



- (51) **International Patent Classification:**
G06F 17/30 (2006.01)
- (21) **International Application Number:**
PCT/US2014/040234
- (22) **International Filing Date:**
30 May 2014 (30.05.2014)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (71) **Applicant:** HEWLETT-PACKARD DEVELOPMENT COMPANY, L.P. [US/US]; 11445 Compaq Center Drive W., Houston, Texas 77070 (US).
- (72) **Inventors:** KOGAN, Hadas; Kiryat Technion \$ Technion, 32000 Haifa (IL). TADESKI, Inbal; Kiryat Technion \$ Technion, 32000 Haifa (IL).
- (74) **Agents:** KIRCHEV, Ivan T. et al.; Hewlett-Packard Company, Intellectual Property Administration, 3404 E. Harmony Road, Mail Stop 35, Fort Collins, Colorado 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, LT, 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, 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).

[Continued on next page]

(54) **Title:** PROCESSING A TABLE OF COLUMNS

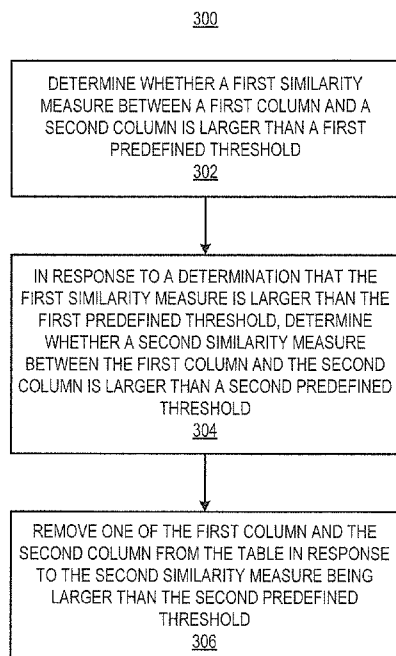


FIG. 3

(57) **Abstract:** According to an example, in a method for processing a table having a plurality of columns, a determination may be made as to whether a first similarity measure between a first column of the plurality of columns and a second column of the plurality of columns is larger than a first predefined threshold. In response to a determination that the first similarity measure is larger than the first predefined threshold, a determination may be made as to whether a second similarity measure between the first column and the second column is larger than a second predefined threshold. In addition, one of the first column and the second column may be removed from the table in response to the second similarity measure being larger than the second predefined threshold.



Declarations under Rule 4.17:

— *as to the identity of the inventor (Rule 4.17(i))*

Published:

— *with international search report (Art. 21(3))*

PROCESSING A TABLE OF COLUMNS

BACKGROUND

[0001] Databases are computerized data storage and retrieval systems that are organized in any of a number of different manners. A typical manner in which databases are organized is through the storage of data in multiple tables composed of columns and rows. The columns typically identify a category or topic and the rows typically include data pertaining to a particular entity or feature. As an example, the category of a column may be “clients” and each of the rows may identify a particular client. Oftentimes, the tables in databases, such as those stored as structured data, contain very large numbers of columns.

BRIEF DESCRIPTION OF THE DRAWINGS

[0002] Features of the present disclosure are illustrated by way of example and not limited in the following figure(s), in which like numerals indicate like elements, in which:

[0003] FIG. 1 is a simplified schematic diagram of a data processing system, which may implement various aspects of the methods disclosed herein, according to an example of the present disclosure;

[0004] FIG. 2 is a simplified block diagram of the data processing system shown in FIG. 1, according to an example of the present disclosure;

[0005] FIG. 3 depicts a flow diagram of a method for processing a table, according to an example of the present disclosure;

[0006] FIGS. 4A-4C, collectively, depict a flow diagram of a method for processing a table, according to another example of the present disclosure;

[0007] FIG. 5 depicts a flow diagram of a method for processing a table, according to another example of the present disclosure; and

[0008] FIG. 6 is schematic representation of a computing device, which may be employed to perform various functions of the computing device depicted in FIG. 2, according to an example of the present disclosure.

DETAILED DESCRIPTION

[0009] For simplicity and illustrative purposes, the present disclosure is described by referring mainly to an example thereof. In the following description, numerous specific details are set forth in order to provide a thorough understanding of the present disclosure. It will be readily apparent however, that the present disclosure may be practiced without limitation to these specific details. In other instances, some methods and structures have not been described in detail so as not to unnecessarily obscure the present disclosure. As used herein, the terms “a” and “an” are intended to denote at least one of a particular element, the term “includes” means includes but not limited to, the term “including” means including but not limited to, and the term “based on” means based at least in part on.

[0010] The tables in databases, such as those stored as structured data, often contain very large numbers of columns. Prior to performing data analytics applications on the tables, the data contained in the tables are often manually pre-processed. That is, an analyst often spends a great deal of time preparing the data for the analytics applications by manually determining which of the columns are informative and which of the columns are non-informative. Typically, only a small subset of the columns is informative. By way of example, analysts often spend upwards of 80% of their time preparing the data for the analysis. As such, the manual pre-processing of the data contained tables is typically time-consuming and inefficient.

[0011] Disclosed herein are methods for processing a table having a plurality of columns and apparatuses for implementing the methods that may overcome many of the inefficiencies common to conventional processing techniques. In the methods, a determination may be made as to whether a first similarity measure between a first column of the plurality of columns and a second column of the plurality of columns is larger than a first predefined threshold. In response to a determination that the first similarity measure is larger than the first predefined threshold, a determination may be made as to whether a second similarity measure between the first column and the second column is larger than

a second predefined threshold, in which the second similarity measure has a greater computational complexity as compared with the first similarity measure. In addition, one of the first column and the second column may be removed from the table in response to the second similarity measure being larger than the second predefined threshold.

[0012] Through implementation of the methods disclosed herein, columns that may contain relatively low entropies, e.g., contain non-informative information, may automatically be identified and removed from a table, which may be a structured data table. In addition, a determination may be made as to which of the columns contain information that is relatively similar to other columns and those columns may also be removed from the table. As disclosed herein, the similarity determination among the columns may be made through implementation of a two-step procedure, in which the computational complexity of the first step is relatively lower than the computational complexity of the second step. That is, the first step may be simpler and faster to implement than the second step, but may provide a relatively less accurate determination of similarity among the columns. In this regard, through performance of the first step, a quick determination that a column is dissimilar from other columns in the table may be made. In addition, the second step may be performed on only those columns that are determined to possibly be similar to other columns, thus reducing the number of sets of columns for which the second step may be performed. Moreover, through the clustering of columns disclosed herein, the columns contained in a cluster may not be compared with other columns, thus also reducing the total number of comparisons that may need to be performed to identify the columns that are to be removed from the table.

[0013] Generally speaking, the methods and apparatuses disclosed herein enable non-informative and duplicative (or similar) columns to automatically be identified and removed from a table. As such, the methods and apparatuses disclosed herein may enable the processing of tables in manners that are relatively faster and more accurate than conventional manual techniques for processing a table for data analytics operations. In addition, as redundant columns may lead to erroneous results in data analytics applications, the

methods and apparatuses disclosed herein may increase the accuracy of the data analytics applications on the tables through removal of the redundant columns. Moreover, the similarity measures disclosed herein may be effective for different column types (e.g., categorical and numerical) because the similarity measures disclosed herein uses mutual information and entropies and thus does not assume a specific dependence, e.g., a linear mapping between two columns.

[0014] With reference first to FIG. 1, there is shown a simplified schematic diagram of a data processing system 100, which may implement various aspects of the methods disclosed herein, according to an example. It should be understood that the data processing system 100 depicted in FIG. 1 may include additional elements and that some of the elements depicted therein may be removed and/or modified without departing from a scope of the data processing system 100.

[0015] As shown in FIG. 1, the data processing system 100 may include a database 110 and a data processing apparatus 120. The database 110 may contain multiple tables (a single table 112 is shown for purposes of simplicity) and may be contained in a single data storage device or across multiple data storage devices. The table 112 is also shown as including a plurality of columns 1-N and a plurality of rows 1-M, in which "N" and "M" may represent values larger than 1 and may differ from each other. The table 112 may be a structured data table and may be stored in the database 110 in any suitable conventional manner. In addition, the table 112 may include various types of data, such as numerical, categorical, character strings, Boolean, etc. According to an example, data may be generalized to any data that contains elements (equivalent to rows) and fields (equivalent to columns) and may be transformed to have the form of the table 112. In this example, for instance, data received in a format other than in a table format, such as an XML file, may be transformed into the form of the table 112.

