

估计差分隐私保护数据中分词频度的方法及装置

技术领域

[01] 本说明书一个或多个实施例涉及数据挖掘领域技术领域，特别涉及用于估计差分隐私保护数据中分词频度的方法及装置。

背景技术

[02] 用户通过终端设备输入或者查看的文本信息（如留言，聊天记录，搜索记录等），能够直接或间接地体现出用户的特征和偏好，这些文本信息对于数据的挖掘与分析具有极其重要的意义。但是，这些文本信息又涉及到用户的个人隐私。因此，通常可以通过终端设备对用户输入或者查看的文本信息进行本地差分隐私处理，得到差分隐私保护数据，并将差分隐私保护数据上报至服务器，由服务器估计差分隐私保护数据中分词频度（分词在文本中出现的次数）。因此，在估计分词频度的过程中，如何能够在计算量较小的情况下更高效和合理的对分词频度进行估计，在数据挖掘领域变得尤为重要。

发明内容

[03] 为了解决上述技术问题之一，本说明书一个或多个实施例提供一种估计差分隐私保护数据中分词频度的方法及装置。

[04] 根据第一方面，提供一种估计差分隐私保护数据中分词频度方法，应用于服务器，包括：获取终端设备上报的、经本地差分隐私处理的各个分词信息；任一分词信息对应于一个分词，并包括表示该分词中包含的词语单元数量的目标个数，该目标个数小于等于预设值 N ；划分出 N 组分词信息，使同组的各个分词信息对应于相同的目标个数；确定各组分词信息各自对应的表示分词频度无偏估计的各组估计数据；基于所述各组估计数据，逐层生成用于记录分词频度的前缀树的各层节点，其中，生成第 n 层节点包括：获取第 $n-1$ 层中的各个节点各自表示的各个 $n-1$ 元分词，第 $n-1$ 层中的任一节点表示的 $n-1$ 元分词通过将根节点到该节点所对应的词语单元顺序排列而形成；基于所述各个 $n-1$ 元分词，确定用于第 n 层节点的多个备选 n 元分词；基于对应于目标个数为 n 的第 n 组估计数据，计算所述备选 n 元分词的频度显著性分布信息；基于所述频度显著性分布信息，选择若干备选 n 元分词作为第 n 层节点表示的 n 元分词，并利用第 n 层中的各个节点记录各自表示的各个 n 元分词的频度； $1 \leq n \leq N$ 。

[05]可选的,所述前缀树的根节点为第0层节点,所述第0层节点表示空字符。

[06]可选的,所述基于所述各个 $n-1$ 元分词,确定用于第 n 层节点的多个备选 n 元分词,包括:将所述各个 $n-1$ 元分词作为前缀,分别与预设词典中的各个预设词语单元构成的多个 n 元分词确定为所述多个备选 n 元分词。

[07]可选的,所述基于对应于目标个数为 n 的第 n 组估计数据,计算所述备选 n 元分词的频度显著性分布信息,包括:基于所述第 n 组估计数据,计算各个备选 n 元分词的各个频度;基于所述各个频度,计算所述各个备选 n 元分词各自对应的各个方差;基于所述各个方差,计算所述备选 n 元分词的频度显著性分布信息。

[08]可选的,所述基于所述各个方差,计算所述备选 n 元分词的频度显著性分布信息,包括:基于所述各个方差,计算所述各个备选 n 元分词各自对应的各个 z 值;基于所述各个 z 值,计算所述各个备选 n 元分词各自对应的各个 p 值,作为所述备选 n 元分词的频度显著性分布信息;其中,所述基于所述频度显著性分布信息,选择若干备选 n 元分词作为第 n 层节点表示的 n 元分词,包括:基于所述各个 p 值,选择若干备选 n 元分词作为第 n 层节点表示的 n 元分词。

[09]可选的,所述基于所述各个 p 值,选择若干备选 n 元分词作为第 n 层节点表示的 n 元分词,包括:将所述各个 p 值按从小到大的顺序进行排列;选取满足预设条件的最大 p 值作为目标 p 值;满足所述预设条件的任一 p 值小于等于该 p 值对应的目标结果,该目标结果为该 p 值在所述排列中的序号和针对第 n 层设定的预设阈值的乘积除以备选 n 元分词的个数所得结果;选取小于所述目标 p 值的各个 p 值各自对应的各个备选 n 元分词,作为所述第 n 层节点表示的 n 元分词。

[10]可选的,所述方法还包括:利用第 n 层中的各个节点记录各自表示的各个 n 元分词的方差以及 p 值。

[11]可选的,所述任一分词信息还包括表示该分词的目标向量,该目标向量经过本地差分隐私处理。

[12]可选的,所述表示该分词的目标向量通过如下方式得到:从多个预设的哈希函数中选取一个哈希函数作为目标哈希函数;利用所述目标哈希函数计算该分词的目标哈希值;

[13]利用满足差分隐私的方式,基于所述目标哈希值确定所述目标向量。

[14]根据第二方面,提供一种估计差分隐私保护数据中分词频度的装置,应用于服务器,包括:获取模块,用于获取终端设备上报的、经本地差分隐私处理的各个分词信息;任

一分词信息对应于一个分词，并包括表示该分词中包含的词语单元数量的目标个数，该目标个数小于等于预设值 N ；分组模块，用于划分出 N 组分词信息，使同组的各个分词信息对应于相同的目标个数；确定模块，用于确定各组分词信息各自对应的表示分词频度无偏估计的各组估计数据；生成模块，用于基于所述各组估计数据，逐层生成用于记录分词频度的前缀树的各层节点，其中，所述生成模块通过如下方式生成第 n 层节点：获取第 $n-1$ 层中的各个节点各自表示的各个 $n-1$ 元分词，第 $n-1$ 层中的任一节点表示的 $n-1$ 元分词通过将从根节点到该节点所对应的词语单元顺序排列而形成；基于所述各个 $n-1$ 元分词，确定用于第 n 层节点的多个备选 n 元分词；基于对应于目标个数为 n 的第 n 组估计数据，计算所述备选 n 元分词的频度显著性分布信息；基于所述频度显著性分布信息，选择若干备选 n 元分词作为第 n 层节点表示的 n 元分词，并利用第 n 层中的各个节点记录各自表示的各个 n 元分词的频度； $1 \leq n \leq N$ 。

