一种基于差分隐私的数据查询分级控制策略

李森有 季新生 游 伟 赵 星
(国家数字交换系统工程技术研究中心 郑州 450002)

摘 要 在数据的查询、发布和共享过程中,保护用户的隐私数据至关重要。现有的隐私保护模型大多未考虑不同信任等级用户的查询结果不同,而为查询数据集的所有用户提供相同隐私保护级别的数据。这种"一刀切"的方法忽略了不同个体之间数据隐私保护要求的差异性。并且多个查询用户可能具有不同的查询权限和信誉值,所查询的数据隐私属性也不尽相同。因此,这些提供相同级别的隐私保护方法无法满足隐私保护的差异化需求。为此,提出一种基于差分隐私的数据查询分级控制策略。当查询用户提交查询请求时,该隐私保护策略可以根据查询者的权限、信誉值和数据隐私属性计算查询安全信任度并量化分级,对不同信任等级的查询返回结果添加服从不同分布特性的 Laplace 噪声以保护数据隐私。为保证高可用性的数据不被低等级查询用户获取,引入可用性评估模块,在保护隐私的同时对数据的可用性进行分析。仿真实验结果表明:所提出的查询分级控制模型能够为不同等级的查询用户提供误差率在0.1%~30%范围内的数据信息,解除了差分隐私仅提供相同级别隐私保护的重要限制,有效解决了多信任等级用户查询的隐私泄露问题。并且,对最终查询返回结果进行可用性分析能够在差分隐私保护范围内最大程度地提高数据的可用性。

关键词 隐私保护,差分隐私,信任度,信任等级,查询控制,数据可用性
中图法分类号 TP309.2 **文献标识码** A **DOI** 10.11896/jsjkx.180901690

Hierarchical Control Strategy for Data Querying Based on Differential Privacy

LI Sen-you JI Xin-sheng YOU Wei ZHAO Xing
(National Digital Switching System Engineering & Technological R&D Center, Zhengzhou 450002, China)

Abstract Protecting users' private data is critical in the process of data querying, publishing and sharing. Most of the existing privacy protection models provide a uniform level of privacy protection for all query users of the dataset without considering the different query results of different trust levels. This "one size fits all" approach ignores the differences of the privacy protection requirements between individuals. Multiple query users may have different query privilege and reputation value, and the data privacy attributes of the queries are also different. Therefore, those methods of providing a uniform level of privacy protection cannot meet the differentiated needs of privacy protection. This paper proposed a hierarchical query control strategy based on differential privacy. When the query user submits a query request, this method can protect data privacy by adding Laplace noise with different distribution characteristics into the returned results for different trust levels queries. The trust levels are based on the query security trust degree according to the privilege, reputation value of users and data privacy attribute. In order to ensure high availability data cannot be obtained by low-level query users, the availability evaluation module is introduced to analyze the data availability while protecting privacy. The simulation experimental results demonstrate that the proposed control model can provide protected data with error rates ranging from 0.1% to 30% for different levels of query users, releasing the important limitation of differential privacy providing only a uniform level of privacy protection, and solving the privacy leakage problem of data query of multi-trust level users. And analyzing the availability of the query results can maximize data availability within the context of differential privacy protection.

Keywords Privacy preserving, Differential privacy, Trust degree, Trust level, Query control, Data availability

---

到稿日期:2018-09-10 修回日期:2019-01-07 本文受国家自然科学基金项目(61521003,61801515)资助。
李森有(1993-),男,硕士,主要研究方向为移动通信安全、隐私保护,E-mail:lisenyou1993@163.com;**季新生**(1968-),男,博士,教授,主要研究方向为移动通信安全、网络通信与网络安全,E-mail:jxs@ndsc.com.cn(通信作者);**游 伟**(1984-),男,博士,讲师,主要研究方向为移动通信网络安全、新一代移动通信网络技术;**赵 星**(1991-),男,博士,主要研究方向为移动边缘计算。

1　引言

随着网络信息技术的快速发展和应用软件的爆炸式增长，人与信息网络越来越密不可分，并产生了大量的个人数据。个人数据主要是个人行为、人际关系、社交媒体等活动中产生的数据信息[1]。政府、医院、学校、银行等组织机构以及我们所使用的各类应用软件的后台中都存储着大量的数据信息。海量数据的产生和使用以及大数据分析给用户敏感数据的隐私性带来了严重的威胁[2-3]。例如，在医疗系统中，主治医生、医院、科研单位以及第三方机构等不同使用者对患者的医疗记录、疾病信息等敏感数据的查询准确性有不同的要求。医生等专业人员需要完全且准确的疾病信息；医院、科研单位在进行疾病的预防、控制等相关研究时只需要相关的统计信息；医疗保健供应商等第三方机构也只需要了解大概的信息就可以为病人提供所需的服务。如果对他们的相关查询服务不加以区分，那么患者的个人敏感信息（如患病情况、就医信息等）将会在各类机构的信息交互使用中不可避免地暴露在更多的数据库中，使得其他不可信机构能够得到某个患者具体的健康信息，从而导致个人敏感信息的泄露。用户敏感数据的泄露将会给数据所有者的信息安全带来严重威胁。因此，如何在数据查询发布和共享过程中保护用户数据的隐私并尽可能地提供高可用性的数据成为近年来隐私保护的研究热点。

