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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 October 2015 (29.10.2015)

WIPO | PCT

(10) International Publication Number
WO 2015/161899 A1

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

(21) International Application Number:
PCT/EP2014/058525

(22) International Filing Date:
25 April 2014 (25.04.2014)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **HEWLETT PACKARD DEVELOPMENT COMPANY L.P.** [US/US]; 11445 Compaq Center Drive West, Houston, Texas, 77070 (US).

(72) Inventor; and
(71) Applicant *(for US only)*: **VAQUERO GONZALEZ, Luis Miguel** [ES/GB]; Longdown Avenue, Stoke Gifford, Bristol BS34 8QZ (GB).

(74) Agents: **SWINDELL & PEARSON LTD** et al.; 48 Friar Gate, Derby Derbyshire DE1 1GY (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, 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).

Declarations under Rule 4.17:

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

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

Published:

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

(54) Title: DETERMINE RELATIONSHIPS BETWEEN ENTITIES IN DATASETS

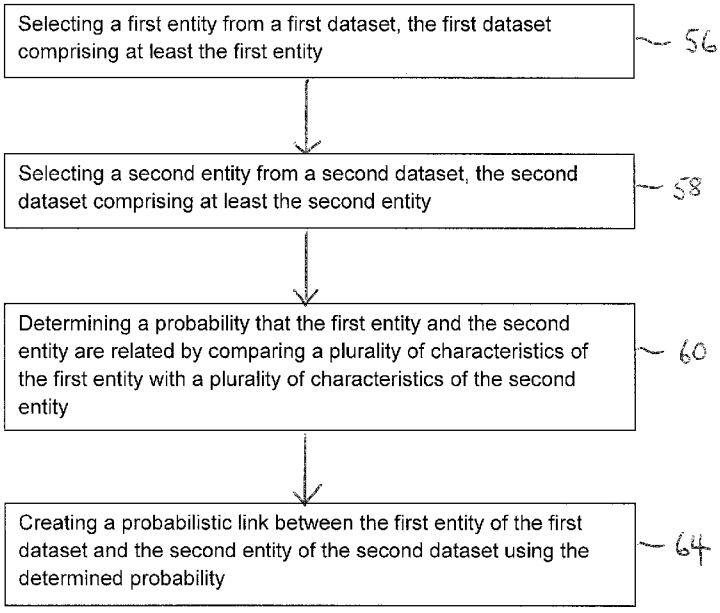


Fig. 4

(57) Abstract: A method to determine relationships between entities in a plurality of datasets in which a first entity is selected from a first dataset, the first dataset comprising at least the first entity. A second entity is selected from a second dataset, the second dataset comprising at least the second entity. A probability that the first entity and the second entity are related is determined by comparing a plurality of characteristics of the first entity with a plurality of characteristics of the second entity. In one example, a probabilistic link between the first entity of the first dataset and the second entity of the second dataset is created using the determined probability.

WO 2015/161899 A1

[0001] TITLE

[0002] Determine relationships between entities in datasets

[0003] BACKGROUND

[0004] Datasets are arranged to store information in a structured manner. For example, a dataset for a social network may store various details concerning a user, such as, user name, user date of birth, user interests, user workplace and so on. Datasets may, at least partially, store substantially the same information but using different wording. For example, in a first dataset, a user's interests may include 'playing musical instruments', whereas, in a second dataset, the same user's interests may include 'musician'. Datasets may also store different information for the same entity. For example, in a first dataset, there may be an entity "Father of two children", and in a second dataset, there may be an entity "computer software engineer" for the same person.

[0005] BRIEF DESCRIPTION

[0006] Reference will now be made by way of example only to the accompanying drawings in which:

[0007] Fig. 1 illustrates a schematic diagram of an apparatus according to an example;

[0008] Fig. 2 illustrates a multiplex graph according to an example;

[0009] Fig. 3 illustrates a flow diagram of a method according to an example;

[0010] Fig. 4 illustrates a flow diagram of another method according to an example; and

[0011] Fig. 5 illustrates a schematic diagram of a controller according to an example.

[0012] DETAILED DESCRIPTION

[0013] A user may find datasets difficult to manage (for example, search, edit), particularly as the size of the datasets increases. To aid the user, a dataset may be visually represented by a graph where nodes represent entities in the dataset, and connections represent the links (also referred to as 'edges') between the nodes. However, the datasets may, at least partially, store substantially the same information but using different wording. This may be challenging for the user to identify.

[0014] As described in the following paragraphs, probabilistic links may be created between entities in different datasets to enable a user to determine relationships between the datasets and to handle the data in the datasets more efficiently.

[0015] Fig. 1 illustrates a schematic diagram of an apparatus 10 including a controller 12, input apparatus 14, and output apparatus 16. The apparatus 10 may also be referred to as a "computer apparatus", a "computer" or a "data storage apparatus". In some examples, the apparatus 10 may be a single machine where the input apparatus 14 and the output apparatus 16 are connected to the controller 12 via a wired or a wireless link, and the controller 12, the input apparatus 14 and the output apparatus 16 are located in close proximity to one another (for example, in the same room as one another). In other examples, the apparatus 10 may be a distributed apparatus where the input apparatus 14 and the output apparatus 16 are located remotely from the controller 12 (for example, in a different room, in a different building, in a different city, or in different country).

[0016] In some examples, the apparatus 10 may be a module. As used herein, 'module' refers to a unit or apparatus that excludes certain parts/components that would be added by an end manufacturer or a user. For example, where the apparatus 10 is a module, the apparatus 10 may comprise the controller 12 and the remaining components (namely, the input apparatus 14 and the output apparatus 16) may be added by an end manufacturer.

[0017] The implementation of the controller 12 can be in hardware alone (for example, a circuit, a processor and so on), have certain aspects in software including firmware alone or can be a combination of hardware and software (including firmware).

[0018] The controller 12 may be implemented using instructions that enable hardware functionality, for example, by using executable computer program instructions 22 in a general-purpose or special-purpose processor 18 that may be stored on a computer readable storage medium 20 (disk, memory and so on) to be executed by such a processor 18.

