



- (51) **International Patent Classification:**
G06F 9/48 (2006.01) *G06F 17/00* (2006.01)
- (21) **International Application Number:**
PCT/US2015/037206
- (22) **International Filing Date:**
23 June 2015 (23.06.2015)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (71) **Applicant:** HEWLETT PACKARD ENTERPRISE DEVELOPMENT LP [US/US]; 11445 Compaq Center Drive West, Houston, TX 77070 (US).
- (72) **Inventors:** DAS, Mahashweta; 1501 Page Mill Rd., Palo Alto, California 94304-1100 (US). OZONAT, Mehmet Kivanc; 1501 Page Mill Rd., Palo Alto, California 94304-1100 (US).
- (74) **Agents:** DAS, Manav et al.; Hewlett Packard Enterprise, 3404 E. Harmony Road, Mail Stop 79, Fort Collins, CO 80528 (US).
- (81) **Designated States** (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM,

DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

- (84) **Designated States** (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

- as to the identity of the inventor (Rule 4.17(i))
- as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))

Published:

- with international search report (Art. 21(3))

(54) **Title:** RECOMMENDING ANALYTIC TASKS BASED ON SIMILARITY OF DATASETS

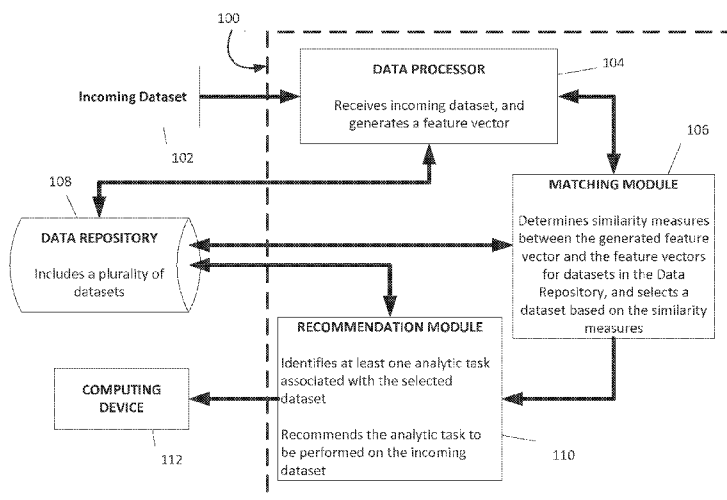


Fig. 1

(57) **Abstract:** Recommending analytic tasks based on similarity of datasets is disclosed. One example is a system including a data processor, a matching module, and a recommendation module. The data processor receives an incoming dataset via a processing system, and generates a feature vector for the incoming dataset. The matching module determines similarity measures between the generated feature vector and representative feature vectors for a plurality of datasets in a data repository, and selects at least one dataset of the plurality of datasets based on the similarity measures. The recommendation module identifies at least one analytic task associated with the selected dataset, and recommends, to a computing device via the processing system, the at least one analytic task to be performed on the incoming dataset.



RECOMMENDING ANALYTIC TASKS BASED ON SIMILARITY OF DATASETS

Background

[0001] A variety of analytic tasks may be performed on datasets. Similar datasets may be generally amenable to similar analytic tasks.

Brief Description of the Drawings

[0002] Figure 1 is a functional block diagram illustrating one example of a system for recommending analytic tasks based on similarity of datasets.

[0003] Figure 2 is a block diagram illustrating one example of a computer readable medium for recommending analytic tasks based on similarity of datasets.

[0004] Figure 3 is a flow diagram illustrating one example of a method for recommending analytic tasks based on similarity of datasets.

Detailed Description

[0005] Big data analytics have enabled data scientists to unlock strategic insights from vast troves of data by employing statistics, computer science, and engineering knowledge. Business analysts, domain experts, scientists like ornithologists, botanists, and so forth, routinely utilize data science toolkits. However, such data science toolkits are not generally equipped to recommend which technique to employ on a given dataset to maximize a return on information. For example, a financial analyst may not know whether to employ regression analysis or a classification algorithm to make stock market predictions. While an optimal choice may be dependent on a variety of factors such as user's objective, application domain, size and type of data, input of the algorithm, etc., the present disclosure provides preliminary guidance to such users.

[0006] Some existing technologies generally integrate heterogeneous geospatial data sources by determining semantic similarity between geospatial data sources that may typically involve matching attributes and attribute values between tables using a semantic similarity measure. Such techniques are generally limited to handling geographic data sources having geographic instance properties, and

may not account for data characteristics like distribution, topic, content, and application domain. Other existing methods match datasets with similar content but that are at different resolutions. Such an approach is generally limited to geographical datasets, and uses similarity measures based on topology, geometry, and semantics. Clustering techniques are generally limited to data elements within a specific dataset, as opposed to a clustering of datasets.

[0007] As disclosed herein, a repository of datasets may be partitioned based on data characteristics such as data type, size, distribution, attributes, topic, content, and application domain. When a new dataset is received, it may be matched to one or more of existing datasets in the repository. Algorithms that are known to have successfully analyzed such datasets (for example, based on historical log information) may be recommended for the new dataset. Consequently, users may rapidly conduct big data analytics in an analytics platform without having to analyze and/or apply the underlying technical and mathematical details of the analytics tasks.

[0008] For example, a data repository may include a group A of structured spatio-temporal earth science datasets such as: (i) World Daily Air Temperature from 2001-2010 by Latitude-Longitude, (ii) US Monthly Precipitation from 1901-2000 by Zipcode, and (iii) California Forest Fire from 1991-2010 by Region. The data repository may also include another group B of time-series datasets such as: (iv) New York Stock Exchange daily closing price from 1951-2000, (v) Euro daily foreign exchange rate from 2001-2014, and (vi) US Bank real estate monthly loan from 1971-2000. A new dataset may be received, where the dataset relates to data on El Nino in the Pacific between 1980 and 1990. The new dataset may be matched to one or more datasets in group A because of spatio-temporal data type, earth science topic, and seasonal pattern in data distribution. Analytics tasks performed on the datasets in group A may be identified. For example, tasks such as predicting the temperature, precipitation, forest fire of a location at a future timestamp, and so forth may be identified. Based on a similarity comparison, it may be determined that such tasks may be applicable to the dataset related to El Nino data as well. Also, for example, the size, i.e., the number of instances and the number of attributes of the datasets,

may be a good indicator of which learning algorithm may be employed on the El Nino dataset. Accordingly, analytics tasks and machine learning algorithms may be automatically recommended to a user, such as, for example, a geoscientist, who may want to extract meaningful patterns from the El Nino data.