目前，所提出的隐私保护方案大多采用匿名模糊或数据失真处理（如添加随机噪声）等技术，运用数学回归分析、数据失真调整以及噪声尺度参数调整等方法减少噪音带来的误差，以提高数据的可用性。但是这些方案也存在一定的不足，即当不同权限、信誉等级的查询用户对敏感度不同的数据进行相关查询时，相同的查询结果将会引起隐私信息的泄露。

针对上述问题，为了在数据查询、发布和共享过程中保证敏感数据的隐私性和可用性，本文提出一种基于差分隐私的数据查询分级控制策略。首先根据查询者的信誉值、访问权限确定查询者信任值；然后根据数据隐私属性的动态性来量化分析信任度等级，并将其映射为不同的信任等级和隐私保护参数；最后根据该隐私保护参数向查询结果中添加符合 Laplace 分布的随机噪声，实现数据的差分隐私保护和不同的可用性。仿真实验证明：所提出的分级控制策略能够在保护数据隐私的前提下，针对查询服务的内在特征发布不同准确率的数据信息。

本文第 2 节描述了相关工作；第 3 节介绍提出的基于差分隐私的数据查询分级控制策略；第 4 节给出仿真实验和分析；最后总结全文。

2　相关工作

已有的关于隐私保护技术的研究主要集中在数据失真、数据加密以及限制发布的隐私保护技术，如通过 k-anonymity[4]，l-diversity[5] 以及 t-closeness[6] 等方法泛化敏感信息的标志属性，从而在一定程度上抵抗部分攻击，但是对于一些新的攻击模型，如背景知识攻击、相似性攻击和倾斜攻击等，仍会面临隐私数据泄露的威胁。此外，这些隐私保护技术在带来某类优势的同时也不可避免地存在某些缺陷，如信息丢失、计算开销大、隐私保护程度无法衡量、无法提供不同安全级别的数据信息等。

差分隐私[7-9] 是 Dwork 针对统计数据库的隐私泄露问题提出的一种新的基于严格数学背景的隐私保护机制，它逐渐成为了隐私保护领域的一个研究热点。差分隐私保护就是保证无论一个具体记录是否在数据集中，对查询结果的影响都微乎其微。对于相差仅为一个记录的相邻数据集，通过向查询结果的返回值中添加服从数学分布的噪声值，可使得相同的查询在两个数据集上产生相同结果的概率比值趋近 1，攻击者无法通过差分攻击对输出结果进行分析得到准确的个体信息，从而达到隐私保护的目的。目前，差分隐私技术的研究主要集中在数据发布、数据挖掘和学习查询处理等方面，也有研究将差分隐私保护技术应用到 K-means 聚类分析[10]、频繁模式挖掘[11] 以及位置大数据服务的位置信息保护[12] 中。文献[13]针对差分隐私数据发布过程中数据分布稀疏、查询偏差过大的问题，提出了一种基于单元合并的差分隐私数据发布方法，在保持数据有效性的同时减少了查询的误差。文献[14]提出了一种与标准差分隐私具有相同隐私保证的个性化差分隐私概念，通过允许数据控制器将实际的数据集进行失真调整，从而使其具有更小的数据失真和更高的分析精度。文献[15]针对隐私保护过程中的数据可用性问题，将高斯过程与差分隐私回归分析相结合，寻找保护敏感数据的最佳协方差噪声，在达到最大准确性的同时提供隐私保证。文献[16]提出了一种以 RBAC 为中心的隐私访问控制模型，给出了服务提供者的信誉度分级方法，为不同的信誉度等级的服务提供者分配不同的角色，以控制其对敏感隐私信息的访问。文献[17]仅针对查询权限对用户进行分级，并对数据查询结果进行隐私保护处理，从而提供隐私保护程度不同的查询结果，在一定程度上避免了隐私泄露。文献[18]针对云数据共享场景下多租户对统一数据的不同级别的访问权限，提出了一种对数据集查询请求的分级访问控制策略，根据访问权限的差异提供细粒度的数据隐私保护。

但是，以上方法并未考虑网络大数据环境下不同信任等级查询用户对隐私保护和数据可用性的具体要求。在对数据进行发布和共享的过程中忽略了查询者之间的差异和查询数据的隐私属性而提供相同的数据查询和发布服务会引发用户敏感信息的泄露。因此，为了更好地保护数据库中的用户敏感数据，需要充分考虑查询用户的信任值和查询数据的隐私属性，以发布不同数据准确率的可用数据信息，从而在保证隐私保护的前提下提供更精确的数据信息。

