

DESCRIPTION

TITLE OF INVENTION: INFORMATION PROCESSING SYSTEM AND INFORMATION PROCESSING METHOD

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This international application claims the priority based on Japanese Patent Application No. 2022-049442 filed on March 25, 2022 with the Japan Patent Office, and the entire disclosure of Japanese Patent Application No. 2022-049442 is incorporated by reference in this international application.

TECHNICAL FIELD

[0002] The present disclosure relates to an information processing system and an information processing method.

BACKGROUND ART

[0003] Data fusion is an already-known technique. The applicant has disclosed a technique for associating a first dataset related to a first group of consumers with a second dataset related to a second group of consumers. The first dataset contains feature data of clusters defined by clustering the first group of consumers. The second dataset contains feature data of clusters defined by clustering the second group of consumers. Based on similarity of the clusters, the feature data of each of the clusters in the first dataset is associated with the feature data of one of the clusters in the second dataset (see, for example, Patent Document 1).

PRIOR ART DOCUMENTS

PATENT DOCUMENTS

[0004] Patent Document 1: Unexamined Japanese Patent Application Publication No. 2016-126609

SUMMARY OF THE INVENTION

PROBLEMS TO BE SOLVED BY THE INVENTION

- [0005] Replacing feature data of individual consumers with feature data of clusters is useful for transmitting and receiving consumer data while protecting personal information.
- [0006] However, it may not be possible to perform association between feature data with high accuracy only by evaluating the similarity of the clusters based on the content of the feature data. In other words, there has been a room for improving the accuracy of association in prior art.
- [0007] Accordingly, it is desirable that one aspect of the present disclosure can provide a novel technique that enables appropriate association, with regard to first clusters defined by clustering a first group of entities and second clusters defined by clustering a second group of entities, between feature data of the first clusters and feature data of the second clusters.

MEANS FOR SOLVING THE PROBLEMS

- [0008] According to one aspect of the present disclosure, an information processing system is provided. The information processing system includes a first acquirer, a second acquirer, and an associator.
- [0009] The first acquirer is configured to acquire a first dataset. The first dataset is a dataset related to multiple first clusters. The multiple first clusters are defined by clustering a first group of entities. The first dataset includes, for each first cluster, a piece of first cluster feature data representing a feature of a corresponding first cluster.
- [0010] The second acquirer is configured to acquire a second dataset. The second dataset is a dataset related to multiple second clusters. The multiple second clusters are defined by clustering a second group of entities. The second dataset includes, for each second cluster, a piece of second cluster feature data representing a feature of a corresponding second cluster.
- [0011] The associator is configured to associate each piece of multiple pieces of

first cluster feature data included in the first dataset with one or more pieces of multiple pieces of second cluster feature data included in the second dataset based on information, with regard to one or more common entities, that enables identification of a first cluster and a second cluster to which each of the one or more common entities belongs. The one or more common entities are one or more entities common to the first group and the second group.

[0012] According to the information processing system configured in this manner, it is possible to accurately perform association between cluster-related feature data by utilizing information about entities common to the first group and the second group.

[0013] According to one aspect of the present disclosure, the associator may be configured to associate each piece of the multiple pieces of first cluster feature data with one or more pieces of the multiple pieces of second cluster feature data so that, out of a first pair that is a pair of a first cluster and a second cluster including an entity common with each other and a second pair that is a pair of a first cluster and a second cluster not including any entity common with each other, association of the piece of first cluster feature data and the piece of second cluster feature data corresponding to the first pair is prioritized over the second pair.

[0014] According to one aspect of the present disclosure, the associator may be configured to, with regard at least to the second pair, determining a pair of the first cluster and the second cluster to be associated with each other based on positional relationship (for example, relative positions) in a feature space between the first cluster and the second cluster determined from the piece of first cluster feature data and the piece of second cluster feature data corresponding to the second pair so as to associate the piece of first cluster feature data with the piece of second cluster feature data.

[0015] According to one aspect of the present disclosure, the associator may be configured to, with regard to one or more specified clusters among the multiple first clusters, associate the piece of first cluster feature data of each of the one

or more specified clusters with the piece of second cluster feature data of the second cluster including a same entity as an entity in a corresponding specified cluster. The one or more specified clusters are one or more first clusters each of which includes at least one of the one or more common entities among the multiple first clusters.

[0016] According to the information processing system configured in this manner, it is possible to associate the piece of first cluster feature data with the piece of second cluster feature data so as to be consistent with the relationship between the first clusters and the second clusters to which the common entities belong.

[0017] According to one aspect of the present disclosure, the associator may be configured to associate, with regard to a non-specified cluster among the multiple first clusters other than the one or more specified clusters, determining a pair of the first cluster and the second cluster to be associated with each other based on positional relationship (for example, relative positions) in a feature space between the first cluster and the second cluster determined from the piece of first cluster feature data and the piece of second cluster feature data so as to associate the piece of first cluster feature data with the piece of second cluster feature data.

[0018] According to one aspect of the present disclosure, the associator may be configured to, with regard to one or more non-common first clusters and one or more non-common second clusters, determine relationship between each of the one or more non-common first clusters and each of the one or more non-common second clusters based on the piece of first cluster feature data of each of the one or more non-common first clusters and the piece of second cluster feature data of each of the one or more non-common second clusters. The associator may be configured to associate the piece of first cluster feature data of each of the one or more non-common first clusters with at least one piece of second cluster feature data of at least one of the one or more non-common second clusters based on the relationship determined. The one or more non-common first clusters are one or more remaining clusters that are the multiple first clusters other than the

one or more specified clusters. The one or more non-common second clusters are one or more remaining clusters that are the multiple second clusters other than one or more clusters associated with the one or more specified clusters.

[0019] According to the information processing system configured in this manner, it is possible to associate the piece of first cluster feature data with the piece of second cluster feature data so as to be consistent with presence or absence of common entities.

[0020] According to one aspect of the present disclosure, the piece of first cluster feature data may be a piece of data that defines a point of the corresponding first cluster in a first feature space. The piece of second cluster feature data may be a piece of data that defines a point of the corresponding second cluster in a second feature space.

[0021] According to one aspect of the present disclosure, the associator may perform a process of searching, using an evaluation function, a mapping that maps a point of each of the multiple first clusters in the first feature space into the second feature space.

[0022] According to one aspect of the present disclosure, the evaluation function may be designed to output a value, with regard to a specified pair that is a pair of the first cluster and the second cluster including a same entity, that varies in a specified direction with a decrease in a distance in the second feature space between the first cluster and the second cluster constituting the specified pair.

[0023] According to one aspect of the present disclosure, the evaluation function may be designed to output a value, with regard to a pair of the first cluster and the second cluster other than the specified pair, that varies in a direction opposite to the specified direction with a decrease in a distance in the second feature space between the first cluster and the second cluster. In this case, the associator may be configured to search a mapping that makes the output of the evaluation function to be the limit in a specified direction.