[15] 可选的，所述前缀树的根节点为第 0 层节点，所述第 0 层节点表示空字符。

[16] 可选的，所述生成模块通过如下方式基于所述各个 $n-1$ 元分词，确定用于第 n 层节点的多个备选 n 元分词：将所述各个 $n-1$ 元分词作为前缀，分别与预设词典中的各个预设词语单元构成的多个 n 元分词确定为所述多个备选 n 元分词。

[17] 可选的，所述生成模块通过如下方式基于对应于目标个数为 n 的第 n 组估计数据，计算所述备选 n 元分词的频度显著性分布信息：基于所述第 n 组估计数据，计算各个备选 n 元分词的各个频度；基于所述各个频度，计算所述各个备选 n 元分词各自对应的各个方差；基于所述各个方差，计算所述备选 n 元分词的频度显著性分布信息。

[18] 可选的，所述生成模块通过如下方式基于所述各个方差，计算所述备选 n 元分词的频度显著性分布信息：基于所述各个方差，计算所述各个备选 n 元分词各自对应的各个 z 值；基于所述各个 z 值，计算所述各个备选 n 元分词各自对应的各个 p 值，作为所述备选 n 元分词的频度显著性分布信息；其中，所述生成模块通过如下方式基于所述频度显著性分布信息，选择若干备选 n 元分词作为第 n 层节点表示的 n 元分词：基于所述各个 p 值，选择若干备选 n 元分词作为第 n 层节点表示的 n 元分词。

[19] 可选的，所述生成模块通过如下方式基于所述各个 p 值，选择若干备选 n 元分词作为第 n 层节点表示的 n 元分词：将所述各个 p 值按从小到大的顺序进行排列；选取满足预设条件的最大 p 值作为目标 p 值；满足所述预设条件的任一 p 值小于等于该 p 值对应的目标结果，该目标结果为该 p 值在所述排列中的序号和针对第 n 层设定的预设阈值的乘积除以备选 n 元分词的个数所得结果；选取小于所述目标 p 值的各个 p 值各自对应的

各个备选 n 元分词，作为所述第 n 层节点表示的 n 元分词。

[20] 可选的，所述生成模块还被配置用于：利用第 n 层中的各个节点记录各自表示的各个 n 元分词的方差以及 p 值。

[21] 可选的，所述任一分词信息还包括表示该分词的目标向量，该目标向量经过本地差分隐私处理。

[22] 可选的，所述表示该分词的目标向量通过如下方式得到：从多个预设的哈希函数中选取一个哈希函数作为目标哈希函数；利用所述目标哈希函数计算该分词的目标哈希值；利用满足差分隐私的方式，基于所述目标哈希值确定所述目标向量。

[23] 根据第三方面，提供一种计算机可读存储介质，所述存储介质存储有计算机程序，所述计算机程序被处理器执行时实现上述第一方面中任一项所述的方法。

[24] 根据第四方面，提供一种电子设备，包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现上述第一方面中任一项所述的方法。

[25] 本说明书的实施例提供的技术方案可以包括以下有益效果：本说明书的实施例提供的估计差分隐私保护数据中分词频度的方法和装置，通过获取终端设备上报的、经本地差分隐私处理的各个分词信息，划分出 N 组分词信息，使同组的各个分词信息对应于相同的目标个数，确定各组分词信息各自对应的表示分词频度无偏估计的各组估计数据，并基于各组估计数据，逐层生成用于记录分词频度的前缀树的各层节点。由于本实施例在生成前缀树的第 n 层节点的过程中，能够基于备选 n 元分词的频度显著性分布信息，选择部分备选 n 元分词作为第 n 层节点表示的 n 元分词，无需遍历由预设的词语单元构成的所有 n 元分词，不仅大大降低了计算量，提高了计算效率，而且基于分词的频度显著性分布信息而筛选出来的第 n 层节点表示的 n 元分词更具合理性。

[26] 应当理解的是，以上的一般描述和后文的细节描述仅是示例性和解释性的，并不能限制本申请。

附图说明

[27] 为了更清楚地说明本申请实施例的技术方案，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申请的一些实施例，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。

[28] 图 1 是本说明书根据一示例性实施例示出的一种估计差分隐私保护数据中分词频度的场景示意图；

[29] 图 2 是本说明书根据一示例性实施例示出的一种估计差分隐私保护数据中分词频度的方法的流程图；

[30] 图 3 是本说明书根据一示例性实施例示出的另一种估计差分隐私保护数据中分词频度的方法的流程图；

[31] 图 4 是本说明书根据一示例性实施例示出的一种估计差分隐私保护数据中分词频度的装置的框图；

[32] 图 5 是本说明书根据一示例性实施例示出的一种前缀树的示意图；

[33] 图 6 是本说明书根据一示例性实施例示出的一种电子设备的结构示意图。

具体实施方式

[34] 这里将详细地对示例性实施例进行说明，其示例表示在附图中。下面的描述涉及附图时，除非另有表示，不同附图中的相同数字表示相同或相似的要素。以下示例性实施例中所描述的实施方式并不代表与本发明相一致的所有实施方式。相反，它们仅是与如所附权利要求书中所详述的、本发明的一些方面相一致的装置和方法的例子。

[35] 在本说明书中使用的术语是仅仅出于描述特定实施例的目的，而非旨在限制本发明。在本说明书和所附权利要求书中所使用的单数形式的“一种”、“所述”和“该”也旨在包括多数形式，除非上下文清楚地表示其他含义。还应当理解，本文中使用的术语“和/或”是指并包含一个或多个相关联的列出项目的任何或所有可能组合。

[36] 应当理解，尽管在本申请可能采用术语第一、第二、第三等来描述各种信息，但这些信息不应限于这些术语。这些术语仅用来将同一类型的信息彼此区分开。例如，在不脱离本申请范围的情况下，第一信息也可以被称为第二信息，类似地，第二信息也可以被称为第一信息。取决于语境，如在此所使用的词语“如果”可以被解释成为“在……时”或“当……时”或“响应于确定”。

[37] 如图 1 所示，在图 1 示出的场景中，用户在持有的终端设备中输入文本信息，终端设备对用户输入的文本信息进行分词处理，得到多个分词，每个分词由一个或多个词语

单元构成（例如，本场景中每个分词最多由 4 个词语单元构成）。终端设备对得到的各个分词进行本地差分隐私处理，得到各个分词各自对应的各个目标向量，并生成各个分词各自对应的各个分词信息。任意一个分词对应的分词信息可以包括该分词对应的目标向量以及构成该分词的词语单元的目标个数。终端设备将得到的各个分词信息上报至服务器。

