(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
23 January 2014 (23.01.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/012576 A1

(51) International Patent Classification:
G06F 17/30 (2006.01)

(21) International Application Number:
PCT/EP2012/063930

(22) International Filing Date:
16 July 2012 (16.07.2012)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant *(for all designated States except US)*: **QATAR FOUNDATION** [QA/QA]; P.O. Box 5825, Doha (QA).

(71) Applicant *(for TT only)*: **HOARTON, Lloyd** [GB/GB]; Forresters, Sherborne House, 119-121 Cannon Street, London, Greater London EC4N 5AT (GB).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **BESKALES, George** [CA/QA]; Qatar Foundation, P.O. Box 5825, Doha (QA). **KALDAS, Ihab Francis Ilyas** [CA/QA]; 294 Wiltshire Place, Waterloo, Ontario (CA).

(74) Agent: **HOARTON, Lloyd Douglas Charles**; Forresters, Sherborne House, 119-121 Cannon Street, London, Greater London EC4N 5AT (GB).

(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, 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, IS, JP, KE, KG, KM, 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, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, 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, ML, MR, NE, SN, TD, TG).

Published:

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

(54) **Title**: A METHOD AND SYSTEM FOR INTEGRATING DATA INTO A DATABASE

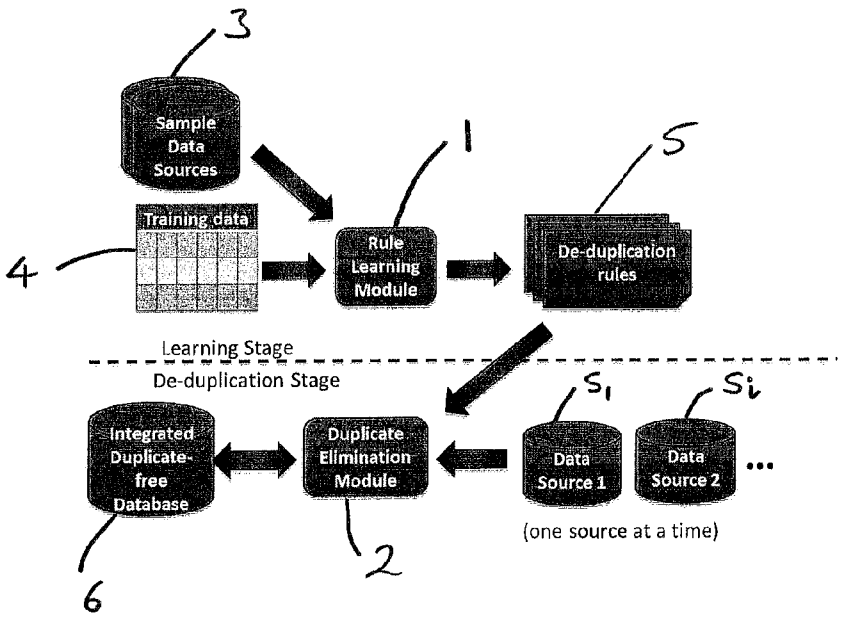

Figure 1

(57) **Abstract**: A method and system for integrating data into a database (6) comprises storing data from a plurality of data sources (S_1,S_i). The system comprises a rule learning module (1) and a duplicate elimination module (2). The rule learning module (1) operates in an initial rule learning stage. The duplicate elimination module (2) then operates in a de-duplication stage using the learnt rules. The de-duplication rules use conditional probability to determine the probability of records in the data sources (S_1,S_i) being duplicates of one another. Duplicate records are integrated and stored in the integrated database (6).

WO 2014/012576 A1

A METHOD AND SYSTEM FOR INTEGRATING DATA INTO A DATABASE

Description of Invention

The present invention relates to a method and system for integrating data into a database and more particularly relates to a method and system for integrating data from multiple data sources into a database whilst minimising data duplication.

The volume of data stored in databases is growing exponentially, as is the rate at which the data becomes available. The data which is to be stored in databases is also becoming more complex since each record often comprises a large number of different attributes.

Data from multiple data sources often needs to be integrated into a central database. With large volumes of data, integration into a central database can result in the data in the central database being plagued with errors and anomalies, such as duplicate records. Duplicate records are database records that refer to the same real-world entity. Duplicate records have a negative impact on the effectiveness of data querying and analysis tasks. The result is poor data analysis efficiency and a higher cost to enterprises using the data.

It has been proposed previously to use de-duplication rules to de-duplicate data. The de-duplication rules are learnt from training data which is either passively collected before the learning process or actively collected during the learning process. Conventional methods which use such de-duplication rules are, however, limited and are not able to handle heterogeneous data representing different types of entities due to the diverse characteristics of each entity type.