[0019] The processor 18 is configured to read from and write to the memory 20. The processor 18 may also comprise an output interface via which data and/or commands are output by the processor 18 and an input interface via which data and/or commands are input to the processor 18.

[0020] The memory 20 stores a computer program 22 comprising computer program instructions that control the operation of the apparatus 10 when loaded into the processor 18. The computer program instructions 22 provide the logic and routines that enables the apparatus 10 to perform the methods illustrated in Fig. 3 and described in the following paragraphs. The processor 18 by reading the memory 20 is able to load and execute the computer program 22.

[0021] The computer program 22 may arrive at the apparatus 10 via any suitable delivery mechanism 24. The delivery mechanism 24 may be, for example, a non-transitory computer-readable storage medium, a computer program product, a memory device, a record medium such as a compact disc read-only memory (CD-ROM) or digital versatile disc (DVD), an article of manufacture that tangibly embodies the computer program 22. The delivery mechanism 24 may be a signal configured to reliably transfer the computer program 22. The apparatus 10 may propagate or transmit the computer program 22 as a computer data signal.

[0022] The memory 20 stores a plurality of datasets 26. The plurality of datasets 26 may be formed from a plurality of separate databases (that is, the plurality of datasets 26 is provided by a plurality of separate database files). In other examples, the plurality of datasets 26 may be formed from separate data matrices within a single database (that is, the plurality of datasets 26 is provided a single database file). The memory 20 may receive new datasets which are then stored in the memory 20.

[0023] The input apparatus 14 may comprise any suitable apparatus for enabling a user to provide an input signal to the controller 12. For example, the input apparatus 14 may include at least one of a keyboard, a keypad, a computer mouse, and a touch screen display. The controller 12 is arranged to receive input signals from the input apparatus 14.

[0024] The output apparatus 16 may comprise any suitable apparatus for providing information to a user. For example, the output apparatus 16 may include at least one display (such as a liquid crystal display or a light emitting diode display). The controller 12 is arranged to control the output apparatus 16 to provide information to the user.

[0025] Fig. 2 illustrates a multiplex graph 28 for a plurality of datasets 26 according to an example. A multiplex graph is a visual representation of a plurality of datasets whereby the nodes of the multiplex graph represent entities of a dataset, and connections between the nodes represent links between the entities. The entities of a dataset are provided in a vertical column (representing the contents of the dataset) and the columns for the datasets are positioned adjacent one another. The plurality of datasets 26 includes a first dataset 261 for 'IT Support', a second dataset 262 for 'Operations', a third dataset 263 for 'Service Management', and a fourth dataset 264 for 'Service Marketing'.

[0026] The first dataset 261 is represented by a first graph 30 and includes the entities: host 32 and log 34. The log entities 34 are connected to the host entities 32 via non-probabilistic links (that is, a link having a probability of 100%, in other words, a certain relationship). As illustrated in Fig. 2, the non-probabilistic links are indicated by solid lines.

[0027] The second dataset 262 is represented by a second graph 36 and includes the entities: host 32, chassis 38, disk image 40 and host monitoring metrics 42. The disk image entities 40 and the host monitoring metric entities 42 are connected to the host entities 32 via non-probabilistic links. The chassis entity 38 interconnects the host entities 32 via non-probabilistic links.

[0028] The third dataset 263 is represented by a third graph 44 and includes the entities: host 32, service 46, service description 48, and service monitoring metrics 50. The service entities 46 are connected to one another and to host entities 32, service description entities 48 and to service monitoring metric entities 50 via non-probabilistic links.

[0029] The fourth dataset 264 is represented by a fourth graph 52 and includes the entities: service 46, service description 48 and service SLA

54. The service entities 46 are connected to one another and to the service description entities 48 and to service SLA entities 54 via non-probabilistic links.

[0030] The host entities 32 of the first graph 30 are connected to the host entities 32 of the second graph 36 via probabilistic links (that is, a link having a probability greater than 0% and less than 100%, in other words, an uncertain relationship). As illustrated in Fig. 2, the probabilistic links are indicated by dashed lines. The host entities 32 of the second graph 30 are connected to the host entities 32 of the third graph 44 via probabilistic links. The service entities 46 of the third graph 44 are connected to the service entities 46 of the fourth graph 52 via probabilistic links. The service description entities 48 of the third graph 44 and connected to the service description entities 48 of the fourth graph 52 via probabilistic links.

[0031] The first, second, third and fourth graphs 30, 36, 44, and 52 share some entities. For example, the first, second and third graphs comprise host entities 32. However, the shared entities may have at least one difference in the different graphs. For example, the host entities 32 may have a different entity name in each of the first, second and third graphs 30, 36 and 44.

[0032] The non-probabilistic links may be defined by the provider (or providers) of the datasets 261, 262, 263, 264. The probabilistic links are defined in accordance with the methods illustrated in Fig. 3 and described in the following paragraphs.

[0033] Fig. 3 illustrates a flow diagram of a method to determine relationships between entities in a plurality of datasets according to an example. In the following paragraphs, the method is described with reference to the datasets illustrated in Fig. 2 as an example.

[0034] At block 56, the controller 12 initiates the method to determine at least one relationship between entities in a plurality of datasets. The method may be initiated by a user. For example, a user may use the input apparatus 14 to provide a control signal to the controller 12 to initiate the method. The method may additionally or alternatively be initiated by the controller 12. For example, the controller 12 may determine that a new dataset has been stored in the memory 20 and may then initiate the method in response.

[0035] At block 56, the controller 12 selects a first entity from a first dataset, where the first dataset comprises at least the first entity. For example, the controller 12 may select a host entity 32 in the IT Support dataset 261.

[0036] At block 58, the controller 12 selects a second entity from a second dataset, where the second dataset comprises at least the second entity. For example, the controller 12 may select a host entity 32 in the operations dataset 262.

[0037] At block 60, the controller 12 determines a probability that the first entity and the second entity are related by comparing a plurality of characteristics of the first entity with a plurality of characteristics of the second entity.

