

## TITLE OF THE INVENTION

DATA MEDIATION APPARATUS AND DATA MEDIATION METHOD

## CLAIM OF PRIORITY

The present application claims priority from Japanese patent application JP 2019-167132 filed on September 13, 2019, the content of which is hereby incorporated by reference into this application.

## BACKGROUND OF THE INVENTION

## 1. Field of the Invention

[0001]

The present invention relates to data mediation between a data provider and a data user.

## 2. Description of the Related Art

[0002]

In recent years, a data distribution service utilizing a platform that mediates between a person (data provider) who provides data and a person (data user) who desires data has been provided. On this platform, a data consumer can browse and search for summary information of registered data, and can order and purchase the desired data. The data provider can register retained data on a data distribution platform, and sell or rent the data. A commission fee for matching between the data provider and the data consumer is used as a profit in some data distribution services. Although relevant data has been

transacted in a specific field (for example, agricultural field), it is considered that data is transacted across fields in the future.

[0003]

On the data distribution platform, when the data user searches for data or understands data, a data catalog that describes summary information and a usage condition of the data is used. In order to encourage transaction of huge amounts of various kinds of data in various fields, it is necessary for the data user to improve discoverability and comprehension of the data, and thus, it is important to improve the quantity and quality of item names and values listed in the data catalog.

[0004]

Meanwhile, in general, the data provider creates the data catalog in many cases, and there is a need to reduce the number of man-hours for creating the data catalog from the perspective of the data provider. In particular, when the quantity and quality of the item names and the values of the data catalog are improved, it is considered that the number of man-hours for creating the data catalog is further increased.

[0005]

Thus, there is a need for a system that supports the creation of the data catalog capable of reducing the number

of man-hours required by the data provider to create the data catalog and improving the quantity and quality of the data catalog.

[0006]

As a technology of generating metadata, JP 2018-151859 A discloses a technology of a metadata generation apparatus, a metadata generation method, and a metadata generation program capable of reducing a calculation load and generating metadata of a sensor. In JP 2018-151859 A discloses that a calculation load and a communication load required for matching between a sensor and an application are suppressed at low values since a calculation load required for generating the metadata is suppressed at a low value and uniformity of the metadata is ensured even when new sensing data is used as input data by generating the metadata of the sensor based on attribute information of learning data representing a cluster to which the input data belongs among clusters of the learning data.

#### SUMMARY OF THE INVENTION

[0007]

JP 2018-151859 A does not disclose a configuration for measuring or calculating the quality of the learning data used for learning. When the technology of JP 2018-151859 A is simply applied to a data mediation system, it

is possible to generate the metadata by machine learning, but when the quality of the data used in the machine learning is low or the data having the low quality is mixed, the quality of the generated metadata is also low. This technology does not disclose a method for calculating and measuring the quality of the learning data used for generating the metadata.

[0008]

Accordingly, in the existing technology, it is not possible to measure or calculate the quality of learning data used by a data catalog generation apparatus, and it is not possible to generate or infer a data catalog having high quality based on the quality of the learning data. Therefore, an object of the present invention is to measure or calculate the quality of learning data for inferring a data catalog to be used in an apparatus that generates the data catalog, and to generate or infer a data catalog having high quality based on the quality of the learning data.

[0009]

In order to solve at least one of the above problems, a representative example of the present invention provides a data mediation apparatus including a processor, and a storage device. The storage device retains a data catalog including metadata of one or more items indicating

at least one of an outline and a usage condition of each of a plurality of data sets to be mediated, and retains a usage history of each data set, and the processor calculates an evaluation value indicating a level of quality of each data catalog based on the usage history, and generates an inference model for inferring a value of the metadata of any item included in the data catalog from at least one of a value of the metadata of another item and each data set based on the evaluation value.

[0010]

According to an aspect of the present invention, in the data mediation system, it is possible to reduce the number of man-hours required for the data provider to register the data catalog, and it is possible to create the data catalog having higher quality.

[0011]

Other objects, configurations, and effects will be made apparent in the following descriptions of the embodiments.

#### BRIEF DESCRIPTION OF THE DRAWINGS

[0012]

Fig. 1 is a block diagram illustrating a configuration of a data mediation system according to a first embodiment;

Fig. 2 is a flowchart illustrating an operation of a learning data generation unit according to the first embodiment;

Fig. 3 is a flowchart illustrating an operation of calculating an evaluation score in the learning data generation unit according to the first embodiment;

Fig. 4 is a sequence diagram of user registration in the first embodiment;

Fig. 5 is an explanatory diagram of a user management table retained by a data mediation apparatus according to the first embodiment;

Fig. 6 is a sequence diagram of data registration by a data provider according to the first embodiment;

Fig. 7 is an explanatory diagram of a data catalog management table retained by the data mediation apparatus according to the first embodiment;

Fig. 8 is a sequence diagram of data browsing, purchasing and acquisition by the data user in the first embodiment;

Fig. 9 is an explanatory diagram of a data usage history management table retained by the data mediation apparatus according to the first embodiment;

Fig. 10 is a block diagram illustrating a configuration of a data mediation system according to a second embodiment;

Fig. 11 is a sequence diagram illustrating a case where an existing data catalog is updated by a data mediation apparatus according to a third embodiment;

Fig. 12 is a flowchart illustrating an operation of a learning data generation unit according to a fourth embodiment; and

Fig. 13 is a block diagram illustrating a hardware configuration of a data mediation apparatus according to the first embodiment.

#### DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0013]

Hereinafter, embodiments of the present invention will be described in detail with reference to the drawings. The same reference numerals are assigned to portions that are common in the drawings.

First embodiment

[0014]

Configuration of data mediation system

A configuration of the data mediation system which is a target of the present invention will be described.

[0015]

Fig. 1 is a block diagram illustrating the configuration of the data mediation system according to a first embodiment.

[0016]

The data mediation system according to the first embodiment includes a data mediation apparatus 10, data provider information processing apparatuses 20 (20-1 to 20-N), and data user information processing apparatuses 30 (30-1 to 30-M). The data provider information processing apparatuses 20 and the data user information processing apparatuses 30 are connected to the data mediation apparatus 10 via a network.

[0017]

The data provider information processing apparatus 20 is an information processing apparatus used by a data provider, and retains data provided by the data provider to a data user for a fee or for free. This data is a mediation target using the data mediation apparatus 10. Meanwhile, the data user information processing apparatus 30 is an information processing apparatus used by a user of the data to be provided, and executes various kinds of processing using the provided data, for example.