[0009] As described in various examples herein, a recommendation system based on a similarity of datasets is disclosed. One example is a system including a data processor, a matching module, and a recommendation module. The data processor receives an incoming dataset via a processing system, and generates a feature vector for the incoming dataset. The matching module determines similarity measures between the generated feature vector and representative feature vectors for a plurality of datasets in a data repository, and selects at least one dataset of the plurality of datasets based on the similarity measures. The recommendation module identifies at least one analytic task associated with the selected dataset, and recommends, to a computing device via the processing system, the at least one analytic task to be performed on the incoming dataset.

[0010] In the following detailed description, reference is made to the accompanying drawings which form a part hereof, and in which is shown by way of illustration specific examples in which the disclosure may be practiced. It is to be understood that other examples may be utilized, and structural or logical changes may be made without departing from the scope of the present disclosure. The following detailed description, therefore, is not to be taken in a limiting sense, and the scope of the present disclosure is defined by the appended claims. It is to be understood that features of the various examples described herein may be combined, in part or whole, with each other, unless specifically noted otherwise.

[0011] Figure 1 is a functional block diagram illustrating one example of a system 100 for recommending analytic tasks based on similarity of datasets. System 100 includes a data processor 104, a matching module 106, and a recommendation module 110. The data processor 104 may receive an incoming dataset 102 via a processing system. In some examples, the data processor 104 may be communicatively linked to a source of data. For example, data processor 104 may receive incoming dataset 102 that is input via a computing device. Also,

for example, data processor 104 may receive incoming dataset 102 from a publicly available source such as a weather data center, a nutrition database, a scientific database, and so forth. The received incoming dataset 102 may include information related to weather, food, drugs, natural phenomena, health, and so forth. For example, the received incoming dataset 102 may include weather data related to El Nino.

[0012] The term “dataset”, as used herein, generally refers to a collection of data. Generally, the dataset may be in structured form. For example, the dataset may be a linked database, a tabular array, a matrix, an excel worksheet, and so forth. In some examples, the dataset may be unstructured, and may be converted to a structured dataset. In some examples, the dataset may be in semi-structured form. For example, the dataset may be a collection of log messages, snippets from text messages, messages from social networking platforms, and so forth. The term “data elements” as used herein generally refers to any data included in a dataset. For example, the dataset may be a matrix, and the data elements may be rows or columns of the matrix. Likewise, the dataset may be a collection of vectors, and the data elements may be the individual vectors.

[0013] In some examples, the data elements may also be any metadata related to a dataset. For example, the data elements may be timestamp data related to log messages in a collection of log messages. As another example, the dataset may be a matrix, and the data elements may be high-dimensional vectors representative of weather-related data. In some examples, the dataset may be a stream of data, and the data elements may be data in the stream of data. For example, the dataset may be a stream of real-time data from a live traffic database. Generally, data elements may be in a variety of different data formats. For example, the data may be textual data, including numeric and/or non-numeric data, video data, image data, audio data, and/or data in mixed format.

[0014] In some examples, the data processor 104 may generate a feature vector for the incoming dataset 102. Generally, a feature vector for a dataset is representative of salient features of an entire dataset. The feature vector for a dataset is different from a feature vector for a data element. For example, if the data set is represented in matrix form, the feature vector of the matrix may be a

vector with components representative of features of the matrix. On the other hand, a feature vector for a data element in the matrix may be a feature vector for a given column in the matrix.

[0015] In some examples, the feature vector of the matrix may be a vector with components including features such as, topic, data type, a number of column groups, a column group header, a column group type, data characteristic, data distribution, a number of rows, a number of columns, and so forth. In some examples, the topic may be earth science, finance, nutrition, and so forth. In some examples, the data type may be multivariate or univariate. In some examples, each column group may be associated with a column group header such as, for example, time, location, temperature values, stock price values, forest fire values, and so forth. In some examples, each column group may be associated with a column group data type such as, for example, discrete, continuous, integer, binary, and so forth. In some examples, the data characteristic may be spatio-temporal, spatial, temporal, and so forth. In some examples, the data distribution may be a probability distribution for the data, including, for example, distributions such as normal, pareto, Gaussian, hypergeometric, and so forth.

	Feature Vector for a dataset for World Daily Air Temperature from 2001-2010 by Latitude- Longitude	Feature Vector for a dataset for US Monthly Precipitation from 1901-2000 by Zipcode	Feature Vector for a dataset for California Forest Fire from 1991-2010 by Region	Feature Vector for a dataset for New York Stock Exchange daily closing price from 1951-2000
<i>Topic</i>	Earth Science	Earth Science	Earth Science	Finance
<i>Data Type</i>	Multivariate	Multivariate	Multivariate	Univariate
<i>No. of column groups</i>	3	3	2	2
<i>Column 1 Group Header</i>	Time	Time	Time	Time
<i>Column 1 Group Data Type</i>	Discrete	Discrete	Discrete	Discrete

<i>Column 2 Group Header</i>	Location	Location	Forest Fire Values	Stock Values
<i>Column 2 Group Data Type</i>	Discrete	Discrete	Continuous	Integer
<i>Column 3 Group Header</i>	Temperature Values	Precipitation Values		
<i>Column 2 Group Data Type</i>	Continuous	Continuous		
<i>Data Characteristic</i>	Spatio-temporal	Spatio-temporal	Temporal	Temporal
<i>Data Distribution</i>	Normal	Normal	Normal	Pareto
<i>No. of Rows</i>	3	3	3	2
<i>No. of columns</i>	650000	40000	1000	100000000

Table 1: Example Feature Vectors for Example Datasets

[0016] Table 1 illustrates example feature vectors for example datasets. For example, the dataset may be related to World Daily Air Temperature from 2001-2010 by Latitude-Longitude. Accordingly, an example feature vector and/or representative feature vector may be generated as illustrated in the second column in Table 1. For example the feature vector may be <earth science, multivariate, 3, time, discrete, location, discrete, temperature values, continuous, spatio-temporal, normal, 3, 650000>.