[0016] The table processing apparatus 120 may access the table 112 to perform processing operations on the table 112. According to an example, a user may implement the table processing apparatus 120 to perform the processing operations on the table 112. In some examples, the table processing apparatus

120 may also be implemented to perform analytical operations on the table 112, such as query operations. A processing operation may include preparing the data in the table 112 for data analytics applications. By way of example, because tables stored in databases, for instance, stored as structured data, often contain a very large number of columns, the processing operation may include removing the columns that may contain unimportant or non-informative information. Reducing the number of columns contained in table 112 may also result in the reduction of the amount of time and computational resources required to perform the data analytics applications, as well as the accuracy of the data analytics applications.

[0017] As discussed in greater detail herein below, the table processing apparatus 120 may determine which of the columns in the table 112 may contain informative information and which of the columns may contain non-informative information. The table processing apparatus 120 may make these determinations regardless of the types of data contained in the columns. That is, in making these determinations, the table processing apparatus 120 is not limited to any specific data type. The columns that include informative information may include those columns that have information that is not duplicated in another column, those columns that have information that is not constant across some or all of the rows of the column, those columns that have information that is not similar to information in another column beyond a predefined similarity measure threshold, etc. The columns that include non-informative information may include those columns that have information that is duplicative with another column, have information that is very similar to another column, have empty rows, have constant values across the column, those columns that have information that is similar to information in another column beyond a predefined similarity measure threshold, etc. A column may be considered to be similar to another column even if the actual values in the columns are completely different from each other. For instance, in an example in which one column contains client identifiers, which may contain numbers, and another column contains client names, which may contain character strings, although these columns appear to be completely different, they may actually be similar because value X in the client identifier and

value A in the client names may always co-occur in the same records. In addition, the occurrences may repeat for all of the different values.

[0018] Turning now to FIG. 2, there is shown a simplified block diagram of the data processing system 100, according to an example. It should be understood that the data processing system 100 depicted in FIG. 2 may include additional elements and that some of the elements depicted therein may be removed and/or modified without departing from a scope of the data processing system 100.

[0019] As shown in FIG. 2, the data processing system 100 is depicted as including a computing device 200. The computing device 200 is depicted as including, in addition to the table processing apparatus 120, a processor 202, an interface 204, and a data store 206. The table processing apparatus 104 is also depicted as including a table accessing module 210, a histogram accessing module 212, an entropy accessing module 214, a first similarity measure determining module 216, a second similarity measure determining module 218, a cluster determining module 220, a column selecting module 222, and a column removing module 224.

[0020] The processor 202, which may be a microprocessor, a micro-controller, an application specific integrated circuit (ASIC), or the like, is to perform various processing functions in the computing device 200. The processing functions may include invoking or implementing the table processing apparatus 120 and particularly, the modules 210-224 of the table processing apparatus 120, as discussed in greater detail herein below. According to an example, the table processing apparatus 120 is a hardware device on which is stored various sets of machine readable instructions. The table processing apparatus 120 may be, for instance, a volatile or non-volatile memory, such as dynamic random access memory (DRAM), electrically erasable programmable read-only memory (EEPROM), magnetoresistive random access memory (MRAM), memristor, flash memory, floppy disk, a compact disc read only memory (CD-ROM), a digital video disc read only memory (DVD-ROM), or other optical or magnetic media, and the like, on which software may be stored. In this example,

the modules 210-224 may be software modules, e.g., sets of machine readable instructions, stored in the table processing apparatus 120.

[0021] In another example, the table processing apparatus 120 may be a hardware component, such as a chip, an integrated circuit, etc., and the modules 210-224 may be hardware modules on the hardware component. In a further example, the modules 210-224 may include a combination of software and hardware modules. In a yet further example, the processor 202 may be an ASIC that is to perform the functions of the modules 210-224. In this example, the processor 202 and the table processing apparatus 120 may be a single processing apparatus.

[0022] The processor 202 may store data in the data store 206 and may use the data in implementing the modules 210-224. For instance, the processor 202 may store data pertaining to the table 112 that is to be processed. In any regard, the data store 206 may be volatile and/or non-volatile memory, such as DRAM, EEPROM, MRAM, phase change RAM (PCRAM), memristor, flash memory, and the like. In addition, or alternatively, the data store 206 may be a device that may read from and write to a removable media, such as, a floppy disk, a CD-ROM, a DVD-ROM, or other optical or magnetic media.

[0023] The interface 204 may include hardware and/or software to enable the processor 202 to communicate with the database 110. The interface 204 may enable a wired or wireless connection to the database 110 (user input/output devices) over a network, such as an intranet, a local area network, the Internet, etc. In other examples, the computing device 200 may be directly connected to the database 110, e.g., the computing device 200 may be a server that is to control access to the database 110. The interface may include a network interface card and/or may also include hardware and/or software to enable the processor 202 to communicate with various input and/or output devices, such as a keyboard, a mouse, a display, another computing device, etc., through which a user may input instructions into the computing device 200.

[0024] Various manners in which the processor 202 in general, and the modules 210-224 in particular, may be implemented are discussed in greater

detail with respect to the methods 300 and 400 respectively depicted in FIGS. 3 and 4A-4C. Particularly, FIGS. 3 and 4A-4C depict flow diagrams of methods 300 and 400 for processing a table 112, according to two examples. It should be apparent to those of ordinary skill in the art that the methods 300 and 400 may represent generalized illustrations and that other operations may be added or existing operations may be removed, modified, or rearranged without departing from the scopes of the methods 300 and 400. Generally speaking, the processor 202 depicted in FIG. 2 may implement the methods 300 and 400 through implementation of at least some of the modules 210-224. In addition, the method 400 is similar to the method 300, but includes more detailed descriptions of the operations described in the method 300.

[0025] The descriptions of the methods 300, 400, and 500 are made with reference to the data processing system 100 illustrated in FIGS. 1 and 2 for purposes of illustration. It should, however, be clearly understood that data processing systems having other configurations may be implemented to perform the methods 300, 400, and 500 without departing from the scopes of the methods 300, 400, and 500. In addition, references to a first column, a second column, a third column, a fourth column, etc., should not be construed as denoting a physical location or a consecutive order of columns. Instead, these references are merely meant to denote different ones of the columns, regardless of their locations with respect to each other.

[0026] With reference first to the method 300 depicted in FIG. 3, at block 302, a determination may be made as to whether a first similarity measure between a first column of the plurality of columns and a second column of the plurality of columns is larger than a first predefined threshold. For instance, the first similarity measure determining module 216 may determine a first similarity measure between the first column and the second column and may determine whether the first similarity measure is above a first predefined threshold. The first similarity measure and the first predefined threshold are discussed in greater detail below with respect to the method 400 below.

[0027] At block 304, in response to a determination that the first similarity measure is larger than the first predefined threshold, a determination may be made as to whether a second similarity measure between the first column and the second column is larger than a second predefined threshold, in which the second similarity measure has a greater computational complexity as compared with the first similarity measure. That is, computation of the second similarity measure may require a greater amount of time and/or computational resources as compared with computation of the first similarity measure. In addition, for instance, the second similarity measure determining module 218 may determine a second similarity measure between the first column and the second column and may determine whether the second similarity measure is above a second predefined threshold. The second similarity measure and the second predefined threshold are discussed in greater detail below with respect to the method 400 below.

[0028] At block 306, one of the first column and the second column may be removed from the table 112. More particularly, for instance, the column removing module 224 may remove one of the first column and the second column from the table 112. In one regard, through implementation of the method 300, a determination may be made that one of the first column and the second column may include non-informative, e.g., duplicative, information and thus, one of the first column and the second column may be removed to reduce the number of columns that may be analyzed during a data analytics operation.

[0029] Turning now to FIGS. 4A-4C, and with reference first to FIG. 4A, at block 402, a table 112 may be accessed. For instance, the table accessing module 210 may access the table 112 directly from the database 110. Alternatively, the table accessing module 210 may access a copy of the table 112 that may have been copied onto the data store 206. According to an example, the table 112 is a copy of an original table.

