(11): 162-169.

[14] HO S S, RUAN S. Differential privacy for location pattern mining[C]//SPRINGL, ACM. 2011: 17-24.

[15] ZHANG J D, GHINITA G, CHOW C Y. Differentially private location recommendations in geosocial networks[C]//IEEE 15th International Conference on MDM, 2014:59-68.

[16] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]//The 3th Theory of Cryptography Conference (TCC). New York, USA, 2006:363-385.

[17] MCSHERRY F, TALWAR K. Mechanism design via differential privacy[C]//The 48th Annual IEEE symposium on Foundations of Computer Science (FOCS). Providence, RI, USA, 2007:94-103.

[18] CORMODE G, PROCOPIUC C, SRIVASTAVA D. Differentially private spatial decompositions[C]//IEEE 28th International Conference on Data Engineering (ICDE). 2012:20-31.

[19] 杨晓春，王雅哲，王斌，等. 数据发布中面向多敏感属性的隐私保护方法[J]. 计算机学报, 2008,31(4):574-587.
YANG X C, WANG Y Z, WANG B, et al. Privacy preserving approaches for multiple sensitive attributes in data publishing[J]. Chinese Journal of Computers, 2008, 31(4): 574-587.

作者简介：

张琳（1980-），女，江苏丰县人，博士，南京邮电大学副教授、硕士生导师，主要研究方向为分布式计算、网络安全、可信计算、隐私保护等。

刘彦（1992-），女，四川成都人，南京邮电大学硕士生，主要研究方向为分布式计算、数据挖掘、隐私保护等。

王汝传（1943-），男，安徽合肥人，南京邮电大学教授、博士生导师，主要研究方向为计算机软件、计算机网络和网格、信息安全、无线传感器网络、移动代理和虚拟现实技术等。