CLAIMS

1. A similarity search method performed on data encrypted for confidentiality, the method comprising:

a first step of encrypting a plurality of units of sampled data for confidentiality by mapping in which a distance is preserved at least locally in a sampled data space;

a second step of clustering and recording a pair of the sampled data encrypted for confidentiality and arbitrary data, based on the sampled data encrypted for confidentiality itself;

a third step of encrypting query data for confidentiality by a method same as the method used for the sampled data;

a fourth step of identifying a cluster in which sampled data similar to the query data is recorded, based on the query data encrypted for confidentiality; and

a fifth step of identifying, from the identified cluster, a unit of sampled data similar to the query data, by calculating a distance between the sampled data encrypted for confidentiality and the query data encrypted for confidentiality, and accessing arbitrary data paired with the unit of the sampled data,

wherein

the encrypting the data for confidentiality in the first step and the third step includes:

a sub step of quantizing the data on a data space thereof;

a sub step of subtracting the quantized data from the data to obtain a residual; and

a sub step of inputting the quantized data to a cryptographic hash function to obtain a hash value,

wherein a set of the hash value and the residual is output as data encrypted for confidentiality, and

the clustering in the second step includes, based on part of the hash value of the sampled data encrypted for confidentiality, associating the hash value and a cluster one to one or one to many, wherein

the encrypting the sampled data for confidentiality in the first step includes:

a first sub step of dividing the sampled data into two by a method of separating a high-order bit and a low-order bit in binary representation of each element;

a second sub step of Gray-coding data of the high-order bit obtained through the division into two for each element;

a third sub step of generating a random number;

a fourth sub step of altering a part of the Gray code under a predetermined rule based on the random number;

a fifth sub step of altering data of the low-order bit obtained through the division into two, based on the random number;

a sixth sub step of supposing the altered data of the low-order bit to be a column vector, and multiplying a random projection (RP) matrix, generated based on the altered Gray code, from a top, to obtain data of the low-order bit to which random projection is applied;

a seventh sub step of applying error correction coding to the altered Gray code to obtain a redundant symbol; and

an eighth sub step of linking the data of the low-order bit, to which the random projection is applied, and the redundant symbol to obtain data encrypted for confidentiality.

2. A similarity search method performed on data encrypted for confidentiality, the method comprising:

a first step of encrypting a plurality of units of sampled data for confidentiality by mapping in which a distance is preserved at least locally in a sampled data space;

a second step of clustering and recording a pair of the sampled data encrypted for confidentiality and arbitrary data, based on the sampled data encrypted for confidentiality itself;

a third step of encrypting query data for confidentiality by a method same as the method used for the sampled data;

a fourth step of identifying a cluster in which sampled data similar to the query data is recorded, based on the query data encrypted for confidentiality; and

a fifth step of identifying, from the identified cluster, a unit of sampled data similar to the query data, by calculating a distance between the sampled data encrypted for confidentiality and the query data encrypted for confidentiality, and accessing arbitrary data paired with the unit of the sampled data,

wherein

the encrypting the data for confidentiality in the first step and the third step includes:

a sub step of quantizing the data on a data space thereof;

a sub step of subtracting the quantized data from the data to obtain a residual; and

a sub step of inputting the quantized data to a cryptographic hash function to obtain a hash value,

wherein a set of the hash value and the residual is output as data encrypted for

confidentiality, and

the clustering in the second step includes, based on part of the hash value of the sampled data encrypted for confidentiality, associating the hash value and a cluster one to one or one to many, wherein

in the encrypting the query data for confidentiality in the third step includes:

a first sub step of dividing the query data into two by a method of separating a high-order bit and a low-order bit in binary representation of each element;

a second sub step of Gray-coding data of the high-order bit obtained through the division into two for each element;

a third sub step of determining, from among numbers of entire sets having a possibility to be generated as random numbers, one on which a trial is to be performed;

a fourth sub step of altering a part of the Gray code under a predetermined rule, based on the one number on which the trial is to be performed;

a fifth sub step of applying error correction coding to the altered Gray code to obtain a redundant symbol;

a sixth sub step of applying error correction decoding to the Gray code before alteration by using the redundant symbol;

a seventh sub step of altering data of the low-order bit obtained by the division into two, based on a position of a bit corrected by the error correction decoding;

an eighth sub step of supposing data of the altered low-order bit to be a column vector, multiplying a random projection (RP) matrix, generated based on the altered Gray code, from a top to obtain data of the lower-order bit to which random projection is applied, and

a ninth sub step of linking the data of the low-order bit, to which the random projection is applied, and the redundant symbol to obtain data encrypted for confidentiality.

3. The similarity search method performed on the data encrypted for confidentiality, according to claim 2, wherein the identifying the cluster in the fourth step includes narrowing down clusters having a possibility that sampled data similar to the query data is recorded, to a number less than a number of times of trials in the third sub step.