[0030] At block 404, histograms for each column in the table 112 may be accessed. For instance, the histogram accessing module 212 may access the histograms by computing the histograms of each of the columns through

implementation of any suitable set of machine readable instructions that is to compute the histograms. As another example, the histograms of the columns may have previously been computed by another module and the histogram accessing module 212 may access the previously computed histograms. According to an example, the histograms of the columns represent the respective averages of the data contained in the columns.

[0031] At block 406, entropies of each of the columns may be accessed. The entropies of the columns, for instance, low entropies, may denote that the columns are empty, almost empty, constant, almost constant, etc. In contrast, high entropies may denote that columns are full, almost full, have different values along the columns, etc. In one regard, therefore, the entropy of a column may denote the amount of change between values contained in the rows of the columns. The entropy accessing module 214 may access the entropies by determining the entropies of the columns based upon the computed histograms of the columns. By way of example, the entropy accessing module 214 may determine the entropies of the columns through computation of the following equation:

$$\text{Equation (1): } H_A = - \sum_{i=1}^k P_{v_i} \log P_{v_i} .$$

[0032] In Equation (1), from the computed histograms of the columns A and B, it may be assumed that $A \in \{v_1, v_2, \dots, v_k\}$, in which "v" represents the values in the columns. In addition, in Equation (1), numerical values may be quantized and handled similarly to categorical values, and the calculations in that equation may also be applied to the numerical values. The probability distribution, which may actually be a normalized histogram, may be denoted through the following equation: $P_A = \{P_{v_1}, P_{v_2}, \dots, P_{v_k}\}$. The probability of each value is $P_{v_i} = \frac{\#(A=v_i)}{N}$, where N is the number of instances (rows) in the table 112. As another example, however, the entropies of the columns may have previously been computed by another module and the entropy accessing module 214 may access the previously computed entropies.

[0033] At block 408, a determination may be made as to whether the entropies of the columns are larger than a predetermined threshold. For instance, the entropy accessing module 214 may determine whether any of the columns have entropies that exceed a predetermined threshold. In addition, the entropy accessing module 214 may determine which of the determined columns have entropies that fall below the predetermined threshold. The predetermined threshold may be set to meet any suitable objective. Thus, for instance, the predetermined threshold value may be set to be a higher value when a higher number of columns are desired to remain in the table 112.

[0034] At block 410, the columns identified as having entropies that fall below the predetermined threshold may be removed from the table 112. That is, for instance, the column removing module 224 may remove the columns that have relatively constant data across their respective rows for containing non-informative information.

[0035] Following blocks 408 and 410, clusters of the remaining columns determined to be similar to each other may be identified and the number of columns in the table 112 may be reduced. At block 412, a cluster A may be set to a first cluster. For instance, the cluster determining module 220 may set the cluster A to be a first cluster that includes a first column. In addition, at block 414, a column A and a column B may be set to be columns that are to be compared with each other. For instance, during a first iteration, the cluster determining module 220 may set column A to be the first column and column B to be the second column in the table 112.

[0036] Following "A" to FIG. 4B, at block 416, a similarity measure (HS) between column A and column B may be determined. For instance, the first similarity measure determining module 218 may determine a histogram similarity (HS) between column A and column B. By way of particular example, the first similarity measure determining module 218 may determine the HS between column A and column B through the following equation:

Equation (2):

$$HS(A, B) = \frac{(P_A - \mu_{P_A})^T (P_B - \mu_{P_B})}{\|P_A - \mu_{P_A}\|^2 + \|P_B - \mu_{P_B}\|^2}$$

[0037] In Equation (2), for a given pair of columns A and B, their histograms are marked by P_A, P_B , their expectations are marked by μ_{P_A}, μ_{P_B} , and “ T ” represents a transpose function. According to an example, the histograms for the columns are sorted. In addition, or alternatively, the shorter histogram is padded with zeroes such that the lengths of the two histograms are identical.

[0038] At block 418, a determination may be made as to whether the HS of column A and column B is larger than a first predefined threshold. For instance, the first similarity measure determining module 216 may compare the HS of column A and column B with the first predefined threshold to determine whether the HS exceeds the first predefined threshold. The first predefined threshold may be set to meet any suitable objective. Thus, for instance, the first predefined threshold may be set to be a higher value when it is desired to have a higher number of columns removed from the table 112.

[0039] In response to a determination that the HS of column A and column B is smaller than the first predefined threshold, column B may be identified as being dissimilar from column A, as indicated at block 420. In addition, at block 422, a determination may be made as to whether an additional column is to be compared with the column A. If yes, the column selecting module 222 may select another column, e.g., the next column adjacent to the previous column B, to be compared with the column A and may set the another column as column B. Moreover, blocks 416-422 may be repeated for the columns A and B as set at block 424. In addition, blocks 416-424 may be repeated until a determination is made at block 422 that no additional columns remain for comparison with column A. In that instance, a determination may be made as to whether an additional cluster is to be identified at block 426.

[0040] At block 426, for instance, the cluster determining module 220 may determine whether there are any remaining columns in the table 112 that have not been included into a cluster. In response to a determination at block 426 that a column in the table 112 has not been included into a cluster, the cluster A may

be set to the next cluster, as indicated at block 428. Particularly, for instance, the cluster determining module 220 may set a current cluster being analyzed to be equal to a next cluster, in which the next cluster includes the next column that has not been included into a cluster. In addition, following "C" back to FIG. 4A, blocks 414-426 may be repeated until a determination that no additional clusters remain to be determined at block 426. Following the "no" condition at block 426, the method 400 may end, as indicated at block 430. In this instance, the method 400 may not have identified any columns that are sufficiently similar to any other columns to be removed from the table 112.

[0041] Returning back to block 418, in response to a determination that the HS of column A and column B is larger than the first predefined threshold, following "B" to FIG. 4C, a mutual information based similarity measure (SMI) between column A and column B may be made, as indicated at block 432. For instance, the second similarity measure determining module 218 may determine the SMI between column A and column B through the following equation:

Equation (3):

$$SMI(A, B) = \frac{MI(A, B)}{0.5(H_A + H_B)}$$

[0042] As noted in Equation (3), the SMI may be based upon the mutual information of two variables, which quantifies the dependency between the two variables. The mutual information (MI) of two variables $A \in \{v_1, v_2, \dots, v_k\}$, $B \in \{w_1, w_2, \dots, w_t\}$ may be defined by the following equation:

Equation (4):

$$MI(A, B) = \sum_{i=1}^k \sum_{j=1}^t P_{v_i, w_j} \log \frac{P_{v_i, w_j}}{P_{v_i} P_{w_j}}$$

[0043] As also noted in Equation (3), the SMI may also be based upon the entropies of the columns A and B. Particularly, the SMI between columns A and B may be higher for dependent columns and zero for independent columns. In addition, the mutual information may be higher when the entropies of the columns are higher and lower when the entropies of the columns are lower. In any regard,

as may be noted by a comparison of Equations (2) and (3), Equation (3) may have a greater computational complexity as compared with Equation (2). In addition, the SMI disclosed in Equation (3) may not be sensitive to the entropies of the columns, and may fall between 0 and 1, i.e., $0 \leq \text{SMI}(A, B) \leq 1$, with high values of SMI indicating similarity among column A and column B.

[0044] At block 434, a determination may be made as to whether the SMI of column A and column B is larger than a second predefined threshold. For instance, the second similarity measure determining module 218 may compare the SMI of column A and column B with the second predefined threshold to determine whether the SMI exceeds the second predefined threshold. The second predefined threshold may be set to meet any suitable objective. Thus, for instance, the second predefined threshold may be set to be a higher value when it is desired to have a higher number of columns removed from the table 112.

[0045] In response to a determination that the SMI of column A and column B is greater than the second predefined threshold at block 434, column B may be added to the cluster of column A, as indicated at block 436. Particularly, for instance, the cluster determining module 220 may add column B to the cluster of column A because column B has been identified as being sufficiently similar to column A. In addition, as discussed below, column B may be omitted from consideration for inclusion into another cluster of another column.

[0046] However, in response to a determination that the SMI of column A and column B is below the second predefined threshold, column B may be identified as being dissimilar from column A, as indicated at block 438. In addition, therefore, the cluster determining module 220 may not include column B in the cluster of column A and column B may be available for consideration for inclusion into another cluster of another column.

