

查询处理方法及装置

5 本申请要求于 2021 年 09 月 23 日提交中国国家知识产权局、申请号为 202111110468.2、
申请名称为“查询处理方法及装置”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

10 本说明书一个或多个实施例涉及数据处理技术领域，尤其涉及一种查询处理方法及装置。

背景技术

 随着大数据时代的到来，如何挖掘数据价值已成为当下的研究热点。在一种挖掘方式
15 中，通过对大量数据进行统计处理，向外界提供统计查询服务。然而，此种查询服务存在
泄露单条数据记录的风险，例如，张三查询到前 500 行数据的均值为 20，之后又查询到前
501 行数据的均值为 20.1，此时，其可以得到第 501 行数据的数值为 70.1。

 差分隐私 (Differential Privacy，简称 DP) 技术被用于统计查询场景，以应对隐私泄
露问题。此项技术的难点在于，如何平衡数据的隐私安全性和查询结果的准确性，因为前
20 者的提高往往会导致后者的降低。因此，迫切需要一种改进的差分隐私方案，可以最小化
敏感数据的隐私泄露风险，同时最大化查询结果的准确性。

发明内容

 本说明书一个或多个实施例描述了一种查询处理方法及装置，针对一批次查询中的每
25 个查询，为其分配确保不会泄露数据隐私的最低噪声功率，从而最大化查询结果的准确性。

 根据第一方面，提供一种保护数据隐私的查询处理方法，包括：确定有待针对目标数
据集进行查询的 L 条查询各自的查询类型，所述目标数据集记录业务对象的数据；确定各
个查询类型针对所述目标数据集的查询敏感度；基于各条查询分别对应的查询敏感度，以
及针对所述 L 条查询的总集预先设定的隐私预算参数，确定各条查询分配得到的噪声功率。

30 在一个实施例中，确定有待针对目标数据集进行查询的 L 条查询各自的查询类型，包
括：接收针对所述目标数据集的 L 条查询请求，其中各条查询请求指示各自的查询类型。

 在一个实施例中，确定有待针对目标数据集进行查询的 L 条查询各自的查询类型，包
括：获取针对所述目标数据集预先配置的可查询次数 L 及其中各次查询的查询类型。

 在一个实施例中，所述查询类型为以下中的任一种：计数、最大值、最小值、均值和
35 方差。

在一个实施例中，所述业务对象为以下中的任一种：用户、商品、业务事件。

在一个具体的实施例中，所述业务事件为以下中的任一种：注册、访问、登录、支付。

在一个实施例中，所述业务对象为用户，所述隐私数据为以下中的任一种：年龄、性别、收入、兴趣爱好、生理指标、操作指标。

- 5 在一个实施例中，确定各个查询类型针对所述目标数据集的查询敏感度，包括：针对所述各个查询类型，根据第一结果和第二结果之间的最大绝对差值，得到该查询类型对应的查询敏感度；其中，所述第一结果是针对所述目标数据集进行该类型查询得到的结果，所述第二结果是针对所述目标数据集的邻近数据集进行该类型查询得到的结果。

- 10 在一个实施例中，所述查询类型包括计数查询；其中，确定各个查询类型针对所述目标数据集的查询敏感度，包括：将所述计数查询的查询敏感度确定为数值 1。

在一个实施例中，所述查询类型包括最大值查询/最小值查询；其中，确定各个查询类型针对所述目标数据集的查询敏感度，包括：确定所述目标数据集中的最大值和最小值；将所述最大值减去所述最小值的结果，确定为所述最大值查询/最小值查询的查询敏感度。

- 15 在一个实施例中，所述查询类型包括均值查询；其中，确定各个查询类型针对所述目标数据集的查询敏感度，包括：确定所述目标数据集中的最大值；将所述最大值的绝对值与所述目标数据集的数据个数加 1 之间的比值，确定为所述均值查询的查询敏感度。

- 20 在一个实施例中，所述查询类型包括方差查询；其中，确定各个查询类型针对所述目标数据集的查询敏感度，包括：确定所述目标数据集中的最大值和最小值；将所述方差查询的查询敏感度确定为以下因子的乘积：所述最大值和最小值之间差值的平方，所述目标数据集的数据个数，以及所述数据个数加 1 后进行平方运算的倒数。

- 25 在一个实施例中，基于各条查询分别对应的查询敏感度，以及针对所述 L 条查询的总集预先设定的隐私预算参数，确定各条查询分配得到的噪声功率，包括：基于各条查询的查询敏感度，确定所述 L 条查询的查询敏感度和值；针对任意一条查询，根据该条查询的查询敏感度，所述查询敏感度和值，以及所述隐私预算参数，确定该条查询分配得到的噪声功率。

- 30 在一个具体的实施例中，根据该条查询的查询敏感度，所述查询敏感度和值，以及所述隐私预算参数，确定该条查询分配得到的噪声功率，包括：获取均值变量的变量值，该变量值基于所述隐私预算参数的参数值，以及差分隐私的高斯机制下所述隐私预算参数和均值变量的约束关系而确定；将该条查询的噪声功率确定为以下因子的乘积：该条查询的查询敏感度，所述查询敏感度和值，以及所述变量值进行平方运算后的倒数。

在一个实施例中，所述隐私预算参数包括预算项参数和松弛项参数。

在一个实施例中，在确定各条查询分配得到的噪声功率之后，所述方法还包括：针对所述 L 条查询中任意的目标查询，将其实际返回结果确定为其原始查询结果叠加从差分隐私的目标噪声分布采样的目标噪声，其中目标噪声分布基于所述目标查询分配得到的噪声

功率而确定。

在一个具体的实施例中，所述目标噪声分布为高斯噪声分布，所述高斯噪声分布以所述目标查询的噪声功率为方差，以 0 为均值。

5 在一个具体的实施例中，所述的方法，还包括：接收针对所述目标数据集的当前查询请求，其对应当前查询类型；判断对应所述当前查询类型的已处理请求数量是否小于预定阈值，所述已处理请求数量对应的查询请求针对所述目标数据集；在判断出小于的情况下，将所述当前查询请求作为所述目标查询。