[0018]

In general, the data mediation system includes a plurality of data provider information processing apparatuses 20 (data provider information processing apparatuses 20-1 to 20-N in the example of Fig. 1) and a plurality of data user information processing apparatuses

30 (data user information processing apparatuses 30-1 to 30-M in the example of Fig. 1). In the following description, when portions common to the plurality of information processing apparatuses are described, these collective common portions are simply referred to as the data provider information processing apparatuses 20 and the data user information processing apparatuses 30. As will be described later, in an actual system, one information processing apparatus may correspond to both the data provider information processing apparatus 20 and the data user information processing apparatus 30.

[0019]

The data mediation apparatus 10 includes a data catalog registration processing unit 101, a data catalog storage unit 102, a data search and catalog browse processing unit 103, a data purchase processing unit 104, a data acquisition processing unit 105, a data usage history log storage unit 106, a data catalog generation support processing unit 110, and other processing units 107. The data catalog generation support processing unit 110 further includes a data catalog inference unit 111, a learning data generation unit 112, a learning data storage unit 113, a model learning unit 114, and an inference model storage unit 115.

[0020]

First, data catalog registration processing in the data mediation apparatus 10 will be described.

[0021]

The data catalog registration processing unit 101 receives a data catalog registration request message from the data provider information processing apparatus, confirms items and values of a data catalog included in the message, and then stores the message in the data catalog storage unit 102. The data catalog registration processing unit receives a metadata inference request message from the data provider information processing apparatus 20, and delivers a part of a metadata item and values included in the message and a metadata item to be inferred to the data catalog inference unit 111.

[0022]

When the metadata inference request message is received from the data catalog registration processing unit 101, the data catalog inference unit 111 infers a value of the metadata item to be inferred based on a part of the received metadata, and returns the result to the data catalog registration processing unit. For example, when the metadata item to be inferred is a classification category of a data set, a title and an explanation of the data catalog input by the data provider are input, and candidate values of the classification category are output

as results to be inferred.

[0023]

Next, data search and catalog browse processing in the data mediation apparatus 10 will be described.

[0024]

When a data search message is received from the data user information processing apparatus 30, the data search and catalog browse processing unit 103 analyzes a query, and inquires of the data catalog storage unit 102 about whether or not there is corresponding data. The data search and catalog browse processing unit 103 creates a message including the result of whether or not there is the corresponding data and the search result when there is the corresponding data, and returns the message to the data user information processing apparatus 30.

[0025]

When a data catalog request message is received from the data user information processing apparatus 30, the data search and catalog browse processing unit 103 acquires the metadata of the corresponding data catalog from an ID of the data catalog included in the message, creates a message including items and values of the acquired metadata, and returns the created message to the data user information processing apparatus 30.

[0026]

Next, data purchase processing in the data mediation apparatus 10 will be described.

[0027]

When data purchasing request message processing is received from the data user information processing apparatus 30, the data purchase processing unit 104 inquires of the data catalog storage unit 102 about the amount of money for the corresponding data, and then requests the other processing unit 107 to perform billing processing of the data. The data purchase processing unit 104 requests the other processing unit 107 to grant access authority of the data user for the relevant data, acquires an issued access token, and returns the acquired access token to the data user information processing apparatus 30.

[0028]

Next, data acquisition processing in the data mediation apparatus 10 will be described.

[0029]

When data acquisition request message processing is received from the data user information processing apparatus 30, the data acquisition processing unit 105 confirms the access token included in the request processing. When the access is OK, the data acquisition processing unit 105 transmits a data acquisition request to the data provider information processing apparatus 20 that

retains the corresponding data, and temporarily acquires the data from the data provider information processing apparatus 20. The data acquisition processing unit returns a response from the data user information processing apparatus 30. The acquired data may be included in a response message, and the message may include a URL accessible to the data. The data user information processing apparatus 30 may access the URL later, and may download the data.

[0030]

Subsequently, processing for creating a data catalog inference model in the data mediation apparatus 10 will be described. The data catalog inference model is created and updated when the number of data catalogs retained in the data catalog storage unit 102 is changed by a certain number or more in the data catalog inference model or regularly at regular intervals. Here, the updating performed at regular intervals will be described.

[0031]

The learning data generation unit 112 collects data used for data catalog inference from the data catalog storage unit 102 at regular intervals. The collected data is stored in the learning data storage unit 113 after being processed or selected as needed. In the case of processing or selecting the data, the learning data generation unit

112 scores each data catalog based on a predetermined calculation expression by using the data usage history log, and selects data based on the score. The model learning unit 114 performs model learning by using the stored data of the learning data storage unit 113, and stores the created model in the inference model storage unit 115. The stored inference model is called and used by the data catalog inference unit 111.

[0032]

Here, the inference model is created and stored for each data catalog item to be inferred. Although a part of the data catalog is assumed as an input of the inference model, other parts may be included. For example, information such as keywords or formats extracted from the contents of the data itself may be used as the input of the inference model.

[0033]

According to the configuration described above, the data user can search, browse, purchase, and acquire the data registered by the data provider via the data mediation apparatus 10. When the data provider registers the data catalog, an inference function of the data catalog can be used. Quality of the learning data used in the data catalog inference can be evaluated based on the data usage history log, and can be selected. By doing so, it is

possible to measure or calculate the quality of the learning data used for inferring the data catalog to be used, and it is possible to generate or infer a high-quality data catalog.

[0034]

Fig. 13 is a block diagram of a hardware configuration of the data mediation apparatus 10 according to the first embodiment.

[0035]

The data mediation apparatus 10 according to the first embodiment includes a processor 1301, a memory 1302, a storage device 1303, an input device 1304, an output device 1305, and a communication device 1306.

[0036]

The processor 1301 controls the data mediation apparatus 10 according to a program stored in the memory 1302.

[0037]

The memory 1302 is, for example, a semiconductor memory, and stores a program executed by the processor 1301, data referenced to by the processor 1301, and data acquired as a result of processing executed by the processor 1301. At least a part of the programs and data stored in the storage device 1303 may be copied to the memory 1302 as needed, or the acquired data may be copied

from the memory 1302 to the storage device 1303 as needed.

[0038]

In the example of Fig. 13, a processing program 1307 is stored in the memory 1302. This program is a program for realizing processing of the data catalog registration processing unit 101, the data search and catalog browse processing unit 103, the data purchase processing unit 104, the data acquisition processing unit 105, the other processing units 107, the data catalog inference unit 111, the learning data generation unit 112, and the model learning unit 114 illustrated in Fig. 1. In the present embodiment, the processing executed by each of the units is actually executed by the processor 1301 according to a command described in the processing program 1307.