A distributed duplicate elimination method has been proposed previously which parallelises the de-duplication process using the MapReduce model. However, the problem with this method is that it is incapable of operating on records which have a sparseness of data, a large number of attributes or heterogeneous attributes/entity types.

A further technique has been proposed previously in which duplicate detection is carried out using structured query language (SQL) queries that are processed using a database management system (DBMS). The problem with this technique is that it must rely on an index to record and retrieve similar records. Building and updating such an index is prohibitively slow for large data sets with thousands of attributes.

There is a need for a method and system for integrating data from multiple data sources into a central database whilst minimising duplication of records in the central database.

The present invention seeks to provide an improved method and system for integrating data into a database.

According to one aspect of the present invention, there is provided a method for integrating data into a database, the method comprising storing data from a first data source, the data comprising a plurality of records which each comprise a plurality of attributes; storing data from a second data source, the data comprising a plurality of records which each comprise a plurality of attributes; analysing the attributes of the records in the data from the first and second data sources; identifying candidate pairs of records, each candidate pair comprising a record from the first data source which has at least one attribute that is similar to at least one attribute in a record from the second data source; generating a similarity value which is indicative of the level of similarity between the two records in each candidate pair; consolidating the two records

of each candidate pair with a similarity value above a predetermined level into one consolidated record; and storing each consolidated record in an integrated database.

Preferably, before the step of identifying candidate pairs of records, the method further comprises identifying true non-duplicate records in the data from the first and second data sources and storing each true non-duplicate record in the integrated database.

Conveniently, before the step of analysing the attributes of the records, the method further comprises learning at least one rule using sample data and then subsequently using each rule in the method to identify the candidate pairs of records and generate the similarity value.

Advantageously, the sample data comprises only duplicate records.

Preferably, the method comprises grouping the records in the sample data into groups of records with records in each group having at least one attribute which is the same as at least one attribute in another record in the group.

Conveniently, the similarity value is a conditional probability value which is indicative of the probability of the two records in a candidate pair being duplicates of one another.

Advantageously, the step of learning each rule comprises calculating the probability distribution by comparing the attributes of known duplicate records with a random sample of pairs of records.

Preferably, the calculation is based on the Bayes' rule.

Conveniently, the method further comprises calculating the conditional probability value from the probability distributions using the Naïve Bayes' rule.

Advantageously, the method comprises using a plurality of rules to identify duplicate data records in the data sources.

Preferably, the method further comprises comparing the similarity value of each candidate pair with the similarity value of each other candidate pair and clustering each candidate pair into a set of disjoint clusters.

Conveniently, the method comprises storing information representing each cluster for use with further data.

Advantageously, the steps of the method are repeated for at least one further data source, one data source at a time.

According to another aspect of the present invention, there is provided a computer readable storage medium storing machine readable instructions that, when executed by a processor, implement a method for integrating data into a database comprising: storing data from a first data source, the data comprising a plurality of records which each comprise a plurality of attributes; storing data from a second data source, the data comprising a plurality of records which each comprise a plurality of attributes; analysing the attributes of the records in the data from the first and second data sources; identifying candidate pairs of records, each candidate pair comprising a record from the first data source which has at least one attribute that is similar to at least one attribute in a record from the second data source; generating a similarity value which is indicative of the level of similarity between the two records in each candidate pair; consolidating the two records of each candidate pair with a similarity value above a predetermined level into one consolidated record; and storing each consolidated record in an integrated database.

Preferably, the computer readable storage medium further stores instructions that, when executed by the processor, implement the method for integrating data into a database further comprising: before the step of identifying candidate pairs of records, identifying true non-duplicate records in the data from the first and second data sources and storing each true non-duplicate record in the integrated database.

Conveniently, the computer readable storage medium further stores instructions that, when executed by the processor, implement the method for integrating data into a database further comprising: before the step of analysing the attributes of the records, learning at least one rule using sample data and then subsequently using each rule in the method to identify the candidate pairs of records and generate the similarity value.

Advantageously, the sample data comprises only duplicate records.

Preferably, the computer readable storage medium further stores instructions that, when executed by the processor, implement the method for integrating data into a database further comprising: grouping the records in the sample data into groups of records with records in each group having at least one attribute which is the same as at least one attribute in another record in the group.

Conveniently, the similarity value is a conditional probability value which is indicative of the probability of the two records in a candidate pair being duplicates of one another.

Advantageously, the step of learning each rule comprises calculating the probability distribution by comparing the attributes of known duplicate records with a random sample of pairs of records.

Preferably, the calculation is based on the Bayes' rule.

