

DESCRIPTION

[Title of Invention]

SIMILAR DATA SEARCH DEVICE, SIMILAR DATA SEARCH METHOD,
5 AND COMPUTER-READABLE STORAGE MEDIUM

[Technical Field]

[0001]

The present invention relates to similar data search devices and, in particular, to a similar data search device and a similar data search method
10 for carrying out a search based on a similarity between sets converted from character strings, and to a computer-readable recording medium that records a similar data search program for implementing such device and method.

[Background Art]

15 [0002]

Similar data searching is the fundamental and important data processing with wide applicability to clustering, redundant data matching, character string soft matching, and the like. Specific methods for similar data searching may include mere calculation of similarities among all the
20 target data pieces and doing a search based on the similarities, but this method needs a tremendously longer time for a greater amount of data.

[0003]

For example, in order to search a database for every pair of data pieces having a similarity no smaller than a certain value, similarities have
25 to be calculated $(N(N - 1) / 2)$ times where N is the number of pieces of data. In other words, for example, assuming that one similarity calculation takes 0.001 milliseconds and the number of pieces of data N is 100,000, similarities have to be calculated about five billion times, which is equivalent to calculating for about 14 days.

[0004]

Thus, NPTL 1 discloses a system for quickly retrieving every pair of data pieces having a similarity no smaller than a certain value to reduce the processing time. The system disclosed in NPTL 1 converts character strings to their feature sets, divides the sets according to their sizes where the size of a set is defined as the number of elements in a set, and generates every inverted index for the sets of the same size. Then, during a search, the system disclosed in NPTL 1 identifies a maximum value and a minimum value of the size for an inverted index to be searched based on the size of the inputted set and a similarity threshold value, and conducts a search on only the inverted indexes having a size falling within the identified range.

[0005]

Specifically, Table 1 in NPTL 1 discloses that, assuming that the requested set for search is denoted by X , it is sufficient to search inverted indexes having a size equal to or greater than $\alpha|X|$ and equal to or less than $|X| / \alpha$ if the Jaccard coefficient $(|X \cap Y| / |X \cup Y|)$ is not smaller than a threshold value α . In this way, the system disclosed in NPTL 1 generates inverted indexes according to the size of a set and identifies which inverted index should be searched by using an upper limit and a lower limit of a size that are determined based on search conditions (search request). Consequently, unnecessary searches are omitted and thus a faster search is achieved.

[Citation List]

[Non Patent Literature]

[0006]

[NPTL 1] Naoaki Okazaki and Jun'ichi Tsujii. A Simple and Fast Algorithm for Approximate String Matching with Set Similarity. Journal of Natural Language Processing, Vol. 18, No. 2, pp. 89–117, 2011.

[Summary of Invention]

5 [Technical Problem]

[0007]

However, the system disclosed in NPTL 1 is problematic in that the retrieval effectiveness is impaired when the search target data contains fewer sets of the same size.

10 [0008]

This is because, due to the fact that every inverted index is created for sets of the same size, if the data of the same size exists in a smaller amount, a smaller number of sets can be searched from an inverted index, and thereby, the number of searches to inverted indexes should be increased to obtain the same result.

15

[0009]

On average, the number of sets of the same size is inversely proportional to the number of searches to inverted indexes required to obtain the same result. For this reason, the retrieval effectiveness is impaired particularly by a high cost of searches to inverted indexes, such as the case where random access is made to an external storage device.

20

[0010]

An object of the present invention is to provide a similar data search device, a similar data search method, and a computer-readable recording medium, solving the above problem and making it possible to suppress lowering the retrieval effectiveness caused by an increased number of searches in inverted indexes even when the search target data contains a small number of sets of the same size.

25

[Solution to Problem]

[0011]

To achieve the above-described object, a similar data search device according to an aspect of the present invention, which is a similar data
5 search device for conducting searches by using sets as search target data and search condition data, includes:

an inverted index generating unit which, for generating inverted indexes used for a search, determines size ranges of sets of search targets for each of the inverted indexes to be generated so that at least a specified
10 number of sets of search targets are included in each of the inverted indexes to be generated, and generates the inverted indexes by dividing the sets of search targets according to the determined size ranges;

an unnecessary inverted index identifying unit which determines, based on a size of a set of search conditions and a specified threshold value
15 of a similarity between the set of search conditions and the set of search targets, a condition of a size of the set of search targets necessary for the similarity to be no smaller than the threshold value, and identifies, as an inverted index unnecessary for searches, from among the inverted indexes, any inverted index other than those inverted indexes which include a set
20 whose minimum size value satisfies the condition; and

a data search unit which conducts a search by applying the set of search conditions to an inverted index other than the identified inverted index unnecessary for searches.

[0012]

25 To achieve the above-described object, a similar data search method according to an aspect of the present invention, which is a similar data search method for conducting searches by using sets as search target data and search condition data, includes the steps of:

(a) for generating inverted indexes used for a search, determining size ranges of sets of search targets for each of inverted indexes to be generated so that at least a specified number of sets of search targets are included in each of the inverted index to be generated, and generating the
5 inverted indexes by dividing the sets of search targets according to the determined size ranges;

(b) determining, based on a size of a set of search conditions and a specified threshold value of a similarity between the set of search conditions and the set of search targets, a condition of a size of the set of
10 search targets necessary for the similarity to be no smaller than the threshold value, and identifying, as an inverted index unnecessary for searches, from among the inverted indexes, any inverted index other than those inverted indexes that include a set whose minimum size value satisfies the condition; and

15 (c) conducting a search by applying the set of search conditions to an inverted index other than the inverted index unnecessary for searches as identified in the step (b).

[0013]

Furthermore, to achieve the above-described object, a
20 computer-readable recording medium according to an aspect of the present invention records a program for conducting a search with a computer by using sets as search target data and search condition data,

wherein the program includes instructions causing the computer to execute the steps of:

25 (a) for generating inverted indexes used for a search, determining size ranges of sets of search targets for each of inverted indexes to be generated so that at least a specified number of sets of search targets are included in each of the inverted index to be generated, and generating the

inverted indexes by dividing the sets of search targets according to the determined size ranges;

(b) determining, based on a size of a set of search conditions and a specified threshold value of a similarity between the set of search
5 conditions and the set of search targets, a condition of a size of the set of search targets necessary for the similarity to be no smaller than the threshold value, and identifying, as an inverted index unnecessary for searches, from among the inverted indexes, any inverted index other than those inverted indexes that include a set whose minimum size value
10 satisfies the condition; and

(c) conducting a search by applying the set of search conditions to an inverted index other than the inverted index unnecessary for searches as identified in the step (b).

[Advantageous Effects of Invention]

15 [0014]

As described above, according to the present invention, it is made possible to suppress lowering the retrieval effectiveness caused by an increased number of searches in inverted indexes even when the search target data contains a small number of sets of the same size.