[38]服务器接收多个终端设备上报的分词信息，并将接收到的分词信息汇总后进行分组，使得同组分词信息所对应的目标个数相等。例如，由 1 个词语单元构成的分词所对应的分词信息被分为一组，作为第 1 组分词信息。由 2 个词语单元构成的分词所对应的分词信息被分为一组，作为第 2 组分词信息，以此类推，本场景中可以得到包括第 3 组分词信息和第 4 组分词信息的共 4 组分词信息。

[39]然后，确定各组分词信息各自对应的表示分词频度无偏估计的各组估计数据。例如，可以基于第 1 组分词信息，确定第 1 组估计数据，第 1 组估计数据可以表示由 1 个词语单元构成的分词的频度无偏估计。可以基于第 2 组分词信息，确定第 2 组估计数据，第 2 组估计数据可以表示由 2 个词语单元构成的分词的频度无偏估计，以此类推，本场景中可以得到包括第 3 组估计数据和第 4 组估计数据的共 4 组估计数据。

[40]最后，可以基于各组估计数据，生成并输出用于记录分词频度的前缀树，作为分词频度估计的结果。具体来说，可以先生成前缀树的根节点，作为第 0 层节点。然后，基于第 n 组估计数据得到第 n 层的各个节点，第 n 层的各个节点各自对应于一个由 n 个词语单元构成的 n 元分词，在第 n 层的各个节点中记录对应的 n 元分词的频度。例如，可以基于第 1 组估计数据得到第 1 层的各个节点，第 1 层的各个节点各自对应于一个由 1 个词语单元构成的 1 元分词，在第 1 层的各个节点中记录对应的 1 元分词的频度。可以基于第 2 组估计数据得到第 2 层的各个节点，第 2 层的各个节点各自对应于一个由 2 个词语单元构成的 2 元分词，在第 2 层的各个节点中记录对应的 2 元分词的频度，以此类推，本场景中前缀树还包括第 3 层和第 4 层节点。

[41]下面将结合具体的实施例对本说明书提供的方案进行详细描述。

[42]如图 2 所示，图 2 示出的该方法可以应用于服务器中，该服务器可以实现为任何具有计算、处理能力的设备、平台、服务器或设备集群。该方法包括以下步骤：在步骤 201 中，获取终端设备上报的、经本地差分隐私处理的各个分词信息。

[43]在本实施例中，所涉及的终端设备可以是任意能够输入或查看文本信息的终端设备。

本领域技术人员可以理解,该终端设备可以包括但不限于诸如智能手机的移动终端设备、智能穿戴式设备、平板电脑、个人数字助理、膝上型便携式电脑以及台式电脑等等。

[44]在本实施例中,服务器可以获取多个终端设备各自上报的多个分词信息。任意一个分词信息对应于由一个或多个词语单元构成的分词,该分词信息可以包括目标向量和目标个数,其中,目标向量表示对应分词且经过本地差分隐私处理,目标个数表示构成该分词的词语单元的个数,该目标个数小于等于预设值 N。

[45]具体来说,在本实施例中,针对任意一个终端设备,首先,可以基于预设词典对用户输入或查看的文本信息进行预设的分词操作,以将该文本信息拆解成预设词典中的词语单元。再根据拆解的词语单元组合出多个分词,每个分词由一个或多个词语单元构成。并且,构成各个分词的词语单元的个数小于等于预设值 N。其中,词语单元可以是字,也可以是词语。

[46]例如,用户输入文本信息“以太坊中有两种不同的数据类型”,预设值 N=3,可以将上述文本信息拆解成以下词语单元:以太坊、中、有、不同的、数据和类型。然后,再根据拆解的词语单元组合出以下分词:以太坊、中、有、不同的、数据、类型、以太坊|中、中|有、有|不同的、不同的|数据、数据|类型、以太坊|中|有、中|有|不同的、有|不同的|数据、不同的|数据|类型,其中,符号“|”表示将词语单元分隔的符号。

[47]接着,该终端设备得到上述各个分词后,对各个分词进行本地差分隐私处理,得到各个分词各自对应的目标向量。任一分词对应的目标向量可以通过如下方式得到:首先,从多个预设的哈希函数中随机选取一个哈希函数,作为目标哈希函数。利用目标哈希函数对该分词进行哈希计算,得到该分词的目标哈希值。最后,利用满足差分隐私的方式,基于目标哈希值确定目标向量。

[48]例如,在一种具体实现方式中,该分词对应的字符串为 S。可以从 k 个预设的哈希函数 $H_1, H_2, \dots, H_k$ 中随机选取一个哈希函数 H_j , 作为目标哈希函数。其中, j 为 H_j 对应的编号。利用 H_j 对该分词的字符串 S 进行哈希计算,得到目标哈希值 $h = H_j(S)$ 。另外,生成一个随机向量 v, 随机向量 v 在各个维度上的取值为 1 或 -1, 其中, 各个维度上的取值为 -1 的概率为:

$$p = \frac{e^{\frac{\epsilon}{2}}}{1 + e^{\frac{\epsilon}{2}}}$$

[49]其中, ε 为预设的隐私预算, 用于表示隐私保护水平。将 v 的第 h 位进行翻转操作, 得到目标向量。例如, 如果 v 的第 h 位为 1, 则将第 h 位的 1 翻转为 -1, 如果 v 的第 h 位为 -1, 则将第 h 位的 -1 翻转为 1。

[50]可以理解, 上述具体实现方式仅仅是示例性的举例说明, 还可以通过其它任意合理的方式对该分词进行本地差分隐私处理, 得到目标向量, 本实施例对目标向量的具体获取方式方面不限定。

[51]在步骤 202 中, 划分出 N 组分词信息, 使同组的各个分词信息对应于相同的目标个数。

[52]在本实施例中, 可以将来自多个终端设备的多个分词信息进行汇总, 并进行分组, 划分出 N 组分词信息, 使同组的各个分词信息对应于相同的目标个数。例如, 预设值 $N=3$, 可以分出 $G1$ 、 $G2$ 、 $G3$ 三组分词信息, 其中, $G1$ 组中的各个分词信息对应的目标个数均为 1, $G2$ 组中的各个分词信息对应的目标个数均为 2, $G3$ 组中的各个分词信息对应的目标个数均为 3。

[53]在步骤 203 中, 确定各组分词信息各自对应的表示分词频度无偏估计的各组估计数据。

[54]在本实施例中, 针对任意一组分词信息, 可以确定该组分词信息对应的表示分词频度无偏估计的估计数据。因为, 划分出 N 组分词信息, 所以, 可以得到 N 组估计数据, 并且同组的各个估计数据也对应于相同的目标个数。