[0039]

The storage device 1303 has a non-volatile storage medium such as a flash memory or a hard disk. The data catalog storage unit 102, the data usage history log storage unit 106, the learning data storage unit 113, and the inference model storage unit 115 illustrated in Fig. 1 correspond to storage regions of the storage device 1303.

[0040]

The input device 1304 receives an input from an operator of the data mediation apparatus 10. Specifically, the input device 1304 may include, for example, a keyboard,

a button, or a pointing device. For example, at least a part of information stored in the storage device 1303 may be input via the input device 1304.

[0041]

The output device 1305 outputs information to the operator of the data mediation apparatus 10. Specifically, the output device 1305 may include, for example, an image display device.

[0042]

The communication device 1306 is a device that enables data exchange between the data mediation apparatus 10 and an external apparatus (for example, the data provider information processing apparatus 20 or the data user information processing apparatus 30) via a wired or wireless network, and may include, for example, a local area network (LAN) adapter. For example, at least a part of the information stored in the storage device 1303 may be input via the communication device 1306, or at least a part of the information generated as a result of the processing according to the processing program 1307 may be output via the communication device 1306.

[0043]

Although the data catalogs are concentrically managed in the above description, the data catalogs may be managed in a distributed manner. Although the data

provider information processing apparatus 20 and the data user information processing apparatus 30 are separately described, a certain person or organization may serve as both the data provider and the data user. For example, the data provider information processing apparatus 20 that provides certain data may operate as the data user information processing apparatus 30 for other data. Although it has been described in this example that the data mediation apparatus 10 acts on behalf of data acquisition from the data provider information processing apparatus 20, the data provider information processing apparatus 20 and the data user information processing apparatus 30 may directly exchange and acquire the data without using the data mediation apparatus 10. Although it has been described that the data provider creates and registers the data catalog, the data provider may create the data catalog of the data requested by the data user, and may provide the data in the data catalog to the data user via the data mediation apparatus 10.

[0044]

Description of operation of learning data generation unit

An operation of the learning data generation unit 112 of the present invention will be described.

[0045]

Fig. 2 is a flowchart illustrating the operation of

the learning data generation unit 112 according to the first embodiment.

[0046]

A trigger for starting the processing of the learning data generation unit 112 may occur, for example, regularly at regular intervals or when the number of registered data catalogs is changed by a certain number or more. When the start trigger occurs, the processing is executed in the following order.

[0047]

First, the learning data generation unit 112 acquires data of a set of data catalogs required for the metadata inference model from the data catalog storage unit 102 (S101). For example, when the title and explanation of the data catalog are used for inferring the classification category, the classification category corresponding to the title and explanation is acquired.

[0048]

Subsequently, the learning data generation unit 112 acquires a data usage history regarding all data sets from the data usage history log storage unit 106 (S102). A data structure of the data usage history will be described later (see Fig. 9).

[0049]

Subsequently, the learning data generation unit 112

calculates an evaluation score indicating a level of quality of the data catalog based on the data usage history corresponding to each data and a browsing history of the data catalog corresponding to each data (S103). A calculation flow of the evaluation score will be described later (see Fig. 3). The learning data generation unit 112 performs this processing on all data. Through this series of processing, a correspondence table of each data and the evaluation score corresponding to each data is created.

[0050]

Subsequently, the learning data generation unit 112 acquires data in the order of evaluation scores until the number of data reaches a target data number (S104). At this time, the data is acquired in descending order of the evaluation scores (that is, in descending order of levels of quality of the data catalogs). Accordingly, the data included in the high-quality data catalog is selected as the learning data. Here, when the data is simply acquired in the order of the evaluation scores, since targets (metadata to be inferred such as classification category or tag) of the acquired data are biased, a ratio of the targets to all original data is adjusted and acquired so as to approach (preferably, so as to become substantially the same) a ratio of the targets to the acquired learning data.

[0051]

Specifically, as will be described later (see Fig. 7), the data catalog includes one or more values for each item such as data set title, data set explanation, classification category, and tag. For example, it is assumed that the values of the classification categories included in all the retained data catalogs are any of "agriculture", "finance", and "travel" and appearance rates thereof are A%, B%, and C%. This means that among all the data catalogs, A% of data catalogs include "agriculture" as the value of the metadata of the classification category, B% of data catalogs include "finance" as the value of the metadata of the classification category, and C% of data catalogs include "travel" as the value of the metadata of the classification category.

[0052]

In this case, when the data is acquired in S104, the learning data generation unit 112 acquires the data such that the appearance rates of the values of the classification categories in the acquired data approaches A%, B%, and C% described above. For example, when an appearance rate of a specific value is high when the data is faithfully obtained according to the evaluation score, among the data corresponding to the specific value, the data having a high evaluation score may be added and acquired from the remaining data corresponding to another

value without acquiring the data having a low evaluation score. The same applies to the metadata of the items other than the classification category. Accordingly, it is possible to reduce the bias of the learning data, and it is possible to generate a highly accurate inference model.

[0053]

Although it has been described in the above example that the target data number is set and the data is acquired in the order of the evaluation scores, a threshold value may be set to the evaluation score, and the data having the evaluation score equal to or greater than the threshold value may be acquired.

[0054]

Finally, the learning data generation unit 112 stores the corresponding data catalog in the learning data storage unit (S105).

[0055]

Thereafter, the model learning unit 114 generates the inference model for inferring the value of the metadata of any item in the data catalog by learning. For example, the model learning unit 114 may learn the inference model for inferring the value of the metadata of the item of "classification category" from the values of the metadata of the items of "data set title" and "data set explanation" included in the data catalog. At this time, the model

learning unit 114 may learn an inference model by using the learning data including, as explanatory variables, the contents of the data set itself instead of (or in addition to) the values of the metadata of the items of "data set title" and "data set explanation". Similarly, it is possible to learn an inference model using not only "classification category" but also any item as the target.

[0056]

The model learning unit 114 also learns the inference model based on the evaluation scores calculated as described above. Specifically, as described above, the model learning unit 114 of the present embodiment learns the inference model by using, as the learning data, at least a part of the data catalog selected based on the evaluation scores and the data sets corresponding to the data catalog. Meanwhile, as another example, the model learning unit 114 may learn the inference model by using the learning data including, as the explanatory variables, the evaluation scores (see a fourth embodiment).

[0057]

Through the above processing, the learning data generation unit 112 can acquire the data from the data catalog storage unit 102, can select the data based on the evaluation score calculated by using the data usage history, can acquire the learning data having a good

evaluation score, and can store the acquired learning data.

[0058]

Description of evaluation score calculation operation in learning data generation unit

Although the evaluation score of each data is calculated in processing S103 of the learning data generation unit 112, an evaluation score calculation operation will be described below.