[0038] A characteristic of an entity (which may also be referred to as a 'feature' of the entity) is information that, at least in part, defines the entity and may enable differentiation between entities. For example, a characteristic of an entity may be information that (at least in part) identifies the entity. For example, such a characteristic may identify the entity by including a name of the entity, or a brief description of the entity. By way of another example, a characteristic of an entity may be information that provides an attribute of the entity. For example, such a characteristic may be a physical location or may be available memory at a

node. By way of a further example, a characteristic of an entity may be information that provides the structure of the dataset connected to that entity. For example, such a characteristic may provide a list of entities that are connected to the entity via non-probabilistic links.

[0039] In one example, an entity has the characteristics: entity name (for example node name or internet protocol (IP)), entity text description (for example, a short plain text description of the entity), entity attributes (for example, available memory in a node or physical location of the entity), text description of an entity attribute (for example, a short text description of an attribute of the entity), entity connections (that is, dataset structure connected to the entity), and links to available online resources (for example, a link to an online encyclopaedia). The entity text description, the text description of an entity attribute and the links to available online resources may be metadata which may be provided by the owner or provider of the dataset. In other examples, an entity may have different characteristics to those described in this paragraph.

[0040] The probability may be determined from matches between the plurality of characteristics of the first entity and the plurality of characteristics of the second entity and a plurality of weighting factors for the plurality of characteristics. In more detail, the probability may be calculated as the sum of weighting factors multiplied by matching result of a characteristic between entities.

[0041] The matching result of a characteristic between entities may be 1 where there is a match, and may be 0 where there is no match. The weighting factors may have values of less than 1 and may be dynamically altered by the controller 12 using feedback (as explained in greater detail in later paragraphs).

[0042] For example, the probability may be determined from the following equation:

[0043] $$p = w_{name} \times s_{name} + w_{entity\ description} \times s_{entity\ description} + w_{attribute\ description} \times s_{attribute\ description} + w_{structure} \times s_{structure}$$

[0044] Where 'w' is the current weight for the characteristic and where 's' is the result of matching that characteristic between entities.

[0045] In some examples, the result of a match is a binary function, *s,* indicating whether there was a match or not. Entity name matching may be performed by case insensitive UTF-8 string matching. However, a name match does not necessarily mean that the entities are the same (e.g. machine 1 can be a name given to a machine by different providers, but they refer to their own machine 1). Such random matches may occur in relatively large datasets, but since most names will not be the same just by chance, most names will not be the same and the weight of that feature may be gradually decreased (as described below with reference to block 66).

[0046] The description of the two entities may be analysed by a Named Entity Extraction (NER) and a Part of Speech (POS) parser. If the entity parser offers the same result, $s_{entity_description}$ is set to 1. NER/POS may also be applied to attribute descriptions. If there is a match between the attribute description and the name of any two attributes in the entity, their values are compared. If there is a match in the value (literal case insensitive character match for text and numeric value for numbers), then $s_{attribute_description}$ is set to 1. If no attribute description was provided in the metadata, only the name of the attribute is used for deciding whether or not the value of the attribute should be compared.

[0047] One approach to matching dataset structural features comprises checking whether two entities in the two separate datasets are connected to the same entities. This approach may require a second pass over the whole dataset: once some potential multiplex have been identified, the dataset structure is analysed to detect if an entity has the same connections across data sets (for example, the service entity 46 is linked to the same service description entity 48 in the datasets 263 and 264.

[0048] At block 62, the controller 12 may determine whether the probability (determined in block 60) exceeds a threshold probability value. In some examples, a user may operate the input apparatus 14 to set the threshold probability value. In other examples, the threshold probability value may be defined during manufacturing and stored in the memory 20.

[0049] At block 64, the controller 12 may create a probabilistic link between the first entity of the first dataset and the second entity of the second dataset using the probability determined in block 60. Where the method includes block 62, block 64 is performed when the determined probability is greater than the threshold probability value, and block 64 is not performed (that is, no probabilistic link is created) when the determined probability is less than the threshold probability value.

[0050] In some examples, if a match is found between entities for the characteristic 'link to available online resources' (for example, the entities both have metadata that includes the same link to an online encyclopaedia), the controller 12 determines that there is a certain relationship between the entities and then creates a non-probabilistic link between the two entities. If no match is found between the entities for the characteristic 'link to available online resources', the controller 12 continues with block 60 by determining whether there are any matches for other characteristics.

[0051] At block 66, the controller 12 may dynamically adjust a weighting factor associated with a first characteristic of the plurality of characteristics in dependence on the frequency of matching of the first characteristic. For example, when relationships are determined for two datasets, the weighting factors for the characteristics may be the same. When there is a match in one of the characteristics (for example, entity name, s_{name}), the weighting factor (for example, w_{name}) for that characteristic is increased.

[0052] At block 68, the controller 12 may receive a user input signal indicative of whether a user confirms or cancels a created probabilistic link between the first entity and the second entity. For example, the controller 12 may control the output apparatus 16 to display the multiplex graph 28 including the newly created probabilistic link. The user may operate the input apparatus 14 to either confirm the newly created probabilistic link or cancel the newly created probabilistic link.

[0053] At block 70, the controller 12 adjusts at least one of the plurality of weighting factors indicating a match between the first entity and the second entity, using the user input signal received in block 68. For example, where the user confirms a probabilistic link between the host entities 32 in the first graph 30 and the host entities 32 in the second graph 36, and the characteristics entity name and entity attributes have been matched, the controller 12 increases the weighting factors for the characteristics: entity name and entity attributes. By way of another example, where the user cancels a probabilistic link between the host entities 32 in the first graph 30 and the host entities 32 in the second graph 36, and the characteristics entity name and entity attributes have been matched, the controller 12 decreases the weighting factors for the characteristics: entity name and entity attributes. In some examples, the probability of the remaining probabilistic links may be recalculated using the adjusted weighting factors.

[0054] In some examples, where the user has confirmed a number of probabilistic relationships between the same type of entities, the controller 12 automatically proceeds to create probabilistic links between the remaining entities. For example, if there are one million multiplex links between datasets, the user may confirm a subset of those links (ten for example) and the controller 12 then automatically creates the remaining probabilistic links, given a target probable error (which may be set by the user using the input apparatus 14 for the automated creation).

