CLAIMS

What is claimed is:

1. A method for estimating a word segment frequency in differential privacy protection data, applied to a server, wherein the method comprises:

obtaining each piece of word segment information that is reported by a terminal device and that is subject to local differential privacy processing, wherein any piece of word segment information corresponds to one word segment, and comprises a target quantity that represents a quantity of word units comprised in the word segment, and the target quantity is less than or equal to a predetermined value N;

obtaining through division N groups of word segment information, so each piece of word segment information of the same group corresponds to the same target quantity;

determining each group of estimated data that is corresponding to each group of word segment information and that represents unbiased word segment frequency estimation; and

generating, layer by layer based on each group of estimated data, each layer of nodes of a prefix tree used to record a word segment frequency, wherein generating an nth layer of nodes comprises: obtaining each (n–1)-tuple word segment represented by each node at an (n–1)th layer, wherein an (n–1)-tuple word segment represented by any node at the (n–1)th layer is formed by sequentially arranging word units corresponding to a root node to the node; determining a plurality of candidate n-tuple word segments for the nth layer of nodes based on each (n–1)-tuple word segment; calculating frequency salient distribution information of the candidate n-tuple word segment based on each variance corresponding to each candidate n-tuple word segment that is calculated based on an nth group of estimated data corresponding to a target quantity n; and selecting, based on the frequency salient distribution information, several candidate n-tuple word segments as n-tuple word segments represented by the nth layer of nodes, and recording, by using each node at the nth layer, a frequency of an n-tuple word segment represented by the node, wherein $1 \leq n \leq N$.

2. The method according to claim 1, wherein the root node of the prefix tree is a 0th-layer node, and the 0th-layer node represents an empty character.

3. The method according to claim 1, wherein the determining a plurality of candidate n-tuple word segments for the nth layer of nodes based on each (n–1)-tuple word segment comprises:

determining, as the plurality of candidate n-tuple word segments, a plurality of n-tuple word segments formed by using each (n–1)-tuple word segment as a prefix and each predetermined word unit in a predetermined dictionary.

4. The method according to claim 1, wherein the calculating frequency salient distribution information of the candidate n-tuple word segment based on each variance further comprises:

calculating each frequency of each candidate n-tuple word segment based on the nth group of estimated data; and

calculating each variance corresponding to each candidate n-tuple word segment based on each frequency.

5. The method according to claim 1, wherein the calculating the frequency salient distribution information of the candidate n-tuple word segment based on each variance comprises:

calculating each z value corresponding to each candidate n-tuple word segment based on each variance; and

calculating each p value corresponding to each candidate n-tuple word segment based on each z value, as the frequency salient distribution information of the candidate n-tuple word segment;

wherein the selecting, based on the frequency salient distribution information, several candidate n-tuple word segments as n-tuple word segments represented by the nth layer of nodes comprises:

selecting, based on each p value, several candidate n-tuple word segments as the n-tuple word segments represented by the nth layer of nodes.

6. The method according to claim 5, wherein the selecting, based on each p value, several candidate n-tuple word segments as the n-tuple word segments represented by the nth layer of nodes comprises:

arranging the p values in ascending order;

selecting a maximum p value that satisfies a predetermined condition as a target p value, wherein any p value that satisfies the predetermined condition is less than or equal to a target result corresponding to the p value, and the target result is a result obtained by dividing a product of a sequence number of the p value in the arrangement

and a predetermined threshold set for the nth layer by a quantity of candidate n-tuple word segments; and

selecting candidate n-tuple word segments corresponding to p values that are less than the target p value as the n-tuple word segments represented by the nth layer of nodes.

7. The method according to any one of claims 4 to 6, wherein the method further comprises:

using each node at the nth layer to record a variance and a p value of an n-tuple word segment represented by the node.

8. The method according to claim 1, wherein the any piece of word segment information further comprises a target vector representing the word segment, and the target vector is subject to local differential privacy processing.

9. The method according to claim 8, wherein the target vector representing the word segment is obtained in the following method:

selecting one hash function from a plurality of predetermined hash functions as a target hash function;

calculating a target hash value of the word segment by using the target hash function; and

determining the target vector based on the target hash value in a method of satisfying differential privacy.