10 根据第二方面，提供一种保护数据隐私的查询处理装置，包括：查询类型确定单元，配置为确定有待针对目标数据集进行查询的 L 条查询各自的查询类型，所述目标数据集记录业务对象的数据；敏感度确定单元，配置为确定各个查询类型针对所述目标数据集的查询敏感度；噪声功率确定单元，配置为基于各条查询分别对应的查询敏感度，以及针对所述 L 条查询的总集预先设定的隐私预算参数，确定各条查询分配得到的噪声功率。

在一个实施例中，所述查询类型确定单元具体配置为：接收针对所述目标数据集的 L 条查询请求，其中各条查询请求指示各自的查询类型。

15 在一个实施例中，所述查询类型确定单元具体配置为：获取针对所述目标数据集预先配置的可查询次数 L 及其中各次查询的查询类型。

在一个实施例中，所述业务对象为以下中的任一种：用户、商品、业务事件。

20 在一个实施例中，所述敏感度确定单元具体配置为：针对所述各个查询类型，根据第一结果和第二结果之间的最大绝对差值，得到该查询类型对应的查询敏感度；其中，所述第一结果是针对所述目标数据集进行该类型查询得到的结果，所述第二结果是针对所述目标数据集的邻近数据集进行该类型查询得到的结果。

在一个实施例中，还包括实际结果确定单元，配置为针对所述 L 条查询中任意的目标查询，将其实际返回结果确定为其原始查询结果叠加从差分隐私的目标噪声分布采样的目标噪声，其中目标噪声分布基于所述目标查询分配得到的噪声功率而确定。

25 在一个具体的实施例中，还包括目标查询确定单元，配置为：接收针对所述目标数据集的当前查询请求，其对应当前查询类型；判断对应所述当前查询类型的已处理请求数量是否小于预定阈值，所述已处理请求数量对应的查询请求针对所述目标数据集；在判断出小于的情况下，将所述当前查询请求作为所述目标查询。

30 根据第三方面，提供了一种计算机可读存储介质，其上存储有计算机程序，当该计算机程序在计算机中执行时，令计算机执行上述第一方面提供的方法。

根据第四方面，提供了一种计算设备，包括存储器和处理器，存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现上述第一方面提供的方法。

采用本说明书实施例提供的方法和装置，采用本说明书实施例披露的保护数据隐私的查询处理方法，可以实现针对一批次查询中的每个查询，为其分配确保不会泄露数据隐私

的最低噪声功率，从而最大化查询结果的准确性。

附图说明

5 为了更清楚地说明本发明实施例的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其它的附图。

图 1 示出根据一个实施例的查询处理方案的实施架构示意图；

图 2 示出根据一个实施例的查询处理方法的流程示意图；

10 图 3 示出根据另一个实施例的查询处理方案的实施架构示意图；

图 4 示出根据又一个实施例的查询处理方案的实施架构示意图；

图 5 示出根据还一个实施例的查询处理方案的实施架构示意图；

图 6 示出根据一个实施例的查询处理装置的结构示意图。

15 具体实施方式

下面结合附图，对本说明书提供的方案进行描述。

如前所述，差分隐私 DP 技术被用于统计查询场景，以应对隐私泄露问题。通常，在原始查询结果上添加随机噪声，以防止实际返回的查询结果泄露原始数据隐私。然而，已有的方案难以实现数据安全性和返回结果准确性的平衡。

20 基于此，本说明书实施例披露一种基于差分隐私实现的查询处理方案，针对一批次查询，可以为其中各条查询（或称各次查询）分配最优的噪声功率，从而实现即使攻击者获取到这一批次查询的返回结果，也无法破解被查询的原始数据，同时，最大化各条查询返回结果的准确性和可用性。

图 1 示出根据一个实施例的查询处理方案的实施架构示意图。如图 1 所示，服务平台提供查询服务。首先，服务平台从终端接收 L 条查询请求，图 1 中示出第 1 条查询请求 Q_1 和第 L 条查询请求 Q_L ；然后，服务平台根据 L 条查询请求中各条请求的查询类型，确定各条查询请求针对目标数据集的查询敏感度；之后，基于各条查询请求分别对应的查询敏感度，以及针对 L 次查询预先设定的总的隐私预算参数，确定各条查询请求分配得到的噪声功率，图 1 中示出查询请求 Q_1 和查询请求 Q_L 分配得到的噪声功率 P_{Q1} 和噪声功率 P_{QL} ；
25 接着，针对 L 条查询请求中任一的目标查询请求（如查询请求 Q_1 ），基于该目标查询请求分配得到的噪声功率（如噪声功率 P_{Q1} ）确定差分隐私的目标噪声分布（如噪声分布 D_{Q1} ），从中采样目标噪声（如噪声 N_{Q1} ），并在该目标查询的原始查询结果（如原始查询结果 R_1 ）上添加该目标噪声，作为该目标查询的实际返回结果（如实际返回结果 R_1' ）。
30

下面，结合具体的实施例，描述上述查询处理方案的具体实施步骤。

图 2 示出根据一个实施例的查询处理方法的流程示意图，所述方法的执行主体可以为任何具有计算、处理能力的服务器、装置、平台或设备集群，例如，可以为图 1 中示出的服务平台，或者，用户终端等。如图 2 所示，所述方法包括以下步骤：

步骤 S210，确定有待针对目标数据集进行查询的 L 条查询各自的查询类型，所述目标数据集记录业务对象的数据；步骤 S220，确定各个查询类型针对所述目标数据集的查询敏感度；步骤 S230，基于各条查询分别对应的查询敏感度，以及针对所述 L 条查询的总集预先设定的隐私预算参数，确定各条查询分配得到的噪声功率。

对以上步骤的展开介绍如下：

首先，在步骤 S210，确定有待针对目标数据集进行查询的 L 条查询各自的查询类型。其中，目标数据集中包含多条数据记录（或称多个数据样本），对应业务对象的数据（或称隐私数据）。需说明，其中任意两条数据记录对应的业务对象可能相同（例如，对应同一用户），也可能不同（例如，对应不同商品），并且，多条数据记录对应针对业务对象的同一特征项（例如，用户年龄）。此外，对上述多条数据记录的存储环境不作限定，在一种可能的存储情况中，上述多条数据记录存储在数据库中，例如，服务平台预先采集用户授权的个人信息，并将之存储在数据库中；在另一种可能的存储情况中，上述多条数据记录存储在用户终端中，例如，用户终端（如智能手机等）采集用户的终端操作数据（输入表情符号的行为数据），并对此进行存储。