[0047] At block 440, following either of blocks 436 and 438, a determination may be made as to whether an additional column is to be compared with column A. If yes, column B may be set to a next column as indicated at block 442. Thus, for instance, the column selecting module 222 may select the next column, e.g., the column adjacent to the previously selected

column, to be column B at block 424. Moreover, blocks 432-440 may be repeated for a next selected column until a determination is made at block 440 that no additional columns remain for comparison with column A. In that instance, a determination may be made as to whether an additional cluster is to be identified at block 444.

[0048] At block 444, for instance, the cluster determining module 220 may determine whether there are any remaining columns in the table 112 that have not been included into a cluster. In response to a determination at block 446 that a column in the table 112 has not been included into a cluster, the cluster A may be set to the next cluster. Particularly, for instance, the cluster determining module may set a current cluster being analyzed to be equal to a next cluster, in which the next cluster includes the next column that has not already been included into a cluster. In addition, following "C" back to FIG. 4A, blocks 414-446 may be repeated until a determination that no additional clusters remain to be determined at block 444.

[0049] Following the "no" condition at block 444, all but one of the columns in each of the clusters may be removed from the table 112, as indicated at block 448. For clusters that include a single column, the columns in those clusters may not be removed since those columns may be sufficiently different from other columns and may contain informative information. In addition, for instance, the column removing module 224 may selectively remove the columns from each of the clusters because the removed columns may be sufficiently similar to the remaining columns, e.g., beyond a predefined threshold, and may thus be construed to contain non-informative, e.g., duplicative, information. For each cluster, the column removing module 224 may determine which of the columns in the clusters to remove. For instance, the column removing module 224 may remove all but the first column in the cluster, the columns having the longest names, columns having the least number of rows of data, etc.

[0050] Generally speaking, through implementation of the method 400, each of the columns in the table 112 may be identified as being similar beyond a predefined threshold to another one of the columns or as being different below

the predefined threshold from each of the other columns. In the former case, the columns in a particular cluster may be construed as being similar to each other and in the latter case, the columns may be construed as being dissimilar from each of the other columns.

[0051] Following block 448, the number of columns contained in the table 112 may substantially be reduced through removal of the columns. The table 112 may also be stored with the reduced set of columns. In addition, the method 400 may end as indicated at block 450.

[0052] Although the methods 300 and 400 are described herein with respect to a first similarity measure and a second similarity measure, these measures may equivalently be construed respectively as a first difference measure and a second difference measure. That is, instead of similarities between a first column and a second column being determined, differences between the first column and the second column may be determined. This example is depicted in the method 500 in FIG. 5. As shown in that figure, at block 502, a determination may be made as to whether a first difference measure between a first column and a second column is lower than a first predefined threshold. In addition, for instance, when the difference between the first column and the second column is below a first predefined threshold, the two columns may be construed as being similar and when the difference between the first column and the second column is above the first predefined threshold, the two columns may be construed as possibly being similar. In response to the first column and the second column having a first difference measure below the first predefined threshold, at block 504, a determination may be made as to whether a second difference measure between the first column and the second column is lower than a second predefined threshold. In response to the second difference measure being below the second predefined threshold, the first column and the second column may be construed as being similar to each other. However, if the second difference measure is above the second predefined threshold, the first column and the second column may be construed as being different from each other. In addition, at block 506, one of first column and the second column may be removed from the table in response the second difference measure being lower

than the second predefined threshold.

[0053] Through implementation of the methods 300, 400, and/or 500, columns that may contain relatively low entropies, e.g., contain non-informative information, may be removed from a table 112. In addition, a determination may be made as to which of the columns contain information that is relatively similar to other columns and those columns may also be removed from the table 112. In one regard, the similarity determination may be made through implementation of a two-step procedure, in which the computational complexity of the first step is relatively lower than the computational complexity of the second step. That is, the first step, e.g., the HS similarity determination, may be simpler and faster to implement than the second step, e.g., the SMI similarity determination. Through performance of the first step first, a quick determination that a column is dissimilar from another column may be made. In addition, the second step may be performed on only those columns that are determined to possibly be similar to other columns. Moreover, through the clustering of columns disclosed herein, the columns contained in a cluster may not be compared with other columns, thus reducing the total number of comparisons that may need to be performed to identify the columns that are to be removed from the table 112.

[0054] Some or all of the operations set forth in the methods 300, 400, and 500 may be contained as utilities, programs, or subprograms, in any desired computer accessible medium. In addition, the methods 300, 400, and 500 may be embodied by computer programs, which may exist in a variety of forms both active and inactive. For example, they may exist as machine readable instructions, including source code, object code, executable code or other formats. Any of the above may be embodied on a non-transitory computer readable storage medium.

[0055] Examples of non-transitory computer readable storage media include computer system RAM, ROM, EPROM, EEPROM, and magnetic or optical disks or tapes. It is therefore to be understood that any electronic device capable of executing the above-described functions may perform those functions enumerated above.

[0056] Turning now to FIG. 6, there is shown a schematic representation of a computing device 600, which may be employed to perform various functions of the computing device 200 depicted in FIG. 2, according to an example. The computing device 600 may include a processor 602, a display 604, such as a monitor; a network interface 608, such as a Local Area Network LAN, a wireless 802.11x LAN, a 3G mobile WAN or a WiMax WAN; and a computer-readable medium 610. Each of these components may be operatively coupled to a bus 612. For example, the bus 612 may be an EISA, a PCI, a USB, a FireWire, a NuBus, or a PDS.

[0057] The computer readable medium 610 may be any suitable medium that participates in providing instructions to the processor 602 for execution. For example, the computer readable medium 610 may be non-volatile media, such as an optical or a magnetic disk; volatile media, such as memory. The computer-readable medium 610 may also store a table processing machine readable instructions 614, which may perform the methods 300, 400, and/or 400 and may include the modules 210-224 of the table processing apparatus 120 depicted in FIG. 2. In this regard, the table processing machine readable instructions 614 may include a table accessing module 210, a histogram accessing module 212, an entropy accessing module 214, a first similarity measure determining module 216, a second similarity measure determining module 218, a cluster determining module 220, a column selecting module 222, and a column removing module 224.

[0058] Although described specifically throughout the entirety of the instant disclosure, representative examples of the present disclosure have utility over a wide range of applications, and the above discussion is not intended and should not be construed to be limiting, but is offered as an illustrative discussion of aspects of the disclosure.

[0059] What has been described and illustrated herein is an example of the disclosure along with some of its variations. The terms, descriptions and figures used herein are set forth by way of illustration only and are not meant as limitations. Many variations are possible within the spirit and scope of the

disclosure, which is intended to be defined by the following claims -- and their equivalents -- in which all terms are meant in their broadest reasonable sense unless otherwise indicated.

What is claimed is:

1. A method for processing a table comprising a plurality of columns, said method comprising:

determining whether a first similarity measure between a first column of the plurality of columns and a second column of the plurality of columns is larger than a first predefined threshold;

in response to a determination that the first similarity measure is larger than the first predefined threshold, determining whether a second similarity measure between the first column and the second column is larger than a second predefined threshold ; and

removing one of the first column and the second column from the table in response to the second similarity measure being larger than the second predefined threshold.

2. The method according to claim 1, further comprising:

accessing a histogram for each column of the plurality of columns; and

comparing the histogram of the first column with the histogram of the second column to determine the first similarity measure.

3. The method according to claim 2, further comprising:

accessing an entropy for each column of the plurality of columns based upon the computed histograms of each of the columns;

determining, for each of the columns, whether the entropy of the column exceeds a predetermined threshold; and

removing the columns from the table that are determined to have entropies below the predetermined threshold prior to determining the first similarity measure.

4. The method according to claim 1, wherein the second similarity measure has a greater computational complexity as compared with the first similarity measure, said method further comprising:

in response to a determination that the first similarity measure is below the first predefined threshold, identifying the second column as being dissimilar from the first column.

5. The method according to claim 1, wherein the second similarity measure comprises a mutual information based similarity measure.

6. The method according to claim 1, further comprising:
prior to removing one of the first column and the second column from the table, including the second column in a first cluster in response to the second similarity measure being larger than the second predefined threshold and omitting the second column from the first cluster in response to the second similarity measure being smaller than the second predefined threshold, wherein the first cluster includes the first column.