[0055] The methods illustrated in Fig. 3 and described in the preceding paragraphs may provide an advantage in that they enable the creation of multiplex graphs including non-probabilistic links and probabilistic links between entities in a plurality of datasets. Furthermore, the weighting of different characteristics may be dynamically adjusted in accordance with feedback (either from successful matches or from a user) to improve the accuracy of the probabilistic links, and to improve the decision as to whether a probabilistic link should be created (where probabilistic links are created when the determined probability is greater than the threshold probability value). Additionally, the methods illustrated in Fig. 3 may be performed automatically by the controller 12 and without user intervention.

[0056] The blocks illustrated in the Fig. 3 may represent steps in a method and/or sections of code in the computer program 22. The illustration of a particular order to the blocks does not necessarily imply that there is a required or preferred order for the blocks and the order and arrangement of the block may be varied in some examples. Furthermore, it may be possible for some blocks to be omitted in some examples. For example, Fig. 4 illustrates a flow diagram of another method according to an example. The flow diagram includes blocks 56, 58, 60 and 64 illustrated in Fig. 3 and omits blocks 62, 66, 68 and 70 illustrated in Fig. 3.

[0057] Although examples have been described in the preceding paragraphs, it should be appreciated that modifications to the examples given can be made without departing from the scope as claimed.

[0058] Although the processor 18 is illustrated as a single component it may be implemented as one or more separate components some or all of which may be integrated/removable and/or may provide permanent/semi-permanent/ dynamic/cached storage.

[0059] Although the memory 20 is illustrated as a single component it may be implemented as one or more separate components some or all of which may be integrated/removable and/or may provide permanent/semi-permanent/ dynamic/cached storage.

[0060] Fig. 5 illustrates a schematic diagram of a controller 12 according to an example. The controller 12 includes a probabilistic link creation module 72, a first dataset 74, a second dataset 76, and a user input module 78. The probabilistic link creation module 72 is to perform blocks 56, 58, 60, 62, 64, 66 and 70. The user input module 78 is to perform block 68. The probabilistic link creation module 72 and the user input module 78 may be software modules that are executable by the controller 12.

[0061] References to 'computer-readable storage medium', 'computer program product', 'tangibly embodied computer program' etc. or a 'controller', 'computer', 'processor' etc. should be understood to encompass not only computers having different architectures such as single /multi- processor architectures and sequential (Von Neumann)/parallel architectures but also specialized circuits such as field-programmable gate arrays (FPGA), application specific circuits (ASIC), signal processing devices and other processing circuitry. References to computer program, instructions, code etc. should be understood to encompass software for a programmable processor or firmware such as,

for example, the programmable content of a hardware device whether instructions for a processor, or configuration settings for a fixed-function device, gate array or programmable logic device etc.

[0062] Features described in the preceding description may be used in combinations other than the combinations explicitly described.

[0063] Although functions have been described with reference to certain features, those functions may be performable by other features whether described or not.

[0064] Although features have been described with reference to certain examples, those features may also be present in other examples whether described or not.

[0065] Whilst endeavoring in the foregoing specification to draw attention to those features believed to be of particular importance it should be understood that the Applicant claims protection in respect of any patentable feature or combination of features hereinbefore referred to and/or shown in the drawings whether or not particular emphasis has been placed thereon.

[0066] What is claimed is:

CLAIMS

1. A method to determine relationships between entities in a plurality of datasets, the method comprising:

selecting a first entity from a first dataset, the first dataset comprising at least the first entity;

selecting a second entity from a second dataset, the second dataset comprising at least the second entity;

determining a probability that the first entity and the second entity are related by comparing a plurality of characteristics of the first entity with a plurality of characteristics of the second entity; and

creating a probabilistic link between the first entity of the first dataset and the second entity of the second dataset using the determined probability.

2. The method as claimed in claim 1, further comprising: determining whether the probability exceeds a threshold probability value; and creating a probabilistic link between the first entity of the first dataset and the second entity of the second dataset where it is determined that the probability exceeds the threshold probability value.

3. The method as claimed in claim 2, wherein the threshold probability value is set by a user.

4. The method as claimed in claim 1, wherein the probability is determined from matches between the plurality of characteristics of the first entity and the plurality of characteristics of the second entity and a plurality of weighting factors for the plurality of characteristics.

5. The method as claimed in claim 4, further comprising dynamically adjusting a weighting factor associated with a first characteristic of the plurality of characteristics in dependence on the frequency of matching the first characteristic.

6. The method as claimed in claim 4, further comprising receiving a user input signal indicative of whether a user confirms or cancels a created probabilistic link between the first entity and the second entity; and adjusting at least one of the plurality of weighting factors indicating a match between the first entity and the second entity, using the received user input signal.

7. An apparatus to determine relationships between entities in a plurality of datasets, the apparatus comprising:

a controller to:

select a first entity from a first dataset, the first dataset comprising at least the first entity;

select a second entity from a second dataset, the second dataset comprising at least the second entity;

determine a probability that the first entity and the second entity are related by comparing a plurality of characteristics of the first entity with a plurality of characteristics of the second entity,

wherein the probability is determined from matches between the plurality of characteristics of the first entity and the plurality of characteristics of the second entity and a plurality of weighting factors for the plurality of characteristics.

8. The apparatus as claimed in claim 7, wherein the controller is to create a probabilistic link between the first entity of the first dataset and the second entity of the second dataset.

9. The apparatus as claimed in claim 7, wherein the controller is to determine whether the probability exceeds a threshold probability value; and to create a probabilistic link between the first entity of the first dataset and the second entity of the second dataset where it is determined that the probability exceeds the threshold probability value.

10. The apparatus as claimed in claim 9, wherein the threshold probability value is set by a user.

11. The apparatus as claimed in claim 7, wherein the controller is to dynamically adjust a weighting associated with a first characteristic of the plurality of characteristics in dependence on the frequency of matching the first characteristic.

12. The apparatus as claimed in claim 7, wherein the controller is to receive a user input signal indicative of whether a user confirms or cancels a created link between the first entity and the second entity; and adjusting at least one of the plurality of weighting factors indicating a match between the first entity and the second entity, using the received user input signal.