在一个实施例中，上述业务对象可以是用户。在一个具体的实施例中，可以是个人用户，相应，上述多个数据样本对应个人用户特征项，如个人用户的姓名、年龄、性别、收入、身份证号、手机号、邮箱、兴趣爱好、生理指标、或操作指标等。在一个示例中，生理指标用于衡量人体健康状况，具体可以是血压、血脂、胆固醇、或血氧浓度等。在一个示例中，操作指标对应特定操作行为，如支付、点击、或浏览等，用于衡量用户做出特定操作行为的时长或频次等。在另一个具体的实施例中，业务对象还可以是企业用户，相应，上述多个数据样本对应企业用户特征项，如企业用户的税务数据、供应商、账面流动资金、或年度营收等。

在另一个实施例中，上述业务对象可以是商品。示例性地，商品具体可以是实体商品，如手办、或纸质书籍等；或者，可以是体验性服务，如旅游产品、按摩服务、视频会员服务、网络课程或支付会员服务等；又或者，还可以是虚拟产品，如游戏皮肤、电子书籍、内容资讯、视频资源、云盘空间、或流量等。相应地，上述多个数据样本对应商品特征项，如商品的成本、库存、销量、或毛利等。

在又一个实施例中，上述业务对象可以是业务事件，如，注册事件、访问事件、登录事件或支付事件等；相应地，上述多个数据样本对应事件特征项，如业务事件的发生时刻、网络环境、地理位置或持续时长等。在一个示例中，其中网络环境可以是 IP 地址、或网络类型等。需理解，网络类型可以包括 Wifi、4G 网络、5G 网络等。

以上，以业务对象是用户、商品、业务事件为示例，对目标数据集进行介绍。需理解，业务对象并不限于这几种。上述针对目标数据集的查询通常是指统计查询，查询目标数据集的统计信息，或者说聚合信息，例如，目标数据集中记录多个用户姓名，则可以查询其中姓黄的用户的人数。

- 5 对于上述 L 条查询及其查询类型，需说明，L 为正整数，取值可以为 1，但通常大于 1；当 L > 1 时，其中任意两条查询对应的查询类型可能相同，也可能不同。示例性地，L 条查询涉及的若干（即一个或多个）查询类型可以包括：计数、最大值、最小值、均值和方差。此外，在一种实施场景中，上述 L 条查询对应针对未来查询请求预先配置的查询信息，包括可查询次数 L 以及其中各次查询的查询类型；在另一种实施场景中，上述 L 条查询依赖接收到的 L 条查询请求，其中各条查询请求指示各自的查询类型。对于这两种实施场景的具体介绍，请参见后文。

由上，可以确定上述 L 条查询各自的查询类型。接着，在步骤 S220，确定该 L 条查询涉及的若干查询类型中，各个查询类型针对上述目标数据集的查询敏感度。

- 15 需说明，某个类型的查询敏感度旨在反映，基于目标数据集进行该类型查询得到的第一结果，以及基于目标数据集的邻近数据集（adjacent datasets）进行该类型查询得到的第二结果之间的最大差异。其中邻近数据集是指，在目标数据集的基础上增加或删除任意的一条数据后可以得到的数据集。对于上述最大差异的度量，其度量标准可以根据实际需求设定，例如，量化为上述第一查询结果和第二查询结果之间的差值绝对值的最大值（文中或简称最大绝对差值）；但实际上，并非所有查询类型对应的最大差异，如最大绝对差值等，都可以被精确求解，此时，可以求解最大差异的近似值或估计值，作为查询敏感度。

下面结合查询类型的具体示例，对查询敏感度的计算进行介绍。

- 25 根据一个实施例，上述若干查询类型包括计数查询（counting query）。对于计数查询，需说明，目标数据集中的多个数据样本对应业务对象的特征项，该特征项的取值空间中包含可能的多个取值，例如，用户性别的可能取值包含男和女，相应，可以将该多个取值作为多个数据类别，使得可以针对其中某个数据类别进行计数查询，即查询该某个数据类别对应的数据样本个数，例如，查询用户数据集中女性用户的个数。

- 30 基于上述对计数查询的介绍可知，在针对某个数据类别进行计数查询的情况下，因目标数据集和其邻近数据集之间的数据条数相差为 1，而相差的这条数据要么属于该某个数据类别，要么不属于，所以，两个查询结果之间差值的绝对值要么为 1，要么为 0，据此，可以将 1 确定为计数查询的查询敏感度，记作：

$$\text{sens}_c=1 \quad (1)$$

如此，可以确定计数查询对应的查询敏感度 sens_c 。

根据另一个实施例，上述若干查询类型包括最大值查询。为清楚描述，文中将上述目标数据集记作 $x = \{x_1, \dots, x_N\}$ ，其中 N 表示数据样本的总数， x_i 表示第 i 个数据样本；

并且，将把 N 个数据样本由小到大进行顺序排列的结果记作 $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N)}$ ，可以理解， $x_{(1)}$ 和 $x_{(N)}$ 分别表示 N 个数据样本中的最小值和最大值。对于目标数据集，其所对应的特征项的取值空间中，存在天然的取值上界和取值下界，或者，工作人员可以针对该特征项设定取值上界和取值下界，不妨将取值上界和取值下界分别记作 l 和 u ，由此可以