[55]例如, 参考步骤 201 提供的一种具体实现方式, 任意一个分词信息可以包括目标向量以及目标个数, 其中, 该分词信息还可以包括得到该目标向量所采用的目标哈希函数的编号 j 。可以利用以下公式计算该组分词信息中每个分词信息对应的参考向量:

$$\overline{x}_i = \frac{k}{2} (c \cdot \overline{v}_i + \overline{e})$$

[56]其中, $\overline{v}_i$ 表示该组分词信息中第 i 个分词信息对应的目标向量, $\overline{x}_i$ 表示第 i 个分词信息对应的参考向量, $\overline{e}$ 为和目标向量维度相同的单位向量, k 为预设的哈希函数的个数, c 为一个常数, c 可以表示为:

$$c = \frac{e^{\frac{\varepsilon}{2}} + 1}{e^{\frac{\varepsilon}{2}} - 1}$$

[57]其中， ε 为预设的隐私预算，和步骤 201 提供的一种具体实现方式中涉及到的隐私预算 ε 相等。

[58]接着，将该组分词信息按照对应的目标哈希函数的编号，将编号相同的分词信息所对应的参考向量相加，得到 k 个向量。将这 k 个向量作为行向量或列向量，按照编号从小到大的顺序进行排列，得到目标矩阵。该目标矩阵即为该组分词信息对应的表示分词频度无偏估计的一组估计数据。

[59]可以理解，上述举例仅仅是针对步骤 201 涉及的一种具体实现方式，而提供的一种确定各组估计数据的实现方式。实际上，针对不同的差分隐私处理方式，可以通过不同的方式方面确定各组估计数据。本实施例对确定各组估计数据的具体方式方面不限定。

[60]在步骤 204 中，基于各组估计数据，逐层生成用于记录分词频度的前缀树的各层节点。

[61]在本实施例中，可以基于各组估计数据，逐层生成用于记录分词频度的前缀树的各层节点，从而得到前缀树。具体来说，首先，生成前缀树的根节点，将前缀树的根节点作为第 0 层节点，该第 0 层节点表示空字符。接着，从第 1 层开始，逐层生成前缀树的各层节点。

[62]具体来说，通过以下步骤 a-步骤 d，生成第 n 层节点， n 为大于等于 1，小于等于 N 的整数：步骤 a：获取第 $n-1$ 层中的各个节点各自表示的各个 $n-1$ 元分词。

[63]在本实施例中，获取第 $n-1$ 层中的各个节点各自表示的各个 $n-1$ 元分词。其中，第 $n-1$ 层中的任一节点表示的 $n-1$ 元分词通过将从根节点到该节点所对应的词语单元顺序排列而形成。其中，若 $n=1$ ，则获取第 0 层节点表示的 0 元分词，即空字符。

[64]步骤 b：基于各个 $n-1$ 元分词，确定用于第 n 层节点的多个备选 n 元分词。

[65]在本实施例中，可以基于各个 $n-1$ 元分词，确定用于第 n 层节点的多个备选 n 元分词。具体来说，可以将各个 $n-1$ 元分词作为前缀，分别与预设词典中的各个预设词语单元组合，将如此构成的多个 n 元分词确定为多个备选 n 元分词。例如，若 $n-1$ 层节点中 2 个节点表示的 $n-1$ 元分词分别为 x 和 w (x 和 w 均为由 $n-1$ 个词语单元构成的 $n-1$ 元分词)，预设词典中包括词语单元 A 、 B 、 C ，则以 x 为前缀，分别与 A 、 B 、 C 组合，

构成的 n 元分词为 x_A 、 x_B 、 x_C ，以 w 为前缀，分别与 A 、 B 、 C 组合，构成的 n 元分词为 w_A 、 w_B 、 w_C 。可以将 x_A 、 x_B 、 x_C 、 w_A 、 w_B 、 w_C 确定为备选 n 元分词。

[66] 步骤 c：基于对应于目标个数为 n 的第 n 组估计数据，计算备选 n 元分词的频度显著性分布信息。

[67] 在本实施例中，可以基于第 n 组估计数据，计算备选 n 元分词的频度显著性分布信息。其中，第 n 组估计数据对应的目标个数为 n 。具体来说，可以通过如下方式计算备选 n 元分词的频度显著性分布信息：首先，可以基于第 n 组估计数据，计算各个备选 n 元分词的各个频度。参考步骤 203 提供的一种具体实现方式，得到目标矩阵作为表示分词频度无偏估计的一组估计数据，第 n 个矩阵为第 n 组估计数据。针对任意一个备选 n 元分词 D ，可以利用上述 k 个预设的哈希函数 $H_1, H_2, \dots, H_k$ ，分别对该备选 n 元分词 D 进行哈希计算，得到目标哈希值 $H_1(D), H_2(D), \dots, H_k(D)$ 。再从第 n 个矩阵中查找以上述目标哈希值为列的各个目标元素，计算目标元素的平均值，并基于该平均值得到该备选 n 元分词 D 的频度。

[68] 接着，可以基于各个备选 n 元分词的各个频度，计算各个备选 n 元分词各自对应的各个方差，从而得到各个方差对应的标准差。基于各个备选 n 元分词各自对应的各个方差，计算各个备选 n 元分词各自对应的各个 z 值。其中，针对任意一个备选 n 元分词，用该备选 n 元分词的频度除以该备选 n 元分词对应的标准差，得到该备选 n 元分词对应的 z 值。

[69] 最后，可以基于各个备选 n 元分词各自对应的各个 z 值，计算各个备选 n 元分词各自对应的各个 p 值（一个样本对应的 p 值为抽取到该样本或抽取到比该样本更极端情况样本的概率），作为各个备选 n 元分词各自对应的频度显著性分布信息。其中，针对任意一个备选 n 元分词，该备选 n 元分词对应的 p 值为标准正态分布中随机变量大于该备选 n 元分词对应的 z 值的概率。

[70] 在其他实施例中，还可以选取其他基于频度的指标作为显著性分布信息，例如，将前述 z 值，或者基于 z 值确定的其他统计分布量作为显著性分布信息。

[71] 步骤 d：基于上述频度显著性分布信息，选择若干备选 n 元分词作为第 n 层节点表示的 n 元分词，并利用第 n 层中的各个节点记录各自表示的各个 n 元分词的频度。