13. A non-transitory computer-readable storage medium encoded with instructions to determine relationships between entities in a plurality of datasets that, when performed by a processor, cause performance of:

select a first entity from a first dataset, the first dataset comprising at least the first entity;

select a second entity from a second dataset, the second dataset comprising at least the second entity;

determine a probability that the first entity and the second entity are related by comparing a plurality of characteristics of the first entity with a plurality of characteristics of the second entity; and

create a probabilistic link between the first entity of the first dataset and the second entity of the second dataset using the determined probability.

14. The non-transitory computer-readable storage medium as claimed in claim 13 encoded with instructions that, when performed by a processor, cause performance of: determine whether the probability exceeds a threshold probability value; and create a probabilistic link between the first entity of the

first dataset and the second entity of the second dataset where it is determined that the probability exceeds the threshold probability value.

15. The non-transitory computer-readable storage medium as claimed in claim 13, wherein the probability is determined from matches between the plurality of characteristics of the first entity and the plurality of characteristics of the second entity and a plurality of weighting factors for the plurality of characteristics.

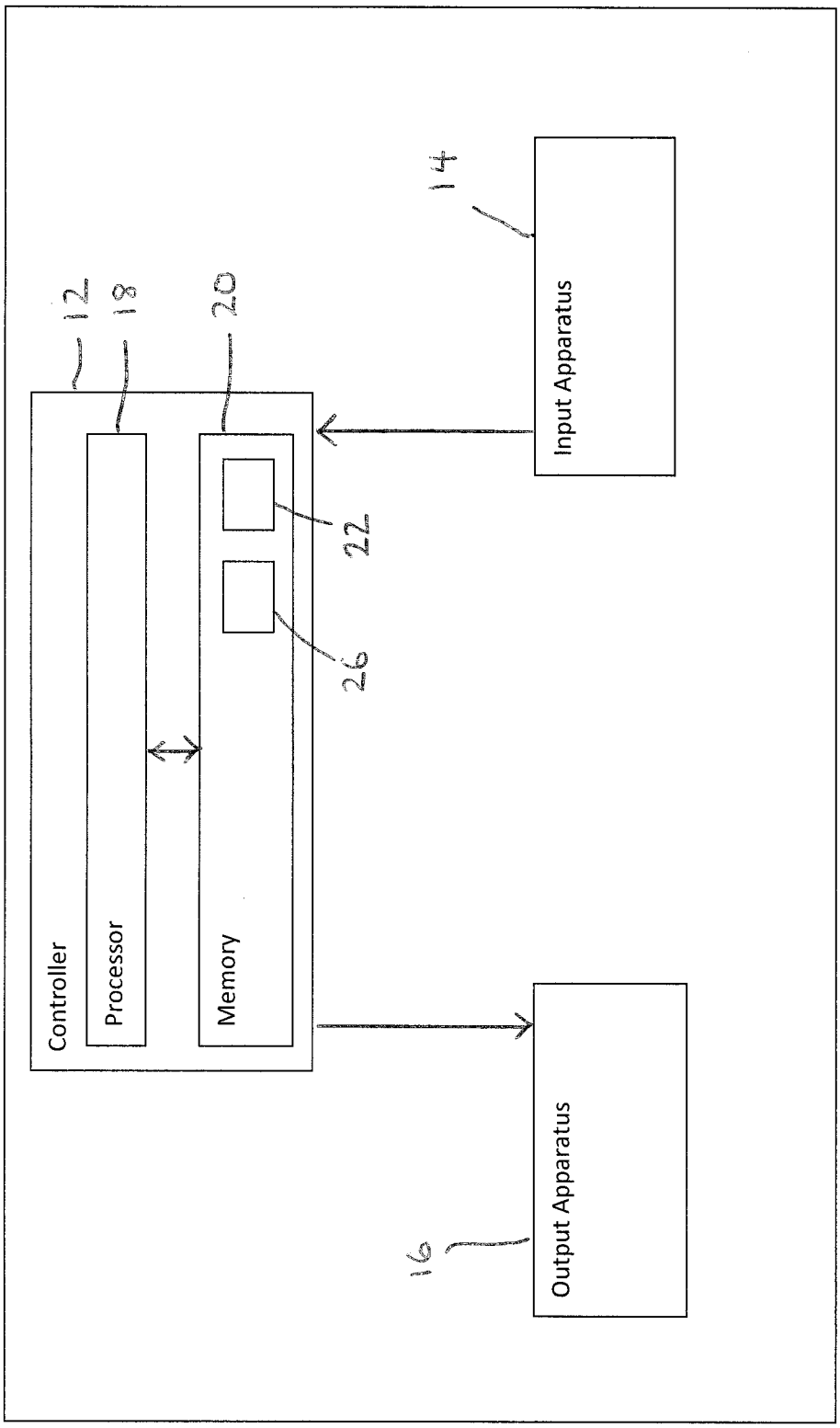

Controller
Processor
Memory
Input Apparatus
Output Apparatus
12
18
20
14
16
22
26
24
10