[0024] According to one aspect of the present disclosure, the evaluation function may be designed to output a value, with regard to the specified pair, that is

smaller for a shorter distance in the second feature space between the first cluster and the second cluster constituting the specified pair. According to one aspect of the present disclosure, the evaluation function may be designed to output a value, with regard to a pair of the first cluster and the second cluster other than the specified pair, that is greater for a shorter distance in the second feature space between the first cluster and the second cluster. In this case the associator may be configured to search a mapping that minimizes the output of the evaluation function.

[0025] According to one aspect of the present disclosure, the associator may be configured to perform a process of associating each piece of the multiple pieces of first cluster feature data with one or more pieces of the multiple pieces of second cluster feature data based on positional relationship (for example, relative positions) between a point of each of the multiple first clusters in the second feature space mapped by the mapping and a point of each of the multiple second clusters in the second feature space.

[0026] According to the information processing system configured in this manner, it is possible to obtain an appropriate mapping from the first feature space to the second feature space focusing on common entities, and then associate the piece of first cluster feature data with the piece of second cluster feature data. Thus, it is possible to achieve highly accurate association.

[0027] According to one aspect of the present disclosure, the associator may be configured to perform a process of associating the piece of first cluster feature data of each of the one or more specified clusters with the piece of second cluster feature data of the second cluster including a same entity. The associator may be further configured to perform a process of associating the piece of first cluster feature data of each of the one or more non-common first clusters with at least one piece of second cluster feature data of at least one of the one or more non-common second clusters based on positional relationship (for example, relative positions) between a point of each of the one or more non-common first clusters in the second feature space mapped by the mapping and a point of each of the

one or more non-common second clusters in the second feature space.

[0028] According to one aspect of the present disclosure, the one or more common entities may be multiple common entities. The multiple first clusters may be a set of clusters defined by clustering the first group so that the multiple common entities are arranged in the multiple first clusters in a dispersed manner. The multiple second clusters may be a set of clusters defined by clustering the second group so that the multiple common entities are arranged in the multiple second clusters in a dispersed manner.

[0029] By dispersing the common entities to multiple clusters, that is, by clustering the first group and the second group so that a single cluster does not include a large number of common entities, it is possible to accurately determine the correspondence relationship between the first clusters and the second clusters through the common entities. Thus, it is possible to accurately associate the piece of first cluster feature data with the piece of second cluster feature data.

[0030] According to one aspect of the present disclosure, the multiple first clusters may be a set of clusters defined by clustering the first group so that respective common entities of the multiple common entities belong to first clusters different from each other. The multiple second clusters may be a set of clusters defined by clustering the second group so that respective common entities of the multiple common entities belong to second clusters different from each other.

[0031] According to one aspect of the present disclosure, the piece of first cluster feature data may be a piece of feature data generated by integrating multiple pieces of feature data of multiple entities belonging to the corresponding first cluster through a statistical process. The piece of second cluster feature data may be a piece of feature data generated by integrating multiple pieces of feature data of multiple entities belonging to the corresponding second cluster through a statistical process.

[0032] To hide information about individual entities, pieces of cluster feature data in which features of entities are statistically compiled may be provided from a

data holder. According to one aspect of the present disclosure, it is possible to appropriately perform association between pieces of cluster feature data that have been subjected to such statistical processing.

[0033] According to one aspect of the present disclosure, the multiple entities may be multiple consumers. The multiple pieces of feature data of the multiple entities may be multiple pieces of feature data describing features of the multiple consumers.

[0034] According to one aspect of the present disclosure, a computer program for at least partially implementing the functions of the aforementioned information processing system may be provided. According to one aspect of the present disclosure, a computer program for causing a computer to function as the first acquirer, the second acquirer, and the associator in the aforementioned information processing system may be provided. The computer program may be stored on a computer-readable non-transitory recording medium.

[0035] According to one aspect of the present disclosure, an information processing method corresponding to the aforementioned information processing system may be provided. According to one aspect of the present disclosure, an information processing method performed by a computer may be provided.

[0036] According to one aspect of the present disclosure, the information processing method may include acquiring a first dataset. The first dataset may be a dataset related to multiple first clusters. The multiple first clusters are defined by clustering a first group of entities. The first dataset may include, for each first cluster, a piece of first cluster feature data representing a feature of a corresponding first cluster.

[0037] The information processing method may include acquiring a second dataset. The second dataset is a dataset related to multiple second clusters. The multiple second clusters may be defined by clustering a second group of entities. The second dataset may include, for each second cluster, a piece of second cluster feature data representing a feature of a corresponding second cluster.

[0038] The information processing method may include associating each piece of

multiple pieces of first cluster feature data included in the first dataset with one or more pieces of multiple pieces of second cluster feature data included in the second dataset based on information, with regard to one or more common entities, that enables identification of a first cluster and a second cluster to which each of the one or more common entities belongs. The one or more common entities may be one or more entities common to the first group and the second group.

[0039] As with the information processing system described above, this information processing method also makes it possible accurately perform association between cluster-related feature data by utilizing information about entities common to the first group of entities corresponding to the first dataset and the second group of entities corresponding to the second dataset.

[0040] According to one aspect of the present disclosure, a computer program may be provided that contains instructions for causing a computer to perform the aforementioned information processing method when the computer program is executed by the computer. A computer-readable recording medium storing the computer program may be provided.

BRIEF DESCRIPTION OF THE DRAWINGS

[0041] FIG. 1 is a block diagram showing a configuration of an information processing system.

FIG. 2A is a diagram illustrating an operation in a data providing system, and FIG. 2B is a diagram showing a configuration of example cluster feature data.

FIG. 3 is an explanatory diagram relating to clustering of a consumer group including common consumers (white circles) and non-common consumers (black circles).

FIG. 4 is a flowchart illustrating an acquisition-related process performed by a processor of the information processing system.

FIG. 5A is a diagram showing a first cluster-consumer relationship table, FIG. 5B is a diagram showing a second cluster- consumer relationship table, and

FIG. 5C is a diagram illustrating a cluster relationship table.

FIG. 6 is a flowchart illustrating an analysis process performed by the processor of the information processing system.

FIG. 7A is a flowchart illustrating a first example of a partial procedure in the analysis process, and FIG. 7B is a flowchart illustrating a second example of the partial procedure in the analysis process.

FIG. 8A is a flowchart illustrating a first example of a clustering process, and FIG. 8B is a flowchart illustrating a second example of the clustering process.

FIG. 9 is a flowchart illustrating an analysis process in a variation.

EXPLANATION OF REFERENCE NUMERALS

[0042] 1...information processing system, 11...processor, 13...memory, 15...storage, 15A...first dataset, 15B...second dataset, 15C...extended dataset, 17...user interface, 19...communication interface, 31...first data providing system, 32...second data providing system, 311...processor, 312...memory, 321...processor, 322...memory, Pr...computer program.

MODE FOR CARRYING OUT THE INVENTION

[0043] Example embodiments of the present disclosure will be described below with reference to the drawings.