[72] 在本实施例中，可以基于上述频度显著性分布信息，从多个备选 n 元分词中选取满足一定条件的至少一个备选 n 元分词，作为第 n 层节点表示的 n 元分词，并利用第 n 层

中的各个节点记录各自表示的各个 n 元分词的频度。在将 p 值作为显著性分布信息的实施例中,还可以利用第 n 层中的各个节点记录各自表示的各个 n 元分词的方差以及 p 值。

[73]例如,备选 n 元分词包括 A、B、C、D、E、F,可以基于上述频度显著性分布信息,从备选 n 元分词中选取满足一定条件的 A、B、C,作为第 n 层节点表示的 n 元分词。然后,生成第 n 层的 3 个节点 a、b、c,这 3 个节点 a、b、c 分别表示 A、B、C。并且,节点 a 记录 A 的频度、方差及 p 值,节点 b 记录 B 的频度、方差及 p 值,节点 c 记录 C 的频度、方差及 p 值。

[74]本说明书的上述实施例提供的估计差分隐私保护数据中分词频度的方法,通过获取终端设备上报的、经本地差分隐私处理的各个分词信息,划分出 N 组分词信息,使同组的各个分词信息对应于相同的目标个数,确定各组分词信息各自对应的表示分词频度无偏估计的各组估计数据,并基于各组估计数据,逐层生成用于记录分词频度的前缀树的各层节点。由于本实施例在生成前缀树的第 n 层节点的过程中,能够基于备选 n 元分词的频度显著性分布信息,选择部分备选 n 元分词作为第 n 层节点表示的 n 元分词,无需遍历由预设的词语单元构成的所有 n 元分词,不仅大大降低了计算量,提高了计算效率,而且基于分词的频度显著性分布信息而筛选出来的第 n 层节点表示的 n 元分词更具合理性。

[75]如图 3 所示,图 3 实施例描述了选择若干备选 n 元分词作为第 n 层节点表示的 n 元分词的过程,该方法可以应用于服务器中,包括以下步骤:在步骤 301 中,将各个备选 n 元分词各自对应的各个 p 值按从小到大的顺序进行排列。

[76]在步骤 302 中,选取满足预设条件的最大 p 值作为目标 p 值。

[77]在本实施例中,对于任意一个 p 值,若该 p 值小于等于该 p 值对应的目标结果,则该 p 值满足预设条件。该目标结果为该 p 值在排列中的序号和针对第 n 层设定的预设阈值的乘积除以备选 n 元分词的个数所得结果。

[78]例如,将各个备选 n 元分词各自对应的各个 p 值按从小到大的顺序进行排列,用 p_i 表示排列中第 i 个 p 值。若 $p_i \leq (i/N) * \alpha_n$,则 p_i 满足预设条件。将最大的满足预设条件的 p 值作为目标 p 值。其中, α_n 为针对第 n 层节点设定的预设阈值,一般来说, n 越大, α_n 越大。

[79]在步骤 303 中,选取小于目标 p 值的各个 p 值各自对应的各个备选 n 元分词,作为第 n 层节点表示的 n 元分词。

[80]本说明书的上述实施例提供的估计差分隐私保护数据中分词频度的方法，通过将各个备选 n 元分词各自对应的各个 p 值按从小到大的顺序进行排列，选取满足预设条件的最大 p 值作为目标 p 值，并选取小于目标 p 值的各个 p 值各自对应的各个备选 n 元分词，作为第 n 层节点表示的 n 元分词。从而使得基于分词的各个 p 值而筛选出来的第 n 层节点表示的 n 元分词进一步具有合理性。

[81]在一些可选实施方式中，上述方法还可以包括：利用第 n 层中的各个节点记录各自表示的各个 n 元分词的方差以及 p 值。

[82]应当注意，尽管在上述实施例中，以特定顺序描述了本说明书实施例的方法的操作，但是，这并非要求或者暗示必须按照该特定顺序来执行这些操作，或是必须执行全部所示的操作才能实现期望的结果。相反，流程图中描绘的步骤可以改变执行顺序。附加地或备选地，可以省略某些步骤，将多个步骤合并为一个步骤执行，和/或将一个步骤分解为多个步骤执行。

[83]下面结合一个完整的应用实例，对本说明书一个或多个实施例的方案进行示意性说明。

[84]应用场景可以为：预设词典中包括词语单元 A、B、C、D，服务器需要基于预设词典估计差分隐私保护数据中 1 元分词，2 元分词和 3 元分词的频度。

[85]具体来说，首先，服务器获取终端设备上报的、经本地差分隐私处理的各个分词信息。并划分出第 1 组分词信息，第 2 组分词信息和第 3 组分词信息。其中，第 1 组分词信息对应的目标个数均为 1，第 2 组分词信息对应的目标个数均为 2，第 3 组分词信息对应的目标个数均为 3。

[86]接着，分别确定第 1 组分词信息对应的表示分词频度无偏估计的第 1 组估计数据，第 2 组分词信息对应的表示分词频度无偏估计的第 2 组估计数据，第 3 组分词信息对应的表示分词频度无偏估计的第 3 组估计数据。

[87]然后，逐层生成用于记录分词频度的前缀树的各层节点。如图 4 所示，可以先生成前缀树的根节点，作为第 0 层节点，该第 0 层节点表示空字符。接着，将预设词典中的词语单元 A、B、C、D 确定为 4 个备选 1 元分词。分别计算各个备选 1 元分词对应的频度显著性分布信息，基于各个备选 1 元分词对应的频度显著性分布信息，选取 A、B、C、D 作为第 1 层节点表示的 1 元分词。构建根节点的子节点 a，节点 a 表示 1 元分词 A，并利用节点 a 记录 1 元分词 A 的频度。构建根节点的子节点 b，节点 b 表示 1 元分词 B，

并利用节点 b 记录 1 元分词 B 的频度。构建根节点的子节点 c, 节点 c 表示 1 元分词 C, 并利用节点 c 记录 1 元分词 C 的频度。构建根节点的子节点 d, 节点 d 表示 1 元分词 D, 并利用节点 d 记录 1 元分词 D 的频度。其中, 节点 a、节点 b、节点 c 和节点 d 即为前缀树第 1 层的节点。