20 [Brief Description of Drawings]

[0015]

Fig. 1 is a block diagram illustrating a configuration of a similar data search device according to a first exemplary embodiment of the present invention.

25 Fig. 2 is a flow diagram illustrating operations of the similar data search device according to the first exemplary embodiment of the present invention.

Fig. 3 is an explanatory diagram illustrating an example of Step A1,

where the Jaccard coefficient is used as a similarity measure.

Fig. 4 is a diagram showing example mathematical expressions for obtaining a size range of a set.

Fig. 5 is a diagram showing example conditions for being
5 equivalent to the case where a similarity is not smaller than a threshold value.

Fig. 6 is a block diagram illustrating a configuration of a similar data search device according to a second exemplary embodiment of the present invention.

10 Fig. 7 is a flow diagram illustrating operations of the similar data search device according to the second exemplary embodiment of the present invention.

Fig. 8 is a block diagram illustrating an example computer implementing the similar data search device according to either of the first
15 and second exemplary embodiments of the present invention.

[Description of Embodiments]

[0016]

(First Exemplary Embodiment)

A similar data search device, a similar data search method, and a
20 program for similar data searches according to a first exemplary embodiment of the present invention will now be described with reference to Figs. 1 to 5.

[0017]

[Device Configuration]

25 To begin with, a configuration of the similar data search device according to the first exemplary embodiment is described with reference to Fig. 1. Fig. 1 is a block diagram illustrating the configuration of the similar data search device according to the first exemplary embodiment of

the present invention.

[0018]

The similar data search device 2 according to the first exemplary embodiment as illustrated in Fig. 1 is a device for conducting searches by using sets as the search target data and the search condition data. As illustrated in Fig. 1, the similar data search device 2 includes an inverted index generating unit 20, an unnecessary inverted index identifying unit 21, and a data search unit 22.

[0019]

The inverted index generating unit 20 generates an inverted index to be used for a search. For this purpose, the inverted index generating unit 20 first determines size ranges of sets of search targets for each of inverted indexes to be generated so that the number of sets of search targets included in each of the inverted indexes to be generated is not smaller than a specified number. Then, the inverted index generating unit 20 generates inverted indexes by dividing the sets of search targets according to the determined size ranges.

[0020]

The unnecessary inverted index identifying unit 21 first determines a condition of the size of a set of search targets required for a similarity to be no smaller than a threshold value, based on the size of the set of search conditions and on a threshold value specified for a similarity between sets of search conditions and search targets. Next, out of the inverted indexes, the unnecessary inverted index identifying unit 21 identifies, as an inverted index unnecessary for searches, any inverted index other than those inverted indexes containing a set whose minimum size value satisfies the condition.

[0021]

The data search unit 22 distinguishes any inverted index from the inverted index(es) unnecessary for searches as identified by the unnecessary inverted index identifying unit 21, and then conducts a search by applying the set of search conditions to the distinguished inverted
5 index.

[0022]

Thus, in the present exemplary embodiment, a size range of sets is determined for each of inverted indexes containing the sets and, based on the size range, an inverted index unsuitable to be searched for a set of
10 search conditions is identified. A search is then conducted on inverted indexes excluding the identified inverted index(es). As a result, even when the search target data contains a small number of sets of the same size, it is made possible to suppress lowering the retrieval effectiveness that would be caused by an increased number of searches in inverted
15 indexes.

[0023]

The configuration of the similar data search device 2 according to the first exemplary embodiment will now be described more specifically. As illustrated in Fig. 1, the similar data search device 2 is connected to a
20 data storage device 1, an input device 3, and an output device 4, constituting a similar data search system 30 with these devices.

[0024]

The data storage device 1 stores the search target data 10 comprised of sets of search targets as well as storing the element importance data 11
25 which is used for identifying importance levels that are pre-assigned to individual elements included in each set of search targets (refer to Fig. 3, which is described later).

[0025]

The input device 3 is used for inputting data, such as a set of search conditions and a similarity threshold value, to the similar data search device 2. Examples of the input device 3 may include a keyboard and other input apparatuses, a terminal device connected to the similar data search device 2 via a network, and the like.

[0026]

The output device 4 is a device to which search results are outputted. Examples of the output device 4 may include not only a display device and a printer but also a terminal device connected to the similar data search device 2 via a network. It should be noted that the input device 3 and the output device 4 may or may not be a single identical terminal device.

[0027]

According to the first exemplary embodiment, a "set" may be comprised of one or more elements and each of the elements may optionally have a pre-assigned importance level as described above (refer to Fig. 3, which is described later). A set may optionally be comprised of character tri-grams as elements, as described in NPTL 1.

[0028]

According to the first exemplary embodiment, a "similarity" between a set of search conditions and a set of search targets is calculated by, for example, solving a mathematical equation having D , Q , and $w(.)$, where D is a set of search targets, Q is a set of search conditions (search request), and $W(.)$ is a function which returns a weight of an importance level. For example, a similarity between sets is calculated by any one of: overlap (Q, D); overlap of Q from the viewpoint of D , overlap (D, Q); overlap of D from the viewpoint of Q , cosine (Q, D); cosine similarity, dice (Q, D); Dice's coefficient, and jaccard (Q, D); the Jaccard coefficient.

[0029]

Specifically, a similarity between sets can be calculated by using any one of the following mathematical expressions 1 to 5. It should be noted, however, that a similarity according to the present exemplary embodiment is not limited to those obtained by solving the following expressions. In the present exemplary embodiment, any similarity measure may be applied without particular limitation as far as it defines a condition by means of the size of a set.

[0030]

10 [Mathematical 1]

$$\text{overlap}(Q, D) = \frac{\sum_{e \in Q \cap D} w(e)}{\sum_{e \in D} w(e)} .$$

[0031]

[Mathematical 2]

$$\text{overlap}(D, Q) = \frac{\sum_{e \in Q \cap D} w(e)}{\sum_{e \in Q} w(e)}$$

15 [0032]

[Mathematical 3]

$$\text{cosine}(Q, D) = \frac{\sum_{e \in Q \cap D} w(e)}{\sqrt{\sum_{e \in Q} w(e)^2} \sqrt{\sum_{e \in D} w(e)^2}}$$

[0033]

[Mathematical 4]

$$dice(Q, D) = \frac{2 \sum_{e \in Q \cap D} w(e)}{\sum_{e \in Q} w(e) + \sum_{e \in D} w(e)}$$

[0034]

[Mathematical 5]

$$jaccard(Q, D) = \frac{\sum_{e \in Q \cap D} w(e)}{\sum_{e \in Q \cup D} w(e)}$$

5 [0035]

[Device Operations]

Now, operations of the similar data search device 2 according to the first exemplary embodiment of the present invention will be described with reference to Figs. 2 to 5. Fig. 2 is a flow diagram illustrating operations of the similar data search device according to the first exemplary embodiment of the present invention. The following description takes Fig. 1 into consideration as may be appropriate. In the first exemplary embodiment, a similar data search method is implemented by operating the similar data search device 2. Hence, the following description about operations of the similar data search device takes the place of description concerning the similar data search method according to the first exemplary embodiment.