[0044] As shown in FIG. 1, an information processing system 1 of the present embodiment includes a processor 11, a memory 13, a storage 15, a user interface 17, and a communication interface 19.

[0045] The processor 11 performs processing in accordance with a computer program Pr stored in the storage 15. The memory 13 is a primary storage device that includes a RAM. The memory 13 is used as a work area when processing is performed by the processor 11.

[0046] The storage 15 is a secondary storage device that includes, for example, a hard disk drive or a solid state drive. The storage 15 stores, in addition to the computer program Pr, various data that are used when processing in accordance

with the computer program Pr is executed.

[0047] The user interface 17 includes an input device for inputting an operation signal from a user operating the information processing system 1 to the processor 11. Examples of the input device include keyboards and pointing devices. The user interface 17 further includes a display for displaying various types of information to the user.

[0048] The communication interface 19 is connected to a Wide Area Network. The information processing system 1 is configured to be capable of communicating with a first data providing system 31 and a second data providing system 32 through the communication interface 19.

[0049] The processor 11, by execution of processing in accordance with the computer program Pr, acquires a first dataset 15A from the first data providing system 31 and acquires a second dataset 15B from the second data providing system 32 through the communication interface 19.

[0050] The processor 11 combines the acquired first dataset 15A with the second dataset 15B to thereby generate an extended dataset 15C that is a data-enriched version of the first dataset 15A.

[0051] The first dataset 15A is a dataset related to first clusters defined by clustering a first group of entities and includes, for each first cluster, cluster feature data representing features of a corresponding first cluster.

[0052] The second dataset 15B is a dataset related to second clusters defined by clustering a second group of entities and includes, for each second cluster, cluster feature data representing features of a corresponding second cluster.

[0053] According to the present embodiment, the entities are consumers. That is, the first dataset 15A related to first clusters defined by clustering a first group of consumers and includes cluster feature data for each first cluster.

[0054] As shown in FIG. 2A, the first data providing system 31 stores consumer feature data of consumers in the first group. Each consumer feature data represents features of a corresponding consumer. The first data providing system 31 performs a clustering process, thereby clustering the first group into

the first clusters.

[0055] Further, the first data providing system 31 statistically processes, for each first cluster, a group of the consumer feature data related to a corresponding first cluster and generates the cluster feature data of each first cluster.

[0056] Each cluster feature data is generated by integrating the consumer feature data of the consumers belonging to the corresponding cluster through a statistical process. The statistical process is performed for protecting personal information, for example. The first data providing system 31 provides the information processing system 1 with the first dataset 15A including the cluster feature data of each first cluster generated in this manner.

[0057] The first data providing system 31 includes a processor 311 and a memory 312. The processor 311 performs the clustering process and the statistical process in accordance with computer programs stored in the memory 312.

[0058] Also, the second data providing system 32 includes a processor 321 and a memory 322. The first data providing system 32 stores the consumer feature data of the consumers in the second group. In the same way as the first data providing system 31, the processor 321 performs the clustering process and the statistical process on a group of the consumer feature data held by the second data providing system 32 in accordance with computer programs stored in the memory 322.

[0059] As a result, the second data providing system 32 generates the second dataset 15B related to the second clusters defined by clustering the second group of consumers and includes, for each second cluster, the cluster feature data representing the features of a corresponding second cluster, and provides the information processing system 1 with the second dataset 15B.

[0060] The consumer feature data represents the features of a corresponding consumer as a vector using values of elements. The cluster feature data also represents the features of a corresponding cluster as a vector using values of elements.

[0061] The value of each element in the cluster feature data is the statistically

representative value of the corresponding element in a consumer set belonging to the corresponding cluster. The representative value may include one or more of the mean, the median, a combination of the mean and the variance, and a combination of the maximum value and the minimum value of the corresponding element in the consumer set.

[0062] The cluster feature data shown in FIG. 2B describes a value of each element representing a feature of the corresponding cluster and a cluster size in association with a cluster ID which is an identification code of the cluster. The cluster size corresponds to the number of consumers constituting the corresponding cluster.

[0063] Further, in the present embodiment, the information processing system 1 communicates with the first data providing system 31 and the second data providing system 32, taking into account the possibility that there may be consumers common to the first group of consumers and the second group of consumers.

[0064] As a result, the information processing system 1 determines consumers who are commonly present in the first group and the second group as common consumers, and transmits a common consumer list to the first data providing system 31 and the second data providing system 32.

[0065] The first data providing system 31 clusters the first group based on the common consumer list so that the common consumers are arranged in the first clusters in a dispersed manner. If the number of common consumers is equal to or less than the number of clusters, the first data providing system 31 clusters the first group so that the respective common consumers belong to the clusters different from each other.

[0066] FIG. 3 illustrates clustering of a set of consumers represented by white circles and black circles in such a manner that a single cluster includes as few common consumers as possible, ideally a single cluster includes one common consumer. An area surrounded by a dashed circle corresponds to one cluster. Each white circle represents one common consumer, and each black circle

represents a consumer who is not a common consumer, that is, one non-common consumer.

[0067] Similarly, the second data providing system 32 clusters the second group so that the common consumers are arranged in the second clusters in a dispersed manner. If the number of common consumers is equal to or less than the number of clusters, the second data providing system 32 clusters the second group so that the respective common consumers belong to the second clusters different from each other.

[0068] The information processing system 1 acquires, from the first data providing system 31, the first dataset 15A containing the cluster feature data of each first cluster generated by such clustering. The information processing system 1 further acquires, from the second data providing system 32, the second dataset 15B containing the cluster feature data of each second cluster.

[0069] FIG. 4 illustrates details of an acquisition-related process performed by the processor 11 of the information processing system 1 to acquire the first dataset 15A and the second dataset 15B. The processor 11 initiates the acquisition-related process based on an instruction from the user via the user interface 17.

[0070] When the acquisition-related process begins, the processor 11 acquires a list of consumers belonging to the first group from the first data providing system 31 as a first consumer list (S110). The first consumer list contains, for each consumer belonging to the first group, information about a consumer ID which is an identification code of the corresponding consumer.

[0071] The processor 11 further acquires a list of consumers belonging to the second group from the second data providing system 32 as a second consumer list (S120). The second consumer list contains, for each consumer belonging to the second group, information about the consumer ID.

[0072] To identify the common consumers, the same type of consumer IDs may be described in the first consumer list and the second consumer list. An example of the consumer IDs may be advertising IDs. Alternatively, the consumer IDs may be IDs given to a large number of consumers, such as those

epitomized by a membership IDs for a reward program with a vast number of members.

[0073] First consumer IDs, which are the consumer IDs used in the first consumer list, may be a type of identification codes different from second consumer IDs, which are the consumer IDs used in the second consumer list. In this case, a table may be required that indicates the correspondence relationship between the first consumer ID and the second consumer ID to identify the same person.

[0074] In subsequent S130, the processor 11 determines one or more common consumers, who are one or more consumers common to the first group and the second group, by comparing the consumer IDs listed in the first consumer list with the consumer IDs listed in the second consumer list, and generates the common consumer list.