Advantageously, the computer readable storage medium further stores instructions that, when executed by the processor, implement the method for integrating data into a database further comprising: calculating the conditional probability value from the probability distributions using the Naïve Bayes' rule.

Preferably, the computer readable storage medium further stores instructions that, when executed by the processor, implement the method for integrating data into a database further comprising: using a plurality of rules to identify duplicate data records in the data sources.

Conveniently, the computer readable storage medium further stores instructions that, when executed by the processor, implement the method for integrating data into a database further comprising: comparing the similarity value of each candidate pair with the similarity value of each other candidate pair and clustering each candidate pair into a set of disjoint clusters.

Advantageously, the computer readable storage medium further stores instructions that, when executed by the processor, implement the method for integrating data into a database further comprising: storing information representing each cluster for use with further data.

Preferably, the computer readable storage medium further stores instructions that, when executed by the processor, implement the method for integrating data into a database further comprising: repeating the steps of the method for at least one further data source, one data source at a time.

In order that the invention may be more readily understood, and so that further features thereof may be appreciated, embodiments of the invention will now be described, by way of example, with reference to the accompanying drawings in which:

Figure 1 is a schematic representation of a system for integrating data into a database according to one embodiment of the invention; and

Figure 2 is a schematic block diagram of an apparatus for implementing an embodiment of the invention.

Referring to figure 1 of the accompanying drawings, a system of an embodiment of the invention incorporates a rule learning module 1 and a duplicate elimination module 2. The system operates in two stages; a learning stage and a de-duplication stage.

The learning stage comprises loading sample data sources 3 and training data 4 and processing the data in the rule learning module 1. The rule learning module 1 is configured to generate rules 5 based on the sample data sources 3 and the training data 4. As will become clear from the description below, the rules 5 are subsequently used by the duplicate elimination module 2 to eliminate duplicates during the de-duplication stage.

The sample data sources 3 comprise data records which each comprise a plurality of entity types and attributes. In a preferred embodiment, the sample data sources 3 comprise duplicate records which are known duplicates of one another.

The rule learning module 1 initially detects the entity types in the records of the sample data sources 3. The rule learning module 1 then groups the sample input records into groups based on the entity types in each record. The rule

learning module 1 analyses each group of records and generates a set of rules by applying a conditional probability algorithm to the records to provide a set of rules which indicate the probability of two records being duplicates of one another.

The rule learning module 1 identifies all distinct entity types from the sample of records. Each entity type in the sample preferably has unique characteristics, such as specific attributes or attribute ranges. The rule learning module 1 learns a separate set of de-duplication rules for each entity type from the training data 4 belonging to each entity type.

For each attribute A, the rule learning module 1 learns the conditional probability of a given similarity value of attribute A of two randomly selected records given that the two records are duplicates, denoted $Pr(Sim_A=s|M)$, where s could be a real value or null (indicating that at least one of the records has A=null). The rule learning module 1 also learns the conditional probability $Pr(Sim_A=s|U)$, where U indicates that the two selected records are non-duplicates. The rule learning module 1 learns these probability distributions by comparing the attributes of known duplicates to the attributes of a random sample of pairs. This is preferably done using the Bayes' rule. Using these probability distributions, the rule learning module 1 computes the probability that two records are duplicates given their attribute similarities, denoted $Pr(M|Sim_A,\ldots,Sim_Z)$. This is preferably done using the Naïve Bayes' rule defined here as Equation 1:

$$Pr(M|Sim_A=s_A,\ldots,Sim_Z=s_Z)= 1/Z\ (Pr(M)\cdot Pr(Sim_A=s_A|M)\ldots Pr(Sim_Z=s_Z|M))$$

where Z is a normalisation constant. For each attribute A, the rule learning module 1 determines a threshold T(A) such that the probability of two records being duplicates is significant given that the similarity of their A values is higher than T(A). For some attributes, there might be no such threshold

satisfying this condition. The attributes with valid thresholds are recorded as "Distinctive Attributes".

The operation of the rule learning module 1 will now be described in terms of a practical example. In this example, the sample data sources 3 comprise three data sources S_1, S_2, S_3 that each contains data about various leisure activities such as movie theatres, golf clubs, downhill skiing, and restaurants. Each record in these sources represents a single activity and it is associated with multiple common attributes such as title, address, rating, latitude, and longitude. Also, each entity type involves a number of unique attributes (e.g., attribute "vertical drop" for downhill skiing entities).