3　基于差分隐私的数据查询分级控制策略

本节从查询者的权限和信誉值以及查询数据的隐私属性两个方面来考虑查询的信任度特征，提出一种用查询信任度来衡量查询者等级的数据查询分级控制策略，将数据收集服

务器中存储的数据信息与差分隐私保护技术相结合，从而进行查询分级控制处理，以为不同的查询用户提供不同准确率的可用数据。

3.1 基于差分隐私保护的查询控制模型

差分隐私保护机制被广泛应用于隐私的保护数据发布、数据挖掘等领域，目的是在数据发布共享时保护数据库中的个体敏感信息。本文利用的 ε-差分隐私的定义[7]如下。

定义 1(ε-差分隐私) 设有定义域为 D，值域为 R 的随机函数 $M:D\to R$，$Range[M]$为 M 所有可能的输出构成的集合，$Sm\subseteq Range[M]$。对于任意两个差别至多为一个记录的相邻数据集 $D1$ 和 $D2$，若算法 M 满足等式(1)，则称算法 M 提供差分隐私保护。

$$Pr[M(D1)\in Sm]\leqslant exp(\varepsilon)\times Pr[M(D2)\in Sm] \tag{1}$$

其中，Pr[M]表示隐私被披露的概率；ε 称为隐私保护预算参数，ε 取值越小，隐私保护程度越高。

而在差分隐私保护机制中考虑查询者权限的差异性和查询数据的隐私性，能有效防止不同权限等级的用户查询敏感数据带来的隐私泄露威胁。为实现针对不同等级查询用户发布不同数据准确率的可用数据信息的目的，提出了如图 1 所示的基于差分隐私保护的数据查询分级控制模型，其主要由原始数据集、查询分级控制、查询者 3 部分构成。其中，查询分级控制是最重要的部分，主要有数据预处理模块、查询分级控制模块和可用性分析模块。

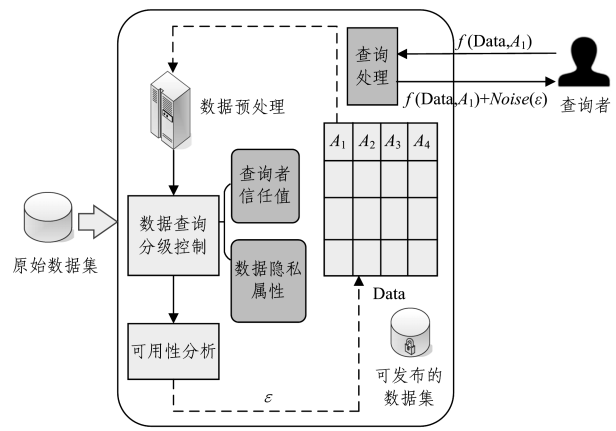

图 1 数据查询分级控制模型

Fig. 1 Hierarchical control model for data query

查询分级控制模块结合了差分隐私保护技术，负责对查询进行分级控制。基于数据研究和使用的查询需要考虑查询权限和查询数据的隐私属性等因素。针对现有差分隐私保护机制中，数据拥有者无法灵活地为不同的查询用户提供不同隐私保护程度的可用数据信息，无法满足数据差异化隐私保护需求的问题，本文所提出方法的解决思路是根据数据查询者的权限、信誉值以及数据的隐私属性等进行信任度量化分析，并划分相应等级来对应不同的隐私保护参数，以达到差异化的差分隐私保护效果。由可用性评估模块衡量差分隐私处理后的数据可用性，实现可控的敏感数据隐私性和可用性保护，能够有效防止恶意攻击者通过信息查询获取用户隐私信息，同时能够在降低隐私泄露的前提下大幅提高发布数据的有效使用率。最后数据管理系统将处理后的结果反馈给查询用户。

3.2 查询安全信任度的计算方法

本文提出查询安全信任度来量化查询的安全可信程度，查询安全信任度由查询主体属性值和查询客体属性值决定，值越大表明查询越可信。本文中查询安全信任度由查询者信任值和数据隐私属性值两部分组成，其计算公式为：

$$QT=\alpha UT+\beta DA \tag{2}$$

其中，QT 为查询安全信任度，UT 为查询者信任值，DA 为数据隐私属性值，α 和 β 分别为查询者信任值和数据隐私属性值的权重，$\alpha+\beta=1$。权重的分配由数据所有者根据数据隐私保护要求确定，对不同的隐私保护需求可以调节查询者信任值和数据隐私属性值的权重来进行安全信任度计算和数据的处理[19]。例如，当安全信任度不考虑数据隐私属性时，$\beta=0$。