[0075] In subsequent S140, the processor 11 transmits the common consumer list generated in S130 to the first and second data providing systems 31, 32, and causes the first and second data providing systems 31, 32 to generate the above-described first dataset 15A and second dataset 15B based on this list.

[0076] In subsequent S150, the processor 11 acquires the first dataset 15A and a first cluster-consumer relationship table from the first data providing system 31. The first cluster consumer relationship table describes, for example as shown in FIG. 5A, a first cluster ID for the consumer ID of each common consumer. The first cluster ID is an identification code of the first cluster to which the corresponding consumer belongs. The first data providing system 31 is configured to transmit the first cluster-consumer relationship table to the information processing system 1 together with the first dataset 15A.

[0077] The processor 11 acquires the second dataset 15B and a second cluster consumer relationship table from the second data providing system 32 (S160). The second cluster consumer relationship table describes, for example as shown in FIG. 5B, a second cluster ID for the consumer ID of each common consumer. The second cluster ID is an identification code of the second cluster to which the corresponding consumer belongs. The second data providing system 32 is

configured to transmit the second cluster-consumer relationship table to the information processing system 1 together with the second dataset 15B.

[0078] In subsequent S170, based on the acquired first and second cluster-consumer relationship tables, the processor 11 generates a cluster relationship table, shown in FIG. 5C, which shows a combination of the first cluster and the second cluster in which the same consumer is included. The cluster relationship table corresponds to information that enables identification of the first cluster and the second cluster to which each common consumer belongs. The cluster relationship table, shown as an example in FIG. 5C, indicates a combination of the first cluster and the second cluster in which the same common consumer is included by a combination of the first cluster ID and the second cluster ID.

[0079] In subsequent S180, the processor 11 saves the cluster relationship table generated in S170 in the storage 15 together with the first dataset 15A and the second dataset 15B, and ends the process illustrated in FIG. 4.

[0080] Subsequently, details of an analysis process performed by the processor 11 will be described using FIG. 6. The analysis process is performed based on an instruction from the user via the user interface 17, for example.

[0081] When the analysis process begins, the processor 11 reads out, from the storage 15, the first dataset 15A designated by the user and generates an array D_f based on the cluster feature data of each first cluster contained in the read first dataset 15A (S210).

[0082] For example, when there are K1 clusters and K1 pieces of cluster feature data each represent the features of the cluster in an L1-dimensional feature vector including L1 elements, the array D_f is a two-dimensional array with K1 rows and L1 columns.

[0083] Each row vector $D_f[i]=(x_i[1], x_i[2], x_i[3], \dots, x_i[L1])$ in the array D_f is the L1-dimensional feature vector of the corresponding i-th first cluster. Here, i corresponds to the row number in the array D_f and takes an integer value ranging from a value of 1 to K1 corresponding to the number of clusters K1.

[0084] Hereinafter, an L1-dimensional space that corresponds to a feature vector $D_f[i]$ represented by the cluster feature data contained in the first dataset 15A will be expressed as a first feature space. The feature vector $D_f[i]$ defines a point at which the corresponding first cluster is located in the first feature space. The first feature space is an L1-dimensional space having dimensions respectively corresponding to the L1 elements of the feature vector $D_f[i]$.

[0085] In subsequent S220, the processor 11 reads out, from the storage 15, the second dataset 15B designated by the user and generates an array D_s based on the cluster feature data of each second cluster contained in the read second dataset 15B.

[0086] For example, when there are K2 clusters and K2 pieces of cluster feature data each represent the features of the cluster in an L2-dimensional feature vector including L2 elements, the array D_s is a two-dimensional array with K2 rows and L2 columns.

[0087] Each row vector $D_s[j]=(y_j[1], y_j[2], y_j[3], \dots, y_j[L2])$ in the array D_s is the L2-dimensional feature vector of the corresponding j-th second cluster. Here, j corresponds to the row number in the array D_s and takes an integer value ranging from a value of 1 to K2 corresponding to the number of clusters K2.

[0088] Hereinafter, an L2-dimensional space corresponding to a feature vector $D_s[j]$ represented by the cluster feature data contained in the second dataset 15B will be expressed as a second feature space. The feature vector $D_s[j]$ defines a point at which the corresponding second cluster is located in the second feature space. The second feature space is an L2-dimensional space having dimensions respectively corresponding to the L2 elements of the feature vector $D_s[j]$.

[0089] In subsequent S230, the processor 11 searches a function g that minimizes the following evaluation function E. In this manner, the processor 11 searches a mapping that best maps each of points corresponding to the respective common consumers in the first feature space identified from the first dataset 15A near the point of the same person in the second feature space identified from the second dataset 15B.

[0090] [Mathematical Formula 1]

$$E = \sum_{m,n \in M1} d(g(D_f[A_f^{-1}[m]]), D_s[A_s^{-1}[n]]) - \tau \sum_{m,n \in M2} d(g(D_f[A_f^{-1}[m]]), D_s[A_s^{-1}[n]])$$

In the above formula, m means the m-th consumer in a consumer group corresponding to the K1 first clusters whose cluster feature data are described in the first dataset 15A.

[0091] The consumer group here may be a consumer group that corresponds to K1 first clusters interpreted when each first cluster is considered to include consumers corresponding to the cluster size thereof. Alternatively, the consumer group may be a group of K1 consumers that corresponds to K1 first clusters interpreted when each cluster is considered to include one representative consumer. $A_f^{-1}[m]$ means that the first cluster to which the m-th consumer belongs is the $A_f^{-1}[m]$ -th cluster among the K1 clusters.

[0092] In the above formula, n means the n-th consumer in a consumer group corresponding to the K2 second clusters whose cluster feature data are described in the second dataset 15B. The consumer group here may be a consumer group that corresponds to K2 second clusters interpreted when each second cluster is considered to include consumers corresponding to the cluster size thereof. Alternatively, the consumer group may be a group of K2 consumers that corresponds to K2 second clusters interpreted when each cluster is considered to include one representative consumer. $A_s^{-1}[n]$ means that the second cluster to which the n-th consumer belongs is the $A_s^{-1}[n]$ -th cluster among the K2 clusters.

[0093] $d(v1, v2)$ is a distance function and represents the distance (for example, Mahalanobis distance) between a vector v1 and a vector v2. τ is a design parameter and is defined by the user, in other words, by an analyst.

[0094] The first term and the second term of the evaluation function E each represent the sum $\sum d(g(D_f[A_f^{-1}[m]]), D_s[A_s^{-1}[n]])$ of a distance $d(g(D_f[A_f^{-1}[m]]), D_s[A_s^{-1}[n]])$ between a point $g(D_f[A_f^{-1}[m]])$ in the second feature space and a point $D_s[A_s^{-1}[n]]$ in the second feature space when a point $D_f[A_f^{-1}[m]]$ in the first feature space is mapped to a point in the second feature space by the function g

as a mapping. The point $D_f[A_f^{-1}[m]]$ is the point $D_f[A_f^{-1}[m]]$ in the first feature space of the $A_f^{-1}[m]$ -th first cluster to which the m -th consumer in the first group belongs. The point $D_s[A_s^{-1}[n]]$ is the point $D_s[A_s^{-1}[n]]$ in the second feature space of the $A_s^{-1}[n]$ -th second cluster to which the n -th consumer in the second group belongs.