[0017] As another example, the dataset may be related to US Monthly Precipitation from 1901-2000 by Zipcode. Accordingly, an example feature vector and/or representative feature vector may be generated as illustrated in the third column in Table 1. For example the feature vector may be <earth science, multivariate, 3, time, discrete, location, discrete, precipitation values, continuous, spatio-temporal, normal, 3, 40000>.

[0018] Also, for example, the dataset may be related to California Forest Fire from 1991-2010 by Region. Accordingly, an example feature vector and/or

representative feature vector may be generated as illustrated in the fourth column in Table 1. For example the feature vector may be <earth science, multivariate, 2, time, discrete, forest fire values, continuous, temporal, normal, 3, 1000>.

[0019] As another example, the dataset may be related to New York Stock Exchange daily closing price from 1951-2000. Accordingly, an example feature vector and/or representative feature vector may be generated as illustrated in the fifth column in Table 1. For example the feature vector may be <finance, univariate, 2, time, discrete, stock values, integer, temporal, pareto, 2, 1000000000>.

[0020] In some examples, the data processor 104 may pre-process a given dataset of the plurality of datasets in the data repository 108 based on information theory concepts to group similar data elements in the given dataset. In some examples, the data processor 104 may employ such data pre-processing techniques to group similar columns in a dataset together and/or to reduce clutter. For example, the data processor 104 may compute mutual information for pairs of columns in a matrix representation of a dataset. The data processor 104 may group columns that have a high mutual information.

[0021] In some examples, the data processor 104 may apply one of text mining and natural language processing to extract properties of the datasets to generate representative features of datasets in the data repository 108. In some examples, the data processor 104 may associate a dataset in the data repository 108 with a feature vector representation based on techniques described herein. Table 1 illustrates example representative feature vectors for example datasets.

[0022] In some examples, datasets may be clustered based on similar column data types. For example, for 'spatio-temporal' type, there may be at least one column for time, and at least one column for location, and at least one column for numeric, categorical, and/or Boolean values. In some examples, datasets may be clustered based on similar column value distributions. For example, a dataset on yearly temperature and yearly precipitation may exhibit seasonal patterns that may be captured via similar data distributions. In some examples, datasets may be clustered based on similar data size. For example, datasets with the same number of rows and/or columns may be clustered together.

[0023] In some examples, the data processor 104 may apply such clustering techniques sequentially. For example, the datasets may be first clustered based on a general topic of content, followed by a clustering based on concepts in column headers, followed by a clustering based on taxonomy terms in column headers, followed by a clustering based on data distributions of columns, followed by a clustering based on data types, and followed by a clustering based on data size.

[0024] In some examples, the data processor 104 may process the data objects and generate structured data from unstructured data. For example, the data processor 104 may receive a collection of log messages, extract information including timestamp data, source data, keywords and/or key phrases, n-grams, video clips, audio clips, etc. from the collection, and generate a structured data set, for example, a matrix that includes such extracted information.

[0025] In some examples, the data processor 104 may apply one of text mining and natural language processing to extract properties of the datasets to generate representative features of datasets in the data repository 108. In some examples, the data processor 104 may associate a dataset in the data repository 108 with a feature vector representation.

[0026] The matching module 106 may determine similarity measures between the generated feature vector and representative feature vectors for a plurality of datasets in the data repository 108. The similarity measure for a pair of vectors is a measure of a degree of similarity of the two vectors. In some examples, the similarity measure may be a comparison of a number of overlaps between components of the vectors. In some examples, the similarity measure may be a semantic comparison of components of the two vectors. In some examples, the similarity measures may be based on at least one of data type, data size, data distribution, data attributes, topic, content, and application domain.

[0027] In some examples, the data processor 104 may be communicatively linked to the data repository 108 and/or to the matching module 106. In some examples, the data processor 104 may cluster the plurality of datasets in the data repository 108 based on the similarity measure. In some examples, datasets may be clustered based on a similar general topic of content. For example,

datasets related to earth sciences may be clustered together. In some examples, datasets may be clustered based on similar column header topics. For example, for topic 'time', column headers may be year, month, etc, and datasets may be clustered based on similar years (and/or range of years). Also, for example, for topic 'location', column headers may be zip code, city name, etc., and datasets may be clustered based on similar zip code, city name, etc. In some examples, the data processor 104 may associate a cluster of datasets in the data repository 108 with a feature vector representation.

[0028] In some examples, the matching module 106 may select at least one dataset of the plurality of datasets based on the similarity measures. In some examples, the matching module 106 may select at least one cluster of datasets based on similarity measures between the generated feature vector and the representative feature vectors. For example, as described herein, the data processor 104 may provide a feature vector representation for the incoming dataset 102 to the matching module 106. Also, for example, the data processor 104 may provide feature vector representations for clusters of datasets in the data repository 108 to the matching module 106. In some examples, the matching module 106 may retrieve the feature vector representations for datasets from the data repository 108.

[0029] In some examples, the matching module 106 may perform a course-grained matching by comparing the feature vector representation for the incoming dataset 102 with the feature vector representations for clusters of datasets and/or feature vector representations for datasets in a cluster, to select the at least one cluster based on similarity measures.

[0030] In some examples, the matching module 106 may select the at least one dataset from the selected at least one cluster based on the similarity measures between the generated feature vector and the representative feature vectors of datasets in the selected at least one cluster. For example, after a course-grained matching identifies the at least one cluster, the matching module 106 may apply a fine-grained matching based on a comparison of the feature vector representation for the incoming dataset 102 with the feature vector representations for datasets in the selected at least one cluster. Accordingly,

whereas the course-grained matching identifies a group of similar datasets in the data repository 108, the fine-grained matching identifies a similar dataset within the identified group.