[88]接着, 获取第 1 层中的节点 a、节点 b、节点 c 和节点 d 各自表示的 1 元分词 A、B、C 和 D, 将 1 元分词 A、B、C 和 D 作为前缀, 分别与预设词典中的词语单元 A、B、C、D 构成多个 2 元分词 AA、AB、AC、AD、BA、BB、BC、BD、CA、CB、CC、CD、DA、DB、DC、DD, 作为多个备选 2 元分词。分别计算各个备选 2 元分词对应的频度显著性分布信息, 基于各个备选 2 元分词对应的频度显著性分布信息, 选取 AB、AC、BC、BD 作为第 2 层节点表示的 2 元分词。构建节点 a 的子节点 ab 和 ac, 构建节点 b 的子节点 bc 和 bd。其中, 节点 ab 表示 2 元分词 AB, 并利用节点 ab 记录 2 元分词 AB 的频度。节点 ac 表示 2 元分词 AC, 并利用节点 ac 记录 2 元分词 AC 的频度。节点 bc 表示 2 元分词 BC, 并利用节点 bc 记录 2 元分词 BC 的频度。节点 bd 表示 2 元分词 BD, 并利用节点 bd 记录 2 元分词 BD 的频度。其中, 节点 ab、节点 ac、节点 bc 和节点 bd 即为前缀树第 2 层的节点。

[89]最后, 获取第 2 层中的节点 ab、节点 ac、节点 bc 和节点 bd 各自表示的 2 元分词 AB、AC、BC、BD, 将 2 元分词 AB、AC、BC、BD 作为前缀, 分别与预设词典中的词语单元 A、B、C、D 构成多个 3 元分词 ABA、ABB、ABC、ABD、ACA、ACB、ACC、ACD、BCA、BCB、BCC、BCD、BDA、BDB、BDC、BDD 作为多个备选 3 元分词。分别计算各个备选 3 元分词对应的频度显著性分布信息, 基于各个备选 3 元分词对应的频度显著性分布信息, 选取 ABA、ABB、ACC、ACD、BDC、BDD 作为第 3 层节点表示的 3 元分词。构建节点 ab 的子节点 aba 和 abb, 构建节点 ac 的子节点 acc 和 acd, 构建节点 bc 的子节点 bdc 和 bdd。其中, 节点 aba 表示 3 元分词 ABA, 并利用节点 aba 记录 3 元分词 ABA 的频度。节点 abb 表示 3 元分词 ABB, 并利用节点 abb 记录 3 元分词 ABB 的频度。节点 acc 表示 3 元分词 ACC, 并利用节点 acc 记录 3 元分词 ACC 的频度。节点 acd 表示 3 元分词 ACD, 并利用节点 acd 记录 3 元分词 ACD 的频度。节点 bdc 表示 3 元分词 BDC, 并利用节点 bdc 记录 3 元分词 BDC 的频度。节点 bdd 表示 3 元分词 BDD, 并利用节点 bdd 记录 3 元分词 BDD 的频度。其中, 节点 aba、节点 abb、节点 acc、节点 acd、节点 bdc 和节点 bdd 即为前缀树第 3 层的节点。

[90]可见, 应用上述方案, 逐层生成用于记录分词频度的前缀树的各层节点, 在生成第

n 层节点的过程中,能够基于备选 n 元分词的频度显著性分布信息,选择部分备选 n 元分词作为第 n 层节点表示的 n 元分词,无需遍历由预设的词语单元构成的所有 n 元分词,不仅大大降低了计算量,提高了计算效率,而且基于分词的频度显著性分布信息而筛选出来的第 n 层节点表示的 n 元分词更具合理性。

[91]与前述估计差分隐私保护数据中分词频度的方法实施例相对应,本说明书还提供了估计差分隐私保护数据中分词频度的装置的实施例。

[92]如图 5 所示,图 5 示出的该装置应用于服务器,该装置可以包括:获取模块 501,分组模块 502,确定模块 503 和生成模块 504。

[93]其中,获取模块 501,用于获取终端设备上报的、经本地差分隐私处理的各个分词信息。任一分词信息对应于一个分词,并包括表示该分词中包含的词语单元数量的目标个数,该目标个数小于等于预设值 N。

[94]分组模块 502,用于划分出 N 组分词信息,使同组的各个分词信息对应于相同的目标个数。

[95]确定模块 503,用于确定各组分词信息各自对应的表示分词频度无偏估计的各组估计数据。

[96]生成模块 504,用于基于各组估计数据,逐层生成用于记录分词频度的前缀树的各层节点,其中,生成模块 504 通过如下方式生成第 n 层节点:获取第 n-1 层中的各个节点各自表示的各个 n-1 元分词,第 n-1 层中的任一节点表示的 n-1 元分词通过将根节点到该节点所对应的词语单元顺序排列而形成。基于各个 n-1 元分词,确定用于第 n 层节点的多个备选 n 元分词,基于对应于目标个数为 n 的第 n 组估计数据,计算备选 n 元分词的频度显著性分布信息,基于上述频度显著性分布信息,选择若干备选 n 元分词作为第 n 层节点表示的 n 元分词,并利用第 n 层中的各个节点记录各自表示的各个 n 元分词的频度, $1 \leq n \leq N$ 。

[97]在一种实施方式中,上述前缀树的根节点为第 0 层节点,第 0 层节点表示空字符。

[98]在另一种实施方式中,生成模块 504 通过如下方式基于各个 n-1 元分词,确定用于第 n 层节点的多个备选 n 元分词:将各个 n-1 元分词作为前缀,分别与预设词典中的各个预设词语单元构成的多个 n 元分词确定为所述多个备选 n 元分词。

[99]在另一种实施方式中,生成模块 504 通过如下方式基于对应于目标个数为 n 的第 n 组估计数据,计算备选 n 元分词的频度显著性分布信息:基于第 n 组估计数据,计算各

个备选 n 元分词的各个频度，基于各个频度，计算各个备选 n 元分词各自对应的各个方差，基于各个方差，计算备选 n 元分词的频度显著性分布信息。

[100] 在另一种实施方式中，生成模块 504 通过如下方式基于各个方差，计算备选 n 元分词的频度显著性分布信息：基于各个方差，计算各个备选 n 元分词各自对应的各个 z 值，基于各个 z 值，计算各个备选 n 元分词各自对应的各个 p 值，作为备选 n 元分词的频度显著性分布信息。

[101] 其中，生成模块 404 通过如下方式基于频度显著性分布信息，选择若干备选 n 元分词作为第 n 层节点表示的 n 元分词：基于各个 p 值，选择若干备选 n 元分词作为第 n 层节点表示的 n 元分词。