5 得到 $l \leq x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N)} \leq u$ 。

基于最大值查询类型和查询敏感度的定义，在一个示例中，可以将最大值查询对应的查询敏感度确定为：

$$\text{sens}_{\max} = u - l \quad (2)$$

在另一个示例中，可以将最大值查询对应的查询敏感度确定为：

$$10 \quad \text{sens}_{\max} = x_{(N)} - x_{(1)} \quad (3)$$

如此，可以确定出最大值查询对应的查询敏感度 $\text{sens}_{\max}$ 。

根据又一个实施例，上述若干查询类型包括最小值查询值，与上述确定最大值查询对应的查询敏感度 $\text{sens}_{\max}$ 相类似地，在一个示例中，可以将最小值查询对应的查询敏感度确定为：

$$15 \quad \text{sens}_{\min} = u - l \quad (4)$$

在另一个示例中，可以将最小值查询对应的查询敏感度确定为：

$$\text{sens}_{\min} = x_{(N)} - x_{(1)} \quad (5)$$

如此，可以确定出最小值查询对应的查询敏感度 $\text{sens}_{\min}$ 。

根据再一个实施例，上述若干查询类型包括均值查询。可以将针对目标数据集进行的

20 均值计算记作：

$$x_{\text{mean}} := f(x) = \frac{\sum_{i=1}^N x_i}{N} \quad (6)$$

假定邻近数据集 x' 相较于目标数据集 x 增加了一条数据 y ，并且， $x_{(1)} \leq y \leq x_{(N)}$ ，也即， $x' = (x_1, \dots, x_N, y)$ 。由此，在一个示例中，可以通过下式计算均值查询对应的查询敏感度：

$$\begin{aligned} |f(x') - f(x)| &= \left| \frac{\sum_{i=1}^N x_i + y}{N+1} - \frac{\sum_{i=1}^N x_i}{N} \right| \\ &\leq \left| \frac{\sum_{i=1}^N x_i + y}{N+1} - \frac{\sum_{i=1}^N x_i}{N+1} \right| \\ &= \left| \frac{\sum_{i=1}^N x_i + y - \sum_{i=1}^N x_i}{N+1} \right| \\ &\leq \frac{|x_{(N)}|}{N+1} \end{aligned} \quad (7)$$

25

据此，可以将均值查询对应的查询敏感度记作：

$$\text{sens}_{\text{mean}} = \frac{|x_{(N)}|}{N+1} \quad (8)$$

如此，可以确定出均值查询对应的查询敏感度 $\text{sens}_{\text{mean}}$ 。

根据另一个实施例，上述若干查询类型包括方差查询。将目标数据集 x 中数据样本的均值记作 μ ，由此，可以将针对目标数据集 x 进行的方差计算记作：

$$\begin{aligned} x_{\text{var}} &:= f(x) = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \\ &= \frac{1}{2N^2} \sum_{i=1}^N \sum_{j=1}^N (x_i - x_j)^2 \end{aligned} \quad (9)$$

假定邻近数据集 x' 相较于目标数据集 x 增加了一条数据 y ，并且， $x_{(1)} \leq y \leq x_{(N)}$ ，也即， $x' = (x_1, \dots, x_N, y)$ 。由此，可以通过下式计算方差查询对应的查询敏感度：

$$\begin{aligned} |f(x') - f(x)| &= \left| \frac{\sum_{i=1}^{N+1} \sum_{j=1}^{N+1} (x_i - x_j)^2}{2(N+1)^2} - \frac{\sum_{i=1}^N \sum_{j=1}^N (x_i - x_j)^2}{2N^2} \right| \\ &\leq \left| \frac{\sum_{i=1}^{N+1} \sum_{j=1}^{N+1} (x_i - x_j)^2}{2(N+1)^2} - \frac{\sum_{i=1}^N \sum_{j=1}^N (x_i - x_j)^2}{2(N+1)^2} \right| \\ &= \left| \frac{2 \sum_{i=1}^N (x_i - y)^2}{2(N+1)^2} \right| \\ &\leq \frac{N(x_{(N)} - x_{(1)})^2}{(N+1)^2} \end{aligned} \quad (10)$$

10 据此，可以将方差查询对应的查询敏感度记作：

$$\text{sens}_{\text{var}} = \frac{N}{(N+1)^2} (x_{(N)} - x_{(1)})^2 \quad (11)$$

如此，可以确定出方差查询对应的查询敏感度 sens_{var} 。

15 以上，以查询类型为计数、最大值、最小值、均值和方差作为示例，对查询敏感度的计算进行介绍。实际上，上述 L 条查询涉及的查询类型不限于这几种，还可以包括其他类型，如分位数查询 (quantile query) 和证据权重查询 (weight of evidence query) 等。对于其他查询类型所对应的查询敏感度的计算，可以结合二者的定义进行计算，在此不作穷举。

由上，可以确定出上述 L 条查询涉及的各个查询类型所对应的查询敏感度。需理解，各条查询具有对应的查询类型，而各个查询类型具有对应的查询敏感度，由此，可以得到各条查询对应的查询敏感度。

20 然后，在步骤 S230，基于各条查询分别对应的查询敏感度，以及针对上述 L 条查询的总集预先设定的隐私预算参数，确定各条查询分配得到的噪声功率。

在一个实施例中，本步骤可以包括：针对 L 条查询中任意的第 ℓ 条查询，根据该条查

查询的查询敏感度 $sens_\ell$ ，L 条查询的查询敏感度和值，以及上述隐私预算参数，确定该条查询分配得到的噪声功率。需说明，查询敏感度和值是基于 L 条查询中各条查询的查询敏感度进行求和而得到。

- 5 在一个具体的实施例中，先根据上述隐私预算参数的参数值，以及差分隐私的高斯机制下该隐私预算参数和均值变量 μ 的约束关系，确定其中均值变量的变量值。其中约束关系是差分隐私的高斯机制中已有的，可以表示为下式：

$$\delta(\epsilon; \mu) = \Phi\left(-\frac{\epsilon}{\mu} + \frac{\mu}{2}\right) - e^\epsilon \Phi\left(-\frac{\epsilon}{\mu} - \frac{\mu}{2}\right) \quad (12)$$

- 在上式中， ϵ 和 δ 分别表示上述隐私预算参数中的预算项参数和松弛项参数，二者的参数值可以是由工作人员根据实际需要人为设定的； μ 表示上述均值变量； $\Phi(t)$ 表示标准高斯分布的概率分布函数， $\Phi(t) = \mathbb{P}[\mathcal{N}(0, 1) \leq t] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-y^2/2} dy$ 。

进一步，针对任意的第 ℓ 条查询，将其噪声功率确定为以下因子的乘积：查询敏感度 $sens_\ell$ ，上述查询敏感度和值，以及上述均值变量的变量值进行平方运算后的倒数。据此，可以将噪声功率的计算式表示为：

$$\sigma_{\ell, opt}^2 = \frac{sens_\ell}{\mu^2} \left(\sum_{k=1}^L sens_k \right), \quad \ell = 1, \dots, L \quad (13)$$

