Office de la Propriété Intellectuelle du Canada
Un organisme d'Industrie Canada

Canadian Intellectual Property Office
An agency of Industry Canada

CA 2366545 C 2005/12/20

(11)(21) 2 366 545

(12) BREVET CANADIEN
CANADIAN PATENT

(13) C

(86) Date de dépôt PCT/PCT Filing Date: 2000/04/06
(87) Date publication PCT/PCT Publication Date: 2000/10/19
(45) Date de délivrance/Issue Date: 2005/12/20
(85) Entrée phase nationale/National Entry: 2001/10/04
(86) N° demande PCT/PCT Application No.: US 2000/009471
(87) N° publication PCT/PCT Publication No.: 2000/062203
(30) Priorité/Priority: 1999/04/09 (09/289,174) US

(51) Cl.Int.^{7}/Int.Cl.^{7} G06F 17/30

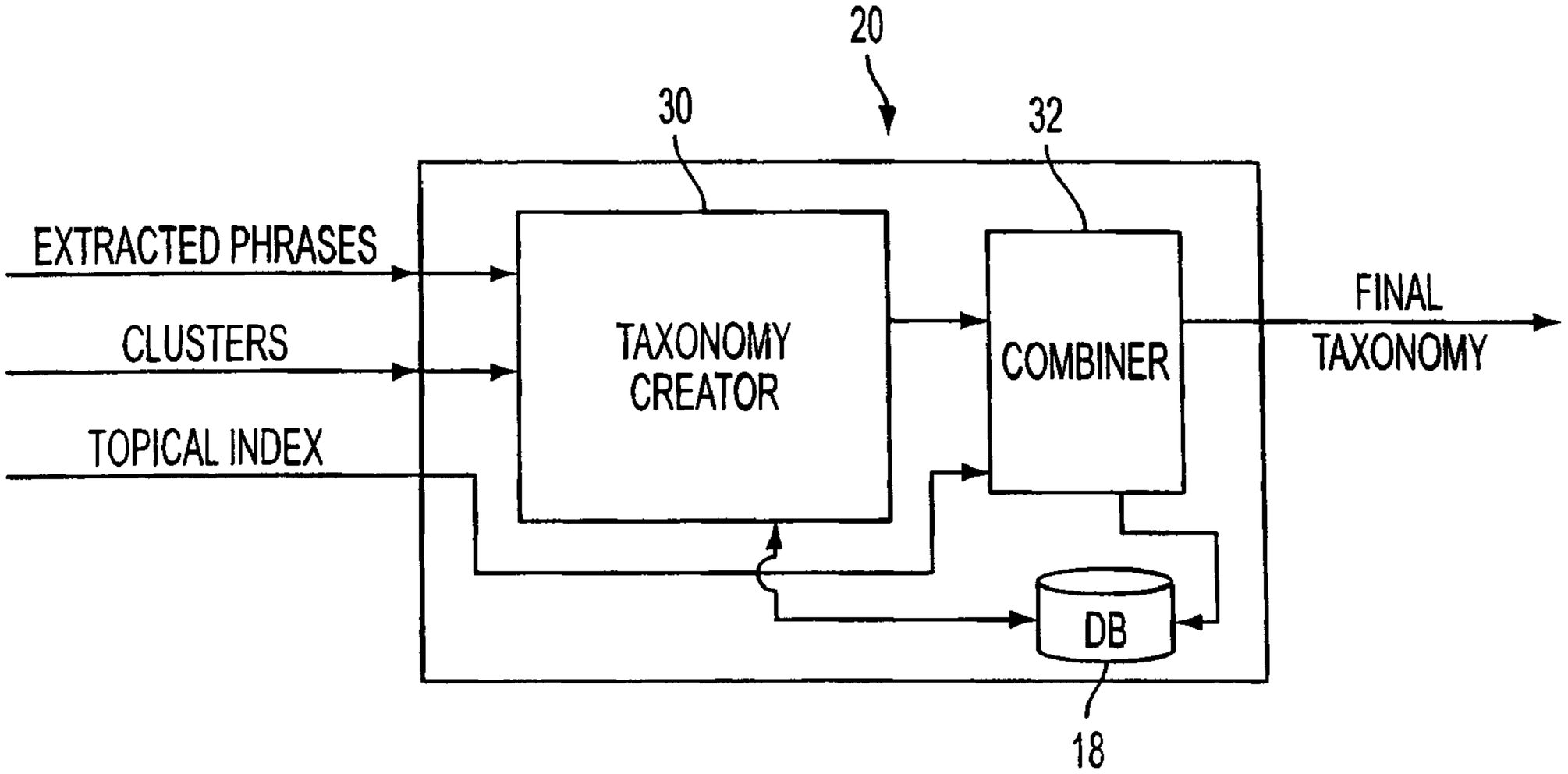
(72) Inventeur/Inventor:
VOGEL, CLAUDE, US