7. The method according to claim 6, further comprising:
in response to the second column being omitted from the first cluster, determining a first similarity measure between the second column and a third column of the plurality of columns;

determining whether the first similarity measure between the second column and the third column is larger than the first predefined threshold;

in response to a determination that the first similarity measure between the second column and the third column is larger than the first predefined threshold, determining whether a second similarity measure between the second column and the third column is larger than the second predefined threshold;

in response to the second similarity measure between the second column and the third column being larger than the second predefined threshold, including the third column in a second cluster, wherein the second cluster includes the second column; and

removing one of the second column and the third column from the table.

8. The method according to claim 6, further comprising:

determining that a third column of the plurality of columns and a fourth column of the plurality of columns are dissimilar from the first column and are outside of the first cluster;

determining a first similarity measure between the third column and the fourth column;

determining whether the first similarity measure between the third column and the fourth column is larger than the first predefined threshold;

in response to a determination that the first similarity measure between the third column and the fourth column is larger than the first predefined threshold, determining whether a second similarity measure between the third column and the fourth column is larger than the second predefined threshold;

in response to the second similarity measure between the third column and the fourth column being larger than the second predefined threshold, including the fourth column in a second cluster, wherein the second cluster includes the third column; and

removing one of the third column and the fourth column from the table.

9. An apparatus for processing a table comprising a plurality of columns, said apparatus comprising:

a processor; and

a memory on which is stored machine readable instructions to cause the processor to:

determine, for each column of the plurality of columns, whether an entropy of the column exceeds a predetermined threshold;

remove, from the table, the columns that are determined to have entropies below the predetermined threshold;

determine whether a first similarity measure between a first column of the plurality of columns and a second column of the plurality of columns is larger than a first predefined threshold;

in response to a determination that the first similarity measure is larger than the first predefined threshold, determine whether a second similarity measure between the first column and the second column is larger than a second

predefined threshold; and

remove one of the first column and the second column from the table in response to the second similarity measure being larger than the second predefined threshold.

10. The apparatus according to claim 9, wherein the second similarity measure has a greater computational complexity as compared with the first similarity measure and wherein the machine readable instructions are further cause the processor to:

determine the first similarity measure through a comparison of a histogram of the first column with the histogram of the second column.

11. The apparatus according to claim 9, wherein the machine readable instructions are further to cause the processor to include the second column in a first cluster in response to the second similarity measure being larger than the second predefined threshold and omit the second column from the first cluster in response to the second similarity measure being smaller than the second predefined threshold, wherein the first cluster includes the first column.

12. The apparatus according to claim 11, wherein the machine readable instructions are further to cause the processor to determine whether the second column is similar to a third column in response to the second column being omitted from the first cluster and to omit further similarity determinations of the second column in response to the second column being included in the first cluster.

13. The apparatus according to claim 11, wherein the machine readable instructions are further to cause the processor to:

determine that a third column of the plurality of columns and a fourth column of the plurality of columns are outside of the first cluster;

determine a first similarity measure between the third column and the fourth column;

determine whether the first similarity measure between the third column and the fourth column is larger than the first predefined threshold;

in response to a determination that the first similarity measure between the third column and the fourth column is larger than the first predefined threshold, determine whether a second similarity measure between the third column and the fourth column is larger than the second predefined threshold;

in response to the second similarity measure between the third column and the fourth column being larger than the second predefined threshold, include the fourth column in a second cluster, wherein the second cluster includes the third column; and

remove one of the third column and the fourth column from the table.

14. A non-transitory computer readable storage medium on which is stored machine readable instructions that when executed by a processor cause the processor to:

determine whether a first difference measure between a first column of the plurality of columns and a second column of the plurality of columns is lower than a first predefined threshold;

in response to a determination that the first difference measure is lower than the first predefined threshold, determine whether a second difference measure between the first column and the second column is lower than a second predefined threshold;; and

remove one of the first column and the second column from the table in response to the second difference measure being lower than the second predefined threshold.

15. The non-transitory computer readable storage medium according to claim 14, wherein the second similarity measure has a greater computational complexity as compared with the first similarity measure and wherein the machine readable instructions are further to cause the processor to:

prior to the determination of the first difference measure, determine, for each column of the plurality of columns, whether an entropy of the column

exceeds a predetermined threshold; and

remove, from the table, the columns that are determined to have entropies below the predetermined threshold.