[0059]

Fig. 3 is a flowchart illustrating an operation of calculating the evaluation score in the learning data generation unit 112 according to the first embodiment.

[0060]

First, the learning data generation unit 112 acquires a field of an organization to which the provider of the corresponding data affiliates (S201). Identification information of the provider of the corresponding data is acquired from the data catalog. The learning data generation unit 112 specifies a user ID of the corresponding data provider, and acquires an affiliation field (here, referred to as a field a) of user management information matching the user ID by referring to a user management table. The user management table will be described later (see Fig. 5).

[0061]

Subsequently, the learning data generation unit 112, from the data usage history log storage unit 106, a search list display number  $S_x$ , a data browsing number  $R_x$ , and a data purchase number  $B_x$  for each affiliation field  $x$  ( $x = 1, \dots$ , and  $L$ ) (S202). Here, the data purchase number  $B_x$  is the number of times each data set is purchased by the user. The data browsing number  $R_x$  is the number of times the data catalog corresponding to each data set is browsed by the user. The search list display number  $S_x$  is the number of times each data set is displayed as the search result of the user. For example, the user searches for the data set. When any of the data sets displayed as the search results is selected, the user displays the data catalog corresponding to the selected data set. The user browses the data catalog, and purchases the corresponding data when it is determined that the corresponding data meets the needs of the user. When such operations are performed, these operations are recorded as a history. These values can be calculated by collecting and counting the corresponding data from the data usage history log storage unit.

[0062]

Subsequently, the learning data generation unit 112 calculates the evaluation score according to a predetermined algorithm or a function based on the search

list display number  $S_x$ , the data browsing number  $R_x$ , and the data purchase number  $B_x$  ( $x = 1, 2, \dots$ , and  $L$ ) calculated in processing S202. Here, when a score calculation function based on evaluation (that is, evaluation calculated from data usage histories of users in the same field) from users in the same field is  $F1$  and a score calculation function based on evaluation from users in different fields (that is, evaluation calculated from data usage histories of users in the different fields) is  $F2$ , the score of the corresponding data can be calculated by the following Expression (1).

[0063]

$$F1(S_a, R_a, B_a) + F2(S_x, R_x, B_x) (x \neq a) \dots (1)$$

[0064]

Here, the score calculation function is divided into  $F1$  based on the evaluation from the users in the same field and  $F2$  based on the evaluation from the users in the different fields, and a comprehensive evaluation score is calculated by combining these functions. At this time, the evaluation from the users in the different fields can be emphasized by increasing a weight of  $F2$ . Since it is considered that the users in the same field are easy to acquire appropriate data based on prior knowledge and experience even though the quality of the catalog is low and it is considered that whether or not the users in the

different fields can acquire appropriate data is likely to be influenced by the quality of the catalog, it is considered that the evaluation based on the usage histories of the users in the different fields strongly reflects the quality of the actual data catalog.

[0065]

However, such an idea is an example. As another example, a weight of F1 may be increased when it is desired to emphasize the evaluation from the users in the same field. Alternatively, the learning data generation unit 112 may calculate the evaluation score regardless of the field of the user.

[0066]

The contents of the score calculation functions F1 and F2 are not limited. In general, setting is performed such that as the search list display number  $S_x$  becomes larger, the evaluation becomes higher, as the data browsing number  $R_x$  becomes larger, the evaluation becomes higher, and the data purchase number  $B_x$  becomes larger, the evaluation becomes higher. Such setting is based on the assumption that as these values become larger, the data catalog becomes more useful (that is, the quality becomes higher). Accordingly, the quality of the data catalog is objectively evaluated. In the score calculation functions F1 and F2, all of the search list display number  $S_x$ , the

data browsing number  $R_x$ , and the data purchase number  $B_x$  may not be necessarily used. For example, the evaluation score may be calculated based on any one or two of these numbers.

[0067]

Through the above processing, the evaluation score of each data can be calculated by using the data usage history log.

[0068]

Other evaluation score calculation methods

[0069]

The score calculation method is merely an example, and another score calculation method may be adopted as long as the score calculation method leads to the "quality" of the data catalog. For example, when a system in which the data user evaluates the data is provided, the evaluation score of the data may be calculated by using the evaluation value (that is, the evaluation value input by the data user). However, the evaluation from the data user may include not only the evaluation regarding the quality of the catalog but also the evaluation regarding the quality of the acquired data itself. Thus, for example, the evaluation of the quality of the data and the evaluation of the quality of the catalog may be distinguished and input by the data user. Accordingly, the feeling of the data

user is reflected in the generation of the inference model.  
[0070]

Alternatively, the score may be calculated by using the contents of the data catalog. For example, in the case of a ratio of the values entered in any input item to the items of the data catalog or the items (for example, tag) of the data catalog in which a plurality of values can be input is used, the score may be calculated by using the number of tags to be assigned. Specifically, for example, as the number of values input to the item such as the tag in which any number of values can be input by the data provider becomes larger, the data provider becomes more enthusiastic in generating the data catalog. Accordingly, it may be estimated that the quality of the generated data catalog is high, and thus, the evaluation score may be calculated such that as the number of input values becomes larger, the evaluation score becomes larger.

[0071]

Sequence of user registration

In the first embodiment, in order to evaluate the quality of the data catalog corresponding to the corresponding data, an evaluation value from each user is calculated by using the usage histories from the users in the same field and the different fields. Here, in order to specify whether the user affiliates to the same field or

the different field as or from the field of the provider, it is necessary to acquire information on which field the provider and the user affiliate to. In the present embodiment, the acquisition of the information is realized by adding a relevant field of the organization to registration information in the case of the user registration. Hereinafter, a sequence of user registration will be described.

[0072]

Fig. 4 is a sequence diagram of the user registration in the first embodiment.

[0073]

When the data provider or the data user perform the user registration, a user registration request message is transmitted to the data provider information processing apparatus 20 or the data user information processing apparatus 30 to the data mediation apparatus 10 (S401). The user registration message includes a username, contact information, an affiliation organization, and a relevant field of the organization.

[0074]

The data mediation apparatus 10 confirms the user registration request message, and determines that the user registration succeeds when the registration information is not defective and can be registered, determines that the

user registration fails when the registration information is defective and cannot be registered, and sends a user registration completion notification message including such a status to a transmission source of a user registration message (S402). The user registration completion notification message includes the username, the contact information, the affiliation organization, and echo information on the relevant field of the organization for confirmation. When the registration information can be registered, the data mediation apparatus 10 adds the received registration information to the user management table (to be described later).

[0075]

User management table

The user management table for managing the users (data provider and data user) who use a service of the data mediation apparatus 10 will be described.