Fig. 1

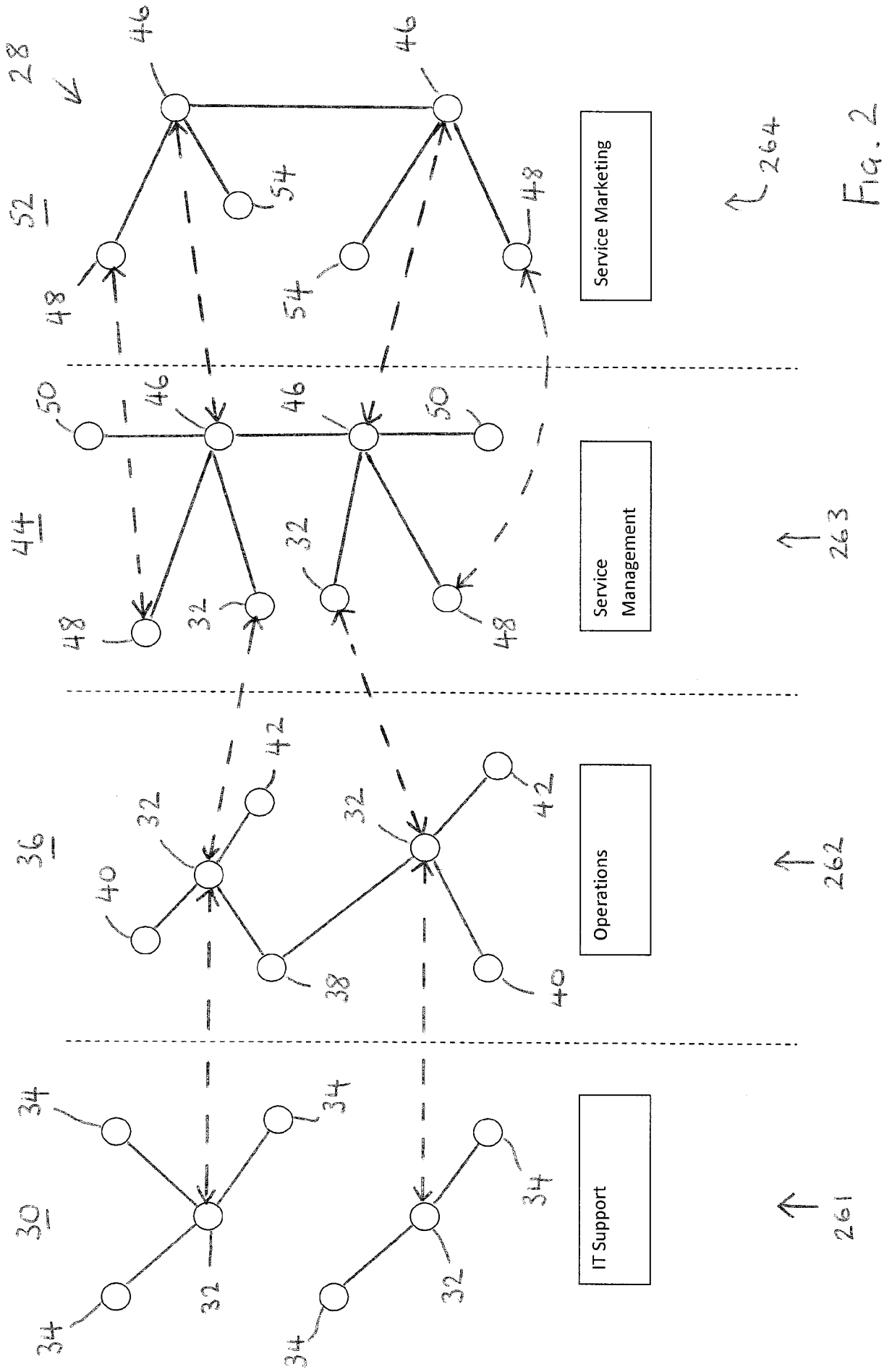

Service Marketing
Service Management
Operations
IT Support
28
52
44
36
30
46
48
50
54
32
34
38
40
42
264
263
262
261