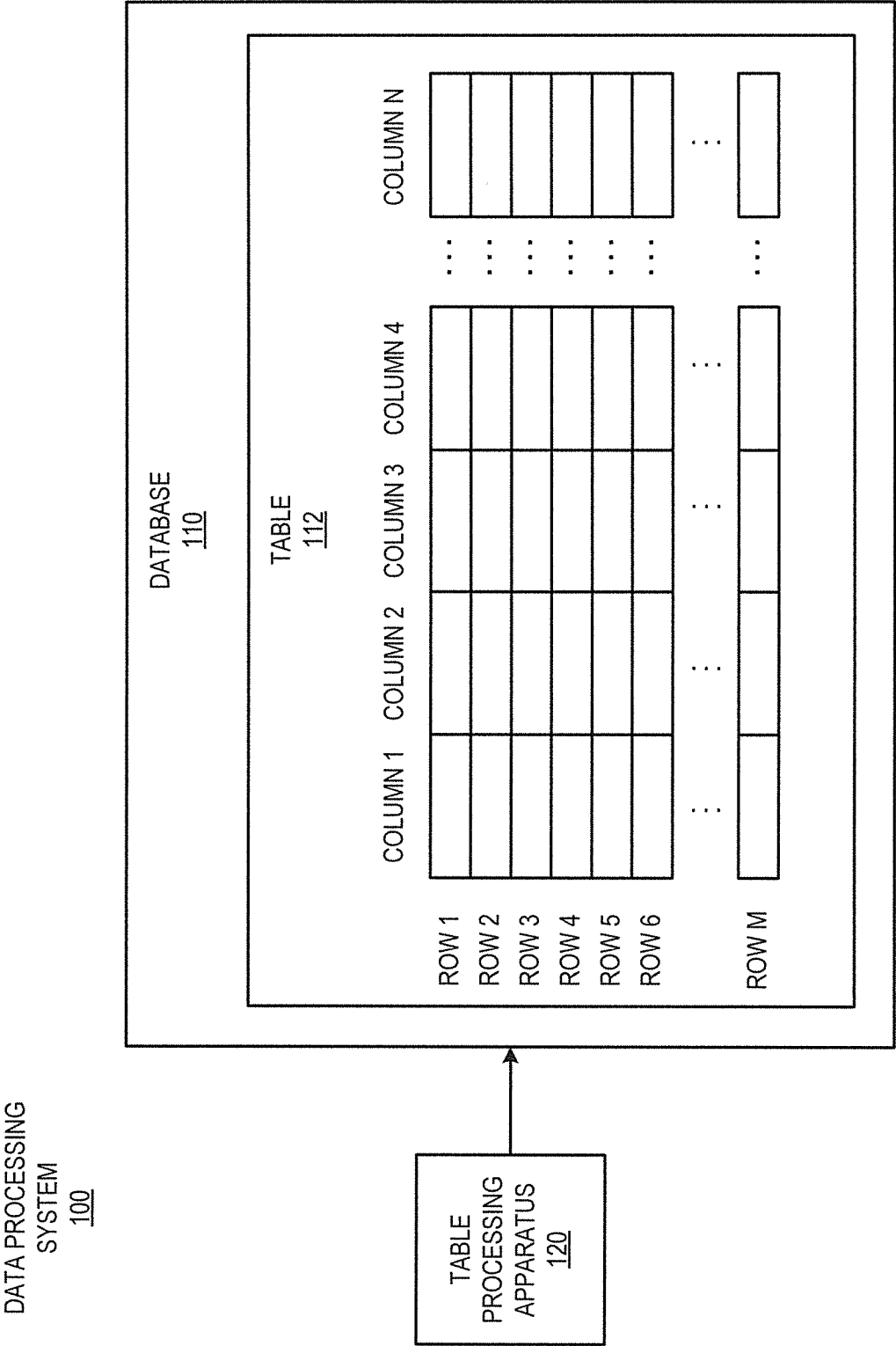


FIG. 1

2 / 8

DATA PROCESSING SYSTEM

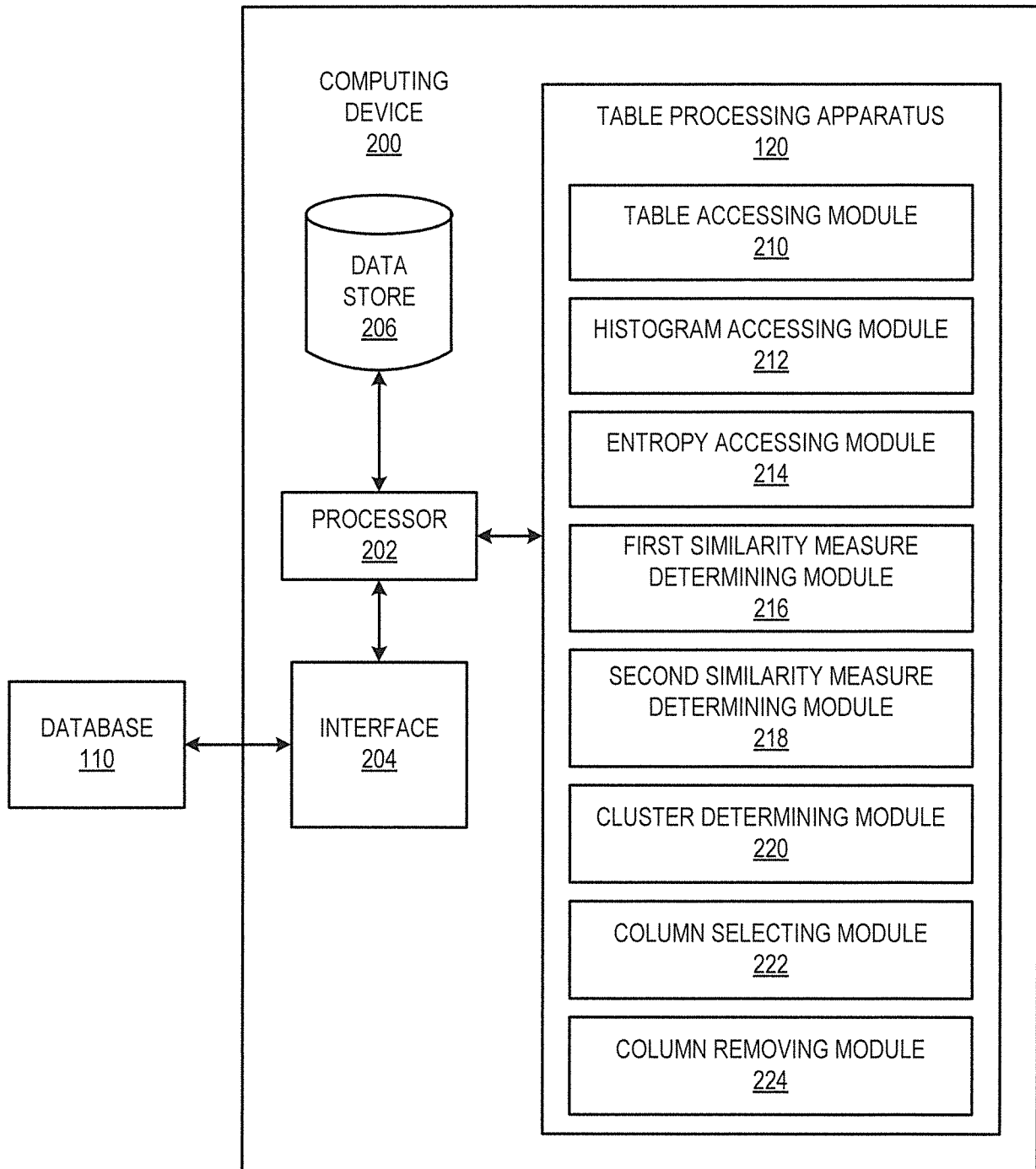
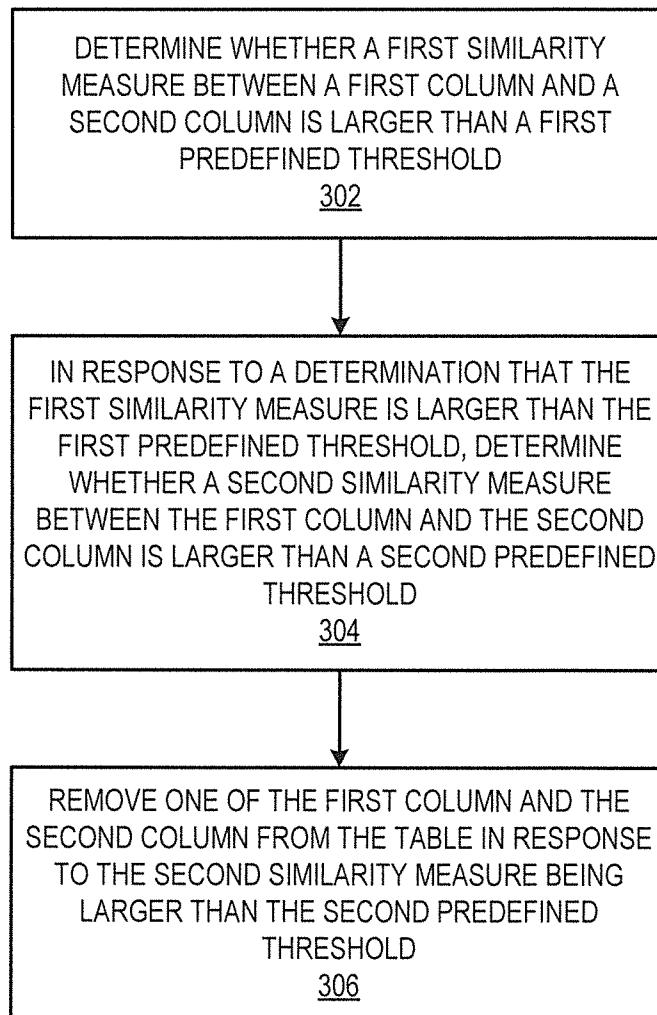
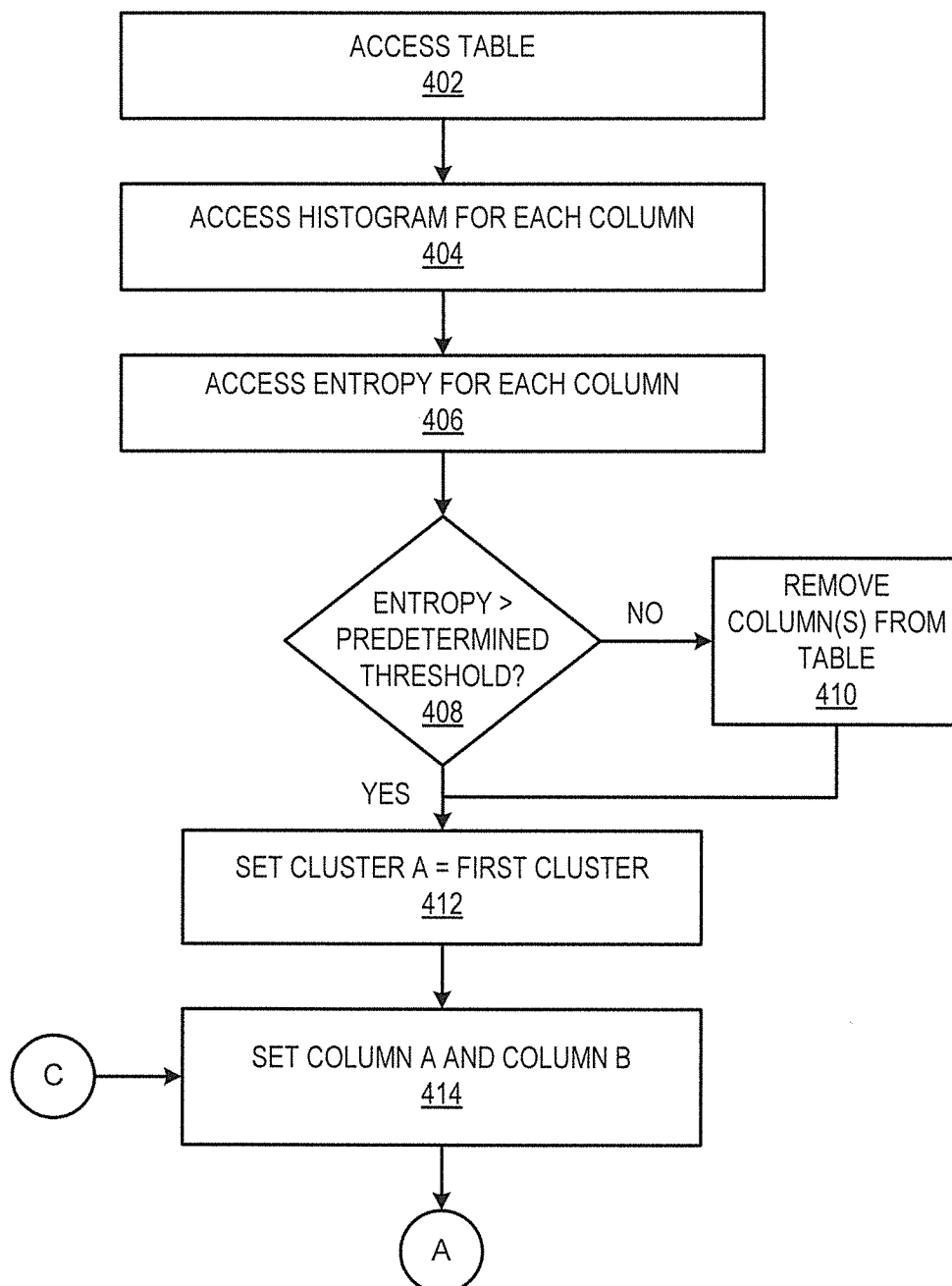
100