[102] 在另一种实施方式中，生成模块 504 通过如下方式基于各个 p 值，选择若干备选 n 元分词作为第 n 层节点表示的 n 元分词：将各个 p 值按从小到大的顺序进行排列，选取满足预设条件的最大 p 值作为目标 p 值。满足预设条件的任一 p 值小于等于该 p 值对应的目标结果，该目标结果为该 p 值在排列中的序号和针对第 n 层设定的预设阈值的乘积除以备选 n 元分词的个数所得结果。选取小于目标 p 值的各个 p 值各自对应的各个备选 n 元分词，作为第 n 层节点表示的 n 元分词。

[103] 在另一种实施方式中，生成模块 504 还被配置用于：利用第 n 层中的各个节点记录各自表示的各个 n 元分词的方差以及 p 值。

[104] 在另一种实施方式中，任一分词信息还可以包括表示该分词的目标向量，该目标向量经过本地差分隐私处理。

[105] 在另一种实施方式中，表示该分词的目标向量通过如下方式得到：从多个预设的哈希函数中选取一个哈希函数作为目标哈希函数，利用目标哈希函数计算该分词的目标哈希值，利用满足差分隐私的方式，基于目标哈希值确定目标向量。

[106] 应当理解，上述装置可以预先设置在服务器中，也可以通过下载等方式而加载到服务器中。上述装置中的相应模块可以与服务器中的模块相互配合以实现估计差分隐私保护数据中分词频度的方案。

[107] 对于装置实施例而言，由于其基本对应于方法实施例，所以相关之处参见方法实施例的部分说明即可。以上所描述的装置实施例仅仅是示意性的，其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根

据实际的需要选择其中的部分或者全部模块来实现本说明书一个或多个实施例方案的目的。本领域普通技术人员在不付出创造性劳动的情况下，即可以理解并实施。

[108] 本说明书一个或多个实施例还提供了一种计算机可读存储介质，该存储介质存储有计算机程序，计算机程序可用于执行上述图 2 至图 3 任一实施例提供的估计差分隐私保护数据中分词频度的方法。

[109] 对应于上述的估计差分隐私保护数据中分词频度的方法，本说明书一个或多个实施例还提出了图 6 所示的根据本说明书的一示例性实施例的电子设备的示意结构图。请参考图 6，在硬件层面，该电子设备包括处理器、内部总线、网络接口、内存以及非易失性存储器，当然还可能包括其他业务所需要的硬件。处理器从非易失性存储器中读取对应的计算机程序到内存中然后运行，在逻辑层面上形成估计差分隐私保护数据中分词频度的装置。当然，除了软件实现方式之外，本说明书一个或多个实施例并不排除其他实现方式，比如逻辑器件抑或软硬件结合的方式等等，也就是说以下处理流程的执行主体并不限定于各个逻辑单元，也可以是硬件或逻辑器件。

[110] 本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于系统实施例而言，由于其基本相似于方法实施例，所以描述地比较简单，相关之处参见方法实施例的部分说明即可。

[111] 上述对本说明书特定实施例进行了描述。其它实施例在所附权利要求书的范围内。在一些情况下，在权利要求书中记载的动作或步骤可以按照不同于实施例中的顺序来执行并且仍然可以实现期望的结果。另外，在附图中描绘的过程不一定要求示出的特定顺序或者连续顺序才能实现期望的结果。在某些实施方式中，多任务处理和并行处理也是可以的或者可能是有利的。

[112] 本领域普通技术人员应该还可以进一步意识到，结合本文中所公开的实施例描述的各示例的单元及算法步骤，能够以电子硬件、计算机软件或者二者的结合来实现，为了清楚地说明硬件和软件的可互换性，在上述说明中已经按照功能一般性地描述了各示例的组成及步骤。这些功能究竟以硬件还是软件方式来执行，取决于技术方案的特定应用和设计约束条件。本领域普通技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能，但是这种实现不应认为超出本申请的范围。其中，软件模块可以置于随机存储器（RAM）、内存、只读存储器（ROM）、电可编程 ROM、电可擦除可编程 ROM、寄存器、硬盘、可移动磁盘、CD-ROM、或技术领域内所公知的任意其它形式的

存储介质中。

[113] 以上所述的具体实施方式，对本申请的目的、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，以上所述仅为本申请的具体实施方式而已，并不用于限定本申请的保护范围，凡在本申请的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

权利要求书

- 1、一种估计差分隐私保护数据中分词频度的方法,应用于服务器,所述方法包括:
获取终端设备上报的、经本地差分隐私处理的各个分词信息;任一分词信息对应于一个分词,并包括表示该分词中包含的词语单元数量的目标个数,该目标个数小于等于预设值 N ;
划分出 N 组分词信息,使同组的各个分词信息对应于相同的目标个数;
确定各组分词信息各自对应的表示分词频度无偏估计的各组估计数据;
基于所述各组估计数据,逐层生成用于记录分词频度的前缀树的各层节点,其中,生成第 n 层节点包括:获取第 $n-1$ 层中的各个节点各自表示的各个 $n-1$ 元分词,第 $n-1$ 层中的任一节点表示的 $n-1$ 元分词通过将从根节点到该节点所对应的词语单元顺序排列而形成;基于所述各个 $n-1$ 元分词,确定用于第 n 层节点的多个备选 n 元分词;基于对应于目标个数为 n 的第 n 组估计数据,计算所述备选 n 元分词的频度显著性分布信息;基于所述频度显著性分布信息,选择若干备选 n 元分词作为第 n 层节点表示的 n 元分词,并利用第 n 层中的各个节点记录各自表示的各个 n 元分词的频度; $1 \leq n \leq N$ 。
- 2、根据权利要求1所述的方法,其中,所述前缀树的根节点为第0层节点,所述第0层节点表示空字符。
- 3、根据权利要求1所述的方法,其中,所述基于所述各个 $n-1$ 元分词,确定用于第 n 层节点的多个备选 n 元分词,包括:
将所述各个 $n-1$ 元分词作为前缀,分别与预设词典中的各个预设词语单元构成的多个 n 元分词确定为所述多个备选 n 元分词。
- 4、根据权利要求1所述的方法,其中,所述基于对应于目标个数为 n 的第 n 组估计数据,计算所述备选 n 元分词的频度显著性分布信息,包括:
基于所述第 n 组估计数据,计算各个备选 n 元分词在各个频度;
基于所述各个频度,计算所述各个备选 n 元分词各自对应的各个方差;
基于所述各个方差,计算所述备选 n 元分词的频度显著性分布信息。
- 5、根据权利要求4所述的方法,其中,所述基于所述各个方差,计算所述备选 n 元分词的频度显著性分布信息,包括:
基于所述各个方差,计算所述各个备选 n 元分词各自对应的各个 z 值;
基于所述各个 z 值,计算所述各个备选 n 元分词各自对应的各个 p 值,作为所述备选 n 元分词的频度显著性分布信息;
其中,所述基于所述频度显著性分布信息,选择若干备选 n 元分词作为第 n 层节点