[0076]

Fig. 5 is an explanatory diagram of a user management table 500 retained by the data mediation apparatus 10 according to the first embodiment.

[0077]

The user management table 500 is retained in the storage device 1303 of the data mediation apparatus 10. For example, the user management table 500 may be included

in the other processing units 107.

[0078]

The user management table 500 includes columns of a username 501, an affiliation organization 502, contact information 503, and a relevant field 504 of the affiliation organization. The relevant field of the affiliation organization 504 is referred to in order to specify the field to which each user affiliates. The information received in the case of the user registration sequence is stored in each column. In this embodiment, both the data user and the data provider are managed as the users in the same table without being distinguished. Thus, the user management table 500 of the present embodiment has a common format for the data user and the data provider. However, the format of the user management table 500 is not necessarily limited thereto. For example, when there is a difference between the information acquired by the data provider and the information acquired by the data user, the user management tables for the data provider and the data user may be separately managed.

[0079]

With the user registration sequence processing and the user management table, a data mediator can acquire and utilize the relevant fields of the organizations to which the data provider and the data user affiliate.

[0080]

Sequence of data registration

Next, a sequence of data registration in the first embodiment will be described. Here, a case where values of items A to L are decided by using the inference function of the item value of the data catalog and the data is registered will be described.

[0081]

Fig. 6 is a sequence diagram of the data registration by the data provider in the first embodiment.

[0082]

First, the data provider information processing apparatus 20 transmits a data catalog metadata inference request message (A) to the data mediation apparatus (S601). The data catalog metadata inference request message includes a designation of an item to be inferred and a set of an item name that is input information used for inference and a value of the item.

[0083]

For example, when the classification category of the data set is inferred from the title of the data set and the explanation of the data set, the data catalog metadata inference request message specifies "classification category" as the item to be inferred, and includes, as the input information, the items of "data set title" and "data

set explanation" and the values thereof.

[0084]

When the data mediation apparatus 10 receives the data catalog metadata inference request message, the data catalog registration processing unit 101 within the data mediation apparatus 10 issues an inference request to the data catalog inference unit 111. The data catalog inference unit 111 outputs candidates of the value of the target item to be inferred by using the model for inferring the target item (the model for inferring the value of "classification category" from the values of "data set title" and "data set explanation" in the above example), and returns the result to the data catalog registration processing unit 101. The data catalog registration processing unit 101 generates a data catalog metadata inference response message (A) including the item A to be inferred and the candidate values thereof, and transmits the generated message to the data provider device (S602).

[0085]

The data mediation apparatus 10 exchanges the data catalog metadata inference request and the response for the items A to L (S603 and S604). For example, the value of "classification category" is first inferred from the values of "data set title" and "data set explanation", and the value of "tag" is then inferred from the values of "data

set title" and "data set explanation". In this manner, the inferences for a desired target are sequentially performed. The set of the item and the value within the data catalog metadata inference request message may be different for each item. For example, item 1 : value 1 and item 2 : value 2 may be used for the inference of the item A, and item 1 : value 1, item 2 : value 2, and item 3 : value 3 may be used for the inference of the item L.

[0086]

When the metadata inference requests and the responses are issued to the plurality of items as in this embodiment, the already acquired response result of the inference may be used for the metadata inference request of another item. For example, when the data catalog metadata inference request of the item B is issued, the inference value A of the item A may be used as the input. Although not illustrated in Fig. 6, an estimated probability may be included in the data catalog metadata inference response message in addition to the value inferred by the model. By doing so, the data provider information processing apparatus 20 can automatically determine whether or not to accept the inference value by using the estimated probability.

[0087]

The data provider information processing apparatus

20 acquires the inference values for the items A to L, generates the data catalog registration request message when the metadata of the data catalog is generated for the item having no the inference function or without using the inference function, and transmits the generated data catalog registration request message to the data mediation apparatus 10 (S605).

[0088]

The data mediation apparatus 10 confirms the contents of the received data catalog registration request message, determines whether or not there is a problem in the content, and transmits a data catalog registration response message to the data provider information processing apparatus 20 (S606). The data catalog registration response message includes information indicating whether or not the registration information can be registered and includes the registered content when the registration information can be registered. When the registration information can be registered, the data mediation apparatus 10 adds the content of the received data catalog to a data catalog management table will be described later.

[0089]

Through the above processing, the data mediating apparatus 10 can present the inference result to the data

provider information processing apparatus 20 by using the metadata inference function of the data catalog, the data provider information processing apparatus 20 can issue a registration request of the data catalog by using the inference result, and the data mediation apparatus 10 can register the data catalog.

[0090]

Data catalog management table

Next, the data catalog management table will be described.

[0091]

Fig. 7 is an explanatory diagram of a data catalog management table 700 retained by the data mediation apparatus 10 according to the first embodiment.

[0092]

The data catalog management table 700 is retained in the storage device 1303 of the data mediation apparatus 10. For example, the data catalog management table 700 may be included in the data catalog storage unit 102.

[0093]

The data catalog management table 700 is added and updated by using the item name and the value of the metadata of the data catalog acquired according to the sequence of the data registration. The metadata includes, for example, information regarding at least one of a

summary and a usage condition of the data set corresponding to the data catalog. The data catalog may include information regarding other data sets.

[0094]

The data catalog management table 700 includes a data set ID 701, a data provider ID 702, a data set title 703, a data set explanation 704, a classification category 705, a tag 706, and a data set access URL 707. Here, although the data catalog management table includes seven items, a table in which other items are added may be used.

[0095]

The data set ID 701 is an ID uniquely assigned in the data mediation system, and is used to identify the data set. The data set ID 701 is decided and assigned by the data mediation apparatus 10 when the data set is registered in the data catalog management table 700. The data set ID 701 may be any ID as long as the data mediation system can avoid duplication. For example, the data set ID may be incremented one by one whenever the data set is added, or a hash value may be used.

[0096]

The data provider ID 702 indicates an identifier of the data provider who provides the corresponding data. The data provider ID 702 may be, for example, a username. The data set title 703 indicates a title name of the data to be

registered. The data set explanation 704 indicates an explanation text of the data to be registered. The classification category 705 is an item for classifying the data into a plurality of pieces in the data mediation system, and indicates, for example, to which field or category the data to be registered corresponds. The tag 706 indicates a keyword related to data, and a plurality of values can be registered. The data set access URL 707 is an access URL of the corresponding data set.

[0097]

The data registered in the data mediation system can be managed by the data catalog management table 700 described above. As will be described later, the data user can search and browse the data by using the data catalog management table 700.

[0098]

Sequence of searching, browsing, purchasing, and evaluating of data

Next, a sequence when the data user searches for the data by using the data mediation system, browses the data catalog, purchases the data, and acquires the data will be described.