FIG. 2

3 / 8

300*FIG. 3*

4 / 8

400**FIG. 4A**

5 / 8

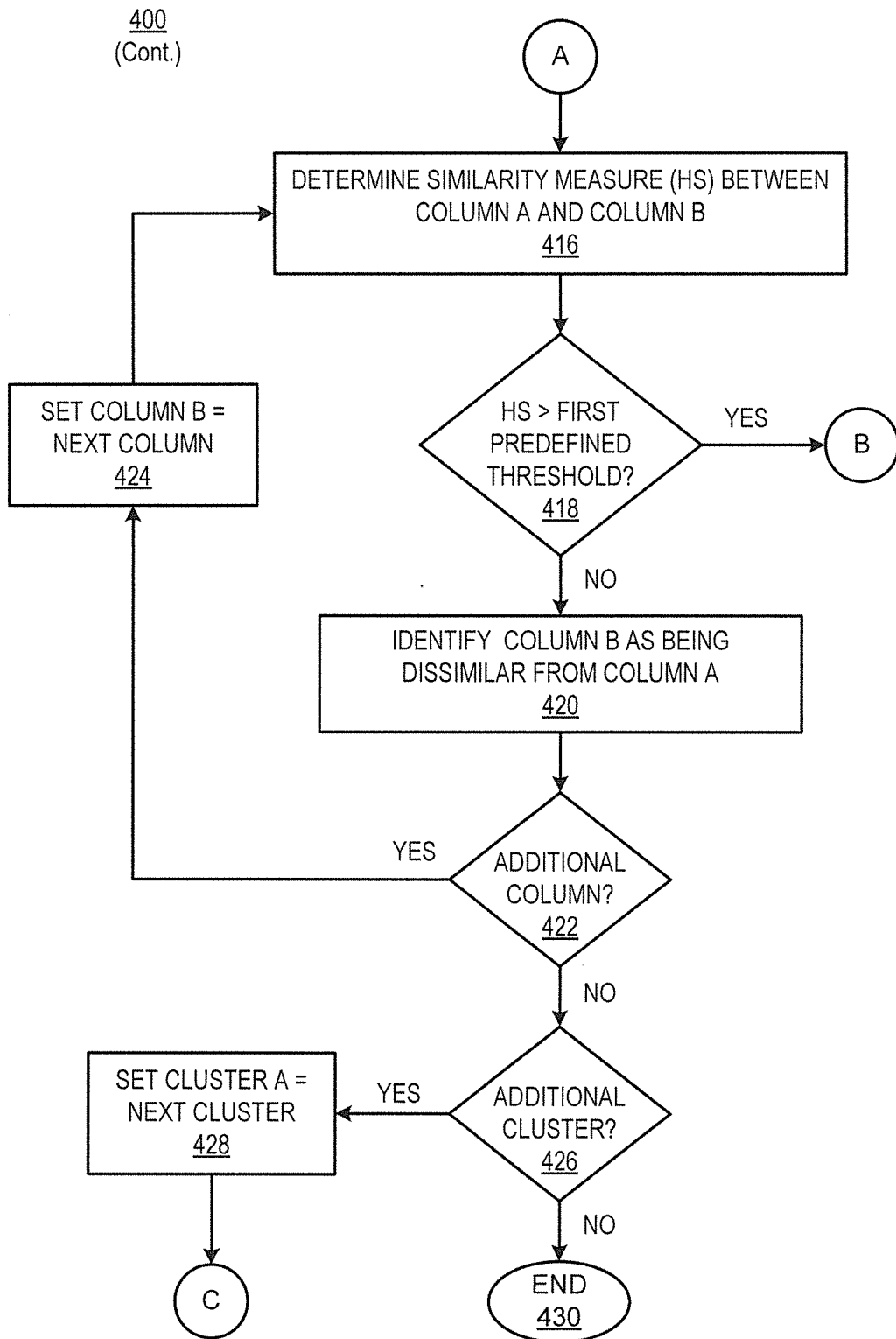


FIG. 4B

6 / 8

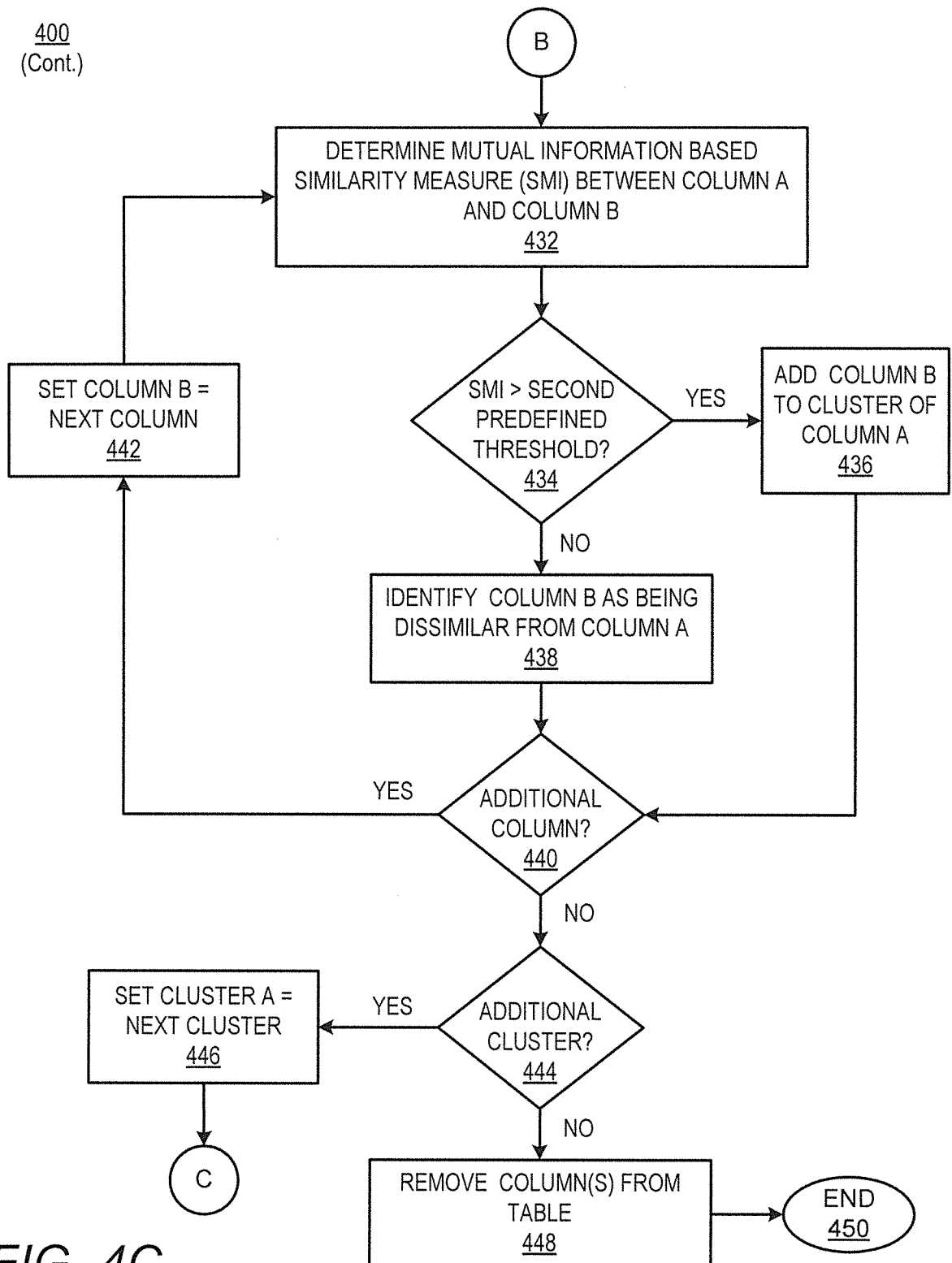
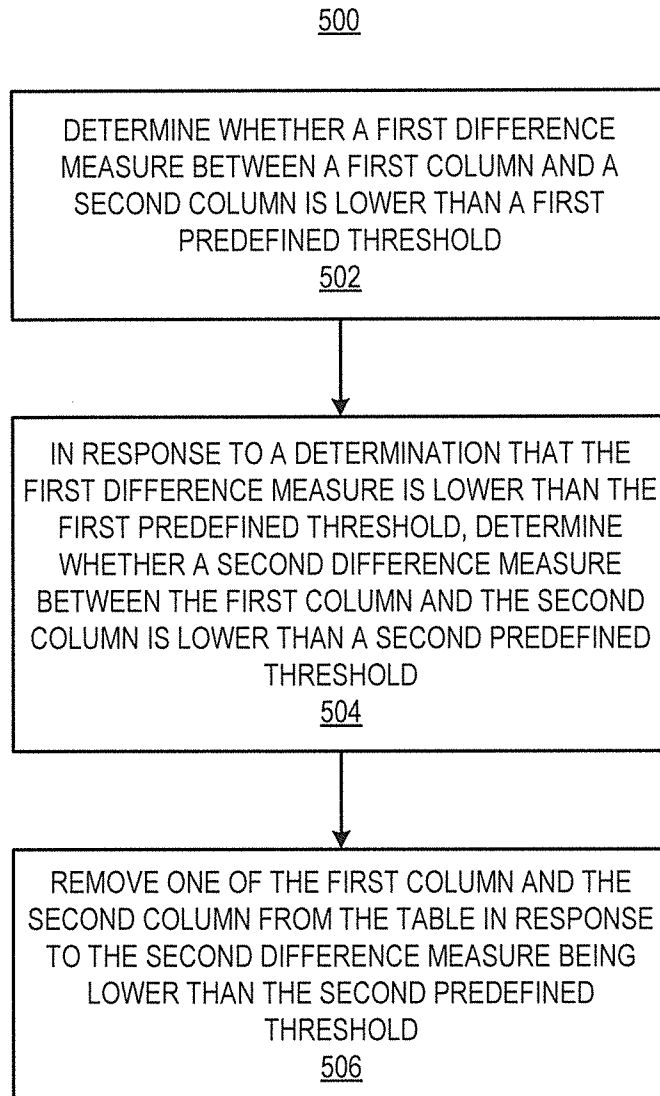


FIG. 4C

7 / 8

*FIG. 5*

8 / 8

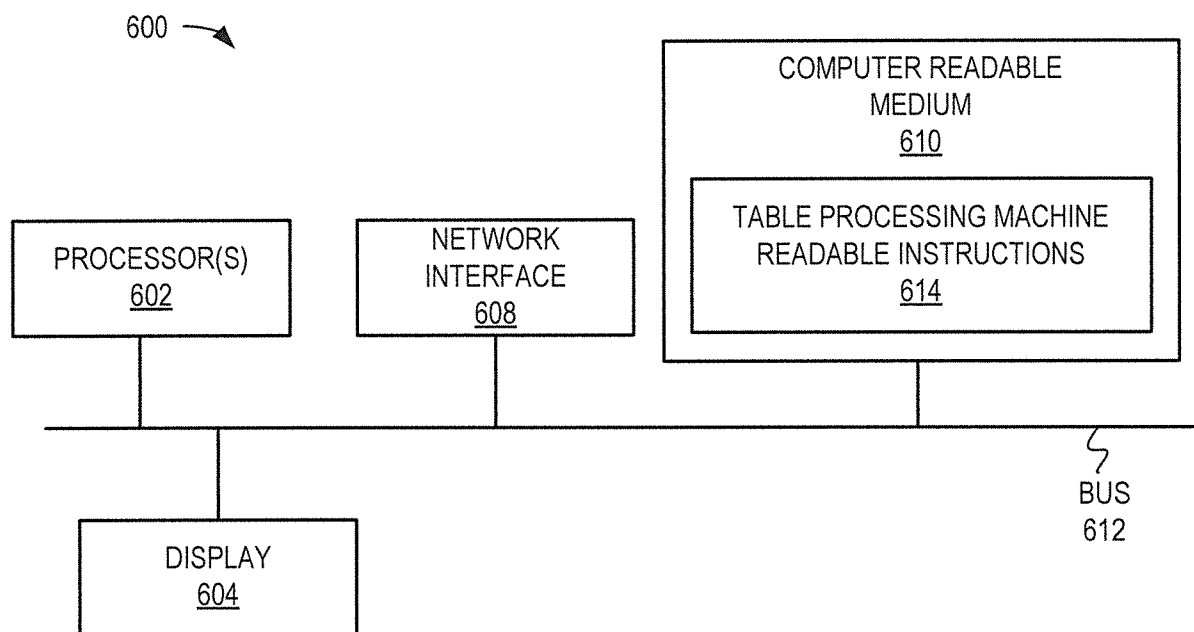


FIG. 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2014/040234**A. CLASSIFICATION OF SUBJECT MATTER****G06F 17/30(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 17/30; G06F 17/00; G06F 7/00; G06F 12/00; G06F 17/21

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: database, table, column, similarity, threshold, removing, comparing, histogram, entropy, and similar terms.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2005-0240615 A1 (BARSNESS, ERIC L. et al.) 27 October 2005 See paragraphs [0050], [0056]-[0057], [0059], [0061], and [0064]-[0067]; claims 1-2; and figures 1-4.	1,4-8,14
Y		2-3,9-13,15
Y	US 2013-0041910 A1 (BETZ, JONATHAN T. et al.) 14 February 2013 See paragraphs [0049]-[0050] and [0058]-[0061]; claim 1; and figure 3.	3,9-13,15
Y	US 2005-0289191 A1 (POPIVANOV, IVAN) 29 December 2005 See paragraphs [0056]-[0065]; claim 1; and figure 9A.	2-3,10