A sample of records in S_1 and S_2 is available to the learning module 1 and some of these records are labelled as duplicates. Entity types could be either given to the system, or automatically learnt. For example, in the latter case, the learning module 1 detects different entity types (e.g., movie theatres, golf clubs, downhill skiing, and restaurants) based on the characteristics of the records (e.g., the present non-null attributes, the ranges of attribute values, and the originating data sources). The learning module 1 detects the entity type of each record, and groups the records based on their entity types. The groups of records are either disjoint or overlapping.

For each group of records, the learning module 1 obtains a set of rules. An example rule might indicate that "any two records with a difference in latitude or longitude greater than 0.1 have a probability of being duplicates equal to 0.001". Another example rule is that "any two records with missing ZIP code have a probability of being duplicates equal to 0.2".

Once the rule learning module 1 has learnt the required de-duplication rules from the sample data sources 3, the system can operate in the de-duplication stage. In the de-duplication stage, the duplicate elimination module 2 uses the

learnt rules to detect duplicate records in data sources $S_1 \ldots S_i$ and consolidate and store the records in a central integrated duplicate-free database 6. At any point in time, the integrated database B contains duplicate-free records from the data sources $S_1,\ldots,S_i$.

The duplicate elimination module 2 accepts data coming from a new source S_{i+1}, detects the duplicate records within S_{i+1}, and duplicate records that span S_{i+1} and any other sources from $S_1,\ldots,S_i$. Duplicates are then consolidated and added to the integrated database B. De-duplication is preferably performed in a holistic way based on all learnt rules. That is, the duplicate detection algorithm considers all relevant rules in order to obtain a single value indicating the probability that certain records are duplicates.

The rules that are used for de-duplicating certain records are preferably selected based on the entity types of such records. For instance, in the example described above, rules involving the attribute “vertical drop” are only used when the involved records represent “downhill skiing” sites.

The de-duplication process is mainly divided into sub-tasks: “similarity join” and “clustering”. Similarity join provides a list of record pairs that have considerable probability of being duplicates. The clustering task uses the list of record pairs to group the records into a set of disjoint clusters such that each cluster of records represents the same real-world entity.

In probabilistic clustering, the outcome is multiple possible clusterings of records rather than a single clustering. The importance of keeping a number of possible clusterings is due to continuously adding new data sources. New data that arrives at a certain point of time can change the systems knowledge about the correct clustering of previously added records. Therefore, the system preferably retains clustering information that can potentially be useful at a later stage.

For each new data source, the duplicate elimination module 2 first obtains a set of "candidate record pairs" that could be potentially duplicates. The records in each pair could belong entirely to the new data source, or to the new data source and an old data source. A candidate pair is a pair of records that has at least one distinctive attribute (call it A) with similarity above a predetermined threshold T(A). Pairs that do not satisfy this criterion are pruned (i.e., assumed non-duplicates). The module 2 uses the learnt rules to compute the probability of each pair being duplicates. This procedure is described in the following algorithm.

Algorithm 1 Sim_Join(T(A),...,T(Z))

$C \leftarrow \varphi$

For each distinctive attribute A_i

Insert into C all tuple pairs (t1,t2) such that $Sim_{Ai} \geq T(A_i)$

End for

For each pair $(t1,t2) \in C$ do

Compute the values of $Sim_A,...,Sim_Z$

Compute $Pr(M = T|Sim= s_A,...,Sim_Z= s_Z)$ based on Equation 1

End for

Return pairs in C

The module 2 removes record pairs with low probability of being duplicates and uses the remaining pairs to cluster the records into groups such that each group (cluster) of records represents one real-world entity.

In order to allow efficient updating to the record clustering when new data arrives, the module 2 avoids computationally intensive operations such as splitting and restructuring existing clusters. To achieve this goal, the module 2 stores multiple possible clusterings, each of which is associated with the probability of being correct. The module 2 updates these clusterings incrementally using simple operations such as appending a new record to a clustering, or merging two existing clusters. At any point of time, the module 2 is able to provide the most probable clustering, which is close to the output of batch-clustering that has all data available from the beginning.

The following example describes an example of the operation of the duplicate elimination module 2. In this example, the following records are found in data sources S_1,S_2 and S_3:

- **r_1** (Entity Type: "Golf", Title: "The Ocean At Olympic Club", City: "San Francisco", State, "CA", architect: "Sam Whiting", holes: 18)
- **r_2** (Entity Type: "Golf", Title: "the Cliffs at Olympic Club", State: "California", holes: 9)
- **r_3** (Entity Type: "Golf and Tennis", Title: "Golfsmith Golf & Tennis", Phone: "(415) 974-6979", City: "San Francisco", State: "CA")

The above records are processed by first obtaining records with at least one similar attribute (e.g., r_1 and r_2 have similar title). Then, the relevant learnt rules (e.g., regarding Golf sites in our example) are used to determine the probability that each candidate pair of records being duplicate. Pairs with small probabilities of being duplicates are discarded, and the remaining pairs are used for clustering the records.