查询者信任值反映了具有不同权限和信誉值的数据查询用户的主观差异，不同的查询权限和信誉值对应的信任值也是不同的。数据隐私属性值则反映了不同敏感数据的隐私程度，通常由领域专家与数据所有者确定。为了方便信任度计算，本文将查询者权限和信誉值以及数据隐私属性值量化为 0 到 1 之间的标量，具体如表 1 所列。

表 1 信任模型的变量和评价

Table 1 Variable and value for trust model

变量	评价水平	参考范围
查询者权限 Q_{pri}	低	[0,0.2]
	一般	(0.2,0.6]
	高	(0.6,1]
查询者信誉值 Q_{rep}	坏	[0,0.2]
	平均	(0.2,0.6]
	好	(0.6,1]
数据隐私属性值 DA	低	[0,0.2]
	一般	(0.2,0.6]
	高	(0.6,1]

查询者信任值计算：查询者信任值由查询权限属性和信誉值两部分组成。在对查询者信任值进行计算时，首先根据用户属性分析查询者的权限值和信誉值，权限值和信誉值为取值在 0 到 1 之间且服从正态分布的随机变量。根据式(2)和数据隐私属性 DA 的取值范围，查询者信任值应在 0～1 之间，选择二维高斯函数[20]作为相应的函数对查询者信任值进行计算，其计算公式为：

$$\begin{aligned} UT &= f(Q_{pri},Q_{rep}) \\ &= \exp\left(-\left(\frac{(Q_{pri}-1)^2}{2}+\frac{(Q_{rep}-1)^2}{2}\right)\right) \end{aligned} \tag{3}$$

本文通过对查询安全信任度的等级划分来描述不同查询的信任等级，从而实现差分隐私保护中的查询分级控制策略。根据表 1 的量化参考值和式(2)、式(3)的计算方法，对信任等级进行量化，并分成 n 个等级，每个信任等级对应一个隐私保护参数 ε，第 i 个信任等级设定的区间为$\left[QT_{\min}+(i-1)\frac{QT_{\max}-QT_{\min}}{n},QT_{\min}+i\frac{QT_{\max}-QT_{\min}}{n}\right)$，相互之间的对应关系如表 2 所列。

表 2　查询信任等级的划分

Table 2　Division of querying trust levels

信任等级	查询安全信任度	隐私保护参数
1 级信任 $L1$	$[QT_{\min}, QT_{\min}+\frac{QT_{\max}-QT_{\min}}{n})$	ε_1
2 级信任 $L2$	$[\frac{QT_{\max}+(n-1)QT_{\min}}{n}, \frac{2QT_{\max}+(n-2)QT_{\min}}{n})$	ε_2
3 级信任 $L3$	$[\frac{2QT_{\max}+(n-2)QT_{\min}}{n}, \frac{3QT_{\max}+(n-3)QT_{\min}}{n})$	ε_3
⋮	⋮	⋮
i 级信任 Li	$[\frac{(i-1)QT_{\max}+(n-i+1)QT_{\min}}{n}, \frac{iQT_{\max}+(n-i)QT_{\min}}{n})$	ε_i
⋮	⋮	⋮
n 级信任 Ln	$[\frac{(n-1)QT_{\max}+QT_{\min}}{n}, QT_{\max}]$	ε_n

信任是评判主体根据某些特性进行的主观评价，根据是否满足一定的信任等级进行交互。信任等级的划分主要体现了对查询者信任程度的高低，在具体应用中不必过于精细。本文中约定：等级越高，信任度越高。

通过分析查询者权限属性的差异和数据隐私属性的动态变化来对查询安全信任度进行计算，查询安全信任度对应不同的信任等级和不同的隐私保护参数 ε，使得敏感数据隐私属性变化时，不同等级的查询用户可以获取不同可用性的隐私保护数据，在保护隐私的前提下在实现查询响应中对敏感数据信息准确率的分级控制。

3.3　查询分级控制算法

为了满足数据所有者的隐私保护需求，我们可以根据查询用户的权限差异以及数据的敏感属性对查询者进行分级控制，以实现个性化的隐私保护。所提出的查询分级控制算法的具体执行过程为：首先根据查询者的查询权限和信誉值分析其信任值；再根据查询数据的隐私属性与查询者的信任值将其映射到不同的信任等级，建立隐私保护预算参数 ε 与查询信任值和数据隐私属性的等级对应关系；然后根据不同的隐私保护预算参数，通过 Laplace 机制随机算法实现差异化差分隐私保护，进而实现数据查询的分级控制。

$$Q(D)=f(D)+Lap(\Delta q/\varepsilon) \tag{4}$$

其中，$Lap(\Delta q/\varepsilon)$为服从尺度参数为 $b=\Delta q/\varepsilon$ 的 Laplace 分布的随机噪声。函数敏感度 Δq 定义为：

$$\Delta q=\max_{(D_1,D_2)}\| f(D_1)-f(D_2)\|_1 =\max_{(D_1,D_2)}\sum_{i=1}^{k}|f(D_1)_i-f(D_2)_i| \tag{5}$$