A	US 2007-0214168 A1 (SANDROCK GRABSKY, JOHN B. et al.) 13 September 2007 See paragraphs [0048]-[0052] and figures 6-7.	1-15
A	US 2012-0173226 A1 (MCEVOY, JOHN et al.) 05 July 2012 See paragraphs [0028]-[0029] 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

30 January 2015 (30.01.2015)

Date of mailing of the international search report

30 January 2015 (30.01.2015)

Name and mailing address of the ISA/KR

International Application Division
Korean Intellectual Property Office
189 Cheongsa-ro, Seo-gu, Daejeon Metropolitan City, 302-701,
Republic of Korea

Facsimile No. ++82 42 472 3473

Authorized officer

NHO, Ji Myong

Telephone No. +82-42-481-8528



INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2014/040234

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2005-0240615 A1	27/10/2005	US 7725498 B2	25/05/2010
US 2013-0041910 A1	14/02/2013	US 2007-198597 A1 US 8244689 B2	23/08/2007 14/08/2012
US 2005-0289191 A1	29/12/2005	US 2008-0052038 A1 US 7299248 B2 US 7774163 B2	28/02/2008 20/11/2007 10/08/2010
US 2007-0214168 A1	13/09/2007	EP 1700237 A1 US 2005-0114365 A1 US 7765247 B2 US 8560500 B2 WO 2005-052817 A1	13/09/2006 26/05/2005 27/07/2010 15/10/2013 09/06/2005
US 2012-0173226 A1	05/07/2012	US 8751216 B2	10/06/2014