- 15 在上式中， $\sigma_{\ell, opt}^2$ 表示第 ℓ 条查询的噪声功率， $sens_\ell$ 表示第 ℓ 条查询的查询敏感度， μ 表示上述均值变量， $\sum_{k=1}^L sens_k$ 表示 L 条查询的查询敏感度和值。

需说明，公式 (13) 的推导过程如下：

在隐私约束下，最小化 L 条查询的总的噪声功率，等同于以下优化问题：

$$\sum_{k=1}^L \frac{S_k^2}{\sigma_k^2} \leq \mu^2$$

约束条件函数：

$$(14)$$

- 20 需求解的原函数： $\underset{\{\sigma_k^2 > 0\}_{k=1}^L}{\text{minimize}} \sum_{k=1}^L \sigma_k^2$ (15)

在公式 (14) 和 (15) 中， S_k 表示第 k 条查询的查询敏感度， μ 表示上述均值变量， σ_k^2 表示第 k 条查询的噪声功率。

- 注意到上述优化问题是凸优化问题并且符合 Slater's 条件，因此，具有强对偶性 (strong duality)。针对此优化问题，引入拉格朗日乘子 λ ，构建如下的拉格朗日 (Lagrangian) $\mathcal{L}$ ，
- 25 将约束条件函数与原函数联系到一起：

$$\mathcal{L}(\{\sigma_k^2\}_{k=1}^L, \lambda) = \sum_{k=1}^L \sigma_k^2 + \lambda \left(\sum_{k=1}^L \frac{S_k^2}{\sigma_k^2} - \mu^2 \right) \quad (16)$$

公式 (16) 的可行解 σ_k^2 需要满足 KKT (The Karush-Kuhn-Tucker) 条件：

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \sigma_\ell^2} = 1 - \lambda \frac{S_\ell^2}{\sigma_\ell^4}, & \ell = 1, \dots, L \\ \lambda \left(\sum_{k=1}^L \frac{S_k^2}{\sigma_k^2} - \mu^2 \right) = 0 \end{cases} \quad (17)$$

由此，在满足 KKT 的条件下极小化拉格朗日 $\mathcal{L}$ ，可以得到如下的最优噪声功率分配策略：

$$\sigma_{\ell, \text{opt}}^2 = \frac{S_\ell}{\mu^2} \left(\sum_{k=1}^L S_k \right) \quad (18)$$

需理解，公式 (18) 等价于公式 (13)，仅是数学符号的记法略有不同。

由上，可以确定出差分隐私的高斯噪声机制下各条查询分配得到的噪声功率。需说明，实际还可以确定出其他噪声机制下各条查询的噪声功率，如拉普拉斯机制、指数机制等。

根据另一方面的实施例，在步骤 S230 之后，上述方法还可以包括：在步骤 S240，针对上述 L 条查询中任意的目标查询，将其实际返回结果确定为其原始查询结果叠加从差分隐私的目标噪声分布采样的目标噪声。其中目标噪声分布基于目标查询分配得到的噪声功率而确定。

在一个实施例中，上述目标噪声分布可以为高斯噪声分布，相应，将目标查询的噪声功率作为高斯分布的方差，并以 0 为均值，生成对应的高斯噪声分布。为清楚描述，以下将目标查询记作第 ℓ 条查询，将其噪声功率记作 $\sigma_{\ell, \text{opt}}^2$ ，由此可以将目标查询对应的高斯噪声分布记作 $\mathcal{N}(0, \sigma_{\ell, \text{opt}}^2)$ 。

基于确定出的高斯噪声分布 $\mathcal{N}(0, \sigma_{\ell, \text{opt}}^2)$ ，可以从中采样目标噪声 z ，叠加在目标查询的原始查询结果 q_ℓ 上，从而得到对应的实际返回结果，计算式如下：

$$\tilde{q}_\ell = q_\ell + z, \quad z \sim \mathcal{N}(0, \sigma_{\ell, \text{opt}}^2) \quad (19)$$

下面，结合查询类型的示例，对公式 (19) 进行具象描述。

根据一个实施例，若目标查询的查询类型是计数查询，则可以基于下式计算实际返回的计数查询结果：

$$\widetilde{\text{Count}}_{c, N} = \lfloor \text{Count}_{c, N} + z_c \rfloor \quad c = 1, \dots, C \quad (20)$$

在上式中，N 表示目标数据集的数据样本总数；c 表示第 c 个数据类别；C 表示数据类别总数； z_c 表示从计数查询所对应的高斯噪声分布中随机采样的噪声； $\text{Count}_{c, N}$ 表示针对第 c 个类别进行计数查询的原始查询结果， $\widetilde{\text{Count}}_{c, N}$ 表示添加噪声后的实际返回结果；运算符 $\lfloor \cdot \rfloor$ 表示四舍五入后进行取整。

如此，可以确定计数查询类型对应的实际返回结果。需理解，在确定每条查询对应的实际返回结果时，都会基于对应的高斯噪声分布进行噪声采样。比如说，对于两条具有相

同类型的查询，会基于二者对应的同一高斯噪声分布进行两次随机抽样，再分别叠加到查询类型对应的原始查询结果中，从而对应得到两个实际返回结果。

根据另一个实施例，若目标查询的查询类型是最大值查询，则可以基于下式计算实际返回的最大值查询结果：

$$\tilde{x}_{\max} = x_{(N)} + z \quad (21)$$

在上式中， z 表示从最大值查询所对应的高斯噪声分布中随机采样的噪声； $x_{(N)}$ 表示最大值查询的原始查询结果，也即目标数据集中的最大值； $\tilde{x}_{\max}$ 表示对最大值查询对应的实际返回结果。

根据又一个实施例，若目标查询的查询类型是最小值查询，则可以基于下式计算实际返回的最小值查询结果：

$$\tilde{x}_{\min} = x_{(1)} + z \quad (22)$$

在上式中， z 表示从最小值查询所对应的高斯噪声分布中随机采样的噪声； $x_{(1)}$ 表示最小值查询的原始查询结果，也即目标数据集中的最小值； $\tilde{x}_{\min}$ 表示最小值查询对应的实际返回结果。