[0036]

[Step A1]

20 As shown in Fig. 2, the inverted index generating unit 20 starts with reading, from the data storage device 1, the search target data 10 comprised of sets of search targets and the element importance data 11 representing importance levels of individual elements of the sets. Then, the inverted index generating unit 20 calculates the size of each of the sets

of search targets by using such data.

[0037]

Next, the inverted index generating unit 20 determines a size range of a set of search targets so that the number of sets of search targets included in each of inverted indexes to be generated is equal to or greater than a specified number. The inverted index generating unit 20 then generates inverted indexes by dividing the sets of search targets according to the determined size ranges (Step A1).

[0038]

If there exist a plurality of sets of search conditions, the inverted index generating unit 20 carries out a test search under each set of search conditions. This allows the inverted index generating unit 20 to determine a minimum number of sets of search targets included in each of inverted indexes to be generated so as to minimize the sum of search times required by the data search unit 22.

[0039]

The following provides a detail description about how to calculate the size of a set. Assuming that X denotes the set whose size is to be calculated, the size of the set X is defined by the following mathematical expression 6 if the above-described two different overlaps, the Dice's coefficient, or the Jaccard coefficient is used as a similarity measure. If cosine similarity is used as a similarity measure, the size of the set X is defined by the mathematical expression 7 below.

[0040]

[Mathematical 6]

$$\sum_{e \in X} w(e)$$

[0041]

[Mathematical 7]

$$\sqrt{\sum_{e \in X} w(e)^2}$$

[0042]

In addition, in Step A1, the inverted index generating unit 20 can
 5 calculate the size of each set of search targets by using importance levels
 that are pre-assigned to individual elements included in a set of search
 targets. Every element may have an importance level of 1; in this case
 the size of the set coincides with the number of elements in the set.

[0043]

10 In contrast, if more specific importance levels are assigned to
 elements of a set, the number of sets of the same size is decreased. Thus,
 the above-described effect of the first exemplary embodiment can be
 provided to a greater extent. Accordingly, in the first exemplary
 embodiment, it is preferable that importance levels are assigned as
 15 specifically as possible.

[0044]

In addition, in Step A1, the inverted index generating unit 20 can
 specify a threshold value for a size range so that every inverted index has
 at least a certain number of sets, and calculate a specified number by
 20 dividing the total number of sets of search targets by a specified value.
 Then, the inverted index generating unit 20 can determine the size range of
 sets of search targets for each of inverted indexes to be generated, based on
 the calculated specified number. In other words, the inverted index
 generating unit 20 may determine sizes by dividing the total number of sets
 25 of search targets by the specified number N to evenly divide the total sets
 into N groups.

[0045]

The threshold for a size range and the specified number N can be determined by actually performing searches with a sample set of candidate search conditions. In this case, it is preferable to determine these values so that the calculation time can be minimized.

5 [0046]

Furthermore, if it is possible to determine criteria for the number of sets that can be retrieved from each inverted index, inverted indexes can be generated by calculating the size of each set of search targets, sorting the sets in ascending order of size, and adding the sets to every inverted index starting from the smallest set in size until a predetermined condition is satisfied.

[0047]

Now a specific example of Step A1 is described with reference to Fig. 3. Fig. 3 is an explanatory diagram illustrating an example of Step A1 using the Jaccard coefficient as a similarity measure. In the example in Fig. 3, the specified number N is set to 50.

15 [0048]

As shown in Fig. 3, each set of search target data is assigned an identifier SID. The size of each set is calculated according to the above mathematical expression 6 because the Jaccard coefficient is used as a similarity measure. For example, the size of the set $SID = 1$ is 6.8.

20 [0049]

As the example in Fig. 3 contains 10,000 sets of search target data, it can be seen that, in order to divide the search target data into 50 groups according to the number of inverted indexes, these inverted indexes will be created so that about 200 sets belong to an inverted index.

25 [0050]

Thus, the inverted index generating unit 20 sorts the individual sets

of search targets in ascending order of size and adds the sets to each inverted index until the number of sets exceeds 200. If there exists another set of the same size as the 200th set, the inverted index generating unit 20 adds such another set to the inverted index where the 200th set is added. In this case, the inverted index generating unit 20 does not add any set to a new inverted index until it encounters a set of a different size. [0051]

Next, the inverted index generating unit 20 identifies a minimum size value β of a set that can be retrieved from each of the inverted indexes. The inverted index generating unit 20 then assigns IDs of inverted indexes to the individual identified values of β in their ascending order.

Assuming that i denotes an ID and D_i denotes a set included in an inverted index of each ID, the relationship between the size range for each inverted index and the size of a set is expressed by the following mathematical expression 8.

[0052]

[Mathematical 8]

$$\beta_i \leq |D_i| < \beta_{i+1}$$

[0053]

According to the above mathematical expression 8, when the inverted index ID = 3 shown in Fig. 3 is used as search targets, any set whose size is equal to or greater than 6.0 and less than 8.0 can be retrieved. Hence, the inverted index of ID = 3 includes the set of SID = 1 whose size is 6.8.

[0054]

[Step A2]

After Step A1, the unnecessary inverted index identifying unit 21

determines a condition of the size of a set of search targets necessary for a similarity to be no smaller than a threshold value, according to a mathematical expression defined for each similarity measure, by using the size of a set of search conditions and the threshold value specified for a similarity.

[0055]

Next, from among the inverted indexes, the unnecessary inverted index identifying unit 21 identifies any inverted index other than those inverted indexes containing a set whose minimum size value satisfies the condition, that is, any inverted index having a similarity that can never be equal or greater than the threshold value, as an inverted index unnecessary for searches (Step A2).

[0056]

Now Step A2 will be described in more detail with reference to Fig. 4. Fig. 4 is a diagram showing example mathematical expressions for obtaining a size range of a set. That is, Fig. 4 shows mathematical expressions for determining a size range of a set that ensures a similarity is no smaller than α , with given Q and α where Q is a set of search conditions and α is a threshold value.

[0057]

Proofs of the mathematical expressions shown in Fig. 4 except overlaps are the same as those disclosed in NPTL 1 for the cases where no rounding up or down to an integer is involved, and thus description of the proofs is omitted in the present first exemplary embodiment. Proofs of overlap measures, $\text{Overlap}(Q, D)$ and $\text{Overlap}(D, Q)$, are as follows.

[0058]

From the definition $\text{Overlap}(Q, D)$ being no smaller than α is expressed by the following mathematical expression 9.