For example, assume that the similarity join process returned two candidate pairs (r_1,r_2) and (r_2,r_3) such that the probability that r_1 and r_2 are duplicates is 0.8 and the probability that r_2 and r_3 are duplicates is 0.1. The clustering

algorithm discards the pairs (r_2,r_3) and returns the clustering {r_1,r_2},{r_3}, which indicates that r_1 and r_2 refer to the same real-world entity and r_3 refers to a different entity. Records r_1 and r_2 are then consolidated into one record r_{12} which could be as follows.

- **r_{12}** (Entity Type: “Golf”, Title: “The Ocean At Olympic Club”, City: “San Francisco”, State, “CA”, architect: “Sam Whiting”, holes: 18)

Embodiments of the invention use learnt de-duplication rules to de-duplicate data originating from a plurality of data sources. The system is operable to integrate data from a plurality of data sources into an integrated central database one source at a time. The system is also operable to learn the rules from sample data by using only samples of duplicate records.

Embodiments of the invention are particularly well suited to de-duplicating and integrating large volumes of data. Embodiments of the invention integrate the data more efficiently and accurately than conventional data de-duplication systems.

Figure 2 is a schematic block diagram of an apparatus according to an embodiment of the invention which is suitable for implementing any of the systems or processes described above. Apparatus 400 includes one or more processors, such as processor 401, providing an execution platform for executing machine readable instructions such as software. Commands and data from the processor 401 are communicated over a communication bus 399. The system 400 also includes a main memory 402, such as a Random Access Memory (RAM), where machine readable instructions may reside during runtime, and a secondary memory 405. The secondary memory 405 includes, for example, a hard disk drive 407 and/or a removable storage drive 430, representing a floppy diskette drive, a magnetic tape drive, a compact disk drive, etc., or a nonvolatile memory where a copy of the machine readable

instructions or software may be stored. The secondary memory 405 may also include ROM (read only memory), EPROM (erasable, programmable ROM), EEPROM (electrically erasable, programmable ROM). In addition to software, data representing any one or more of updates, possible updates or candidate replacement entries, and listings for identified tuples may be stored in the main memory 402 and/or the secondary memory 405. The removable storage drive 430 reads from and/or writes to a removable storage unit 409 in a well-known manner.

A user interfaces with the system 400 with one or more input devices 411, such as a keyboard, a mouse, a stylus, and the like in order to provide user input data. The display adaptor 415 interfaces with the communication bus 399 and the display 417 and receives display data from the processor 401 and converts the display data into display commands for the display 417. A network interface 419 is provided for communicating with other systems and devices via a network (not shown). The system can include a wireless interface 421 for communicating with wireless devices in the wireless community.

It will be apparent to one of ordinary skill in the art that one or more of the components of the system 400 may not be included and/or other components may be added as is known in the art. The system 400 shown in figure 2 is provided as an example of a possible platform that may be used, and other types of platforms may be used as is known in the art. One or more of the steps described above may be implemented as instructions embedded on a computer readable medium and executed on the system 400. The steps may be embodied by a computer program, which may exist in a variety of forms both active and inactive. For example, they may exist as software program(s) comprised of program instructions in source code, object code, executable code or other formats for performing some of the steps. Any of the above may be embodied on a computer readable medium, which include storage devices

and signals, in compressed or uncompressed form. Examples of suitable computer readable storage devices include conventional computer system RAM (random access memory), ROM (read only memory), EPROM (erasable, programmable ROM), EEPROM (electrically erasable, programmable ROM), and magnetic or optical disks or tapes. Examples of computer readable signals, whether modulated using a carrier or not, are signals that a computer system hosting or running a computer program may be configured to access, including signals downloaded through the Internet or other networks. Concrete examples of the foregoing include distribution of the programs on a CD ROM or via Internet download. In a sense, the Internet itself, as an abstract entity, is a computer readable medium. The same is true of computer networks in general. It is therefore to be understood that those functions enumerated above may be performed by any electronic device capable of executing the above-described functions.

In one embodiment, equivalence classes 405 can reside in memory 402 having been derived from records of a database 209. One or more of algorithms of blocks 300, 305 or 307 can reside in memory 402 such as to provide respective engines 403 for cleaning, merging and selecting records of a database, including a modified instance of a database for example. That is, engine 403 can be a cleaning engine or a merge engine which is operable to perform the processes associated with the tasks of blocks 300, 305, 307 for example.