根据还有一个实施例，若目标查询的查询类型是均值查询，则可以基于下式计算实际返回的均值查询结果：

$$\tilde{x}_{\text{mean}} = \frac{\sum_{i=1}^N x_i}{N} + z \quad (23)$$

在上式中， z 表示从均值查询所对应的高斯噪声分布中随机采样的噪声， $\frac{\sum_{i=1}^N x_i}{N}$ 表示目标数据集的均值， $\tilde{x}_{\text{mean}}$ 表示均值查询对应的实际返回结果。

根据再一个实施例，若目标查询的查询类型是方差查询，则可以基于下式计算实际返回的方差查询结果：

$$\tilde{x}_{\text{var}} = \sum_{i=1}^N (x_i - \mu)^2 + z \quad (24)$$

在上式中， z 表示从方差查询所对应的高斯噪声分布中采样的噪声， $\sum_{i=1}^N (x_i - \mu)^2$ 表示目标数据集的方差， $\tilde{x}_{\text{var}}$ 表示方差查询对应的实际返回结果。

以上，以差分隐私的高斯机制为示例，对目标噪声的采样和添加进行介绍。在另一个实施例中，上述目标噪声功率的确定基于差分隐私的拉普拉斯机制，相应，可以将目标查询的噪声功率作为拉普拉斯分布的尺度参数，并以0为位置参数，生成对应的拉普拉斯噪声分布，从而实现目标噪声的采样和添加，得到实际返回结果。在另一个实施例中，可以基于差分隐私的指数机制进行目标噪声的采样和添加，进而得到实际返回结果。

由上，可以确定L条查询中任意目标查询对应的实际返回结果。

需说明，本说明书实施例披露的上述查询处理方法，可以应用于多种实施场景。在一种典型的场景中，如图1所示，先接收一批次待处理的查询请求，于是，在步骤S210中，

待处理的查询条数 L 以及其中各条查询的查询类型基于这批次查询请求而确定, 在一个实施例中, 可以获取在预设时段内 (如前一天内或最近 10min 内) 接收到的多条查询请求, 将其数量记作 L , 并确定各条查询请求指示的查询类型。之后在步骤 S240 中, 将 L 条查询请求分别作为目标查询, 从而对应进行噪声采样和基于原始查询结果的噪声叠加, 进而得到各条查询请求的实际返回结果。如此, 可以完成对一批次查询请求的差分隐私处理。

在另一种实施场景中, 可以针对未来可能接收到的查询请求, 预先配置可查询次数 L 以及其中各次查询的查询类型, 由此, 在步骤 S210 中可以直接获取预先配置的这部分信息。需说明, 查询次数和查询类型的配置, 可以由工作人员基于目标数据集对应的历史查询数据进行分析而确定。

10 进一步, 在一个细化的实施场景中, 如图 3 所示, 在预备阶段通过执行上述步骤 S210、S220 和 S230, 确定各条查询分配得到的噪声功率。然后, 在在线处理阶段, 基于接收的当前查询请求 Q_i 执行步骤 S240, 具体, 将当前查询请求 Q_i 作为目标查询, 其对应当前查询类型 C_i ; 在判断出当前查询类型 C_i 对应的针对目标数据集的已处理请求数量小于预置的可查询次数 L_{C_i} 的情况下, 将当前查询请求 Q_i 作为目标查询, 从而基于目标数据集确定查
15 询类型 C_i 对应的原始查询结果 F_{C_i} , 并获取预备阶段中确定的对应当前查询类型 C_i 的噪声功率 P_{C_i} , 进而将从基于噪声功率 P_{C_i} 确定的噪声分布中采样的噪声 N_j 叠加在原始查询结果 F_{C_i} 上, 得到实际返回结果 F_j ; 否则, 在判断出上述已处理请求数量等于可查询次数 L_{C_i} 的情况下, 抛弃当前查询请求 Q_i 。

在另一个细化的实施场景中, 如图 4 所示, 在预备阶段通过执行上述步骤 S210、S220
20 和 S230, 确定各条查询分配得到的噪声功率, 并且, 还基于目标数据集确定各个查询类型对应的原始查询结果; 然后在在线处理阶段, 基于接收的当前查询请求 Q_i 执行步骤 S240, 具体, 将当前查询请求 Q_i 作为目标查询, 其对应当前查询类型 C_i ; 在判断出此类型对应的针对目标数据集的已处理请求数量小于预置的可查询次数 L_{C_i} 的情况下, 将当前查询请求 Q_i 作为目标查询, 从而获取预备阶段确定的对应当前查询类型 C_i 的原始查询结果 F_{C_i} 和
25 噪声功率 P_{C_i} , 进而将从基于噪声功率 P_{C_i} 确定的噪声分布中采样的噪声 N_j 叠加在原始查询结果 F_{C_i} 上, 得到实际返回结果 F_j ; 否则, 抛弃当前查询请求 Q_i 。

在又一个细化的实施场景中, 如图 5 所示, 在预备阶段通过执行上述步骤 S210 至步骤 S240, 确定 L 条查询中各条查询对应的实际返回结果。之后, 在在线处理阶段, 接收当前查询请求 Q_i , 其对应当前查询类型 C_i ; 在判断出此类型对应的针对目标数据集的已处理
30 请求数量小于预置的可查询次数的情况下, 从预备阶段确定出的 L 个实际返回结果中, 获取对应当前查询类型 C_i 且尚未被使用的某个实际返回结果, 作为当前查询请求 Q_i 的实际返回结果 R_i' ; 否则, 抛弃当前查询请求 Q_i 。

以上, 对本说明书实施例披露的查询处理方法的实施场景进行介绍。

综上, 采用本说明书实施例披露的保护数据隐私的查询处理方法, 可以实现针对一批

次查询中的每个查询，为其分配确保不会泄露数据隐私的最低噪声功率，从而最大化查询结果的准确性。

与上述查询处理方法相对应的，本说明书实施例还披露查询处理装置。图 6 示出根据一个实施例的查询处理装置的结构示意图，所述装置可以实现为任何具有计算、处理能力的计算单元、平台、服务器、设备集群等，例如，用户终端或图 1 中示出的服务平台。如图 6 所示，所述装置 600 包括以下单元：