[0031] In some examples, based on data attributes, the course-grained matching may be sufficient to identify the at least one dataset similar to the incoming dataset 102. Accordingly, a fine-grained matching may not be performed. For example, the data repository 108 may include datasets related to spatio-temporal earth sciences data. Data attributes may include location, latitude, longitude, zipcode, timestamp, and so forth. The incoming dataset 102 may include soil moisture data for various locations in the United States. Based on the techniques described herein, a course-grained similarity match based on a numerical comparison may be performed to identify a cluster of datasets that include data related to temperature, rainfall, and humidity for various locations in the United States. Accordingly, a course-grained matching may be sufficient for the incoming soil data. In some examples, based on data attributes, the course-grained matching may be identical to the fine-grained matching.

[0032] However, in some examples, the incoming dataset 102 may include more categorical data such as country name, city name, region, type of vegetation, etc. In such a situation, the course-grained matching based on numerical data may not be sufficient, and a fine-grained matching may be needed to account for the categorical data. For example, the data related to zip codes may have been used as a numerical data in the course-grained matching. However, the categorical nature of the incoming dataset 102 may suggest that the data related to zip codes may be utilized as a categorical data. In some examples, semantic comparisons may be performed for such categorical data.

[0033] In some examples, the recommendation module 110 may identify at least one analytic task associated with the selected dataset. For example, the recommendation module 110 may be communicatively linked to the data repository 108. In addition to the plurality of datasets, the data repository 108 may also include data related to analytic tasks performed on the datasets. For example, the data repository 108 may be a linked database where a dataset may be linked to an analytic task that has been performed on the dataset.

[0034] For example, a first dataset may relate to financial data, and the data repository 108 may also include a link to a regression analysis technique associated with the first dataset. Also, for example, a second dataset may include census data related to housing and education, based on age, gender, house value, population data, number of cars, and so forth, and the data repository 108 may include a link to analytic tasks performed on the second dataset. For example, the link may indicate that regression analyses have been performed on the second dataset based on house value with respect to a number of residents, house value with respect to local climate, house value with respect to a level of education of the owner, and so forth.

[0035] In some examples, the at least one analytic task may include one of a machine learning algorithm and an analytics algorithm. The machine learning algorithm may be supervised learning, unsupervised learning, and/or semi-supervised learning. For example, a dataset may include image data for a plurality of human faces, and a machine learning algorithm may have been employed to extract one or more features associated with the dataset. The analytics algorithm may be any algorithm that analyzes a given dataset. For example, the analytics algorithm may be a pattern processing algorithm, an anomaly detection algorithm, and so forth.

[0036] In some examples, the recommendation module 110 may recommend, to a computing device 112 via the processing system, the at least one analytic task to be performed on the incoming dataset 102. For example, the incoming dataset 102 may be new census data. Based on the techniques described herein, the matching module may identify that the new census data is similar to the second dataset described herein. Accordingly, the recommendation module 110 may recommend that a regression analysis may be performed on the new census data. For example, the recommendation module 110 may recommend that a regression analysis may be performed on the new census data for education with respect to a number of cars in the household.

[0037] In some examples, the recommendation module 110 may recommend the at least one analytic task based on log information. Log information may generally be log data related to analytic tasks that may be performed on datasets.

In some examples, the log information may be historical log information. As described herein, such log data may be stored via a link in a linked database, and associated with a dataset in the data repository 108. In some examples, log information may be extracted from a cloud-based analytics platform system. In some examples, log information may be extracted from a stack overflow.

[0038] In some examples, the recommendation module 110 may recommend an analytic task that may be inferred from analytic tasks that were performed in the past. For example, the recommendation module 110 may recommend an analytic task that may be inferred from historical log information. For example, if past analyses on the second dataset included regression analyses for education cost with respect to local temperature, then the recommendation module 110 may recommend that a regression analysis be performed on the new census data based on housing value with respect to local temperature.

[0039] In some examples, the recommendation module 110 may provide the at least one analytic task to a computing device via an interactive graphical user interface. For example, the recommendation module 110 may provide a message indicative of the analytic task to be performed. In some examples, the recommendation module 110 may provide a selectable icon to accept and/or reject the recommended analytic task. For example, a user may identify that the recommended analytic task is not appropriate for the incoming dataset 102, and may reject the recommended analytic task. In some examples, another recommended analytic task may be provided to the user.

[0040] In some examples, the recommendation module 110 may store an association between the at least one analytic task and the selected dataset in the data repository 108. Such a stored association may be utilized later to recommend analytic tasks to additional incoming datasets.

[0041] The components of system 100 may be computing resources, each including a suitable combination of a physical computing device, a virtual computing device, a network, software, a cloud infrastructure, a hybrid cloud infrastructure that may include a first cloud infrastructure and a second cloud infrastructure that is different from the first cloud infrastructure, and so forth. The components of system 100 may be a combination of hardware and programming

for performing a designated visualization function. In some instances, each component may include a processor and a memory, while programming code is stored on that memory and executable by a processor to perform a designated function.

[0042] For example, the matching module 106 may be a combination of hardware and programming to match the incoming dataset 102 to similar datasets of the plurality of datasets. Also, for example, the matching module 106 may include software programming to determine similarity measures. The matching module 106 may include hardware to physically store, for example, such determined quantities, and/or maintain a dynamically updated database that stores the similarity measures.

[0043] Likewise, the recommendation module 110 may be a combination of hardware and programming to identify and recommend analytic tasks. Also, for example, the recommendation module 110 may include programming to provide the recommended analytic task to a computing device. The recommendation module 110 may include hardware to physically store, for example, an association of the incoming dataset 102 with a recommended analytic task. Also, for example, the recommendation module 110 may include software programming to dynamically interact with the other components of system 100.

[0044] Generally, the components of system 100 may include programming and/or physical networks to be communicatively linked to other components of system 100. In some instances, the components of system 100 may include a processor and a memory, while programming code is stored and on that memory and executable by a processor to perform designated functions.

[0045] A computing device, as used herein, may be, for example, a web-based server, a local area network server, a cloud-based server, a notebook computer, a desktop computer, an all-in-one system, a tablet computing device, a mobile phone, an electronic book reader, or any other electronic device suitable for provisioning a computing resource to perform a unified visualization interface. The computing device may include a processor and a computer-readable storage medium.