[0059]

[Mathematical 9]

$$\frac{\sum_{e \in Q \cap D} w(e)}{\sum_{e \in D} w(e)} \geq \alpha$$

[0060]

5 The above mathematical expression 9 is transformed into the following mathematical expression 10. A maximum value of $|D|$ in the mathematical expression 10 is expressed by the mathematical expression 11 below.

[0061]

10 [Mathematical 10]

$$\frac{|Q|}{\alpha} \geq \frac{\sum_{e \in Q} w(e)}{\alpha} \geq \frac{\sum_{e \in Q \cap D} w(e)}{\alpha} \geq \sum_{e \in D} w(e) = |D|$$

[0062]

[Mathematical 11]

$$\frac{|Q|}{\alpha} \geq |D|$$

15 [0063]

Similarly, Overlap (D, Q) is expressed by the following mathematical expression 12 from the definition.

[0064]

[Mathematical 12]

$$\frac{\sum_{e \in Q \cap D} w(e)}{\sum_{e \in Q} w(e)} \geq \alpha$$

[0065]

The above mathematical expression 12 is transformed into the following mathematical expression 13. A minimum value of $|D|$ in the mathematical expression 13 is expressed by the mathematical expression 14 below.

[0066]

[Mathematical 13]

$$|D| = \sum_{e \in D} w(e) \geq \sum_{e \in Q \cap D} w(e) \geq \alpha \sum_{e \in Q} w(e) = \alpha |Q|$$

10 [0067]

[Mathematical 14]

$$|D| \geq \alpha |Q|$$

[0068]

For example, when Q , $|Q|$, and α are given where Q is a set of search conditions, $|Q|$ is the size of the set, and α is a threshold, to ensure that the similarity is not smaller than the threshold α , the size of the set of search targets D needs to be equal to or greater than $\alpha|Q|$ and equal to or less than $|Q| / \alpha$, because the present example uses the Jaccard coefficient. Specifically, if the set of search conditions Q are comprised of elements e and f and the threshold α is 0.6, then $|Q|$ is 2.2 and thus the minimum and maximum values of the size are 1.32 ($= 2.2 \times 0.6$) and 3.667 ($\approx 2.2 / 0.6$), respectively.

20 [0069]

Now, referring to lower limits β for inverted indexes listed in Fig. 3, the size of a set contained in either of the inverted index ID 1 and ID 2 is

25

represented by $\beta_1 (= 0.5) \leq |D| < \beta_3 (= 6.0)$, which already includes the minimum and maximum values. This indicates that ID 3 and subsequent inverted indexes are unnecessary for searches. In this way, the unnecessary inverted index identifying unit 21 identifies any inverted
 5 index unnecessary for searches by using a minimum size value for each of inverted indexes.

[0070]

[Step A3]

Finally, with respect to the inverted indexes other than any
 10 identified inverted index unnecessary for searches, the data search unit 22 calculates a similarity between the set of search conditions and the individual sets that include applicable elements, and then outputs, as a search result, any set whose similarity is not smaller than the threshold value, to the output device 4 (Step A3).

15 [0071]

For example, given that the above-described set of search conditions Q includes elements e and f , the data search unit 22 retrieves, for example, $SID = 3$ from the inverted index whose ID is 1 shown in Fig. 3. The data search unit 22 also retrieves, for example, $SID = 10000$ from the
 20 inverted index whose ID is 2. Then, the data search unit 22 can calculate a similarity between the individual sets as obtained above and the set of search conditions to output, as a search result, the data whose similarity is actually not smaller than the threshold α .

[0072]

25 In Step A3, the data search unit 22 may also handle searches as the τ -overlap problem, similarly to NPTL 1. That is, the data search unit 22 identifies any element common to those elements in the set of search conditions for each of the sets included in an inverted index (hereinafter

denoted as a "non-identified inverted index") other than any identified inverted index, and then calculates the sum of importance levels of the identified elements. If the calculated sum satisfies the condition for being equivalent to the case where a similarity is not smaller than a threshold α , the data search unit 22 presents, as a search result, the set included in the non-identified inverted index subjected to the calculation. [0073]

Specifically, when $|D|$, $|Q|$, and α are given where $|D|$ is the size of a set of search targets, $|Q|$ is the size of a set of search conditions, and α is a threshold value, the case where the sum of importance levels of elements common to sets Q and D is equal to or greater than τ , which is calculated according to any of the expressions listed in Fig. 5, is deemed to be equivalent to a similarity between the sets D and Q being no smaller than α . In this case, calculation of the size $|D|$ of a set of search targets need only be performed once at the time of generating an inverted index; moreover, calculation of the size $|Q|$ of a set of search conditions need only be performed once as well, which eliminates the need for calculating similarities every time, thus increasing efficiency of calculation. Fig. 5 is a diagram showing example conditions for being equivalent to the case where a similarity is not smaller than a threshold value.

[0074]

Furthermore, in the present exemplary embodiment, the data search unit 22 can check every element in a set of search conditions, on a one-by-one basis, against each of the sets included in a non-identified inverted index in sequence. In this case, if the sum of importance levels of the unchecked elements becomes neither equal to nor greater than τ (that is, less than τ), the data search unit 22 carries out the checking by using the unchecked elements only against the sets that have already been

checked by that time, from among the sets included in an inverted index. The data search unit 22 then calculates the sum of importance levels of the common elements with regard to only the sets that have been checked by that time.

5 [0075]

In other words, the first exemplary embodiment can optionally utilize the property as disclosed in NPTL 1: when the sum of unsearched elements in a set of search conditions Q becomes less than τ , the sum of importance levels of common elements in both a set of any SID that is subsequently to be first retrieved and the set Q becomes to be equal to or greater than τ .

[0076]

Specifically, the data search unit 22 considers that a minimum size value β of a set included in each inverted index is $|D|$ and calculates τ so as to satisfy a minimum requirement with respect to a set in the inverted index. Once the sum of unsearched elements in the set of search conditions Q is less than τ , the data search unit 22, presuming that the only SIDs that have already been searched by that time are candidates, checks for any remaining unsearched element by performing a binary search on a list of elements obtained from the inverted index for each SID. While a linear search has computational complexity of $O(n)$, a binary search for checking existence has computational complexity of $O(\log n)$, where n is the number of sets that contain elements, which means the efficiency can be improved.

25 [0077]

It should be noted that after the switching to the binary search, the size of each set SID is used for τ on each set. Additionally, to efficiently determine that the sum of the unsearched elements in the set of search

conditions Q is less than τ , the data search unit 22 preferably searches (checks) elements in descending order of importance level.