表示的 n 元分词, 包括:

基于所述各个 p 值, 选择若干备选 n 元分词作为第 n 层节点表示的 n 元分词。

6、根据权利要求 5 所述的方法, 其中, 所述基于所述各个 p 值, 选择若干备选 n 元分词作为第 n 层节点表示的 n 元分词, 包括:

将所述各个 p 值按从小到大的顺序进行排列;

选取满足预设条件的最大 p 值作为目标 p 值; 满足所述预设条件的任一 p 值小于等于该 p 值对应的目标结果, 该目标结果为该 p 值在所述排列中的序号和针对第 n 层设定的预设阈值的乘积除以备选 n 元分词的个数所得结果;

选取小于所述目标 p 值的各个 p 值各自对应的各个备选 n 元分词, 作为所述第 n 层节点表示的 n 元分词。

7、根据权利要求 4-6 中任一项所述的方法, 其中, 所述方法还包括:

利用第 n 层中的各个节点记录各自表示的各个 n 元分词的方差以及 p 值。

8、根据权利要求 1 所述的方法, 其中, 所述任一分词信息还包括表示该分词的目标向量, 该目标向量经过本地差分隐私处理。

9、根据权利要求 8 所述的方法, 其中, 所述表示该分词的目标向量通过如下方式得到:

从多个预设的哈希函数中选取一个哈希函数作为目标哈希函数;

利用所述目标哈希函数计算该分词的目标哈希值;

利用满足差分隐私的方式, 基于所述目标哈希值确定所述目标向量。

10、一种估计差分隐私保护数据中分词频度的装置, 应用于服务器, 所述装置包括:

获取模块, 用于获取终端设备上报的、经本地差分隐私处理的各个分词信息; 任一分词信息对应于一个分词, 并包括表示该分词中包含的词语单元数量的目标个数, 该目标个数小于等于预设值 N ;

分组模块, 用于划分出 N 组分词信息, 使同组的各个分词信息对应于相同的目标个数;

确定模块, 用于确定各组分词信息各自对应的表示分词频度无偏估计的各组估计数据;

生成模块, 用于基于所述各组估计数据, 逐层生成用于记录分词频度的前缀树的各层节点, 其中, 所述生成模块通过如下方式生成第 n 层节点: 获取第 $n-1$ 层中的各个节点各自表示的各个 $n-1$ 元分词, 第 $n-1$ 层中的任一节点表示的 $n-1$ 元分词通过将从根节

点到该节点所对应的词语单元顺序排列而形成；基于所述各个 $n-1$ 元分词，确定用于第 n 层节点的多个备选 n 元分词；基于对应于目标个数为 n 的第 n 组估计数据，计算所述备选 n 元分词的频度显著性分布信息；基于所述频度显著性分布信息，选择若干备选 n 元分词作为第 n 层节点表示的 n 元分词，并利用第 n 层中的各个节点记录各自表示的各个 n 元分词的频度； $1 \leq n \leq N$ 。

11、根据权利要求 10 所述的装置，其中，所述前缀树的根节点为第 0 层节点，所述第 0 层节点表示空字符。

12、根据权利要求 10 所述的装置，其中，所述生成模块通过如下方式基于所述各个 $n-1$ 元分词，确定用于第 n 层节点的多个备选 n 元分词：

将所述各个 $n-1$ 元分词作为前缀，分别与预设词典中的各个预设词语单元构成的多个 n 元分词确定为所述多个备选 n 元分词。

13、根据权利要求 10 所述的装置，其中，所述生成模块通过如下方式基于对应于目标个数为 n 的第 n 组估计数据，计算所述备选 n 元分词的频度显著性分布信息：

基于所述第 n 组估计数据，计算各个备选 n 元分词在各个频度；

基于所述各个频度，计算所述各个备选 n 元分词各自对应的各个方差；

基于所述各个方差，计算所述备选 n 元分词的频度显著性分布信息。

14、根据权利要求 13 所述的装置，其中，所述生成模块通过如下方式基于所述各个方差，计算所述备选 n 元分词的频度显著性分布信息：

基于所述各个方差，计算所述各个备选 n 元分词各自对应的各个 z 值；

基于所述各个 z 值，计算所述各个备选 n 元分词各自对应的各个 p 值，作为所述备选 n 元分词的频度显著性分布信息；

其中，所述生成模块通过如下方式基于所述频度显著性分布信息，选择若干备选 n 元分词作为第 n 层节点表示的 n 元分词：

基于所述各个 p 值，选择若干备选 n 元分词作为第 n 层节点表示的 n 元分词。

15、根据权利要求 14 所述的装置，其中，所述生成模块通过如下方式基于所述各个 p 值，选择若干备选 n 元分词作为第 n 层节点表示的 n 元分词：

将所述各个 p 值按从小到大的顺序进行排列；

选取满足预设条件的最大 p 值作为目标 p 值；满足所述预设条件的任一 p 值小于等于该 p 值对应的目标结果，该目标结果为该 p 值在所述排列中的序号和针对第 n 层设定的预设阈值的乘积除以备选 n 元分词的个数所得结果；

选取小于所述目标 p 值的各个 p 值各自对应的各个备选 n 元分词，作为所述第 n 层

节点表示的 n 元分词。

16、根据权利要求 13-15 中任一项所述的装置，其中，所述生成模块还被配置用于：

利用第 n 层中的各个节点记录各自表示的各个 n 元分词的方差以及 p 值。

17、根据权利要求 10 所述的装置，其中，所述任一分词信息还包括表示该分词的目标向量，该目标向量经过本地差分隐私处理。

18、根据权利要求 17 所述的装置，其中，所述表示该分词的目标向量通过如下方式得到：

从多个预设的哈希函数中选取一个哈希函数作为目标哈希函数；

利用所述目标哈希函数计算该分词的目标哈希值；

利用满足差分隐私的方式，基于所述目标哈希值确定所述目标向量。

19、一种计算机可读存储介质，所述存储介质存储有计算机程序，所述计算机程序被处理器执行时实现上述权利要求 1-9 中任一项所述的方法。

20、一种电子设备，包括存储器、处理器及存储在所述存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时实现上述权利要求 1-9 中任一项所述的方法。