[0032] Figure 2 is a block diagram illustrating one example of a computer readable medium for recommending analytic tasks based on similarity of datasets. Processing system 200 includes a processor 202, a computer readable medium 208, input devices 204, and output devices 206. Processor 202, computer readable medium 208, input devices 204, and output devices 206 are coupled to each other through a communication link (e.g., a bus).

[0033] Processor 202 executes instructions included in the computer readable medium 208. Computer readable medium 208 includes incoming dataset receipt instructions 210 to receive an incoming dataset via the processor 202.

[0034] Computer readable medium 208 includes feature vector generation instructions 212 to generate a feature vector for the incoming dataset.

[0035] Computer readable medium 208 includes representative feature vector identification instructions 214 to match, via the processor 202, representative feature vectors for a plurality of datasets in a data repository.

[0036] Computer readable medium 208 includes similarity measure determination instructions 216 to determine, via the processor 202, similarity measures between the generated feature vector and the representative feature vectors for the plurality of datasets.

[0037] Computer readable medium 208 includes dataset selection instructions 218 to select at least one dataset of the plurality of datasets based on the similarity measures.

[0038] Computer readable medium 208 includes analytic task identification instructions 220 to identify, via the processor 202, at least one analytic task associated with the selected dataset, wherein the at least one analytic task includes one of a machine learning algorithm and an analytics algorithm.

[0039] Computer readable medium 208 includes analytic task recommendation instructions 222 to recommend, to a computing device via the processor 202, the at least one analytic task to be performed on the incoming dataset.

[0040] In some examples, the similarity measure determination instructions 216 include instructions to determine the similarity measures based on a comparison of at least one of data type, data size, data distribution, data attributes, topic, content, and application domain.

[0041] Input devices 204 include a keyboard, mouse, data ports, and/or other suitable devices for inputting information into processing system 200. In some examples, input devices 204, such as a computing device, are used to receive the incoming dataset. Output devices 206 include a monitor, speakers, data ports, and/or other suitable devices for outputting information from processing system 200. In some examples, output devices 206 are used to recommend the at least one analytic task to be performed on the incoming dataset.

[0042] As used herein, a “computer readable medium” may be any electronic, magnetic, optical, or other physical storage apparatus to contain or store information such as executable instructions, data, and the like. For example, any computer readable storage medium described herein may be any of Random Access Memory (RAM), volatile memory, non-volatile memory, flash memory, a storage drive (e.g., a hard drive), a solid state drive, and the like, or a combination thereof. For example, the computer readable medium 208 can include one of or multiple different forms of memory including semiconductor memory devices such as dynamic or static random access memories (DRAMs or SRAMs), erasable and programmable read-only memories (EPROMs), electrically erasable and programmable read-only memories (EEPROMs) and flash memories; magnetic disks such as fixed, floppy and removable disks; other magnetic media including tape; optical media such as compact disks (CDs) or digital video disks (DVDs); or other types of storage devices.

[0043] As described herein, various components of the processing system 200 are identified and refer to a combination of hardware and programming configured to perform a designated visualization function. As illustrated in Figure 2, the programming may be processor executable instructions stored on tangible computer readable medium 208, and the hardware may include processor 202 for executing those instructions. Thus, computer readable medium 208 may store program instructions that, when executed by processor 202, implement the various components of the processing system 200.

[0044] Such computer readable storage medium or media is (are) considered to be part of an article (or article of manufacture). An article or article of manufacture can refer to any manufactured single component or multiple

components. The storage medium or media can be located either in the machine running the machine-readable instructions, or located at a remote site from which machine-readable instructions can be downloaded over a network for execution.

[0045] Computer readable medium 208 may be any of a number of memory components capable of storing instructions that can be executed by Processor 202. Computer readable medium 208 may be non-transitory in the sense that it does not encompass a transitory signal but instead is made up of one or more memory components configured to store the relevant instructions. Computer readable medium 208 may be implemented in a single device or distributed across devices. Likewise, processor 202 represents any number of processors capable of executing instructions stored by computer readable medium 208. Processor 202 may be integrated in a single device or distributed across devices. Further, computer readable medium 208 may be fully or partially integrated in the same device as processor 202 (as illustrated), or it may be separate but accessible to that device and processor 202. In some examples, computer readable medium 208 may be a machine-readable storage medium.

[0046] Figure 3 is a flow diagram illustrating one example of a method for recommending analytic tasks based on similarity of datasets. In some examples, such an example method may be implemented by a system such as, for example, system 100 of Figure 1.

[0047] At 300, an incoming dataset may be received via a processing system. In some examples, the process at 300 may be implemented by the data processor 104 of system 100 described with reference to Figure 1.

[0048] At 302, a feature vector for the incoming dataset may be generated. In some examples, the process at 302 may be implemented by the data processor 104 of system 100 described with reference to Figure 1.

[0049] At 304, representative feature vectors may be identified for a plurality of datasets in a data repository. In some examples, the process at 304 may be implemented by the matching module 106 of system 100 described with reference to Figure 1.

[0050] At 306, at least one cluster of datasets of the plurality of datasets may be selected based on similarity measures between the generated feature vector and

the representative feature vectors. In some examples, the process at 306 may be implemented by the matching module 106 of system 100 described with reference to Figure 1.

[0051] At 308, at least one dataset from the selected at least one cluster may be selected based on similarity measures between the generated feature vector and the representative feature vectors of datasets in the selected at least one cluster. In some examples, the process at 308 may be implemented by the matching module 106 of system 100 described with reference to Figure 1.

[0052] At 310, at least one analytic task associated with the selected dataset may be identified. In some examples, the process at 310 may be implemented by the recommendation module 110 of system 100 described with reference to Figure 1.

[0053] At 312, the at least one analytic task may be recommended to be performed on the incoming dataset. In some examples, the process at 312 may be implemented by the recommendation module 110 of system 100 described with reference to Figure 1.

[0054] In some examples, the at least one analytic task may include at least one of a machine learning algorithm and an analytics algorithm.

[0055] In some examples, identifying the at least one analytic task may be based on log information.

[0056] In some examples, the method may further include generating the clusters of datasets of the plurality of datasets based on similarity measures, and wherein the similarity measures are based on a comparison of at least one of data type, data size, data distribution, data attributes, topic, content, and application domain.