[0078]

(Effect of First Exemplary Embodiment)

5 As described above, in the first exemplary embodiment, the inverted index generating unit 20 generates each inverted index so that the number of sets of search targets is not reduced. The unnecessary inverted index identifying unit 21 then identifies, based on search conditions and a minimum size value of a set in each of inverted indexes, any unnecessary
10 inverted index for finding a set having a similarity no smaller than a threshold value. Next, the data search unit 22 performs a search on inverted indexes other than the unnecessary inverted index(es). Consequently, according to the first exemplary embodiment, it is made possible to find all the sets having similarities no smaller than a threshold
15 value efficiently due to the fact that the number of references to inverted indexes is decreased on the whole, even when there are a small number of sets of the same size as that of the set of search conditions.

[0079]

[Program]

20 A program according to the present exemplary embodiment may be any program causing a computer to execute Steps A1 to A3 shown in Fig. 2. The similar data search device 30 and the similar data search method according to the present exemplary embodiment can be implemented by installing the program on a computer and executing it. In this case, a
25 central processing unit (CPU) in the computer acts as the inverted index generating unit 20, the unnecessary inverted index identifying unit 21, and the data search unit 22 to handle processes.

[0080]

(Second Exemplary Embodiment)

Now a similar data search device, a similar data search method, and a program for similar data searches according to a second exemplary embodiment of the present invention will be described below with

5 reference to Figs. 6 and 7.

[0081]

[Device Configuration]

To begin with, a configuration of the similar data search device according to the second exemplary embodiment is described with reference
10 to Fig. 6. Fig. 6 is a block diagram illustrating the configuration of the similar data search device according to the second exemplary embodiment of the present invention.

[0082]

As shown in Fig. 6, the similar data search device 5 according to
15 the second exemplary embodiment is different from the similar data search device 2 of the first exemplary embodiment illustrated in Fig. 1 in that: first, the similar data search device 5 of the second exemplary embodiment includes a synonymous element converting unit 23, in addition to an inverted index generating unit 20, an unnecessary inverted index
20 identifying unit 21, and a data search unit 22. The synonymous element converting unit 23 converts, among elements included in a set of search targets or search conditions, any element belonging to a set of determined synonymous elements into a representative element of the synonymous elements.

25 [0083]

Additionally, in the second exemplary embodiment, the similar data search device 5 utilizes synonymous element data 12, in addition to search target data 10 and element importance data 11. The synonymous element

data 12 is the data for defining apparently synonymous elements, being stored in the data storage device 1 along with the search target data 10 and the element importance data 11.

[0084]

5 Specifically, the synonymous element converting unit 23 reads the search target data 10, the element importance data 11, and the synonymous element data 12 to generate a set of synonymous elements. Then, with respect to each of the sets of search targets and search conditions, the synonymous element converting unit 23 replaces elements belonging to a
10 set of synonymous elements with the representative element of the set of synonymous elements and outputs any set that has been subjected to the replacement to the inverted index generating unit 20.

[0085]

[Operation of the Device]

15 Now, operations of the similar data search device 5 according to the second exemplary embodiment of the present invention will be described with reference to Fig. 7. Fig. 7 is a flow diagram illustrating operations of the similar data search device according to the second exemplary embodiment of the present invention. The following description takes
20 Fig. 6 into consideration as may be appropriate. In the second exemplary embodiment, a similar data search method is implemented by operating the similar data search device 5. Hence, the following description about operations of the similar data search device takes the place of description concerning the similar data search method according to the second
25 exemplary embodiment.

[0086]

[Step B1]

As shown in Fig. 7, the synonymous element converting unit 23

starts with reading the synonymous element data and the element importance data to create a set of synonymous elements, with respect to each set of search targets and each set of search conditions.

[0087]

5 Next, the synonymous element converting unit 23 selects a representative element of each set of synonymous elements and replaces the elements belonging to a set of synonymous elements with the representative element, with respect to each set of search targets and each set of search conditions (Step B1).

10 [0088]

In Step B1, a set of synonymous elements is created by drawing an undirected edge between nodes, which are regarded as a pair of elements apparently being synonymous, and by interpreting the nodes all along link components outgoing from an element as synonymous elements.

15 [0089]

The representative element may be selected from an element of the highest importance, an element of the lowest importance, an element of the median importance, a first element in the case of totally ordered elements, and the like. It should be noted that no particular limitation is imposed
20 on the method of selecting a representative element.

[0090]

[Steps B2 to B4]

Next, Steps B2 to B4 are carried out through the use of a set of search targets where elements have been converted to a representative
25 element as well as a set of search conditions where elements have been converted to a representative element. Steps B2 to B4 are identical to Steps A1 to A3 in Fig. 2, respectively, and thus they are carried out as illustrated in Fig. 2 and finally a search result is outputted.

[0091]

(Effect of Second Exemplary Embodiment)

As described above, in the second exemplary embodiment, the synonymous element converting unit 23 replaces synonymous elements with a representative element prior to the search processing. Similar data searches are thus conducted by equating different but synonymous elements with one element, achieving searches of higher accuracy.

[0092]

(Computer)

The following describes a computer which implements the similar data search device according to either of the first and second exemplary embodiments by executing a program, referring to Fig. 8. Fig. 8 is a block diagram illustrating an example computer implementing the similar data search device according to either of the first and second exemplary embodiments of the present invention.

[0093]

As illustrated in Fig. 8, a computer 110 includes a CPU 111, main memory 112, a storage device 113, an input interface 114, a display controller 115, a data reader/writer 116, and a communication interface 117. These individual units are connected to a bus 121 so that data can be communicated with one another via the bus.

[0094]

The CPU 111 performs various computations by deploying programs (code) according to an exemplar embodiment of the present invention stored in the storage device 113 into the main memory 112 and by executing these programs in a predetermined order. The main memory 112 is typically a volatile storage device such as dynamic random access memory (DRAM). The programs according to the present exemplary

embodiment are provided in the state where they are contained in a computer-readable recording medium 120. The programs according to the present exemplary embodiment may optionally be distributed on the Internet which is connected via the communication interface 117.

5 [0095]

Specific examples of the storage device 113 may include a semiconductor storage device, such as flash memory, in addition to a hard disk drive. The input interface 114 provides an interface for data transmission between the CPU 111 and an input apparatus 118 such as a keyboard or mouse. The display controller 115, which is connected to a display device 119, controls display on the display device 119.

[0096]

The data reader/writer 116 provides an interface for data transmission between the CPU 111 and the recording medium 120, reads programs out of the recording medium 120, and writes results of processing carried out in the computer 110 into the recording medium 120. The communication interface 117 provides an interface for data transmission between the CPU 111 and another computer.

[0097]

20 Specific examples of the recording medium 120 may include a general-purpose semiconductor storage device such as CompactFlash® (CF) and Secure Digital (SD), a magnetic storage medium such as a flexible disk, and an optical storage medium such as Compact Disk Read-Only Memory (CD-ROM).

25 [0098]

The whole or part of the above-described exemplary embodiments can be described as, but is not limited to, the following Supplementary Notes 1 to 30.