[0095] However, the first term of the evaluation function E is, according to the cluster relationship table, the sum $\sum d(g(D_f[A_f^{-1}[m]]), D_s[A_s^{-1}[n]])$ for a set $M1$ of combinations of m and n in which the same consumer is present in the $A_f^{-1}[m]$ -th first cluster and the $A_s^{-1}[n]$ -th second cluster. That is, the first term of the evaluation function E is the sum $\sum d(g(D_f[A_f^{-1}[m]]), D_s[A_s^{-1}[n]])$ for same consumer combinations (m, n) .

[0096] The second term of the evaluation function E is, according to the cluster relationship table, the sum $\sum d(g(D_f[A_f^{-1}[m]]), D_s[A_s^{-1}[n]])$ for a set $M2$ of combinations of m and n in which the same consumer is not present in the $A_f^{-1}[m]$ -th first cluster and the $A_s^{-1}[n]$ -th second cluster. That is, the second term of the evaluation function E is the sum $\sum d(g(D_f[A_f^{-1}[m]]), D_s[A_s^{-1}[n]])$ for different consumer combinations (m, n) .

[0097] This evaluation function E is a function that outputs a value: with regard to specified pairs each of which is a pair of the first cluster and the second cluster that include the same consumer, that is smaller for a shorter distance in the second feature space between the first cluster and the second cluster constituting the specified pair; and, with regard to pairs of the first cluster and the second cluster other than the specified pairs, that is greater for a smaller distance in the second feature space between the first cluster and the second cluster.

[0098] Thus, searching the function g that minimizes the evaluation function E enables searching a mapping that best maps each of the points corresponding to the respective common consumers in the first feature space identified from the first dataset 15A to a position near the point of the same person in the second feature space identified from the second dataset 15B. According to this mapping, the distance between non-identical persons is large in the second

feature space. The function g may be a transformation orthogonal matrix or a multi-layer perceptron.

[0099] After searching the function g that minimizes the evaluation function E in S230, the processor 11 associates each of the first clusters to one or more second clusters based on the positional relationship (in other words, the relative positions) between the points $g(D_f[A_f^{-1}[m]])$ in the second feature space of the $K1$ first clusters mapped by the searched function g and the points $D_s[A_s^{-1}[n]]$ in the second feature space of the $K2$ second clusters (S240).

[0100] For example, the processor 11 assumes that at each point there are a certain number of consumers that matches the cluster size of the corresponding cluster, and associates each of the first clusters to one or more second clusters by solving an optimal transportation problem.

[0101] For example, in S240, the processor 11 may perform the processing shown in FIG. 7A. According to this processing, based on the cluster relational table, the processor 11 sets, with regard to pairs of the first cluster and the second cluster in each pair of which the first cluster and the second cluster include the same consumer, the distance between the first cluster and the second cluster to a specified value of ε (S241). The value of ε may be, for example, zero or close to zero.

[0102] With regard to pairs of the first cluster and the second cluster not including the same consumer, the processor 11 uses the distance in the second feature space between the first cluster and the second cluster identified from the feature vector $g(D_f[A_f^{-1}[m]])$ and the feature vector $D_s[A_s^{-1}[n]]$ unchanged to solve the aforementioned optimal transportation problem, and thereby associates each of the first clusters to one or more second clusters (S243).

[0103] Alternatively, in S240, the processor 11 may perform the processing shown in FIG. 7B. According to this processing, with regard to the pairs of the first cluster and the second cluster each pair of which the first cluster and the second cluster include the same consumer, the processor 11 equally associates on a pair-by-pair basis the first cluster with the pairwise-corresponding second cluster

without referring to the cluster feature data (S245).

[0104] With regard to a group of pairs of the first cluster and the second cluster not including the same consumer, each of the first clusters is associated with one or more second clusters based on the positional relationship between the point g ($D_f[A_f^{-1}[m]]$) in the second feature space of the first cluster identified from the cluster feature data and the point $D_s[A_s^{-1}[n]]$ in the second feature space of the second cluster, on a pair-by-pair basis (S247).

[0105] For example, only for the group of pairs of first cluster and the second cluster not including the same consumer, the processor 11 associates each of the first clusters with one or more second clusters by solving an optimal transportation problem for the group (S247).

[0106] Upon completing the processing in S240 in this manner, the processor 11 combines the first dataset 15A and the second dataset 15B so that the cluster feature data of each first cluster contained in the first dataset 15A is combined with the cluster feature data of the one or more second clusters that are contained in the second dataset 15B and have been associated by the processing of S240. Through this combining (that is, data fusion), the processor 11 generates the extended dataset 15C and stores the same in the storage 15 (S250). Subsequently, the processor 11 ends the analysis process.

[0107] According to the information processing system 1 of the present embodiment described above, it is possible to accurately perform association between cluster-related feature data by utilizing information about consumers common to the first group of consumers corresponding to the first dataset 15A and the second group of consumers corresponding to the second dataset 15B. In particular, it is possible to associate the first cluster feature data with the second cluster feature data so that the first cluster and the second cluster including the same consumer are associated with each other.

[0108] To improve the accuracy of association, it is preferable that the clustering process be performed in the first data providing system 31 and the second data providing system 32 so that each cluster includes as few common consumers as

possible, ideally so that each cluster includes one common consumer.

[0109] For this purpose, the first data providing system 31 and the second data providing system 32 may perform a clustering process shown in FIG. 8A which is an improvement of the k-means method.

[0110] That is, the processor 311 of the first data providing system 31 may perform, as the clustering process, processing (S300) to cluster the first group into K clusters where K is a specified number so that the following penalized evaluation function H is minimized.

[0111] [Mathematical Formula 2]

$$H = \sum_m \sum_k I(m \in A[k]) W[m] d(C[k], D[m]) + \lambda \sum_{i,j \in Q} I(A^{-1}[i] == A^{-1}[j])$$

The first term of the evaluation function H is an evaluation function in accordance with the k-means method, and the second term of the evaluation function H is a term in which a penalty λ is added when two or more common consumers are present in one cluster.

[0112] In the evaluation function H, m is an index of a consumer and k is an index of a cluster. $A[k]$ is a consumer set constituting the k-th cluster. The indicator function $I(m \in A[k])$ is a function that returns a value of 1 when the m-th consumer is present in the consumer set constituting the k-th cluster, and returns a value of 0 when it is otherwise. $W[m]$ is the weight of the m-th consumer. The weight W may be understood to be one.