查询类型确定单元 610，配置为确定有待针对目标数据集进行查询的 L 条查询各自的查询类型，所述目标数据集记录业务对象的数据；敏感度确定单元 620，配置为确定各个查询类型针对所述目标数据集的查询敏感度；噪声功率确定单元 630，配置为基于各条查询分别对应的查询敏感度，以及针对所述 L 条查询的总集预先设定的隐私预算参数，确定各条查询分配得到的噪声功率。

在一个实施例中，所述查询类型确定单元 610 具体配置为：接收针对所述目标数据集的 L 条查询请求，其中各条查询请求指示各自的查询类型。

在一个实施例中，所述查询类型确定单元 610 具体配置为：获取针对所述目标数据集预先配置的可查询次数 L 及其中各次查询的查询类型。

在一个实施例中，所述查询类型为以下中的任一种：计数、最大值、最小值、均值和方差。

在一个实施例中，所述业务对象为以下中的任一种：用户、商品、业务事件。

在一个实施例中，所述业务事件为以下中的任一种：注册、访问、登录、支付。

在一个实施例中，所述业务对象为用户，所述隐私数据为以下中的任一种：年龄、性别、收入、兴趣爱好、生理指标、操作指标。

在一个实施例中，敏感度确定单元 620 具体配置为：针对所述各个查询类型，根据第一结果和第二结果之间的最大绝对差值，得到该查询类型对应的查询敏感度；其中，所述第一结果是针对所述目标数据集进行该类型查询得到的结果，所述第二结果是针对所述目标数据集的邻近数据集进行该类型查询得到的结果。

在一个实施例中，所述查询类型包括计数查询；敏感度确定单元 620 具体配置为：将所述计数查询的查询敏感度确定为数值 1。

在一个实施例中，所述查询类型包括最大值查询/最小值查询；敏感度确定单元 620 具体配置为：确定所述目标数据集中的最大值和最小值；将所述最大值减去所述最小值的结果，确定为所述最大值查询/最小值查询的查询敏感度。

在一个实施例中，所述查询类型包括均值查询；敏感度确定单元 620 具体配置为：确定所述目标数据集中的最大值；将所述最大值的绝对值与所述目标数据集的数据个数加 1 之间的比值，确定为所述均值查询的查询敏感度。

在一个实施例中，所述查询类型包括方差查询；敏感度确定单元 620 具体配置为：确定所述目标数据集中的最大值和最小值；将所述方差查询的查询敏感度确定为以下因子的乘积：所述最大值和最小值之间差值的平方，所述目标数据集的数据个数，以及所述数据个数加 1 后进行平方运算的倒数。

- 5 在一个实施例中，噪声功率确定单元 630 具体配置为：基于各条查询的查询敏感度，确定所述 L 条查询的查询敏感度和值；针对任意一条查询，根据该条查询的查询敏感度，所述查询敏感度和值，以及所述隐私预算参数，确定该条查询分配得到的噪声功率。

- 10 在一个具体的实施例中，噪声功率确定单元 630 进一步配置为：获取均值变量的变量值，该变量值基于所述隐私预算参数的参数值，以及差分隐私的高斯机制下所述隐私预算参数和均值变量的约束关系而确定；将该条查询的噪声功率确定为以下因子的乘积：该条查询的查询敏感度，所述查询敏感度和值，以及所述变量值进行平方运算后的倒数。

在一个实施例中，所述隐私预算参数包括预算项参数和松弛项参数。

- 15 在一个实施例中，所述装置 600 还包括实际结果确定单元 640，配置为针对所述 L 条查询中任意的目标查询，将其实际返回结果确定为其原始查询结果叠加从差分隐私的目标噪声分布采样的目标噪声，其中目标噪声分布基于目标查询分配得到的噪声功率而确定

在一个具体的实施例中，所述目标噪声分布为高斯噪声分布，所述高斯噪声分布以所述目标查询的噪声功率为方差，以 0 为均值。

- 20 另一方面，在一个具体的实施例中，所述装置还包括目标查询确定单元 650，配置为：接收针对所述目标数据集的当前查询请求，其对应当前查询类型；判断对应所述当前查询类型的已处理请求数量是否小于预定阈值，所述已处理请求数量对应的查询请求针对所述目标数据集；在判断出小于的情况下，将所述当前查询请求作为所述目标查询。

根据另一方面的实施例，还提供一种计算机可读存储介质，其上存储有计算机程序，当所述计算机程序在计算机中执行时，令计算机执行结合图 2 所描述的方法。

- 25 根据再一方面的实施例，还提供一种计算设备，包括存储器和处理器，该存储器中存储有可执行代码，所述处理器执行所述可执行代码时，实现结合图 2 所描述的方法。

- 30 本领域技术人员应该可以意识到，在上述一个或多个示例中，本发明所描述的功能可以用硬件、软件、固件或它们的任意组合来实现。当使用软件实现时，可以将这些功能存储在计算机可读介质中或者作为计算机可读介质上的一个或多个指令或代码进行传输。

以上所述的具体实施方式，对本发明的目的、技术方案和有益效果进行了进一步详细说明，所应理解的是，以上所述仅为本发明的具体实施方式而已，并不用于限定本发明的保护范围，凡在本发明的技术方案的基础之上，所做的任何修改、等同替换、改进等，均应包括在本发明的保护范围之内。

权 利 要 求 书

1. 一种查询处理方法，包括：

确定有待针对目标数据集进行查询的 L 条查询各自的查询类型，所述目标数据集记录业务对象的数据；

确定各个查询类型针对所述目标数据集的查询敏感度；

基于各条查询分别对应的查询敏感度，以及针对所述 L 条查询的总集预先设定的隐私预算参数，确定各条查询分配得到的噪声功率。

2. 根据权利要求 1 所述的方法，其中，确定有待针对目标数据集进行查询的 L 条查询各自的查询类型，包括：

接收针对所述目标数据集的 L 条查询请求，其中各条查询请求指示各自的查询类型。

3. 根据权利要求 1 所述的方法，其中，确定有待针对目标数据集进行查询的 L 条查询各自的查询类型，包括：

获取针对所述目标数据集预先配置的可查询次数 L 及其中各次查询的查询类型。

4. 根据权利要求 1 所述的方法，其中，所述查询类型为以下中的任一种：计数、最大值、最小值、均值和方差。

5. 根据权利要求 1 所述的方法，其中，所述业务对象为以下中的任一种：用户、商品、业务事件。

6. 根据权利要求 5 所述的方法，其中，所述业务事件为以下中的任一种：注册、访问、登录、支付。

7. 根据权利要求 1 所述的方法，其中，所述业务对象为用户，所述业务对象的数据为以下中的任一种：年龄、性别、收入、兴趣爱好、生理指标、操作指标。

8. 根据权利要求 1 所述的方法，其中，确定各个查询类型针对所述目标数据集的查询敏感度，包括：

针对所述各个查询类型，根据第一结果和第二结果之间的最大绝对差值，得到该查询类型对应的查询敏感度；其中，所述第一结果是针对所述目标数据集进行该类型查询得到的结果，所述第二结果是针对所述目标数据集的邻近数据集进行该类型查询得到的结果。