[0099]

(Supplementary Note 1)

A similar data search device for conducting a search by using sets as search target data and search condition data, the device comprising:

5 an inverted index generating unit which, for generating inverted indexes used for a search, determines size ranges of sets of search targets for each of the inverted indexes to be generated so that at least a specified number of sets of search targets are included in each of the inverted indexes to be generated, and generates the inverted indexes by dividing the
10 sets of search targets according to the determined size ranges;

 an unnecessary inverted index identifying unit which determines, based on a size of a set of search conditions and a specified threshold value of a similarity between the set of search conditions and the set of search targets, a condition of a size of the set of search targets necessary for the
15 similarity to be no smaller than the threshold value, and identifies, as an inverted index unnecessary for searches, from among the inverted indexes, any inverted index other than those inverted indexes which include a set whose minimum size value satisfies the condition; and

 a data search unit which conducts a search by applying the set of
20 search conditions to an inverted index other than the identified inverted index unnecessary for searches.

[0100]

(Supplementary Note 2)

 The similar data search device according to Supplementary Note 1,
25 wherein the inverted index generating unit calculates the specified number by dividing a total number of sets of the search targets by a specified value and determines, based on the specified number as calculated, the size ranges of sets of the search targets for each of the inverted indexes to be

generated.

[0101]

(Supplementary Note 3)

The similar data search device according to Supplementary Note 1
or 2, wherein the inverted index generating unit determines, if a plurality
of sets of the search conditions exist, a minimum number of sets of search
targets included in each of the inverted indexes to be generated so as to
minimize the sum of search times required for a search conducted by the
data search unit.

[0102]

(Supplementary Note 4)

The similar data search device according to any one of
Supplementary Notes 1 to 3,

wherein the unnecessary inverted index identifying unit
calculates the condition by using a mathematical expression and the
threshold value, the mathematical expression being defined by any one of
an overlap of a set of the search conditions with a set of the search targets,
an overlap of a set of the search targets with a set of the search conditions,
the Jaccard coefficient, Dice's coefficient, and cosine similarity.

[0103]

(Supplementary Note 5)

The similar data search device according to any one of
Supplementary Notes 1 to 4,

wherein the data search unit identifies, from among sets that are
included in an inverted index other than the identified inverted index
unnecessary for searches, any set that includes an element of a set of the
search conditions, and presents, as a search result, the identified set if the
similarity between the identified set and the set of the search conditions is

not smaller than the threshold value.

[0104]

(Supplementary Note 6)

The similar data search device according to any one of

5 Supplementary Notes 1 to 5,

wherein the inverted index generating unit further calculates the size of each set of the search targets by using importance levels that are pre-assigned to individual elements included in a set of the search targets.

[0105]

10 (Supplementary Note 7)

The similar data search device according to Supplementary Note 6,

wherein the similarity is calculated by using a mathematical expression defined by any one of an overlap of a set of the search conditions with a set of the search targets, an overlap of a set of the search targets with a set of the search conditions, the Jaccard coefficient, Dice's

15 coefficient, and cosine similarity,

and wherein the data search unit identifies, for each set included in an inverted index other than the identified inverted index unnecessary for searches, any elements common to elements of a set of the search

20 conditions, calculates the sum of the importance levels of the identified elements, and, if the calculated sum satisfies a condition for being equivalent to a case where the similarity is not smaller than the threshold value, presents the set subjected to calculation as a search result.

[0106]

25 (Supplementary Note 8)

The similar data search device according to Supplementary Note 7,

wherein the data search unit

checks every element in a set of the search conditions, on a

one-by-one basis, against each set included in an inverted index other than the identified inverted index unnecessary for searches in sequence;

if the sum of importance levels of unchecked elements no longer satisfies the condition, carries out checking by using the unchecked

5 elements only against sets that have already been checked by that time from among the sets included in the inverted index; and

calculates the sum of the importance levels of the common elements with regard to only the sets that have been checked by that time.

[0107]

10 (Supplementary Note 9)

The similar data search device according to Supplementary Note 8, wherein the data search unit checks elements of a set of the search conditions in descending order of importance level.

[0108]

15 (Supplementary Note 10)

The similar data search device according to any one of Supplementary Notes 1 to 9, further comprising:

a synonymous element converting unit which converts, among elements included in a set of the search targets and a set of the search conditions, any element belonging to a set of determined synonymous elements into a representative element of the synonymous elements.

20

[0109]

(Supplementary Note 11)

A method for conducting a search by using sets as search target data and search condition data, the method comprising the steps of:

25

(a) for generating inverted indexes used for a search, determining size ranges of sets of search targets for each of inverted indexes to be generated so that at least a specified number of sets of search targets are

included in each of the inverted index to be generated, and generating the inverted indexes by dividing the sets of search targets according to the determined size ranges;

(b) determining, based on a size of a set of search conditions and a
5 specified threshold value of a similarity between the set of search
conditions and the set of search targets, a condition of a size of the set of
search targets necessary for the similarity to be no smaller than the
threshold value, and identifying, as an inverted index unnecessary for
searches, from among the inverted indexes, any inverted index other than
10 those inverted indexes that include a set whose minimum size value
satisfies the condition; and

(c) conducting a search by applying the set of search conditions to
an inverted index other than the inverted index unnecessary for searches as
identified in the step (b).

15 [0110]

(Supplementary Note 12)

The similar data search method according to Supplementary Note
11, wherein, in the step (a), the specified number is calculated by dividing
a total number of sets of the search targets by a specified value and the size
20 ranges of sets of the search targets are determined for each of the inverted
indexes to be generated, based on the specified number as calculated.

[0111]

(Supplementary Note 13)

The similar data search method according to Supplementary Note
25 11 or 12, wherein, in the step (a), if a plurality of sets of the search
conditions exist, a minimum number of sets of search targets included in
each of the inverted indexes to be generated is determined so as to
minimize the sum of search times required for a search in the step (c).

[0112]

(Supplementary Note 14)

The similar data search method according to any one of
Supplementary Notes 11 to 13,

5 wherein, in the step (b), the condition is calculated by using a
mathematical expression and the threshold value, the mathematical
expression being defined by any one of an overlap of a set of the search
conditions with a set of the search targets, an overlap of a set of the search
targets with a set of the search conditions, the Jaccard coefficient, Dice's
10 coefficient, and cosine similarity.

[0113]

(Supplementary Note 15)

The similar data search method according to any one of
Supplementary Notes 11 to 14,

15 wherein, in the step (c), from among sets that are included in an
inverted index other than the inverted index unnecessary for searches as
identified in the step (b), any set that includes an element of a set of the
search conditions is identified, and the identified set is presented as a
search result if the similarity between the identified set and the set of the
20 search conditions is not smaller than the threshold value.