[0057] Examples of the disclosure provide a generalized system for recommending analytic tasks based on similarity of datasets. The generalized system automatically recommends an analytics task and/or machine learning algorithm to be performed on an incoming dataset, based on a similarity between the incoming dataset and an existing dataset that was amenable to such an analytic task and/or machine learning algorithm in the past. The similarity measures compare a wide range of features of the data. Also, a course-grained

matching identifies a cluster of similar datasets, while a fine-grained matching identifies a dataset from the cluster of similar datasets.

[0058] Although specific examples have been illustrated and described herein, a variety of alternate and/or equivalent implementations may be substituted for the specific examples shown and described without departing from the scope of the present disclosure. This application is intended to cover any adaptations or variations of the specific examples discussed herein. Therefore, it is intended that this disclosure be limited only by the claims and the equivalents thereof.

CLAIMS

1. A system comprising:
 - a data processor to:
 - receive an incoming dataset via a processing system, and
 - generate a feature vector for the incoming dataset;
 - a matching module to:
 - determine similarity measures between the generated feature vector and representative feature vectors for a plurality of datasets in a data repository, and
 - select at least one dataset of the plurality of datasets based on the similarity measures; and
 - a recommendation module to:
 - identify at least one analytic task associated with the selected dataset, and
 - recommend, to a computing device via the processing system, the at least one analytic task to be performed on the incoming dataset.
2. The system of claim 1, wherein the at least one analytic task includes at least one of a machine learning algorithm and an analytics algorithm.
3. The system of claim 1, wherein the recommendation module is to recommend the at least one analytic task based on log information.
4. The system of claim 1, wherein the data processor is to further cluster the plurality of datasets in the data repository based on similarity measures, and wherein the similarity measures are based on at least one of data type, data size, data distribution, data attributes, topic, content, and application domain.

5. The system of claim 4, wherein the matching module is to select at least one cluster of datasets based on similarity measures between the generated feature vector and the representative feature vectors.
6. The system of claim 5, wherein the data processor is to further apply at least one of text mining and natural language processing to extract properties of the datasets to generate the representative features of the datasets.
7. The system of claim 5, wherein the matching module is to further select the at least one dataset from the selected at least one cluster based on similarity measures between the generated feature vector and the representative feature vectors of datasets in the selected at least one cluster.
8. The system of claim 1, wherein the data processor is to further pre-process a given dataset of the plurality of datasets based on information theory concepts to group similar data elements in the given dataset.
9. The system of claim 1, wherein the data processor is to generate the feature vector for the incoming dataset based on at least one of topic, data type, number of column groups, column group header, column group data type, data characteristic, data distribution, number of rows, and number of columns.
10. A method comprising:
 - receiving an incoming dataset via a processing system;
 - generating a feature vector for the incoming dataset;
 - identifying representative feature vectors for a plurality of datasets in a data repository;
 - selecting at least one cluster of datasets of the plurality of datasets based on similarity measures between the generated feature vector and the representative feature vectors;

selecting at least one dataset from the selected at least one cluster based on similarity measures between the generated feature vector and the representative feature vectors of datasets in the selected at least one cluster;

identifying at least one analytic task associated with the selected dataset; and

recommending, to a computing device via the processing system, the at least one analytic task to be performed on the incoming dataset.

11. The method of claim 10, wherein the at least one analytic task includes at least one of a machine learning algorithm and an analytics algorithm.

12. The method of claim 10, wherein identifying the at least one analytic task is based on log information.

13. The method of claim 10, further comprising generating the clusters of datasets of the plurality of datasets based on similarity measures, and wherein the similarity measures are based on a comparison of at least one of data type, data size, data distribution, data attributes, topic, content, and application domain.

14. A non-transitory computer readable medium comprising executable instructions to:

receive an incoming dataset via a processor;

generate a feature vector for the incoming dataset;

identify, via the processor, representative feature vectors for a plurality of datasets in a data repository;

determine, via the processor, similarity measures between the generated feature vector and the representative feature vectors for the plurality of datasets;

select at least one dataset of the plurality of datasets based on the similarity measures;

identify, via the processor, at least one analytic task associated with the selected dataset, wherein the at least one analytic task includes one of a machine learning algorithm and an analytics algorithm; and

recommend, to a computing device via the processor, the at least one analytic task to be performed on the incoming dataset.

15. The computer readable medium of claim 14, wherein the similarity measures are based on a comparison of at least one of data type, data size, data distribution, data attributes, topic, content, and application domain.