9. 根据权利要求 1 或 8 所述的方法，其中，所述查询类型包括计数查询；其中，确定各个查询类型针对所述目标数据集的查询敏感度，包括：

将所述计数查询的查询敏感度确定为数值 1。

10. 根据权利要求 1 或 8 所述的方法，其中，所述查询类型包括最大值查询/最小值查询；其中，确定各个查询类型针对所述目标数据集的查询敏感度，包括：

确定所述目标数据集中的最大值和最小值；

将所述最大值减去所述最小值的结果，确定为所述最大值查询/最小值查询的查询敏感

度。

11. 根据权利要求 1 或 8 所述的方法，其中，所述查询类型包括均值查询；其中，确定各个查询类型针对所述目标数据集的查询敏感度，包括：

确定所述目标数据集中的最大值；

将所述最大值的绝对值与所述目标数据集的数据个数加 1 之间的比值，确定为所述均值查询的查询敏感度。

12. 根据权利要求 1 或 8 所述的方法，其中，所述查询类型包括方差查询；其中，确定各个查询类型针对所述目标数据集的查询敏感度，包括：

确定所述目标数据集中的最大值和最小值；

将所述方差查询的查询敏感度确定为以下因子的乘积：所述最大值和最小值之间差值的平方，所述目标数据集的数据个数，以及所述数据个数加 1 后进行平方运算的倒数。

13. 根据权利要求 1 所述的方法，其中，基于各条查询分别对应的查询敏感度，以及针对所述 L 条查询的总集预先设定的隐私预算参数，确定各条查询分配得到的噪声功率，包括：

基于各条查询的查询敏感度，确定所述 L 条查询的查询敏感度和值；

针对任意一条查询，根据该条查询的查询敏感度，所述查询敏感度和值，以及所述隐私预算参数，确定该条查询分配得到的噪声功率。

14. 根据权利要求 13 所述的方法，其中，根据该条查询的查询敏感度，所述查询敏感度和值，以及所述隐私预算参数，确定该条查询分配得到的噪声功率，包括：

获取均值变量的变量值，该变量值基于所述隐私预算参数的参数值，以及差分隐私的高斯机制下所述隐私预算参数和均值变量的约束关系而确定；

将该条查询的噪声功率确定为以下因子的乘积：该条查询的查询敏感度，所述查询敏感度和值，以及所述变量值进行平方运算后的倒数。

15. 根据权利要求 1 或 14 所述的方法，其中，所述隐私预算参数包括预算项参数和松弛项参数。

16. 根据权利要求 1 所述的方法，其中，在确定各条查询分配得到的噪声功率之后，所述方法还包括：

针对所述 L 条查询中任意的目标查询，将其实际返回结果确定为其原始查询结果叠加从差分隐私的目标噪声分布采样的目标噪声，其中目标噪声分布基于所述目标查询分配得到的噪声功率而确定。

17. 根据权利要求 16 所述的方法，其中，所述目标噪声分布为高斯噪声分布，该高斯噪声分布以所述目标查询的噪声功率为方差，以 0 为均值。

18. 根据权利要求 16 所述的方法，还包括：

接收针对所述目标数据集的当前查询请求，其对应当前查询类型；

判断对应所述当前查询类型的已处理请求数量是否小于预定阈值，所述已处理请求数量对应的查询请求针对所述目标数据集；

在判断出小于的情况下，将所述当前查询请求作为所述目标查询。

19. 一种查询处理装置，包括：

查询类型确定单元，配置为确定有待针对目标数据集进行查询的 L 条查询各自的查询类型，所述目标数据集记录业务对象的数据；

敏感度确定单元，配置为确定各个查询类型针对所述目标数据集的查询敏感度；

噪声功率确定单元，配置为基于各条查询分别对应的查询敏感度，以及针对所述 L 条查询的总集预先设定的隐私预算参数，确定各条查询分配得到的噪声功率。

20. 根据权利要求 19 所述的装置，其中，所述查询类型确定单元具体配置为：

接收针对所述目标数据集的 L 条查询请求，其中各条查询请求指示各自的查询类型。

21. 根据权利要求 19 所述的装置，其中，所述查询类型确定单元具体配置为：

获取针对所述目标数据集预先配置的可查询次数 L 及其中各次查询的查询类型。

22. 根据权利要求 19 所述的装置，其中，所述敏感度确定单元具体配置为：

针对所述各个查询类型，根据第一结果和第二结果之间的最大绝对差值，得到该查询类型对应的查询敏感度；其中，所述第一结果是针对所述目标数据集进行该类型查询得到的结果，所述第二结果是针对所述目标数据集的邻近数据集进行该类型查询得到的结果。

23. 根据权利要求 19 所述的装置，还包括：

实际结果确定单元，配置为针对所述 L 条查询中任意的目标查询，将其实际返回结果确定为其原始查询结果叠加从差分隐私的目标噪声分布采样的目标噪声，其中目标噪声分布基于所述目标查询分配得到的噪声功率而确定。

24. 根据权利要求 23 所述的装置，还包括目标查询确定单元，配置为：

接收针对所述目标数据集的当前查询请求，其对应当前查询类型；

判断对应所述当前查询类型的已处理请求数量是否小于预定阈值，所述已处理请求数量对应的查询请求针对所述目标数据集；

在判断出小于的情况下，将所述当前查询请求作为所述目标查询。

25. 一种计算机可读存储介质，其上存储有计算机程序，其中，当所述计算机程序在计算机中执行时，令计算机执行权利要求 1-18 中任一项所述的方法。