[0114]

(Supplementary Note 16)

The similar data search method according to any one of
Supplementary Notes 11 to 15,

25 wherein, additionally in the step (a), the size of each set of the
search targets is calculated by using importance levels that are
pre-assigned to individual elements included in a set of the search targets.

[0115]

(Supplementary Note 17)

The similar data search method according to Supplementary Note 16,

wherein the similarity is calculated by using a mathematical expression defined by any one of an overlap of a set of the search conditions with a set of the search targets, an overlap of a set of the search targets with a set of the search conditions, the Jaccard coefficient, Dice's coefficient, and cosine similarity,

and wherein, in the step (c), for each set included in an inverted index other than the inverted index unnecessary for searches as identified in the step (b), any elements common to elements of a set of the search conditions are identified, and the sum of the importance levels of the identified elements is calculated, and, if the calculated sum satisfies a condition for being equivalent to a case where the similarity is not smaller than the threshold value, the set subjected to calculation is presented as a search result.

[0116]

(Supplementary Note 18)

The similar data search method according to Supplementary Note 17,

wherein, in the step (c), every element in a set of the search conditions is checked, on a one-by-one basis, against each set included in an inverted index other than the identified inverted index unnecessary for searches in sequence;

if the sum of importance levels of unchecked elements no longer satisfies the condition, checking is carried out by using the unchecked elements only against sets that have already been checked by that time from among the sets included in the inverted index; and

the sum of the importance levels of the common elements is calculated with regard to only the sets that have been checked by that time.
[0117]

(Supplementary Note 19)

5 The similar data search method according to Supplementary Note 18, wherein, in the step (c), elements of a set of the search conditions are checked in descending order of importance level.

[0118]

(Supplementary Note 20)

10 The similar data search method according to any one of Supplementary Notes 11 to 19, further comprising the step of:

(d) converting, among elements included in a set of the search targets and a set of the search conditions, any element belonging to a set of determined synonymous elements into a representative element of the
15 synonymous elements.

[0119]

(Supplementary Note 21)

A computer-readable recording medium which records a program for conducting a search with a computer by using sets as search target data
20 and search condition data,

wherein the program comprises instructions causing the computer to execute the steps of:

(a) for generating inverted indexes used for a search, determining size ranges of sets of search targets for each of inverted indexes to be
25 generated so that at least a specified number of sets of search targets are included in each of the inverted index to be generated, and generating the inverted indexes by dividing the sets of search targets according to the determined size ranges;

(b) determining, based on a size of a set of search conditions and a specified threshold value of a similarity between the set of search conditions and the set of search targets, a condition of a size of the set of search targets necessary for the similarity to be no smaller than the threshold value, and identifying, as an inverted index unnecessary for searches, from among the inverted indexes, any inverted index other than those inverted indexes that include a set whose minimum size value satisfies the condition; and

(c) conducting a search by applying the set of search conditions to an inverted index other than the inverted index unnecessary for searches as identified in the step (b).

[0120]

(Supplementary Note 22)

The computer-readable recording medium according to Supplementary Note 21, wherein, in the step (a), the specified number is calculated by dividing a total number of sets of the search targets by a specified value and the size ranges of sets of the search targets are determined for each of the inverted indexes to be generated, based on the specified number as calculated.

[0121]

(Supplementary Note 23)

The computer-readable recording medium according to Supplementary Note 21 or 22, wherein, in the step (a), if a plurality of sets of the search conditions exist, a minimum number of sets of search targets included in each of the inverted indexes to be generated is determined so as to minimize the sum of search times required for a search in the step (c).

[0122]

(Supplementary Note 24)

The computer-readable recording medium according to any one of Supplementary Notes 21 to 23,

wherein, in the step (b), the condition is calculated by using a mathematical expression and the threshold value, the mathematical
5 expression being defined by any one of an overlap of a set of the search conditions with a set of the search targets, an overlap of a set of the search targets with a set of the search conditions, the Jaccard coefficient, Dice's coefficient, and cosine similarity.

[0123]

10 (Supplementary Note 25)

The computer-readable recording medium according to any one of Supplementary Notes 21 to 24,

wherein, in the step (c), from among sets that are included in an inverted index other than the inverted index unnecessary for searches as
15 identified in the step (b), any set that includes an element of a set of the search conditions is identified, and the identified set is presented as a search result if the similarity between the identified set and the set of the search conditions is not smaller than the threshold value.

[0124]

20 (Supplementary Note 26)

The computer-readable recording medium according to any one of Supplementary Notes 21 to 25,

wherein, additionally in the step (a), the size of each set of the search targets is calculated by using importance levels that are
25 pre-assigned to individual elements included in a set of the search targets.

[0125]

(Supplementary Note 27)

The computer-readable recording medium according to

Supplementary Notes 26,

wherein the similarity is calculated by using a mathematical expression defined by any one of an overlap of a set of the search conditions with a set of the search targets, an overlap of a set of the search targets with a set of the search conditions, the Jaccard coefficient, Dice's coefficient, and cosine similarity,

and wherein, in the step (c), for each set included in an inverted index other than the inverted index unnecessary for searches as identified in the step (b), any elements common to elements of a set of the search conditions are identified, and the sum of the importance levels of the identified elements is calculated, and, if the calculated sum satisfies a condition for being equivalent to a case where the similarity is not smaller than the threshold value, the set subjected to calculation is presented as a search result.

[0126]

(Supplementary Note 28)

The computer-readable recording medium according to Supplementary Notes 27,

wherein, in the step (c), every element in a set of the search conditions is checked, on a one-by-one basis, against each set included in an inverted index other than the identified inverted index unnecessary for searches in sequence;

if the sum of importance levels of unchecked elements no longer satisfies the condition, checking is carried out by using the unchecked elements only against sets that have already been checked by that time from among the sets included in the inverted index; and

the sum of the importance levels of the common elements is calculated with regard to only the sets that have been checked by that time.

[0127]

(Supplementary Note 29)

The computer-readable recording medium according to
Supplementary Note 28, wherein, in the step (c), elements of a set of the
5 search conditions are checked in descending order of importance level.

[0128]

(Supplementary Note 30)

The computer-readable recording medium according to any one of
Supplementary Notes 21 to 29,

10 wherein the program further comprises an instruction causing the
computer to execute the step of:

(d) converting, among elements included in a set of the search
targets and a set of the search conditions, any element belonging to a set of
determined synonymous elements into a representative element of the
15 synonymous elements.

[0129]

The present invention has been described with reference to
exemplary embodiments, but the invention is not limited to these
embodiments. Various modification of the present invention that could
20 be understood by those skilled in the art may be made to configurations or
details of the invention within the scope of the invention.

[0130]

The present application claims priority based on Japanese Patent
Application No. 2013-045566 filed on March 7, 2013, the entire disclosure
25 of which is herein incorporated.