(73) Propriétaire/Owner:
ENTRIEVA, INC., US

(74) Agent: SMART & BIGGAR

(54) Titre : SYSTEME ET PROCEDE PERMETTANT DE GENERER UNE TAXONOMIE A PARTIR D'UNE PLURALITE DE DOCUMENTS

(54) Title: SYSTEM AND METHOD FOR GENERATING A TAXONOMY FROM A PLURALITY OF DOCUMENTS



(57) Abrégé/Abstract:
A system and method for generating a taxonomy (30) is provided in which the taxonomy is generated based on clusters of phrases and a topical library (52). The taxonomy permits a user of a text processing system to rapidly search through a database (18) and find relevant documents since the classifications in the taxonomy are narrow enough to limit the number of documents classified in each classification.

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : G06F 17/30

A1

(11) International Publication Number: WO 00/62203

(43) International Publication Date: 19 October 2000 (19.10.00)

(21) International Application Number: PCT/US00/09471

(22) International Filing Date: 6 April 2000 (06.04.00)

(30) Priority Data:
09/289,174 9 April 1999 (09.04.99) US

(71) Applicant: SEMIO CORPORATION [US/US]; Suite 101, 1730 S. Amphlett Boulevard, San Mateo, CA 94402 (US).

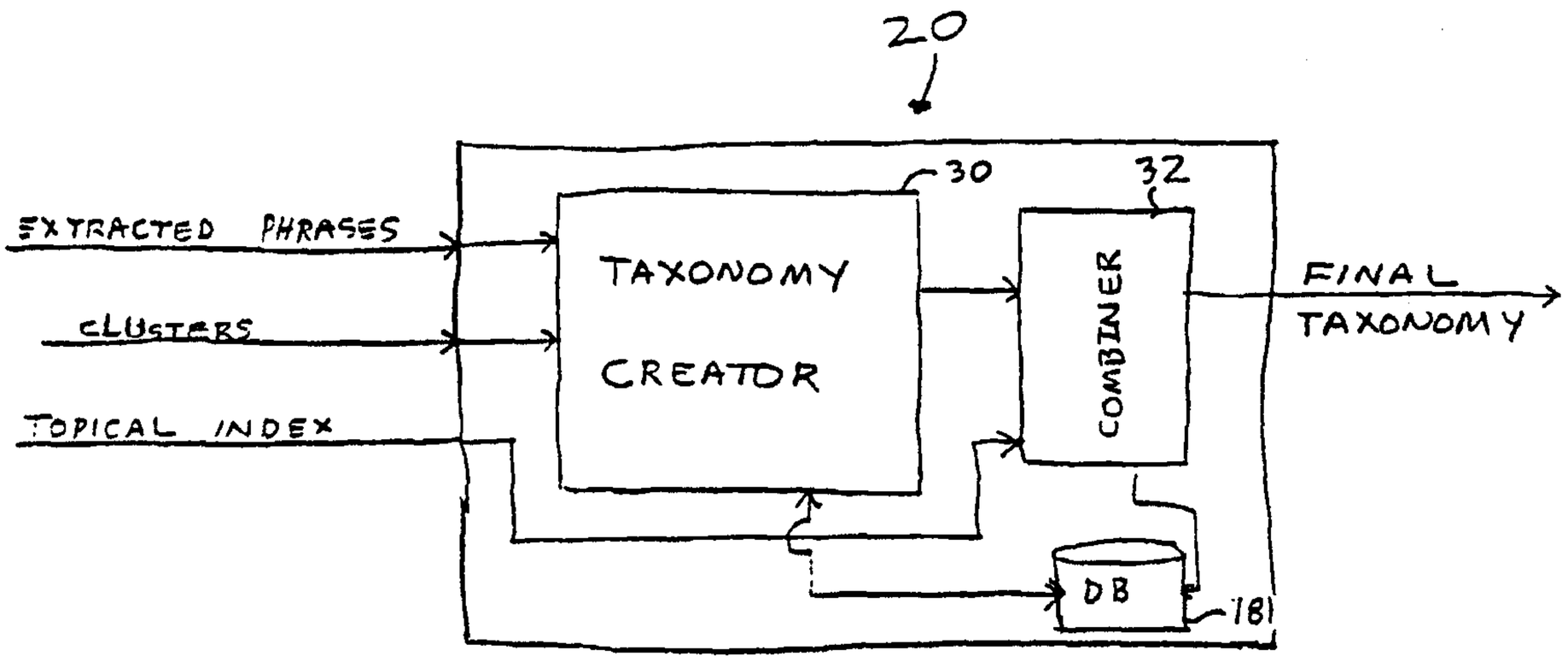
(72) Inventor: VOGEL, Claude; 733 Bounty Drive #203, Foster City, CA 94404 (US).

(74) Agent: LOHSE, Timothy, W.; Gray Cary Ware & Freidenrich LLP, 3340 Hillview Avenue, Palo Alto, CA 94304 (US).

(81) Designated States: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.

(54) Title: SYSTEM AND METHOD FOR GENERATING A TAXONOMY FROM A PLURALITY OF DOCUMENTS



(57) Abstract

A system and method for generating a taxonomy (30) is provided in which the taxonomy is generated based on clusters of phrases and a topical library (52). The taxonomy permits a user of a text processing system to rapidly search through a database (18) and find relevant documents since the classifications in the taxonomy are narrow enough to limit the number of documents classified in each classification.

SYSTEM AND METHOD FOR GENERATING A TAXONOMY FROM A PLURALITY OF DOCUMENTS

Background of the Invention