Fig. 2

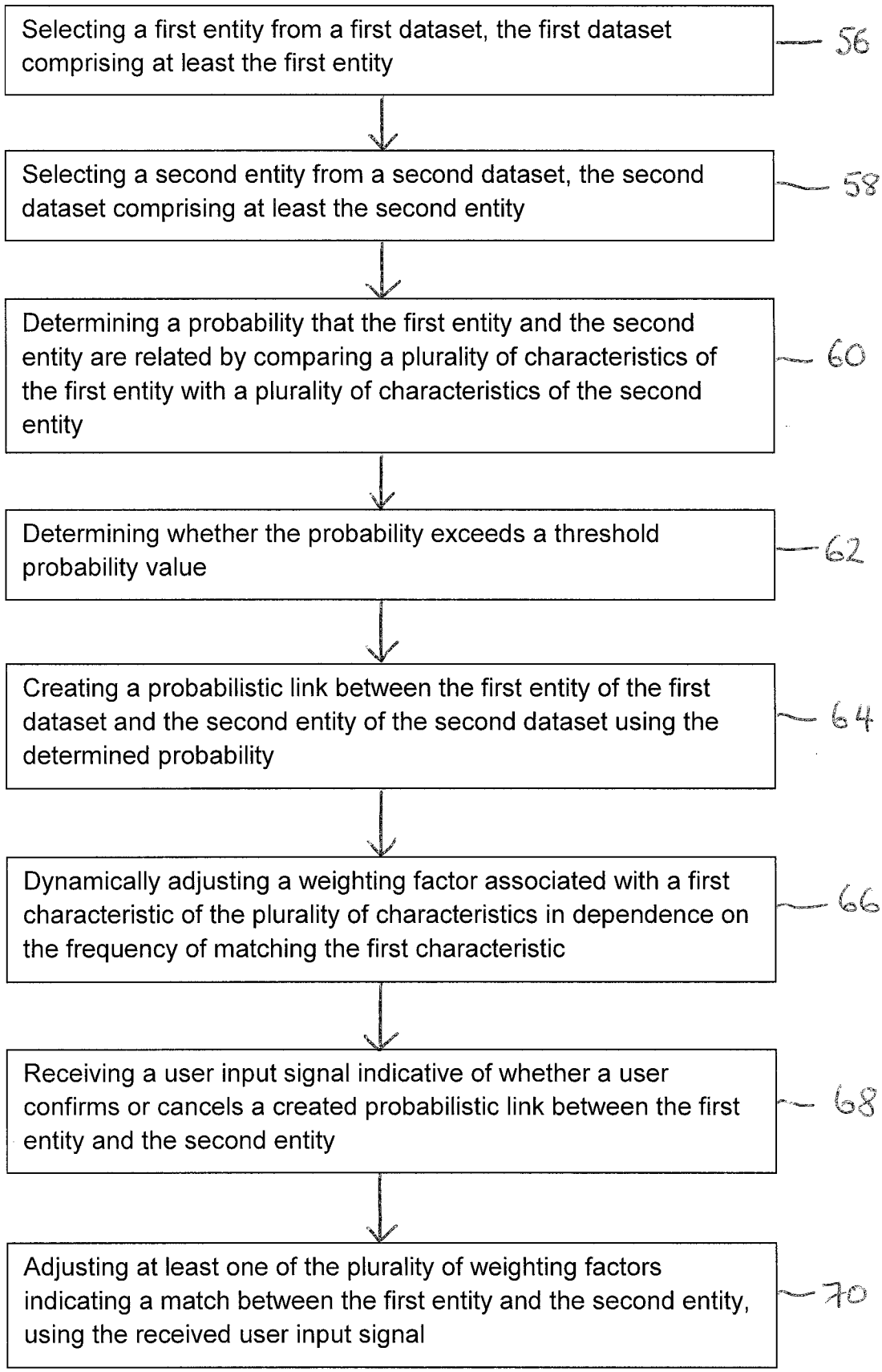

Selecting a first entity from a first dataset, the first dataset comprising at least the first entity
56
Selecting a second entity from a second dataset, the second dataset comprising at least the second entity
58
Determining a probability that the first entity and the second entity are related by comparing a plurality of characteristics of the first entity with a plurality of characteristics of the second entity
60
Determining whether the probability exceeds a threshold probability value
62
Creating a probabilistic link between the first entity of the first dataset and the second entity of the second dataset using the determined probability
64
Dynamically adjusting a weighting factor associated with a first characteristic of the plurality of characteristics in dependence on the frequency of matching the first characteristic
66
Receiving a user input signal indicative of whether a user confirms or cancels a created probabilistic link between the first entity and the second entity
68
Adjusting at least one of the plurality of weighting factors indicating a match between the first entity and the second entity, using the received user input signal
70

Fig. 3

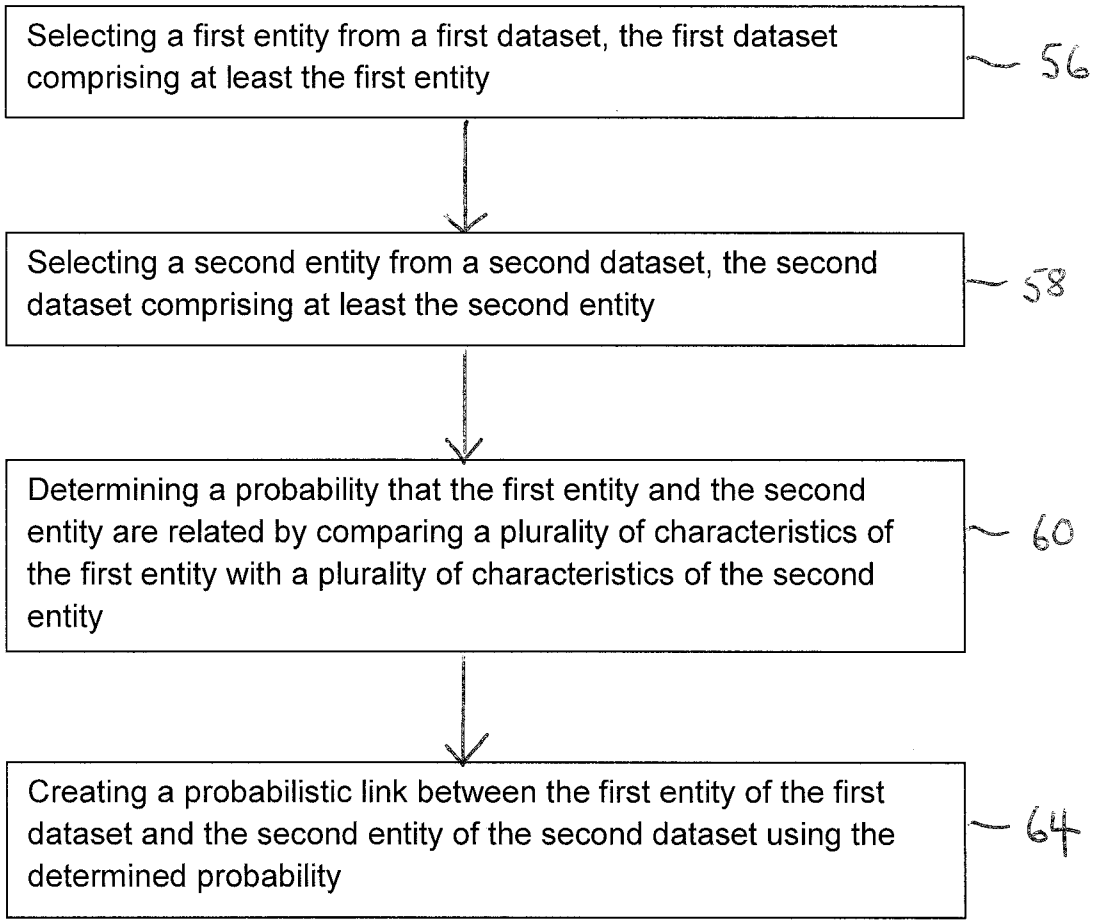

Selecting a first entity from a first dataset, the first dataset comprising at least the first entity
56
Selecting a second entity from a second dataset, the second dataset comprising at least the second entity
58
Determining a probability that the first entity and the second entity are related by comparing a plurality of characteristics of the first entity with a plurality of characteristics of the second entity
60
Creating a probabilistic link between the first entity of the first dataset and the second entity of the second dataset using the determined probability
64