[Advantageous Effects of Invention]

[0131]

As described above, according to the present invention, it is made

possible to suppress lowering the retrieval effectiveness caused by an increased number of searches in inverted indexes even when the search target data contains a small number of sets of the same size. The present invention is particularly useful for data clustering systems which handle
 5 matching redundant data to delete redundant data and grouping similar data, systems which handle dictionary soft matching through soft matching with dictionary entries, and the like.

[Reference signs List]

[0132]

- | | | |
|----|-----|--|
| 10 | 1 | Data storage device |
| | 2 | Similar data search device (first exemplary embodiment) |
| | 3 | Input device |
| | 4 | Output device |
| | 5 | Similar data search device (second exemplary embodiment) |
| 15 | 10 | Search target data |
| | 11 | Element importance data |
| | 12 | Synonymous element data |
| | 20 | Inverted index generating unit |
| | 21 | Unnecessary inverted index for search identifying unit |
| 20 | 22 | Data search unit |
| | 23 | Synonymous element converting unit |
| | 30 | Similar data search system |
| | 110 | Computer |
| | 111 | CPU |
| 25 | 112 | Main memory |
| | 113 | Storage device |
| | 114 | Input interface |
| | 115 | Display controller |

	116	Data reader/writer
	117	Communication interface
	118	Input apparatus
	119	Display device
5	120	Recording medium
	121	Bus

CLAIMS

[Claim 1]

A similar data search device for conducting a search by using sets
5 as search target data and search condition data, the device comprising:

an inverted index generating unit which, for generating inverted
indexes used for a search, determines size ranges of sets of search targets
for each of the inverted indexes to be generated so that at least a specified
number of sets of search targets are included in each of the inverted
10 indexes to be generated, and generates the inverted indexes by dividing the
sets of search targets according to the determined size ranges;

an unnecessary inverted index identifying unit which determines,
based on a size of a set of search conditions and a specified threshold value
of a similarity between the set of search conditions and the set of search
15 targets, a condition of a size of the set of search targets necessary for the
similarity to be no smaller than the threshold value, and identifies, as an
inverted index unnecessary for searches, from among the inverted indexes,
any inverted index other than those inverted indexes which include a set
whose minimum size value satisfies the condition; and

20 a data search unit which conducts a search by applying the set of
search conditions to an inverted index other than the identified inverted
index unnecessary for searches.

[Claim 2]

The similar data search device according to claim 1, wherein the
25 inverted index generating unit calculates a specified number by dividing a
total number of sets of the search targets by a specified value and
determines, based on the specified number as calculated, the size ranges of
sets of the search targets for each of the inverted indexes to be generated.

[Claim 3]

The similar data search device according to claim 1 or 2, wherein the inverted index generating unit determines, if a plurality of sets of the search conditions exist, a minimum number of sets of search targets
5 included in each of the inverted indexes to be generated so as to minimize the sum of search times required for a search conducted by the data search unit.

[Claim 4]

The similar data search device according to any one of claims 1 to
10 3,

wherein the unnecessary inverted index identifying unit calculates the condition by using a mathematical expression and the threshold value, the mathematical expression being defined by any one of an overlap of a set of the search conditions with a set of the search targets,
15 an overlap of a set of the search targets with a set of the search conditions, the Jaccard coefficient, Dice's coefficient, and cosine similarity.

[Claim 5]

The similar data search device according to any one of claims 1 to
4,

20 wherein the data search unit identifies, from among sets that are included in an inverted index other than the identified inverted index unnecessary for searches, any set that includes an element of a set of the search conditions, and presents, as a search result, the identified set if the similarity between the identified set and the set of the search conditions is
25 not smaller than the threshold value.

[Claim 6]

The similar data search device according to any one of claims 1 to
5,

wherein the inverted index generating unit further calculates the size of each set of the search targets by using importance levels that are pre-assigned to individual elements included in a set of the search targets.
[Claim 7]

5 The similar data search device according to claims 6,
 wherein the similarity is calculated by using a mathematical expression defined by any one of an overlap of a set of the search conditions with a set of the search targets, an overlap of a set of the search targets with a set of the search conditions, the Jaccard coefficient, Dice's
10 coefficient, and cosine similarity,

 and wherein the data search unit
 checks every element in a set of the search conditions, on a one-by-one basis, against each set included in an inverted index other than the identified inverted index unnecessary for searches in sequence in
15 descending order of the importance level;

 if the sum of importance levels of unchecked elements no longer satisfies the condition, carries out checking by using the unchecked elements only against sets that have already been checked by that time from among the sets included in the inverted index;

20 calculates the sum of the importance levels of the elements common to elements of a set of the search conditions with regard to only the sets that have been checked by that time; and

 if the calculated sum satisfies a condition for being equivalent to a case where the similarity is not smaller than the threshold value, presents
25 the set subjected to calculation as a search result.

[Claim 8]

 The similar data search device according to any one of claims 1 to 7, further comprising:

a synonymous element converting unit which converts, among elements included in a set of the search targets and a set of the search conditions, any element belonging to a set of determined synonymous elements into a representative element of the synonymous elements.

5 [Claim 9]

A method for conducting a search by using sets as search target data and search condition data, the method comprising the steps of:

(a) for generating inverted indexes used for a search, determining size ranges of sets of search targets for each of inverted indexes to be
10 generated so that at least a specified number of sets of search targets are included in each of the inverted index to be generated, and generating the inverted indexes by dividing the sets of search targets according to the determined size ranges;

(b) determining, based on a size of a set of search conditions and a
15 specified threshold value of a similarity between the set of search conditions and the set of search targets, a condition of a size of the set of search targets necessary for the similarity to be no smaller than the threshold value, and identifying, as an inverted index unnecessary for searches, from among the inverted indexes, any inverted index other than
20 those inverted indexes that include a set whose minimum size value satisfies the condition; and

(c) conducting a search by applying the set of search conditions to an inverted index other than the inverted index unnecessary for searches as identified in the step (b).

25 [Claim 10]

A computer-readable recording medium which records a program for conducting a search with a computer by using sets as search target data and search condition data,

wherein the program comprises instructions causing the computer to execute the steps of:

(a) for generating inverted indexes used for a search, determining size ranges of sets of search targets for each of inverted indexes to be generated so that at least a specified number of sets of search targets are included in each of the inverted index to be generated, and generating the inverted indexes by dividing the sets of search targets according to the determined size ranges;

(b) determining, based on a size of a set of search conditions and a specified threshold value of a similarity between the set of search conditions and the set of search targets, a condition of a size of the set of search targets necessary for the similarity to be no smaller than the threshold value, and identifying, as an inverted index unnecessary for searches, from among the inverted indexes, any inverted index other than those inverted indexes that include a set whose minimum size value satisfies the condition; and

(c) conducting a search by applying the set of search conditions to an inverted index other than the inverted index unnecessary for searches as identified in the step (b).