[0099]

Fig. 8 is a sequence diagram of the browsing, purchasing, and acquisition of the data by the data user in

the first embodiment.

[0100]

First, the data user inputs a keyword via the data user information processing apparatus 30, and transmits a data catalog search message requesting the search of the data catalog to the data mediation apparatus 10 (S801).

[0101]

The data mediation apparatus 10 receives the data catalog search message, extracts the corresponding data by using the data catalog management table 700, and transmits, as a data catalog search result message, the result to the data user information processing apparatus 300 (S802). The search result message includes a part or all of the data catalog information corresponding to the extracted data set.

[0102]

The data user browses the data catalog search result, and extracts data of interest. Here, as an example, a case where two pieces of data of interest are extracted will be described. The data user transmits a data catalog browsing request message for one of the pieces of data of interest via the data user information processing apparatus (S803). The data browsing request message includes a data set ID of the data of interest obtained in the search result message.

[0103]

When the data catalog browsing request message is received, the data mediation apparatus 10 confirms the message content, acquires the data set ID, acquires the data catalog information corresponding to the corresponding data set ID by using the data catalog management table 700, and transmits the information to the data user information processing apparatus (S804). This exchange is repeated by the number of pieces of data of interest (S805 and S806).

[0104]

When there is data desired to be purchased after the data catalog is browsed, the data user transmits the data purchasing request message to the data mediation apparatus via the data user information processing apparatus 30 (S807). Information regarding payment is included in the data purchasing request message.

[0105]

The data mediation apparatus 10 determines whether or not the data can be purchased when the data purchasing request message is received, and executes payment processing by using the information regarding the payment when the data can be purchased. When the payment processing is completed, the data mediation apparatus 10 notifies the data user information processing apparatus 30 of the result of the payment processing (S808). When the

payment is performed with no problem, the access token required to acquire the corresponding data is issued, and the access token is added to the notification.

[0106]

Although it has been described in this embodiment that the determination of whether or not the data can be purchased and access control of the data acquisition are performed by the data mediation apparatus 10, the determination and the control may be performed by the data provider information processing apparatus 20. In this case, the data purchasing request message may be transferred to the data provider information processing apparatus 20. The data provider information processing apparatus 20 may decide whether or not the data can be purchased, may transmit the result to the data mediation apparatus 10, and may transfer the result to the data user.

[0107]

The data user information processing apparatus 30 transmits a data acquisition request message including the data set ID to be acquired and the access token to the data mediation apparatus 10 (S809).

[0108]

When the data acquisition request is received, the data mediation apparatus 10 confirms the access token, and transmits the data acquisition request to the data provider

information processing apparatus 20 that retains the corresponding data when the data can be accessed (S810). Thereafter, when a data acquisition request response (including the data) is acquired from the data provider information processing apparatus 20 (S811), the data mediation apparatus 10 transmits the data acquisition request response including the acquired data to the data user information processing apparatus 30 (S812).

[0109]

Through the above processing, the data user can search for the data by using the data mediation system, can browse the data catalog, can purchase the data, and can acquire the data.

[0110]

When the data mediation system has a mechanism for posting evaluation regarding the data purchased by the data user, the data user may post the evaluation (for example, a quantitative score or a qualitative evaluation result) in the acquisition and use of the data to the data mediation apparatus 10 (S813). In this case, the data mediation apparatus 10 may manage the evaluation result for each data, and may add the evaluation result to the data catalog. In this case, the evaluation value may be used for the evaluation score calculation described in Fig. 3.

[0111]

The data mediation system reflects the log in the procedure of the above processing in the data usage history management table to be described later.

[0112]

Data usage history management table

An example of the data usage history management table will be described.

[0113]

Fig. 9 is an explanatory diagram of a data usage history management table 900 retained by the data mediation apparatus 10 according to the first embodiment.

[0114]

The data usage history management table 900 is retained in the storage device 1303 of the data mediation apparatus 10. For example, the data usage history management table 900 may be included in the data usage history log storage unit 106.

[0115]

The data usage history management table 900 records and manages logs of various events such as searching, catalog browsing, purchasing, acquiring, and evaluating for each data set.

[0116]

The data usage history management table 900 includes a data set ID 901, an event log type 902, a username 903,

an access user affiliation field 904, an event occurrence time 905, and detailed information 906.

[0117]

The data set ID 901 is an ID for identifying the data set. As an example, a data set of which the data set ID 901 is Dataset00001 is displayed in Fig. 9.

[0118]

The event log type 902 indicates a type of an event occurred. For example, the event log type 902 includes values such as search list display, catalog browsing, data purchasing, data acquiring, or data evaluating. The username 903 is a name of the user who accesses the corresponding data.

[0119]

The access user affiliation field 904 indicates the field of the affiliation organization of the user who accesses the corresponding data.

[0120]

The event occurrence time 905 indicates a time at which an event related to data usage occurs.

[0121]

The detailed information 906 indicates detailed information on various events. For example, when the corresponding data set is displayed in a search list, the detailed information 906 may include a search expression

(keyword) used for the search and an order of the corresponding data set in the list displayed as the search result. Alternatively, for example, when the event is the catalog browsing, the detailed information 906 may include a browsing time.

[0122]

The data usage history management table 900 is updated whenever various events occur. The evaluation score is calculated for the data set by using this data usage history. The evaluation score is calculated mainly based on a count value (the number of times the searching is performed, the number of times the catalog is browsed, and the number of times the catalog is purchased) obtained by counting the number of times of various event logs for each access user affiliation field.

[0123]

Although it has been described in the present embodiment that a value obtained by counting the number of times is used to calculate the evaluation score, detailed information of the data usage history may be used. For example, the order in the search list may be considered, the browsing time of the catalog may be considered, or the data evaluation value may be considered.

[0124]

The data usage history management table 900 can

manage a history of events regarding the data use occurred in the data mediation apparatus 10, and can evaluate the data.

[0125]

According to the first embodiment, it is possible to select the data used for inferring the item value of the data catalog by using the usage history of the data. Since the model of the data catalog item value inference can be constructed by using the selected data, the data catalog item value can be inferred with higher quality.

[0126]

Modification example of first embodiment

It has been described in the first embodiment that when the learning data is generated, the model is constructed by supervised learning by calculating the evaluation score of each data, selecting data having a high evaluation score, and using only the data having the high evaluation score as the learning data. However, in this case, since the number of learning data is reduced by the selection, there is a concern that inference accuracy is accordingly reduced.