10. An apparatus for estimating a word segment frequency in differential privacy protection data, applied to a server, wherein the apparatus comprises:

an acquisition module, configured to obtain each piece of word segment information that is reported by a terminal device and that is subject to local differential privacy processing, wherein any piece of word segment information corresponds to one word segment, and comprises a target quantity that represents a quantity of word units comprised in the word segment, and the target quantity is less than or equal to a predetermined value N;

a grouping module, configured to obtain through division N groups of word segment information, so each piece of word segment information of the same group corresponds to the same target quantity;

a determining module, configured to determine each group of estimated data that is corresponding to each group of word segment information and that represents unbiased word segment frequency estimation; and

a generation module, configured to generate, layer by layer based on each group of estimated data, each layer of nodes of a prefix tree used to record a word segment frequency, wherein the generation module generates an nth layer of nodes in the following method: obtaining each (n–1)-tuple word segment represented by each node at an (n–1)th layer, wherein an (n–1)-tuple word segment represented by any node at the (n–1)th layer is formed by sequentially arranging word units corresponding to a root node to the node; determining a plurality of candidate n-tuple word segments for the nth layer of nodes based on each (n–1)-tuple word segment; calculating frequency salient distribution information of the candidate n-tuple word segment based on each variance corresponding to each candidate n-tuple word segment that is calculated based on an nth group of estimated data corresponding to a target quantity n; and selecting, based on the frequency salient distribution information, several candidate n-tuple word segments as n-tuple word segments represented by the nth layer of nodes, and recording, by using each node at the nth layer, a frequency of an n-tuple word segment represented by the node, wherein $1 \leq n \leq N$.

11. The apparatus according to claim 10, wherein the root node of the prefix tree is a 0th-layer node, and the 0th-layer node represents an empty character.

12. The apparatus according to claim 10, wherein the generation module determines the plurality of candidate n-tuple word segments for the nth layer of nodes based on each (n–1)-tuple word segment in the following method:

determining, as the plurality of candidate n-tuple word segments, a plurality of n-tuple word segments formed by using each (n–1)-tuple word segment as a prefix and each predetermined word unit in a predetermined dictionary.

13. The apparatus according to claim 10, wherein the generation module is further configured to calculate the frequency salient distribution information of the candidate n-tuple word segment based on each variance in the following method:

calculating each frequency of each candidate n-tuple word segment based on the nth group of estimated data; and

calculating each variance corresponding to each candidate n-tuple word segment based on each frequency.

14. The apparatus according to claim 10, wherein the generation module calculates the frequency salient distribution information of the candidate n-tuple word segment based on each variance in the following method:

calculating each z value corresponding to each candidate n-tuple word segment based on each variance; and

calculating each p value corresponding to each candidate n-tuple word segment based on each z value, as the frequency salient distribution information of the candidate n-tuple word segment;

wherein the generation module selects, based on the frequency salient distribution information, several candidate n-tuple word segments as the n-tuple word segments represented by the nth layer of nodes in the following method:

selecting, based on each p value, several candidate n-tuple word segments as the n-tuple word segments represented by the nth layer of nodes.

15. The apparatus according to claim 14, wherein the generation module selects, based on each p value, several candidate n-tuple word segments as the n-tuple word segments represented by the nth layer of nodes in the following method:

arranging the p values in ascending order;

selecting a maximum p value that satisfies a predetermined condition as a target p value, wherein any p value that satisfies the predetermined condition is less than or equal to a target result corresponding to the p value, and the target result is a result obtained by dividing a product of a sequence number of the p value in the arrangement and a predetermined threshold set for the nth layer by a quantity of candidate n-tuple word segments; and

selecting candidate n-tuple word segments corresponding to p values that are less than the target p value as the n-tuple word segments represented by the nth layer of nodes.

16. The apparatus according to any one of claims 10 to 15, wherein the generation module is further configured to:

use each node at the nth layer to record a variance and a p value of an n-tuple word segment represented by the node.

17. The apparatus according to claim 10, wherein the any piece of word segment information further comprises a target vector representing the word segment, and the target vector is subject to local differential privacy processing.

18. The apparatus according to claim 17, wherein the target vector representing the word segment is obtained in the following method:

selecting one hash function from a plurality of predetermined hash functions as a target hash function;

calculating a target hash value of the word segment by using the target hash function; and

determining the target vector based on the target hash value in a method of satisfying differential privacy.

19. A computer readable storage medium on which a computer program is stored, wherein the method according to any one of claims 1 to 9 is implemented when the computer program is executed by a processor.

20. An electronic device, comprising a memory, a processor, and a computer program that is stored in the memory and that is capable of running on the processor, the processor implementing the method according to any one of claims 1 to 9 when executing the program.