[0113] $C[k]$ is the center vector of the k-th cluster, and $D[m]$ is the feature vector of the m-th consumer identified from the consumer feature data of the m-th consumer. $A^{-1}[i]$ indicates that the cluster to which the i-th consumer belongs is the $A^{-1}[i]$ -th cluster among the K clusters.

[0114] The indicator function $I(A^{-1}[i] == A^{-1}[j])$ is a function that returns a value of 1 when the i-th consumer and the j-th consumer are present in the same cluster in a common consumer set Q, and returns a value of 0 when the i-th consumer and the j-th consumer are not present in the same cluster.

[0115] That is, the second term of the evaluation function H is a term obtained by

multiplying the sum of values of $I(A^{-1}[i]==A^{-1}[j])$ by the penalty λ where values of $I(A^{-1}[i]==A^{-1}[j])$ are for combinations of consumers who are in the common consumer set Q and $i \neq j$. The penalty λ may be defined to be a positive value by the designer.

[0116] As with the first data providing system 31, the processor 321 of the second data providing system 32 is also capable of clustering the second group into second clusters in accordance with the penalized evaluation function H .

[0117] In another example, the processor 311 of the first data providing system 31 may perform a clustering process shown in FIG. 8B.

[0118] According to the clustering process shown in FIG. 8B, the processor 311 sets K empty clusters (S410) where K is a specified number. Then, the processor 311 randomly selects, from the first group of consumers, one of the common consumers who is common to the second group as a target consumer (S420).

[0119] Subsequently, the processor 311 determines whether there are any empty clusters to which no common consumer is assigned among the set K clusters (S430). If the processor 311 determines that there are one or more empty clusters (Yes in S430), the processor 311 assigns the target consumer to one empty cluster (S440).

[0120] In subsequent S450, the processor 311 determines whether all the common consumers have each been assigned to any one of the K clusters. If the processor 311 makes a negative determination here (No in S450), the processor 311 selects another common consumer as the target consumer in S420, and performs the processing of S430 and following steps for the selected target consumer.

[0121] In this manner, the processor 311 performs the processing of S430 and the following steps for each common consumer, and assigns each common consumer to one of the K clusters.

[0122] In a case where the number of common consumers is smaller than the number of clusters, no negative determination is made in S430, and the common

consumers are each assigned to one of the K clusters. In contrast, in a case where the number of common consumers is greater than the number of clusters, the processor 311, after assigning each of K common consumers to an empty cluster, determines in S430 that there are no empty clusters when selecting another common consumer from the remaining common consumers as the target consumer.

[0123] The processor 311, if determining in S430 that there is no empty cluster, assigns the target consumer to a cluster for which the following evaluation function $R=d(C[k], D[i])$ is minimum (S445).

[0124] In the evaluation function R , $C[k]$ is the center vector of the k -th cluster, and $D[i]$ is the feature vector of the i -th consumer, who is the target consumer, identified from the consumer feature data of the i -th consumer. $d()$ is a distance function.

[0125] After the cluster assignment in S445, the processor 311 performs the processing of S450. The processor 311, when determining in S450 that all the common consumers have each been assigned to one of the K clusters, assigns each of remaining non-common consumers in the first group to a nearby cluster so as to minimize the sum of the distances between the feature vectors of the consumers and the center vectors of the clusters as with the case of the k -means method (S460).

[0126] The processor 311 clusters the first group of consumers into K clusters by assigning each consumer belonging to the first group to one of the K clusters in this manner. Then, the processor 311 ends the clustering process shown in FIG. 8B.

[0127] According to this clustering process, when the number of common consumers is smaller than the number of clusters, it is possible to assign the common consumers to multiple clusters in a dispersed manner so that each cluster includes one common consumer. Thus, it is possible to perform association of the first clusters with the second clusters in the information processing system 1 with high accuracy based on the information about the

common consumer.

[0128] The technology of the present disclosure described above is not limited to the above-described embodiment, and may be implemented in various manner.

[0129] In the above-described embodiment, even if there is no common element between the cluster feature data contained in the first dataset 15A and the cluster feature data contained in the second dataset 15B as an element that describes a feature of the clusters, it is possible to associate the first clusters with the second clusters by searching the mapping (function g).

[0130] If there is a common element, that is, if there is a common variable, association between the first clusters and the second clusters may be performed by solving the optimal transportation problem based only on this common variable. There may be cases where the cluster feature data contained in the first dataset 15A and the cluster feature data contained in the second dataset 15B have a common element that describes the feature of the clusters. That is, there may be cases where there is a common variable between the feature vector $(x_i[1], x_i[2], x_i[3], \dots, x_i[L1])$ of the cluster feature data contained in the first dataset 15A and the feature vector $(y_j[1], y_j[2], y_j[3], \dots, y_j[L2])$ of the cluster feature data contained in the second dataset 15B.

[0131] In this case, the processor 11 may perform an analysis process shown in FIG. 9 in place of the analysis process shown in FIG. 6. When the analysis process shown in FIG. 9 begins, the processor 11 reads out the designated first dataset 15A from the storage 15 in S510. Based on the cluster feature data of each first cluster contained in the read first dataset 15A, the processor 11 generates, for each first cluster, a feature vector $D_f^*[i]=(x_i[1], x_i[2], x_i[3], \dots, x_i[L0])$ consisting only of common variables.

[0132] It is assumed here that $L0$ elements $x_i[1], x_i[2], x_i[3], \dots, x_i[L0]$ out of $L1$ elements describing the features of the first cluster in the first dataset 15A, and $L0$ elements $y_j[1], y_j[2], y_j[3], \dots, y_j[L0]$ out of $L2$ elements describing the features of the second cluster in the second dataset 15B are common variables.

[0133] In subsequent S520, the processor 11 reads out the designated second

dataset 15B from the storage 15. Based on the cluster feature data of each second cluster contained in the read second dataset 15B, the processor 11 generates, for each second cluster, a feature vector $D_s^*[j]=(y_j[1], y_j[2], y_j[3], \dots, y_j[L0])$ consisting only of common variables.

[0134] In subsequent S540, based on the distance in the feature space between the feature vector $D_f^*[i]$ and the feature vector $D_s^*[j]$ for each combination of the first cluster and the second cluster, the processor 11 associates each of the first clusters to one or more second clusters by solving the optimal transportation problem.

[0135] In S540, as with the processing shown in FIG. 7A, the processor 11 may set the distance between the first cluster and the second cluster to the specified value of ε and solve the optimal transportation problem with regard to pairs of the first cluster and the second cluster in each pair of which the first cluster and the second cluster includes the same consumer.

[0136] Alternatively, with regard to the pairs of the first cluster and the second cluster in each pair of which the first cluster and the second cluster include the same consumer, the processor 11 may equally associate on a pair-by-pair basis the first cluster with the pairwise-corresponding second cluster without referring to the cluster feature data. For a group of pairs of first cluster and second cluster not including the same consumer, the processor 11 may associate each of the first clusters to one or more second clusters by solving the optimal transportation problem based on the distance in the feature space between the feature vector $D_f^*[i]$ and the feature vector $D_s^*[j]$.