[0127]

Alternatively, a model may be constructed by using data having a low evaluation score by semi-supervised learning. For example, the data mediation apparatus 10 may

construct a model by semi-supervised learning by setting supervised data to which metadata to be inferred is given to the data having the high evaluation score, setting unsupervised data to metadata to be inferred is not given to the data having the low evaluation score, and using the supervised data and the unsupervised data. By doing so, it is possible to construct a highly accurate model even when the total number of data is small.

#### Second embodiment

[0128]

Next, a second embodiment of the present invention will be described. Units of the system according to the second embodiment have the same functions as the functions of the units according to the first embodiment denoted by the same reference numerals except for differences to be described below, and thus, the description thereof will be omitted.

[0129]

Although it has been described in the first embodiment that the learning data is generated and selected by using the data catalog registered within the data mediation system in the first embodiment, the learning data may be collected from outside the data mediation system. In the second embodiment, the learning data is generated

and selected by collecting data catalogs in other data mediation systems in addition to the data catalogs within the data mediation system.

[0130]

Fig. 10 is a block diagram illustrating a configuration of the data mediation system according to the second embodiment.

[0131]

There are differences from the first embodiment in that the data mediation apparatus 10 includes an external learning data acquisition unit 116 and the data mediation apparatus can communicate with other data mediation apparatuses 11 via a network.

[0132]

The data mediation apparatus 10 according to the second embodiment can be realized by a computer system illustrated in Fig. 13 as in the first embodiment. However, the processing program 1307 according to the second embodiment includes a program for realizing processing of the external learning data acquisition unit 116. The communication between the data mediation apparatus 10 and the other data mediation apparatus 11 is performed via the network connected to the communication device 1306.

[0133]

The learning data generation unit 112 acquires a set of data catalogs retained by the data catalog storage unit 102, acquires a usage history log for each data catalog retained by the data usage history log storage unit 106, and acquires a set of data catalogs and a data usage history retained by the other data mediation apparatus 11 via the external learning data acquisition unit 116. Here, there may be a plurality of other data mediation apparatuses 11. The other data mediation apparatus 11 may be an open data catalog site or may be a data mediation apparatus that handles data of a specific field.

[0134]

A method of calculating the evaluation score when the learning data generation unit 112 selects the learning data may be the same or may be different between a case where the learning data retained by the host data mediation apparatus 10 is used as the target and a case where the learning data acquired from the other data mediation apparatus 11 is used as the target. For example, when the data usage history capable of being acquired from the other data mediation apparatus 11 is only the number of browsed data, the scoring of the data catalog retained by the other data mediation apparatus 11 is calculated by using only the number of browsed data.

[0135]

When the learned pieces of data acquired from the host data mediation apparatus 10 and the other data mediation apparatus 11 is selected, the learning data generation unit 112 stores these pieces of data in the learning data storage unit 113. Thereafter, an operation of learning the model by using the learning data, storing the inference model in the inference model storage unit, and inferring the data catalog by using the inference model is the same as the operation in the first embodiment.

[0136]

According to the second embodiment, it is possible to collect and select the learning data of more diverse data catalogs which are not limited to the data catalogs of the host data mediation apparatus. It is possible to infer data catalogs corresponding to more diverse data by using the learning data, and it is also possible to improve the inference accuracy of the data catalogs by increasing the number of learning data.

Third embodiment

[0137]

Next, a third embodiment of the present invention will be described. Units of the system according to the third embodiment have the same functions as the functions of the units according to the first or second embodiment

denoted by the same reference numerals except for differences to be described below, and thus, the description thereof will be omitted.

[0138]

As the number of data registered in the data mediation system is increased and the usage of the registered data is increased, performance of the model used in the data catalog inference function is improved. For example, although one type of tag is recommended even for the same data set in the case of the data registration, three types of tags may be currently recommended. Thus, it is desirable that the data catalog is appropriately updated in the data mediation system.

[0139]

It has been described in the first embodiment that the data catalog inference function is operated only when the data provider registers the data. In the third embodiment, the data catalog inference function is operated in the existing data catalog not only when the data is registered but also when the model is regularly updated.

[0140]

In the third embodiment, when the model of the data catalog inference is updated, the items and values of the existing data catalogs are also inferred.

[0141]

Fig. 11 is a sequence diagram illustrating a case where the existing data catalog is updated by the data mediation apparatus 10 according to the third embodiment. [0142]

The data catalog inference unit 111 of the data mediation apparatus 10 performs inference processing on the existing data catalog stored in the data catalog storage unit 102 at a timing when the model is updated. When the inferred value and the value of the existing data catalog are different, the data mediation apparatus 10 issues a data catalog metadata recommendation message, and transmits the data catalog metadata recommendation message to the data provider information processing apparatus 20 (S1101). The data catalog metadata recommendation message includes a target data set ID, a pair (for example, item A : inferred value A) of a metadata item to be changed and an existing value of the item, and a recommendation value (for example, item A : inference value A') of the item.

[0143]

When the data catalog metadata recommendation message is received, the data provider information processing apparatus 20 determines whether to accept the recommendation. When the recommendation is accepted, the data provider information processing apparatus 20 transmits a data catalog change request message to the data mediation

apparatus 10 (S1102). The data catalog change request message includes a target data set ID, a pair (item A : inference value A) of an item and a value before the change, and a pair (for example, item A : inference value A') of an item and a value after the change.

[0144]

When the data catalog request message is received, the data mediation apparatus 10 updates the value of the data catalog management table 700 corresponding to the corresponding data set ID 701. When the data catalog management table 700 is updated, the data mediation apparatus 10 generates a data catalog change response, and transmits a data catalog change response message (S1103).

[0145]

The sequence of updating the data catalog described above may be executed whenever the model used by the data catalog inference unit 111 is updated. The processing is executed in this manner, and thus, the existing data catalog can also be updated by using the data catalog inference unit.

[0146]

According to the third embodiment, it is possible to update the data catalog to a data catalog having higher quality in the data mediation system.

Fourth embodiment

[0147]

Next, a fourth embodiment of the present invention will be described. Units of the system according to the fourth embodiment have the same functions as the functions of the units according to the first to third embodiments denoted by the same reference numerals except for differences to be described below, and thus, the description thereof will be omitted.

[0148]

It has been described in the first embodiment that the evaluation score of each data set is calculated by using the data usage history and the data used for learning is selected by using the evaluation score. In the fourth embodiment, a model for inferring the item and the value of the data catalog by using the evaluation score of each data set as a feature value.

[0149]

Fig. 12 is a flowchart illustrating an operation of the learning data generation unit 112 in the fourth embodiment.