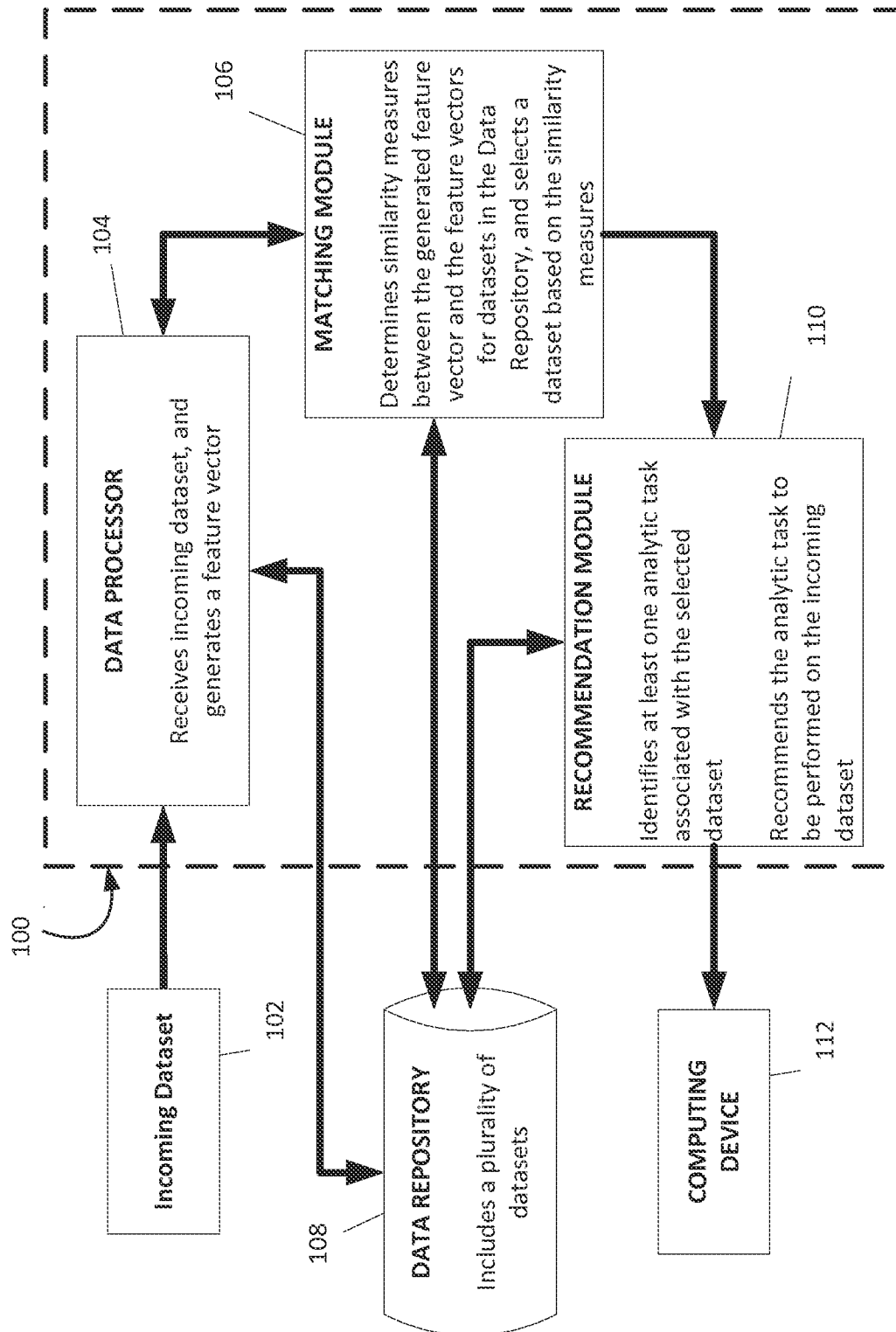


Fig. 1

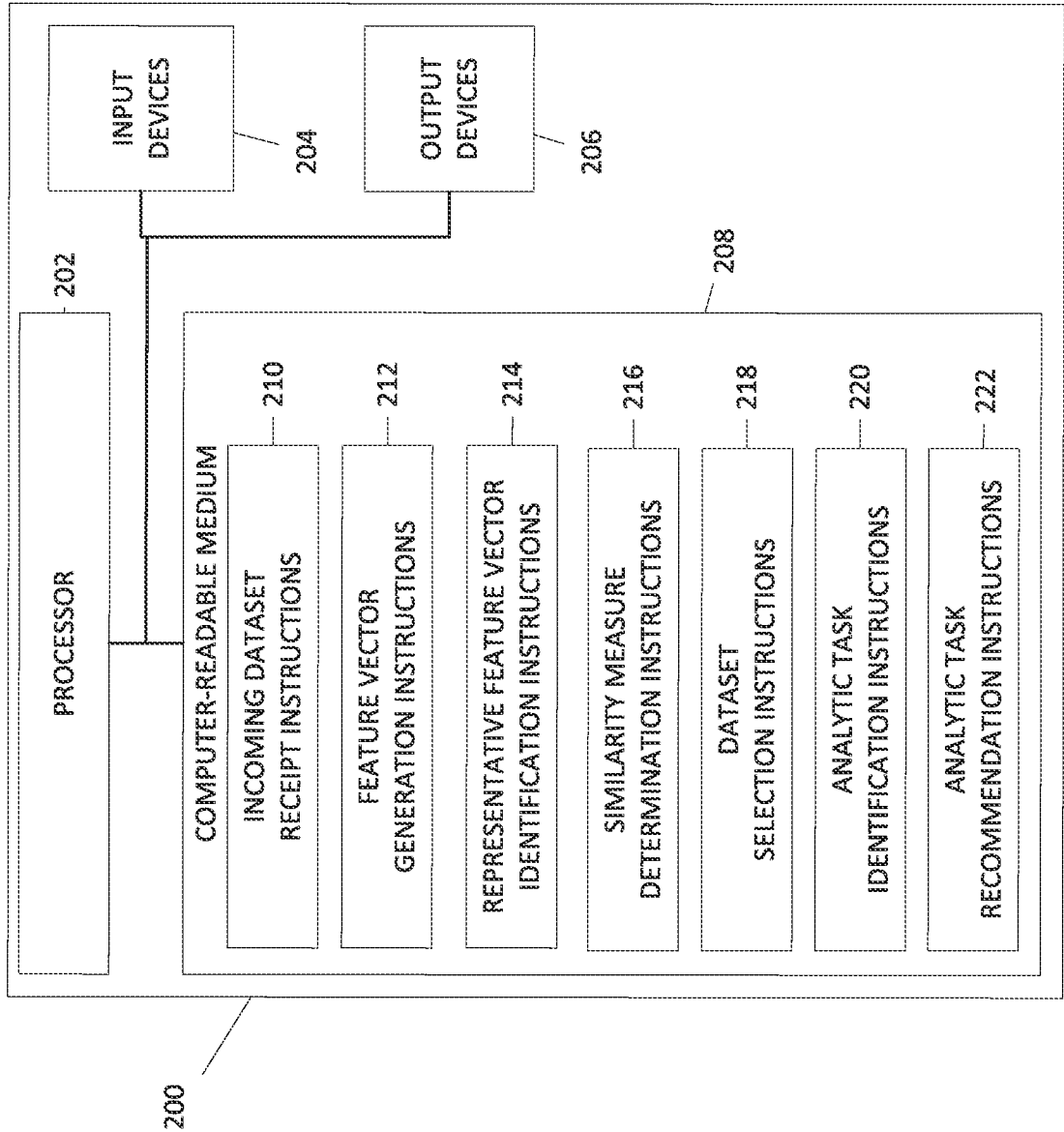
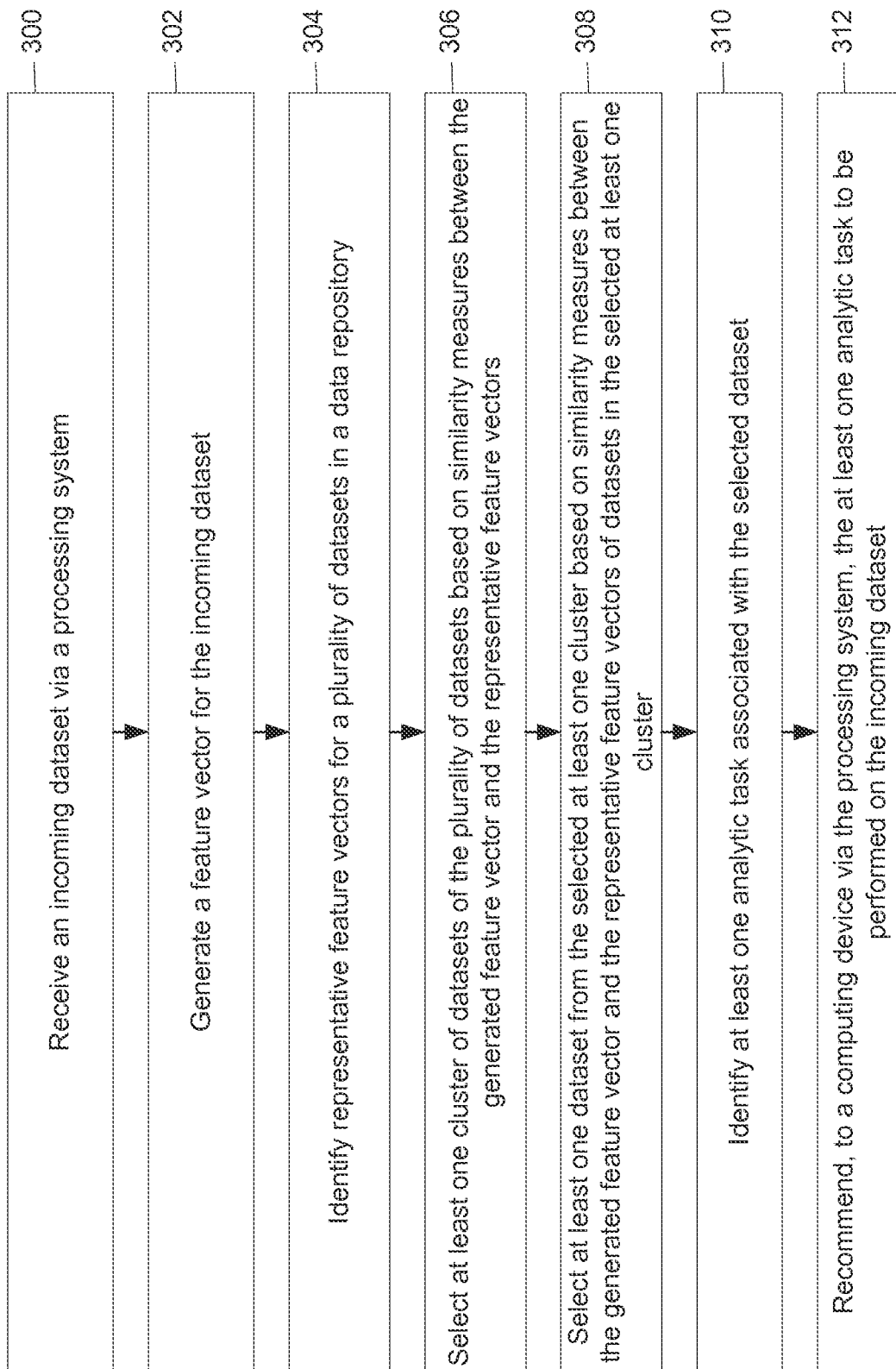


