CLAIMS

1. A method for identifying private data, comprising:

obtaining first statistic information of a first data set, the first data set being a set of private data samples, and the first statistic information including a first length distribution vector and first character distribution statistical information of the first data set;

obtaining second statistic information of a second data set, the second data set being a set of data samples corresponding to a to-be-identified field, and the second statistic information including a second length distribution vector and second character distribution statistical information of the second data set;

determining a length distribution similarity between the first data set and the second data set based on the first length distribution vector and the second length distribution vector, wherein values in the first length distribution vector and the second length distribution vector represents numbers of data samples of different lengths in a respective data set;

determining a character distribution similarity between the first data set and the second data set based on the first character distribution statistical information and the second character distribution statistical information in response to determining that the length distribution similarity meets a first predetermined condition, wherein each of the first character distribution statistical information and the second character distribution statistical information comprises a plurality of character distribution vectors of a number corresponding to the sample length of the respective data set, wherein for a character distribution vector corresponding to a length position of a data sample, values in the character distribution vector represent statistics of predetermined characters that are of all data samples in the respective data set and that appear at the length position; and

determining data corresponding to the to-be-identified field as private data of a same type as the first data set in response to determining that the character distribution similarity meets a second predetermined condition, wherein the determining the character distribution similarity between the first data set and the second data set specifically includes:

determining at least one data group, each data group includes a first data subset and a second data subset, a data sample length of the first data subset being the same as a data sample length of the second data subset in a same data group, data samples in the first data subset being selected from the first data set, and data samples in the second data subset being selected from the second data set;

for each data group in the at least one data group, calculating a similarity sub-value between first character distribution statistical information of the first data subset and second character distribution statistical information of the second data subset;

determining a similarity value between the first data set and the second data set based on the similarity sub-value corresponding to each data group in the at least one data group.

2. The method according to claim 1, further comprising:

before the obtaining the first statistic information of the first data set, obtaining the first data set selected by a user; and

determining the first statistic information of the first data set.

3. The method according to claim 1, wherein the obtaining the second statistic information of the second data set includes determining the second length distribution vector of the second data set by:

initializing the second length distribution vector, a value at an i-th position in the second length distribution vector representing a number of data samples with a length i, and i being a positive integer less than or equal to a length of the second length distribution vector;

traversing data samples in the second data set to obtain length information of each data sample; and

determining the second length distribution vector based on the length information of each data sample.

4. The method according to claim 1, wherein the obtaining the second statistic information of the second data set includes:

determining at least one data subset from the second data set based on lengths of data samples in the second data set, data samples in a same data subset having a same length; and

for each data subset from the second data set, determining second character distribution statistical information of data samples in the data subset.

5. The method according to claim 4, wherein the determining the second character distribution statistical information of the data samples in the data subset includes:

for each length position of the data samples in the data subset, determining a character distribution vector corresponding to the length position based on characters included in the data samples in the data subset; and

obtaining the second character distribution statistical information of the data samples in the data subset based on the character distribution vectors corresponding to the length positions.

6. The method according to claim 1, wherein the determining the length distribution similarity between the first data set and the second data set based on the first length distribution vector and the second length distribution vector specifically includes:

calculating a first KL distance between the first length distribution vector and the second length distribution vector; and

wherein the determining that the length distribution similarity meets the first predetermined condition includes:

determining that the first KL distance is less than or equal to a first KL threshold.

7. The method according to claim 1, wherein the calculating the similarity sub-value between the first character distribution statistical information of the first data subset and the second character distribution statistical information of the second data subset includes:

for each length position of the first data subset and the second data subset, calculating a second KL distance between a first character distribution vector and a second character distribution vector to obtain a plurality of second KL distances;

determining a first number of second KL distances that are in the plurality of second KL distances and that are less than or equal to a second KL threshold; and

calculating a first ratio between the first number and a total number of the plurality of second KL distances.

8. The method according to claim 1, wherein the determining the similarity value between the first data set and the second data set includes:

determining a similarity sub-value corresponding to one data group meeting a predetermined condition as the similarity value between the first data set and the second data set; and

wherein the determining that the character distribution similarity meets the second predetermined condition includes:

determining that the similarity value is greater than a predetermined similarity threshold.

9. The method according to claim 1, wherein the determining the similarity value between the first data set and the second data set includes:

calculating the similarity value between the first data set and the second data set based on similarity sub-values corresponding to two or more data groups from the at least one data group; and

wherein the determining that the character distribution similarity meets the second predetermined condition includes:

determining that the similarity value is greater than a predetermined similarity threshold.

10. The method according to claim 1, further comprising:

before the determining the length distribution similarity between the first data set and the second data set, determining that a length corresponding to the largest number of data samples in the first data set is the same as a length corresponding to the largest number of data samples in the second data set.

11. The method according to claim 1, wherein a language type of the data samples in the first data set and a language type of the data samples in the second data set include at least one of English, French, or Chinese, and the language type of the data samples in the first data set is the same as the language type of the data samples in the second data set.

12. The method according to claim 1, further comprising:

after the determining the data corresponding to the to-be-identified field as the private data of the same type as the first data set, anonymizing the data corresponding to the to-be-identified field.

13. A non-transitory computer readable medium storing contents that, when executed by one or more processors, cause the one or more processors to implement a method according to any one of claims 1-12.

14. A system, comprising:

one or more processors; and

memory storing contents that, when executed by the one or more processors, cause the system to implement a method according to any one of claims 1-12.