[0150]

The learning data generation unit 112 according to the fourth embodiment acquires the data from the data catalog storage unit 102 (S1201), acquires the data usage

history from the data usage history log storage unit 106 (S1202), and calculates the evaluation score of each data set (S1203). Thereafter, the learning data generation unit 112 handles the evaluation score as the feature value, and stores the evaluation score in the learning data storage unit 113 (S1204). The calculation of the evaluation score is the same as the calculation in the first embodiment.

[0151]

The model learning unit 114 learns the model by adding the evaluation score to the same explanatory variable as the explanatory variable in the first embodiment.

[0152]

When a part of the metadata and the evaluation score is input, the data catalog inference unit 111 outputs a target metadata recommendation value. Since the evaluation score is unknown for newly registered data, a value higher than an average value of the evaluation scores of the existing data catalogs is set. The value higher than the average value is set, and thus, it is considered that the data having the high evaluation score is reflected and metadata having good quality is recommended.

[0153]

Through the above processing, it is possible to construct the inference model using the evaluation score as

the feature value instead of selecting the data catalog used for the learning data by using the evaluation score, and it is possible to infer the metadata having high quality for the newly registered data catalog.

[0154]

The present invention is not limited to the aforementioned embodiments, and includes various modification examples. For example, the aforementioned embodiments are described in detail in order to facilitate easy understanding of the present invention, and are not limited to necessarily include all the described components. Some of the components of a certain embodiment can be substituted into the components of another embodiment, and the components of another embodiment can be added to the component of a certain embodiment. The components of another embodiment can be added, removed, and substituted to, from, and into some of the components of the aforementioned embodiments.

[0155]

A part or all of the aforementioned configurations, functions, processing units, and processing means may be realized by hardware by designing an integrated circuit, for example. Each of the aforementioned configurations and functions may be realized by software by interpreting and executing a program that realizes each function by the

processor. Information of programs, tables, and files for realizing the functions can be stored in a storage device such as a memory, a hard disk drive, or a solid state drive (SSD), or a computer-readable non-transitory data storage medium such as an IC card, an SD card, or a DVD.

[0156]

The drawings illustrate control lines and information lines considered to be necessary for describing the embodiments, and do not necessarily illustrate all the control lines and the information lines included in an actual product to which the present invention is applied. Almost all the configurations may be considered to be actually connected to each other.

What is claimed is:

1. A data mediation apparatus comprising:

a processor; and

a storage device, wherein

the storage device retains a data catalog including metadata of one or more items indicating at least one of a summary and a usage condition of each of a plurality of data sets to be mediated, and retains a usage history of each data set, and

the processor calculates an evaluation value indicating a level of quality of each data catalog based on the usage history, and generates an inference model for inferring a value of the metadata of any item included in the data catalog from at least one of a value of the metadata of another item and each data set based on the evaluation value.

2. The data mediation apparatus according to claim 1, wherein

the storage device retains information indicating a field of each data set and information indicating a field to which a user of each data set affiliates, and

the processor calculates a first evaluation value based on the usage history regarding the user who affiliates to the same field as the field of each data set

and a second evaluation value based on the usage history regarding the user who affiliates to a field different from the field of each data set, among the usage histories of the data sets, such that the first and second evaluation values have different weights, and calculates the evaluation value indicating the level of quality of each data catalog by combining the first evaluation value and the second evaluation value.

3. The data mediation apparatus according to claim 2, wherein

the weight of the second evaluation value is larger than the weight of the first evaluation value.

4. The data mediation apparatus according to claim 1, wherein

the evaluation value includes an evaluation value of the quality of the data catalog corresponding to the data set, which is input by the user of the data set.

5. The data mediation apparatus according to claim 1, wherein

the usage history of the data set includes at least one of the number of times the data set is purchased, the number of times the data catalog corresponding to the data set is browsed, and the number of times the data set is displayed as a search result, and

the processor calculates the evaluation value based

on at least one of the number of times the data set is purchased, the number of times the data catalog is browsed, and the number of times the data set is displayed as the search result.

6. The data mediation apparatus according to claim 5, wherein

the processor calculates the evaluation value such that as the number of times the data set is purchased becomes larger, the evaluation value becomes a value indicating that the quality of the data catalog corresponding to the data set is high, calculates the evaluation value such that as the number of times the data catalog is browsed becomes larger, the evaluation value becomes a value indicating that the quality of the data catalog is high, and calculates the evaluation value such that the number of times the data catalog is displayed as the search result becomes larger, the evaluation value becomes a value indicating that the quality of the data catalog is high.

7. The data mediation apparatus according to claim 1, further comprising:

a communication device connected to a network,  
wherein

the processor generates the inference model by using learning data including the data catalog retained in the

storage device and a data catalog acquired via the communication device.

8. The data mediation apparatus according to claim 1, wherein

the processor infers the value of the metadata included in the data catalog retained in the storage device by using the generated inference model,

outputs the inferred value of the metadata when the inferred value of the metadata is not included in the data catalog retained in the storage device, and

adds the inferred value of the metadata to the data catalog when an instruction to add the inferred value of the metadata to the data catalog is input.

9. The data mediation apparatus according to claim 8, wherein

the processor generates the inference model regularly or when the number of data catalogs retained in the storage device is changed by a predetermined number after the previous generation of the inference model.

10. The data mediation apparatus according to claim 1, wherein

the processor selects the data catalog for which it is determined that the quality is high based on the evaluation value, and generates the inference model by using, as learning data, at least a part of the selected

data catalog and a data set corresponding to the selected data catalog.

11. The data mediation apparatus according to claim 10, wherein

the processor selects the data catalog such that an appearance rate of the value of the metadata included in the selected data catalog approaches an appearance rate of the value of the metadata included in all the data catalogs before being selected.

12. The data mediation apparatus according to claim 10, wherein

the processor generates the inference model by performing semi-supervised learning by using the selected data catalog and the data set corresponding to the selected data catalog as a supervised learning data to which the metadata to be inferred is given and using an unselected data catalog and a data set corresponding to the unselected data catalog as unsupervised learning data to which the metadata to be inferred is not given.

13. The data mediation apparatus according to claim 1, wherein

the processor generates the inference model by using learning data including the evaluation value as an explanatory variable.

14. A data mediation method executed by a computer

system that includes a processor and a storage device,  
wherein

the storage device retains a data catalog including metadata of one or more items indicating at least one of a summary and a usage condition of each of a plurality of data sets to be mediated, and retains a usage history of each data set, and

the data mediation method includes

calculating, by the processor, an evaluation value indicating a level of quality of each data catalog based on the usage history, and

generating, by the processor, an inference model for inferring a value of the metadata of any item included in the data catalog from at least one of a value of the metadata of another item and each data set based on the evaluation value.