A database 209 is shown in figure 2 as a standalone database connected to bus 399. However, it can be a database which can be queried and have data written to it from a remote location using the wired or wireless network connections mentioned above. Alternatively, database 209 may be stored in memory 405, such as on a HDD of system 400 for example.

When used in this specification and claims, the terms "comprises" and "comprising" and variations thereof mean that the specified features, steps or integers are included. The terms are not to be interpreted to exclude the presence of other features, steps or components.

5

CLAIMS

1. A method for integrating data into a database, the method comprising:

storing data from a first data source, the data comprising a plurality of records which each comprise a plurality of attributes;

storing data from a second data source, the data comprising a plurality of records which each comprise a plurality of attributes;

analysing the attributes of the records in the data from the first and second data sources;

identifying candidate pairs of records, each candidate pair comprising a record from the first data source which has at least one attribute that is similar to at least one attribute in a record from the second data source;

generating a similarity value which is indicative of the level of similarity between the two records in each candidate pair;

consolidating the two records of each candidate pair with a similarity value above a predetermined level into one consolidated record; and

storing each consolidated record in an integrated database.

2. A method according to claim 1, wherein, before the step of identifying candidate pairs of records, the method further comprises identifying true non-duplicate records in the data from the first and second data sources and storing each true non-duplicate record in the integrated database.

3. A method according to claim 1 or claim 2, wherein, before the step of analysing the attributes of the records, the method further comprises learning at least one rule using sample data and then subsequently using each rule in the method to identify the candidate pairs of records and generate the similarity value.

4. A method according to claim 3, wherein the sample data comprises only duplicate records.

5. A method according to claim 3 or claim 4, wherein the method comprises grouping the records in the sample data into groups of records with records in each group having at least one attribute which is the same as at least one attribute in another record in the group.

6. A method according to any one of the preceding claims, wherein the similarity value is a conditional probability value which is indicative of the probability of the two records in a candidate pair being duplicates of one another.

7. A method according to claim 6 as dependent on any one of claims 3 to 5, wherein the step of learning each rule comprises calculating the probability distribution by comparing the attributes of known duplicate records with a random sample of pairs of records.

8. A method according to claim 7, wherein the calculation is based on the Bayes' rule.

9. A method according to claim 7 or claim 8, wherein the method further comprises calculating the conditional probability value from the probability distributions using the Naïve Bayes' rule.

10. A method according to any one of claims 3 to 9, wherein the method comprises using a plurality of rules to identify duplicate data records in the data sources.

11. A method according to any one of the preceding claims, wherein the method further comprises comparing the similarity value of each candidate pair with the similarity value of each other candidate pair and clustering each candidate pair into a set of disjoint clusters.

12. A method according to claim 11, wherein the method comprises storing information representing each cluster for use with further data.

13. A method according to any one of the preceding claims, wherein the steps of the method are repeated for at least one further data source, one data source at a time.

14. A computer readable storage medium storing machine readable instructions that, when executed by a processor, implement a method for integrating data into a database comprising:

storing data from a first data source, the data comprising a plurality of records which each comprise a plurality of attributes;

storing data from a second data source, the data comprising a plurality of records which each comprise a plurality of attributes;

analysing the attributes of the records in the data from the first and second data sources;

identifying candidate pairs of records, each candidate pair comprising a record from the first data source which has at least one attribute that is similar to at least one attribute in a record from the second data source;

generating a similarity value which is indicative of the level of similarity between the two records in each candidate pair;

consolidating the two records of each candidate pair with a similarity value above a predetermined level into one consolidated record; and

storing each consolidated record in an integrated database.

15. A computer readable storage medium according to claim 14 further storing instructions that, when executed by the processor, implement the method for integrating data into a database further comprising: before the step of identifying candidate pairs of records, identifying true non-duplicate records

in the data from the first and second data sources and storing each true non-duplicate record in the integrated database.

16. A computer readable storage medium according to claim 14 or claim 15 further storing instructions that, when executed by the processor, implement the method for integrating data into a database further comprising: before the step of analysing the attributes of the records, learning at least one rule using sample data and then subsequently using each rule in the method to identify the candidate pairs of records and generate the similarity value.

17. A computer readable storage medium according to claim 16, wherein the sample data comprises only duplicate records.

18. A computer readable storage medium according to claim 16 or claim 17 further storing instructions that, when executed by the processor, implement the method for integrating data into a database further comprising: grouping the records in the sample data into groups of records with records in each group having at least one attribute which is the same as at least one attribute in another record in the group.

19. A computer readable storage medium according to any one of claims 14 to 18, wherein the similarity value is a conditional probability value which is indicative of the probability of the two records in a candidate pair being duplicates of one another.