函数敏感度 Δq 是决定加入噪声量大小的重要参数，$\| f(D_1)-f(D_2)\|_1$ 表示在相邻数据集 D_1 和 D_2 上函数输出值的一阶范数距离。函数敏感度是指在数据集中删除或添加任意一条记录对输出结果产生的最大影响，其由函数本身性质决定，与其他因素无关。基于差分隐私的查询分级控制算法如算法 1 所示。

算法 1　基于差分隐私的查询分级控制算法

输入：查询者权限 Q_{pri}，信誉值 Q_{rep} 以及所查询数据的隐私属性 DA，真实的查询输出值 f(D)

输出：查询分级控制下的差分隐私保护结果 Q(D)

1. $Q_{pri}\leftarrow 0$；$Q_{rep}\leftarrow 0$；DA$\leftarrow 0$
2. For each $Q_i\in Q$，values$\leftarrow Q_i$，根据式(3)：$UT=f(Q_{pri},Q_{rep})=\exp(-(\frac{(Q_{pri}-1)^2}{2}+\frac{(Q_{rep}-1)^2}{2}))$计算查询信任值 UT_i
3. Updata DA_i
4. $QT=\alpha UT+\beta DA$，计算查询信任度 QT_i
5. for $QT_i\in[QT_{min}+(i-1)\frac{QT_{max}-QT_{min}}{n}, QT_{min}+i\frac{QT_{max}-QT_{min}}{n})$，get G_i and ε_i according to Table 2
6. end for
7. $\text{LapNoise}_\varepsilon(\frac{\Delta q}{\varepsilon_i})=-\frac{\Delta q}{\varepsilon_i}\text{sgn}(U)\ln(1-2|U|)$
8. $Q'(D)\leftarrow f(D)+\text{LapNoise}_\varepsilon(\frac{\Delta q}{\varepsilon})$
9. if $Q'(D)$satisfy (α,β)- useful
   $Q(D)=Q'(D)$
10. else return (7)
11. end if
12. end for
13. return Q(D)

算法 1 中，第 2－5 步是通过计算公式根据查询者的权限、信誉值和数据隐私属性值得出信任度等级，并确定不同等级的隐私保护参数 ε_i；第 7－8 步是对不同隐私保护程度的数据查询结果根据确定的隐私保护参数 ε_i 添加服从 Laplace 分布的随机噪声；第 9—10 步是进行差分隐私处理，对所得的结果进行可用性分析，以得到既保障数据隐私性又具有高可用性的数据信息。

3.4　算法分析

3.4.1　算法性能分析

算法的性能分析是根据差分隐私 Laplace 机制下查询返回结果的反馈信息以及数据拥有者对数据隐私的要求以及查询者权限的差异，动态调整发布数据的查询精确度。

(1)隐私性分析

数据的隐私保护效果主要是通过对数据集处理前后数据分布的差异来侧面反映的，差异程度直接影响着数据处理的隐私性。由式(1)可以得到，当随机算法满足 ε-差分隐私保护时，ε 越小，不同数据集通过算法 M 的输出值的差异也越小；当 ε 取 0 时，数据集处理前后以相同的概率输出相同的值，此时隐私保护程度最高，但同时数据的可用性也最低。

(2)算法的有效性分析

差分隐私保护技术中，在数据查询的真实结果中添加符合 Laplace 分布的噪声的算法复杂度会直接影响算法的运行时间，进而影响差分隐私查询分级控制模型的运算效率。设 $Q=\{Q_n\}_{n=1,2,3,\cdots}$ 为全体查询集，$X=\{X_n\}_{n=1,2,3,\cdots}$ 为总体数据域。算法的复杂度与所需要添加噪声的数据集大小成正比。在对查询结果进行添加噪声处理时，如果差分隐私保护算法的运行时间是多项式$(n,\log(|Q_n|),\log(|X_n|))$，则可以说该算法是有效的。

3.4.2 数据的可用性分析

在使用差分隐私对数据库进行隐私保护的研究中，需要考虑差分隐私保护算法中所添加的噪声对数据可用性的影响。从理论角度来看，通常采用(α,β)-$useful$[9,21]技术来衡量。

(α,β)-$useful$：对于差分隐私算法 M，设 Q 是一组查询函数，X 是数据域。如果对于任何 n 项数据库 x，Q 和 X 满足等式(6)，则算法 M 满足(α,β)-$useful$。

$$Pr[|M(D_1)-M(D_2)|\leqslant\alpha]>1-\beta \tag{6}$$

在实际应用中，常根据不同的使用场景选择不同的差度量方法，如相对误差、绝对误差、欧拉函数以及 F-measure 等。根据所采用的数据计量操作，本文选择平均相对误差作为数据库处理前后数据可用性分析的标准。

$$ARE(Q(D))=\sum_{i=1}^{n}\frac{1}{n}\frac{|Q_i(D)-Q_i(D')|}{\max(Q(D),s)} \tag{7}$$