[0137] In subsequent S550, the processor 11 combines the first dataset 15A and the second dataset 15B in accordance with the association in S540, so as to combine the cluster feature data of each first cluster contained in the first dataset 15A with the cluster feature data of the one or more associated second clusters contained in the second dataset 15B. Through this combining, the processor 11 generates the extended dataset 15C and stores the same in the storage 15. Subsequently, the processor 11 ends the analysis process.

- [0138] The analysis process shown in FIG. 9 focuses on common variables and enables data fusion between the first dataset 15A and the second dataset 15B with a low-load procedure.
- [0139] The above has described an example in which the technology of the present disclosure is applied to combining of datasets containing consumer feature data. However, the technology of the present disclosure may be applied to examples of various entities other than consumers, including information terminals and automobiles. That is, the entities may be other than consumers and may be non-human.
- [0140] One or a plurality of functions of one component in the aforementioned embodiments may be distributed as a plurality of components, or one or a plurality of functions of a plurality of components may be integrated into one component. A part of the configurations of the aforementioned embodiments may be omitted. At least a part of the configurations of the aforementioned embodiments may be added to or replaced with another configuration of the aforementioned embodiments. Any and all embodiments encompassed by the technical idea defined by the language of the claims are embodiments of the present disclosure.

CLAIMS

1. An information processing system comprising:

a first acquirer configured to acquire a first dataset related to multiple first clusters defined by clustering a first group of entities, the first dataset including, for each first cluster, a piece of first cluster feature data representing a feature of a corresponding first cluster;

a second acquirer configured to acquire a second dataset related to multiple second clusters defined by clustering a second group of entities, the second dataset including, for each second cluster, a piece of second cluster feature data representing a feature of a corresponding second cluster; and

an associator configured to associate each piece of multiple pieces of first cluster feature data included in the first dataset with one or more pieces of multiple pieces of second cluster feature data included in the second dataset based on information, with regard to one or more common entities that are one or more entities common to the first group and the second group, that enables identification of a first cluster and a second cluster to which each of the one or more common entities belongs.

2. The information processing system according to claim 1,

wherein the associator is configured to, with regard to one or more specified clusters among the multiple first clusters, associate the piece of first cluster feature data of each of the one or more specified clusters with the piece of second cluster feature data of the second cluster including a same entity, and

wherein the one or more specified clusters are one or more first clusters each of which includes at least one of the one or more common entities among the multiple first clusters.

3. The information processing system according to claim 2,

wherein the associator is configured to:

with regard to one or more non-common first clusters and one or more non-

common second clusters, determine relationship between each of the one or more non-common first clusters and each of the one or more non-common second clusters based on the piece of first cluster feature data of each of the one or more non-common first clusters and the piece of second cluster feature data of each of the one or more non-common second clusters; and

associate the piece of first cluster feature data of each of the one or more non-common first clusters with at least one piece of second cluster feature data of at least one of the one or more non-common second clusters based on the relationship determined,

wherein the one or more non-common first clusters are one or more remaining clusters that are the multiple first clusters other than the one or more specified clusters, and

wherein the one or more non-common second clusters are one or more remaining clusters that are the multiple second clusters other than one or more clusters associated with the one or more specified clusters.

4. The information processing system according to claim 1 or 2,

wherein the piece of first cluster feature data defines a point of the corresponding first cluster in a first feature space,

wherein the piece of second cluster feature data defines a point of the corresponding second cluster in a second feature space,

wherein the associator is configured to perform:

a process of searching, using an evaluation function, a mapping that maps a point of each of the multiple first clusters in the first feature space into the second feature space; and

a process of associating each piece of the multiple pieces of first cluster feature data with one or more pieces of the multiple pieces of second cluster feature data based on positional relationship between a point of each of the multiple first clusters in the second feature space mapped by the mapping and a point of each of the multiple second clusters in the second feature space, and

wherein the evaluation function is designed to output a value:

with regard to a specified pair that is a pair of the first cluster and the second cluster including a same entity, that varies in a specified direction with a decrease in a distance in the second feature space between the first cluster and the second cluster constituting the specified pair; and

with regard to a pair of the first cluster and the second cluster other than the specified pair, that varies in a direction opposite to the specified direction with a decrease in a distance in the second feature space between the first cluster and the second cluster.

5. The information processing system according to claim 3,

wherein the piece of first cluster feature data defines a point of the corresponding first cluster in a first feature space,

wherein the piece of second cluster feature data defines a point of the corresponding second cluster in a second feature space,

wherein the associator is configured to perform:

a process of searching, using an evaluation function, a mapping that maps a point of each of the multiple first clusters including the one or more specified clusters in the first feature space into the second feature space;

a process of associating the piece of first cluster feature data of each of the one or more specified clusters with the piece of second cluster feature data of the second cluster including a same entity; and

a process of associating the piece of first cluster feature data of each of the one or more non-common first clusters with at least one piece of second cluster feature data of at least one of the one or more non-common second clusters based on positional relationship between a point of each of the one or more non-common first clusters in the second feature space mapped by the mapping and a point of each of the one or more non-common second clusters in the second feature space, and

wherein the evaluation function is designed to output a value:

with regard to a specified pair that is a pair of the first cluster and the second cluster including a same entity, that varies in a specified direction with a decrease

in a distance in the second feature space between the first cluster and the second cluster constituting the specified pair; and

with regard to a pair of the first cluster and the second cluster other than the specified pair, that varies in a direction opposite to the specified direction with a decrease in a distance in the second feature space between the first cluster and the second cluster.

6. The information processing system according to claim 4 or 5,
wherein the evaluation function is designed to output a value:

with regard to the specified pair, that is smaller for a shorter distance in the second feature space between the first cluster and the second cluster constituting the specified pair; and

with regard to a pair of the first cluster and the second cluster other than the specified pair, that is greater for a shorter distance in the second feature space between the first cluster and the second cluster, and

wherein the associator is configured to search the mapping that minimizes an output of the evaluation function.

7. The information processing system according to any one of claims 1 to 6,
wherein the one or more common entities are multiple common entities,
wherein the multiple first clusters are a set of clusters defined by clustering the first group so that the multiple common entities are arranged in the multiple first clusters in a dispersed manner, and

wherein the multiple second clusters are a set of clusters defined by clustering the second group so that the multiple common entities are arranged in the multiple second clusters in a dispersed manner.

8. The information processing system according to any one of claims 1 to 6,
wherein the one or more common entities are multiple common entities,
wherein the multiple first clusters are a set of clusters defined by clustering the

first group so that respective common entities of the multiple common entities belong to first clusters different from each other, and

wherein the multiple second clusters are a set of clusters defined by clustering the second group so that respective common entities of the multiple common entities belong to second clusters different from each other.

9. The information processing system according to any one of claims 1 to 8, wherein the piece of first cluster feature data is a piece of feature data generated by integrating multiple pieces of feature data of multiple entities belonging to the corresponding first cluster through a statistical process, and

wherein the piece of second cluster feature data is a piece of feature data generated by integrating multiple pieces of feature data of multiple entities belonging to the corresponding second cluster through a statistical process.