20. A computer readable storage medium according to claim 19 as dependent on any one of claims 16 to 18, wherein the step of learning each rule comprises calculating the probability distribution by comparing the attributes of known duplicate records with a random sample of pairs of records.

21. A computer readable storage medium according to claim 20, wherein the calculation is based on the Bayes' rule.

22. A computer readable storage medium according to claim 20 or claim 21 further storing instructions that, when executed by the processor, implement the method for integrating data into a database further comprising: calculating the conditional probability value from the probability distributions using the Naïve Bayes' rule.

23. A computer readable storage medium according to any one of claims 16 to 22 further storing instructions that, when executed by the processor, implement the method for integrating data into a database further comprising: using a plurality of rules to identify duplicate data records in the data sources.

24. A computer readable storage medium according to any one of claims 14 to 23 further storing instructions that, when executed by the processor, implement the method for integrating data into a database further comprising: comparing the similarity value of each candidate pair with the similarity value of each other candidate pair and clustering each candidate pair into a set of disjoint clusters.

25. A computer readable storage medium according to claim 24 further storing instructions that, when executed by the processor, implement the method for integrating data into a database further comprising: storing information representing each cluster for use with further data.

26. A computer readable storage medium according to any one of claims 14 to 25 further storing instructions that, when executed by the processor, implement the method for integrating data into a database further comprising: repeating the steps of the method for at least one further data source, one data source at a time.

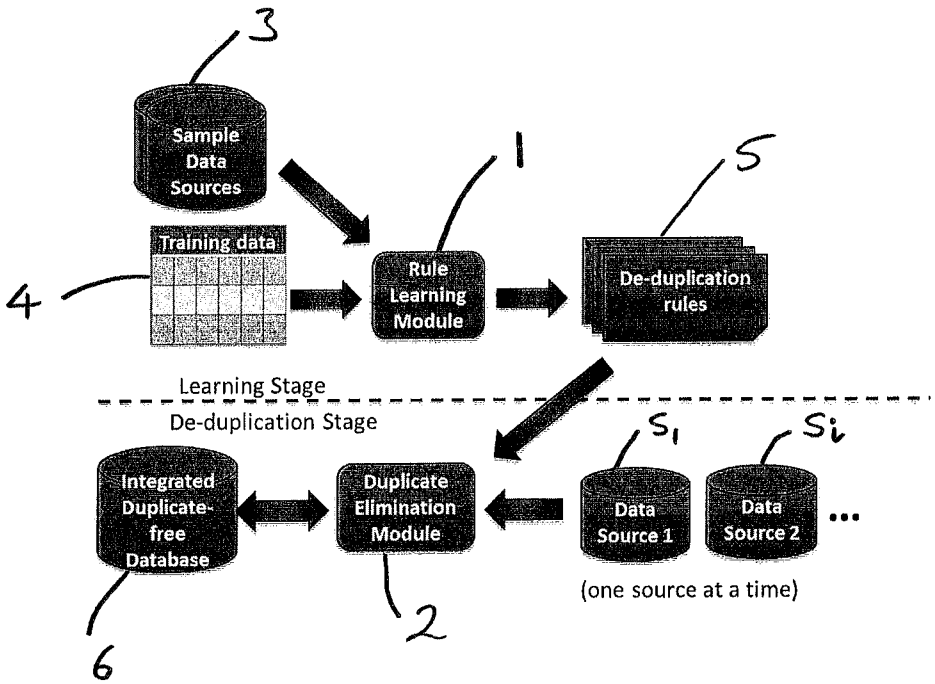
3
Sample Data Sources
1
5
Training data
4
Rule Learning Module
De-duplication rules
Learning Stage
De-duplication Stage
S1
Si
Integrated Duplicate-free Database
Duplicate Elimination Module
Data Source 1
Data Source 2
...
(one source at a time)
2
6