其中，$Q_i(D)$为数据集第 i 次查询的真实输出结果，$Q_i(D')$为相对应的隐私保护处理后的查询输出结果，s 是当 $Q(D)$出现特别小的情况时的数值下限。

对于查询函数 q，数据在差分隐私 Laplace 机制保护前后输出的查询结果与真实结果的相似度体现了差分隐私保护算法对数据可用性的影响。本文所采用的平均相对误差 $ARE(Q(D))$的衡量标准与数据的可用性呈负相关。可用性分析模块判断差分隐私保护的查询返回结果的数据隐私性和可用性是否满足用户的需求，并通过反馈机制向隐私保护模型反馈误差信息。当查询处理结果的隐私保护程度低于用户隐私要求时，重新进行噪声添加处理，直至得到满足用户要求的结果。当满足隐私保护的前提时，分析数据的准确性和可用性，从而适当降低数据添加噪声造成的干扰，降低差分隐私对数据集的泛化程度，进而向查询者提供可用性较高的数据结果。数据的可用性分析是在确保数据隐私安全性的前提下减少噪声添加所带来的误差，同时平衡数据的隐私性和可用性。

3.4.3 安全性分析

(1)基于差分隐私的保护

差分隐私保护是一种新的基于数据失真的具有严格数学背景的隐私保护技术，其提供了隐私保护程度可量化、可评估、可证明的方法。本文提出的基于差分隐私的数据查询分级控制策略是基于差分隐私 Laplace 机制的数据隐私保护，在满足差分隐私保护的基础上对数据集的隐私保护处理水平进行量化分析，以实现隐私度量可控的隐私保护。

(2)差异化的数据隐私保护

所提出的查询分级控制方法通过分析查询者信任值和查询数据的隐私属性值，将查询信任度进行量化分级并设置对应的隐私保护预算参数，从而在查询返回结果中添加服从不同分布特性的 Laplace 噪声，实现对数据发布共享中细粒度差异化的隐私保护。

(3)数据可用性隐私保护

在本文提出的数据查询分级控制模型中引入可用性评估模块，对差分隐私处理后的数据可用性进行分析，使不同隐私保护参数对应的不同信任等级的查询获取的数据可用性不同，保证了高准确性的查询结果不会被低等级的查询用户获取，能够有效防止恶意攻击者通过信息查询获取用户隐私信息，实现了对可控的敏感数据的隐私性和可用性的保护，大大降低了隐私泄露的风险。

4 仿真实验与分析

4.1 实验环境与数据

为了对差分隐私查询分级控制算法的可用性和有效性进行分析，本节通过仿真实验进行验证。实验的硬件环境为 Intel(R) Core(TM) i5-3230M CPU_@_2.60 GHz，操作系统为 Microsoft Windows10，内存为 8 GB。仿真平台为 MATLAB2016a。实验数据集采用 UCI[1)] 中的 Adult test Dataset，Car Evaluation Dataset，Breast Cancer Wisconsin Dataset 以及 Heart Disease Cleveland Dataset，主要涉及美国人口普查数据测试集、车辆评估、威斯康星州乳腺癌数据以及克利夫兰市心脏病数据信息，其在数据发布隐私保护研究中被广泛使用，具体信息如表 3 所列。

表 3 数据集属性信息

Table 3 Dataset attribute information

数据集名称	数据集类型	数据集记录数	属性数	数据集别名
Adult test	Multivariate	16281	14	D_1
Car Evaluation	Multivariate	1728	6	D_2
Breast Cancer	Multivariate	569	32	D_3
Heart Disease	Multivariate	303	5	D_4

4.2 实验分析

在本文提出的差分隐私保护中，对数据集中不同的敏感值查询结果添加符合 Laplace 分布的噪声。为了对数据的精确度进行测量分析，我们对 D_1 数据集进行了相对误差值的实验。假设有 100 个查询者进行某一敏感数据结果的查询，本实验分析差分隐私算法下 $\varepsilon=0.01,0.02,0.05,0.1,0.2,1$ 时查询结果与真实结果之间的误差值，实验结果如图 2 所示。然后针对数据集 D_1,D_2,D_3,D_4 在不同隐私预算 ε 下进行多组实验，利用差分隐私动态查询分级控制算法对查询数据结果进行添加噪声处理，每组实验中分别对每个数据集特定 ε 值下的差分隐私分级算法进行多次实验并取平均值，分析实验的平均相对误差。实验结果如图 3 所示。