This invention relates generally to a system and method for processing a document and in particular to a system and method for generating a list of classifications (a taxonomy) from a plurality of phrases which are extracted from a set of documents.

Various factors have contributed to the storage of vast amounts of textual data information in computer systems and computer databases. The dramatic increase in the storage capacity of computer storage devices, such as hard disks, tape drives and the like and a decrease in the cost of these higher capacity computer hard drives are factors. Other factors include an increase in the transmission speed of computer communications, an increase in the processing speed of personal computers and an expansion of various computer communications networks, such as a bulletin board or the Internet. People therefore currently have access to the large amounts of textual data stored in these databases. However, although the current technology facilitates the storage of and the access to this textual data, there are new problems that have been created by the vast amount of textual data that is now available.

In particular, a person trying to access the textual data in these computer databases needs a system for analyzing and processing the data in order to retrieve the desired information quickly and efficiently without retrieving extraneous information.

In addition, the person trying to access the information needs an efficient system for condensing each large document into a plurality of phrases (one or more words) which characterize the document so that the person can browse the phrases and understand the document without actually viewing the entire document. It is also desirable to be able to automatically generate a classification system based on the extracted phrases stored in the computer database so that a person may use the classification system to browse through the database and focus in on the most relevant documents or pieces of textual data in the computer database. The classification system may be known as a taxonomy which consists of one or more subject matter headings with sub-headings reflecting the phrases extracted from the documents in the database.