Figure 1

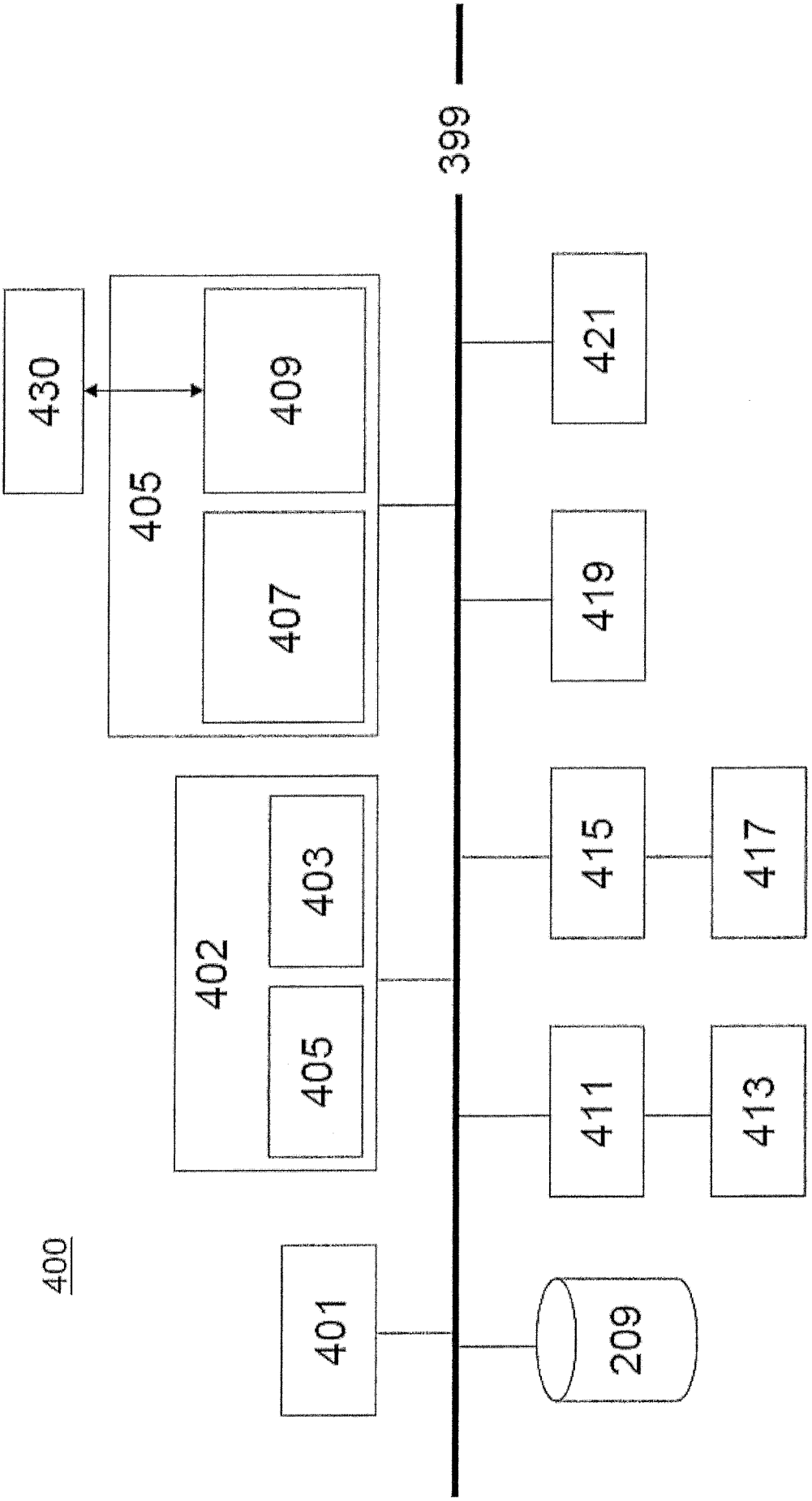
400
399
401
209
402
403
405
405
407
409
411
413
415
417
419
421
430

Figure 2

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2012/063930

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F17/30
ADD.

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

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2004/010497 A1 (BRADLEY PAUL S [US] ET AL) 15 January 2004 (2004-01-15) paragraphs [0028] - [0056]; figures 1-4	1-26
A	LIU XING: "Mineralization Information Mining from GIS Map & Attribution Database and Multi-sources Data Intergration", DATABASE TECHNOLOGY AND APPLICATIONS, 2009 FIRST INTERNATIONAL WORKSHOP ON, IEEE, PISCATAWAY, NJ, USA, 25 April 2009 (2009-04-25), pages 217-220, XP031515231, ISBN: 978-0-7695-3604-0 the whole document	1-26

-/--

[X] Further documents are listed in the continuation of Box C.

[X] 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	Date of mailing of the international search report
25 October 2012	02/11/2012

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Moon, Timothy

2

Form PCT/ISA/210 (second sheet) (April 2005)

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	NADKARNI P M ET AL: "Data extraction and ad hoc query of an entity-attribute-value database", JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION (JAMIA), HANLEY AND BELFUS, PHILADELPHIA, PA, US, vol. 5, no. 6, 1 November 1998 (1998-11-01), pages 511-527, XP008113615, ISSN: 1067-5027 the whole document -----	1-26

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2012/063930

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2004010497 A1	15-01-2004	NONE	