10. The information processing system according to claim 9, wherein the multiple entities are multiple consumers, and wherein the multiple pieces of feature data of the multiple entities are multiple pieces of feature data describing features of the multiple consumers.

11. An information processing method performed by a computer, the method comprising:

acquiring a first dataset related to multiple first clusters defined by clustering a first group of entities, the first dataset including, for each first cluster, a piece of first cluster feature data representing a feature of a corresponding first cluster;

acquiring a second dataset related to multiple second clusters defined by clustering a second group of entities, the second dataset including, for each second cluster, a piece of second cluster feature data representing a feature of a corresponding second cluster; and

associating each piece of multiple pieces of first cluster feature data included in the first dataset with one or more pieces of multiple pieces of second cluster feature data

included in the second dataset based on information, with regard to one or more common entities that are one or more entities common to the first group and the second group, that enables identification of a first cluster and a second cluster to which each of the one or more common entities belongs.

12. The information processing method according to claim 11,

wherein the associating includes associating, with regard to one or more specified clusters among the multiple first clusters, the piece of first cluster feature data of each of the one or more specified clusters with the piece of second cluster feature data of the second cluster including a same entity, and

wherein the one or more specified clusters are one or more first clusters each of which includes at least one of the one or more common entities among the multiple first clusters.

13. The information processing method according to claim 12,

wherein the associating includes:

determining, with regard to one or more non-common first clusters and one or more non-common second clusters, relationship between each of the one or more non-common first clusters and each of the one or more non-common second clusters based on the piece of first cluster feature data of each of the one or more non-common first clusters and the piece of second cluster feature data of each of the one or more non-common second clusters; and

associating the piece of first cluster feature data of each of the one or more non-common first clusters with at least one piece of second cluster feature data of at least one of the one or more non-common second clusters based on the relationship determined,

wherein the one or more non-common first clusters are one or more remaining clusters that are the multiple first clusters other than the one or more specified clusters, and

wherein the one or more non-common second clusters are one or more remaining clusters that are the multiple second clusters other than one or more clusters associated

with the one or more specified clusters.

14. The information processing method according to claim 11 or 12,
wherein the piece of first cluster feature data defines a point of the corresponding first cluster in a first feature space,

wherein the piece of second cluster feature data defines a point of the corresponding second cluster in a second feature space,

wherein the associating includes:

searching, using an evaluation function, a mapping that maps a point of each of the multiple first clusters in the first feature space into the second feature space;
and

associating each piece of the multiple pieces of first cluster feature data with one or more pieces of the multiple pieces of second cluster feature data based on positional relationship between a point of each of the multiple first clusters in the second feature space mapped by the mapping and a point of each of the multiple second clusters in the second feature space, and

wherein the evaluation function is designed to output a value:

with regard to a specified pair that is a pair of the first cluster and the second cluster including a same entity, that varies in a specified direction with a decrease in a distance in the second feature space between the first cluster and the second cluster constituting the specified pair; and

with regard to a pair of the first cluster and the second cluster other than the specified pair, that varies in a direction opposite to the specified direction with a decrease in a distance in the second feature space between the first cluster and the second cluster.

15. The information processing method according to claim 13,
wherein the piece of first cluster feature data defines a point of the corresponding first cluster in a first feature space,

wherein the piece of second cluster feature data defines a point of the

corresponding second cluster in a second feature space,

wherein the associating includes:

searching, using an evaluation function, a mapping that maps a point of each of the multiple first clusters including the one or more specified clusters in the first feature space into the second feature space;

associating the piece of first cluster feature data of each of the one or more specified clusters with the piece of second cluster feature data of the second cluster including a same entity; and

associating the piece of first cluster feature data of each of the one or more non-common first clusters with at least one piece of second cluster feature data of at least one of the one or more non-common second clusters based on positional relationship between a point of each of the one or more non-common first clusters in the second feature space mapped by the mapping and a point of each of the one or more non-common second clusters in the second feature space, and

wherein the evaluation function is designed to output a value:

with regard to a specified pair that is a pair of the first cluster and the second cluster including a same entity, that varies in a specified direction with a decrease in a distance in the second feature space between the first cluster and the second cluster constituting the specified pair; and

with regard to a pair of the first cluster and the second cluster other than the specified pair, that varies in a direction opposite to the specified direction with a decrease in a distance in the second feature space between the first cluster and the second cluster.

16. The information processing method according to claim 14 or 15,

wherein the evaluation function is designed to output a value:

with regard to the specified pair, smaller for a shorter distance in the second feature space between the first cluster and the second cluster constituting the specified pair; and

with regard to a pair of the first cluster and the second cluster other than

the specified pair, greater a shorter distance in the second feature space between the first cluster and the second cluster, and

wherein the associating includes searching the mapping that minimizes an output of the evaluation function.

17. The information processing method according to any one of claims 11 to 16,

wherein the one or more common entities are multiple common entities,

wherein the multiple first clusters are a set of clusters defined by clustering the first group so that the multiple common entities are arranged in the multiple first clusters in a dispersed manner, and

wherein the multiple second clusters are a set of clusters defined by clustering the second group so that the multiple common entities are arranged in the multiple second clusters in a dispersed manner.

18. The information processing method according to any one of claims 11 to 16,

wherein the one or more common entities are multiple common entities,

wherein the multiple first clusters are a set of clusters defined by clustering the first group so that respective common entities of the multiple common entities belong to first clusters different from each other, and

wherein the multiple second clusters are a set of clusters defined by clustering the second group so that respective common entities of the multiple common entities belong to second clusters different from each other.

19. The information processing method according to claim 11,

wherein the associating includes:

associating each piece of the multiple pieces of first cluster feature data with one or more pieces of the multiple pieces of second cluster feature data so that, out of a first pair that is a pair of a first cluster and a second cluster including an entity

common with each other and a second pair that is a pair of a first cluster and a second cluster not including any entity common with each other, association of the piece of first cluster feature data and the piece of second cluster feature data corresponding to the first pair is prioritized over the second pair; and

with regard at least to the second pair, determining a pair of the first cluster and the second cluster to be associated with each other based on positional relationship in a feature space between the first cluster and the second clusters determined from the piece of first cluster feature data and the piece of second cluster feature data corresponding to the second pair so as to associate the piece of first cluster feature data with the piece of second cluster feature data.

20. The information processing method according to claim 12,

wherein the associating includes, with regard to a non-specified cluster among the multiple first clusters other than the one or more specified clusters, determining a pair of the first cluster and the second cluster to be associated with each other based on positional relationship in a feature space between the first cluster and the second cluster determined from the piece of first cluster feature data and the piece of second cluster feature data so as to associate the piece of first cluster feature data with the piece of second cluster feature data.

21. A computer-readable recording medium storing a computer program that contains instructions for causing a computer to perform an information processing method according to any one of claims 11 to 20 when the computer program is executed by the computer.