从对图 2、图 3 的分析中可以看出，当 ε 值很小时，由于算法对敏感数据信息提供了很强的隐私保护能力，所添加的噪音量也比较大，因此误差率很大。这时查询者只能大概知道真实结果的范围，对于数据精度要求不高的查询，该隐私预算参数 ε 既能保证敏感数据的隐私性，又能提供满足要求的查询结果。而当 ε 值增大时，根据 Laplace 概率密度分布图，噪声值越来越小且越来越集中，输出数据的精度越来越高。当 ε 取值为 1 或者更大时，所添加的平均噪音量为 1 或者更小，输出结果几乎与原始数据结果相同，使得查询结果的错误率

1) http://archive.ics.uci.edu/ml/

大大降低，数据的精确度和可用性更高。在数据集大小方面，$D_1 > D_2 > D_3 > D_4$，从图 3 可知，在相同的数据精确度要求下，数据集越大，所需的 ε 值越小，这说明对于大型数据集，在较高的隐私保护级别下，数据可用性更高。

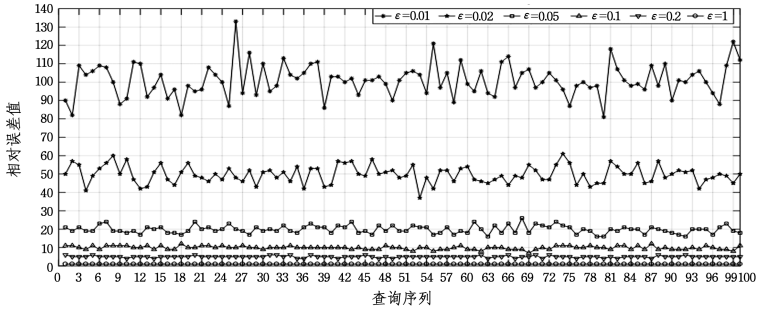

图 2　数据集 D_1 中不同隐私预算下的相对误差值

Fig. 2　Relative error value under different privacy budgets in D_1

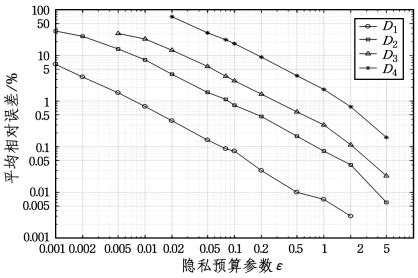

图 3　不同隐私预算下的平均相对误差

Fig. 3　Average relative error under different privacy budgets

为了更好地说明不同查询级别下的隐私保护，我们对数据集 D_2 进行了分析。从图 4 可知，随着 ε 取值的增大，数据查询结果的误差率大幅下降，隐私保护效果减弱，数据的准确性和可用性升高。因此，对于不同查询等级的用户，所提出的查询分级控制策略可以提供不同误差率的查询结果。在图 4 所示的数据集 D_2 中，通过对查询者的权限、信誉值以及数据的隐私属性等进行动态分析得出相应的等级权限，对于等级权限低的用户，在添加噪声处理时采用 ε 取值在 0.001 3～0.007 的 Laplace 机制，使其获得的数据结果的误差率为 10%～30%；对于等级权限较高的用户，为了向其提供准确率较高的结果，可以选择 ε=0.09，0.17，0.8 等值时对应的 Laplace 差分隐私保护算法产生的结果。在实际应用中，由于数据集大小的不同以及敏感数据隐私要求的差异，不同等级权限的结果误差率也不尽相同，因此用户等级的划分是由数据所有者决定并在算法的输出中进行限定的，从而实现查询分级控制，在保证数据隐私性的同时尽可能地提高数据的可用性和准确性。

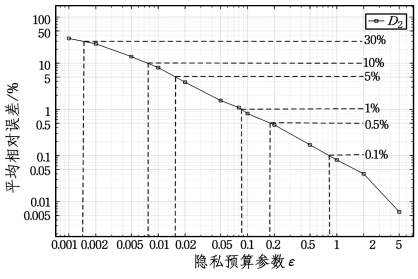

图 4　数据集 D_2 中的查询结果误差率

Fig. 4　Error rate of query result in D_2

结束语　本文针对大数据环境下不同用户的多样性隐私保护需求，提出了一种基于差分隐私的数据查询分级控制策略。通过对数据查询者的访问权限、信誉值以及查询数据的隐私属性进行信任度量化评估划分不同的信任等级，并使其对应不同的隐私保护参数。在差分隐私保护过程中通过添加不同的噪声尺度实现输出结果准确性的分级控制，通过可用性分析在有效降低数据查询中的隐私泄露风险的同时提高可用数据的准确性，为海量数据的查询服务提供细粒度和差异化的隐私保护。仿真实验针对不同的数据集和不同隐私保护预算分析数据的可用性和准确性，通过分析可以看出，提出的策略能够在保护用户敏感数据的前提下，对不同的用户等级和数据隐私要求提供精准的且具有差异化的可用数据信息。在未来的工作中将继续深入研究其他扰动机制和查询机制的差分隐私保护技术、优化算法和等级划分等，以提供满足更高要求的差分隐私保护方法。

参 考 文 献

[1] LIANG F，YU W，AN D，et al. A Survey on Big Data Market：Pricing，Trading and Protection [J]. IEEE Access，2018，6：15132-15154.