Fig. 4

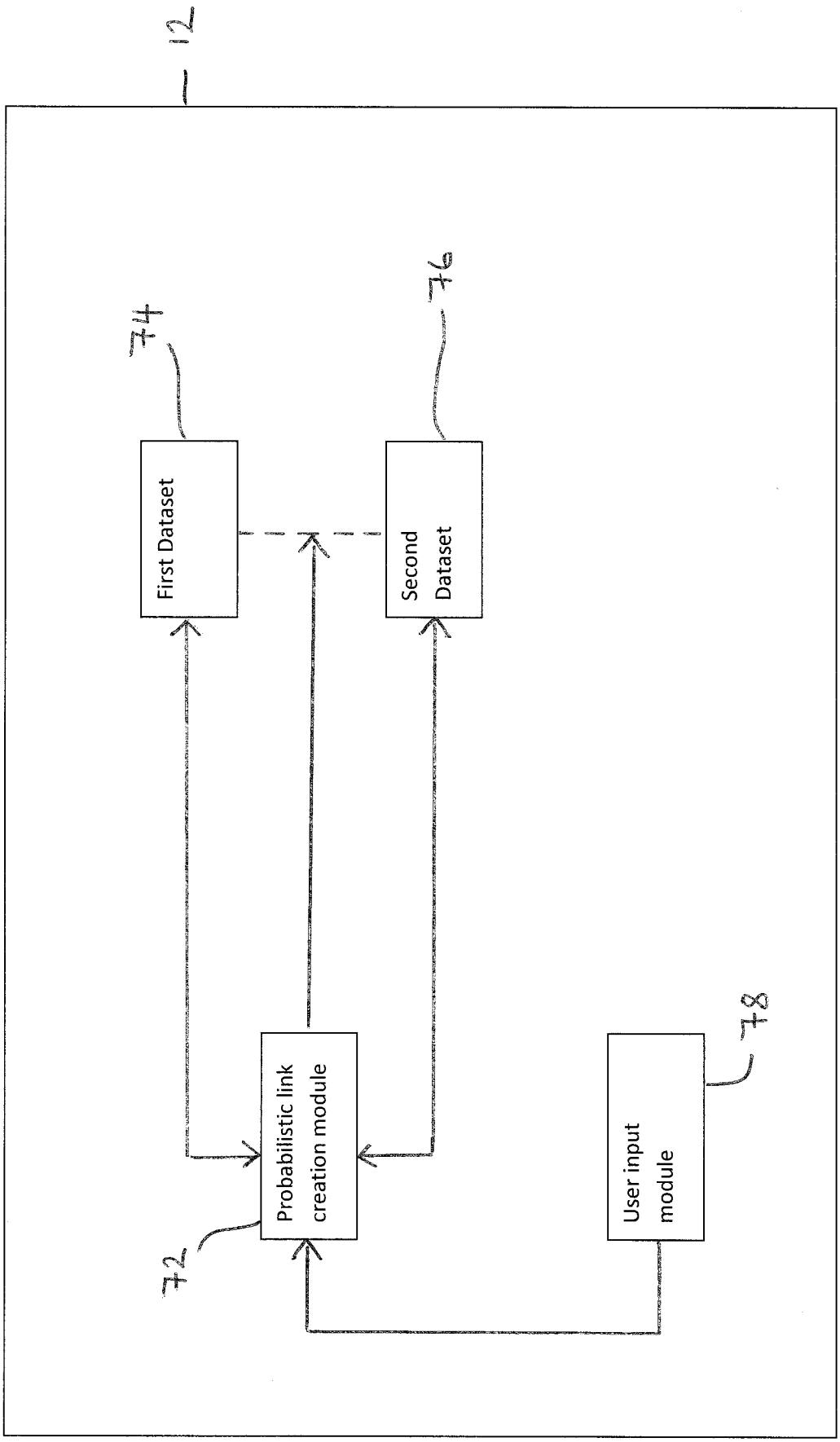

12
74
76
First Dataset
Second Dataset
72
Probabilistic link creation module
78
User input module

Fig. 5

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2014/058525

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, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 7 912 842 B1 (BAYLISS DAVID [US]) 22 March 2011 (2011-03-22) abstract column 1, line 17 - line 21 column 2, line 65 - column 3, line 61 column 4, line 28 - line 34 -----	1-15
X	US 7 792 864 B1 (RICE SCOTT GREGORY ROBERT [US]) 7 September 2010 (2010-09-07) abstract column 1, line 25 - line 26 column 2, line 4 - line 17 column 2, line 36 - line 56 ----- -/--	1-15

[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
9 January 2015	19/01/2015

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	de Castro Palomares

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2014/058525

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 98/55947 A1 (MADISON INFORMATION TECHNOLOGI [US]) 10 December 1998 (1998-12-10) page 5, line 6 - line 15 page 6, line 2 - line 4 -----	1-15
X	US 2009/271359 A1 (BAYLISS DAVID ALAN [US]) 29 October 2009 (2009-10-29) abstract figure 1 paragraph [0040] - paragraph [0042] paragraph [0066] -----	1-15

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2014/058525

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 7912842 B1	22-03-2011	NONE	
US 7792864 B1	07-09-2010	US 7792864 B1	07-09-2010
		US 2011119291 A1	19-05-2011
WO 9855947 A1	10-12-1998	AU 764321 B2	14-08-2003
		AU 8057298 A	21-12-1998
		CA 2292494 A1	10-12-1998
		EP 0993641 A1	19-04-2000
		JP 2002504248 A	05-02-2002
		US 5991758 A	23-11-1999
		WO 9855947 A1	10-12-1998
US 2009271359 A1	29-10-2009	CA 2723167 A1	29-10-2009
		CA 2749080 A1	29-10-2009
		CA 2749310 A1	29-10-2009
		GB 2472335 A	02-02-2011
		US 2009271359 A1	29-10-2009
		US 2009271363 A1	29-10-2009
		US 2009271397 A1	29-10-2009
		US 2009271404 A1	29-10-2009
		US 2009271405 A1	29-10-2009
		US 2009271424 A1	29-10-2009
		US 2009271694 A1	29-10-2009
		US 2009287689 A1	19-11-2009
		US 2009292694 A1	26-11-2009
		US 2009292695 A1	26-11-2009
		US 2012173545 A1	05-07-2012
		US 2012173546 A1	05-07-2012
		US 2012173548 A1	05-07-2012
		US 2012278340 A1	01-11-2012
		US 2012290585 A1	15-11-2012
		US 2013297635 A1	07-11-2013
		US 2014032556 A1	30-01-2014
		US 2014032557 A1	30-01-2014
		US 2014032594 A1	30-01-2014
		US 2014250111 A1	04-09-2014
		WO 2009132263 A2	29-10-2009