Fig. 2

**Fig. 3**

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2015/037206**A. CLASSIFICATION OF SUBJECT MATTER****G06F 9/48(2006.01)i, G06F 17/00(2006.01)i**

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F 9/48; G06Q 30/02; G06F 17/30; G06Q 10/00; G06F 17/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models

Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS(KIPO internal) & keywords: big data, recommendation, characteristic, matching, feature vector, dataset, similarity, processor**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2012-0296701 A1 (HANS C. BREITER) 22 November 2012 See paragraphs [0010], [0012], [0208], [0220], [0285]; claim 1 and figure 22.	1-15
A	US 2014-0006338 A1 (APPLIED MATERIALS, INC.) 02 January 2014 See paragraphs [0052]-[0056] and figure 5.	1-15
A	US 2010-0281025 A1 (DOROTHEA TSATSOU et al.) 04 November 2010 See paragraphs [0050]-[0064] and figure 3.	1-15
A	WO 2013-085709 A1 (BEYONDCORE, INC.) 13 June 2013 See page 6, line 24 - page 8, line 11 and figure 1.	1-15
A	WO 2014-100290 A1 (APPLANGO SYSTEMS LTD. et al.) 26 June 2014 See paragraphs [0034]-[0036] and figure 3.	1-15



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

18 March 2016 (18.03.2016)

Date of mailing of the international search report

21 March 2016 (21.03.2016)

Name and mailing address of the ISA/KR

International Application Division
Korean Intellectual Property Office
189 Cheongsu-ro, Seo-gu, Daejeon Metropolitan City, 35208,
Republic of Korea

Facsimile No. +82-42-472-7140

Authorized officer

CHIN, Sang Bum

Telephone No. +82-42-481-8398



INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2015/037206

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2012-0296701 A1	22/11/2012	WO 2012-161740 A2 WO 2012-161740 A3	29/11/2012 07/11/2013
US 2014-0006338 A1	02/01/2014	KR 10-2015-0027277 A TW 201403353 A WO 2014-005073 A1	11/03/2015 16/01/2014 03/01/2014
US 2010-0281025 A1	04/11/2010	WO 2010-129165 A2 WO 2010-129165 A3	11/11/2010 24/03/2011
WO 2013-085709 A1	13/06/2013	US 2013-0144813 A1	06/06/2013
WO 2014-100290 A1	26/06/2014	AU 2013-361457 A1 KR 10-2015-0096762 A US 2014-173105 A1 US 2015-172400 A1	26/06/2014 25/08/2015 19/06/2014 18/06/2015