[2] VIJI D，SARAVANAN K，HEMAVATHI D. A journey on privacy protection strategies in big data[C]// International Conference on Intelligent Computing and Control Systems. IEEE，2018.

[3] YU L，WANG H Y. Application of Big Data Technology in Network Security Analysis[J]. Modern Information Technology，2018，2(2)：158-161.(in Chinese)
刘瑜，王洪艳. 大数据技术在网络安全分析中的应用初探[J]. 现代信息技术，2018，2(2)：158-161.

[4] LATANYA S. k-anonymity：a model for protecting privacy [J]. International Journal on Uncertainty，Fuzziness and Knowledge based Systems，2002，10(5)：557-570.

[5] ASHWIN M，JOHANNES G，DANIEL K. ℓ-Diversity：Privacy Beyond k-Anonymity[J]. International Conference on Data Engineering，2006，1(1)：24-35.

[6] LI N H，LI T C，SURESH V. t-Closeness：Privacy Beyond k-Anonymity and l-Diversity[C]// IEEE International Conference

on Data Engineering. 2007:106-115.

[7] XIONG P,ZHU T Q,WANG X F. A Survey on Differential privacy protection and application [J]. Chinese Journal of Computers,2014,37(1):101-122. (in Chinese)
熊平,朱天清,王晓峰. 差分隐私保护及其应用[J]. 计算机学报,2014,37(1):101-122.

[8] DWORK C,ROTH A. The Algorithmic Foundations of Differential Privacy[M]. Hanover:Now Publishers Inc. 2014.

[9] ZHU T Q,LI G,ZHOU W,et al. Differentially Private Data Publishing and Analysis:A Survey [J]. IEEE Transactions on Knowledge & Data Engineering,2017,29(8):1619-1638.

[10] REN J,XIONG J,YAO Z,et al. DPLK-Means:A Novel Differential Privacy K-Means Mechanism[C]// IEEE Second International Conference on Data Science in Cyberspace. IEEE,2017:133-139.

[11] CHENG X,SU S,XU S,et al. A Two-Phase Algorithm for Differentially Private Frequent Subgraph Mining [J]. IEEE Transactions on Knowledge & Data Engineering,2018,30(8):1411-1425.

[12] YIN C,XI J,SUN R,et al. Location Privacy Protection based on Differential Privacy Strategy for Big Data in Industrial Internet-of-Things [J]. IEEE Transactions on Industrial Informatics,2018,14(8):3628-3636.

[13] LI Q,LI Y,ZENG G,et al. Differential privacy data publishing method based on cell merging[C]// IEEE,International Conference on Networking,Sensing and Control. IEEE,2017:778-782.

[14] SORUIA-COMAS J,DOMINGO-FERRER J,SANCHEZ D,et al. Individual Differential Privacy:A Utility-Preserving Formulation of Differential Privacy Guarantees [J]. IEEE Transactions on Information Forensics & Security,2017,12(6):1418-1429.

[15] SMITH M,LOPEZ MAA,ZWIESSELE M,et al. Differentially private regression with Gaussian processes[C]// International Conference on Artificial Intelligence and Statistics. 2018.

[16] ZHANG X M,HUANG Z Q,SUN Y. Research on Privacy Access Control Based on RBAC [J]. Compute Science,2016,43(1):166-171. (in Chinese)
张学明,黄志球,孙艺. 基于 RBAC 的隐私访问控制研究[J]. 计算机科学,2016,43(1):166-171.

[17] ZHANG W J,LI H. A differentially-private mechanism for multi-level data publishing [J]. Chinese Journal of Network and Information Security,2015:219-223. (in Chinese)
张文静,李晖. 差分隐私保护下的数据分级发布机制[J]. 网络与信息安全学报,2015,1(1):58-65.

[18] JIA Z,WU W,GUO Y,et al. A privacy-preserving multi-levels access control protocol for sensitive data sharing[C]// International Conference on Communication Technology. IEEE,2016:883-887.

[19] DUBEY J,TOKEKAR V. Bayesian network based trust model with time window for Pure P2P computing systems[C]// Wireless Computing and NETWORKING. IEEE,2015:219-223.

[20] WIKIPEDIA. Gaussian function [EB/OL]. https://en.wikipedia.org/wiki/Gaussian_function.

[21] BONOMI L,XIONG L. A two-phase algorithm for mining sequential patterns with differential privacy[C]// Proceedings of the 22nd ACM international conference on Information & Knowledge Management. ACM,2013:269-278.