To generate a typical subject matter classification system, a person must manually generate a classification system into which the one or more pieces of textual data in the database may then be manually or automatically classified. Thus, the typical classification system is manually generated and does not use the extracted phrases from the documents in the database to create the classification system. Therefore, the typical classifications are generally very broad reflecting the inability to more accurately classify the documents since the exact contents of the documents are not known. Thus, the typical classification may permit the user to select only a broad category which probably then contains too many documents to easily review. In addition, for each different database, a new classification system must be manually generated which is slow and time consuming.

Thus, it is desirable to provide a taxonomy generation system and method which solves the above problems and limitations with conventional classification systems and it is to this end that the present invention is directed.

Summary of the Invention

A system and method for generating a taxonomy in accordance with the invention is provided in which the phrases extracted from the plurality of documents or pieces of textual data stored in the database may be used to automatically generate a taxonomy for the database. The generated taxonomy is unique to the particular database (since it is based on the phrases extracted from the database) and permits a searcher in the database to rapidly find documents within the database since the taxonomy has more detail than a typical classification system. Thus, a search through the taxonomy may reduce the number of documents which the user must review. In addition, the taxonomy generation system may automatically generate a taxonomy for any database provided that key phrases from documents within the database are available.

In more detail, a broad topical library is manually generated and used to provide a structure for the taxonomy. The topical library may be of the same broad level of detail as a typical classification system. Next, the phrases extracted from the documents within the database may be clustered and mapped to form maps illustrating the relationships between the phrases within the documents. The phrases in the maps which are connected to the most other phrases in the map, known as leaders, may be

used to form the first level of the taxonomy underneath the topical index. Next, phrases which are related to the leaders based on the clustering form the next level of the taxonomy and so on until the taxonomy has been automatically generated. In this manner, the taxonomy generated reflects the actual phrases contained in the database so that it is more accurate than typical classifications.

Thus, in accordance with the invention, a system for generating a taxonomy for a database is provided including a database having a plurality of pieces of text and a plurality of phrases extracted from the pieces of text wherein the phrases are associated with one or more other phrases. A leader phrase from the phrases in the database is identified wherein the leader phrase is associated with a predetermined number of other phrases in the database. A first level of a taxonomy is generated based on the identified leader phrases wherein the leader phrases form a first level of headings in a hierarchical topical outline. A second level of the taxonomy is generated based on phrases in the database associated with the leader phrases wherein the phrases are sub-headings underneath the leader phrases with which they are associated. The generated taxonomy reflects the phrases extracted from the pieces of text in the database so that a user searches through the database using the final taxonomy.

Brief Description of the Drawings

Figure 1 is a block diagram of a text processing system;

Figure 2 is a block diagram of a taxonomy generation system in accordance with the invention;

Figure 3 is a flowchart illustrating a method for generating a taxonomy in accordance with the invention;

Figure 4 is a diagram illustrating a map containing clustered phrases;

Figure 5 illustrates an example of a topical library;

Figure 6 illustrates an example of a taxonomy generated in accordance with the invention;

Figure 7 illustrates an example of a final taxonomy incorporating the taxonomy of Figure 6 and the topical library of Figure 4 in accordance with the invention;

Figure 8 illustrates an example of a user interface for searching a top level of the taxonomy in accordance with the invention; and

Figure 9 illustrates an example of a user interface for searching a lower level of the taxonomy in accordance with the invention and for displaying documents relating to the selected topics.

Detailed Description of a Preferred Embodiment

The invention is particularly applicable to a system for generating a taxonomy from English language documents and it is in this context that the invention will be

described. It will be appreciated, however, that the system and method in accordance with the invention has greater utility, such as to document in other languages and to any other types of pieces of text that may be stored in a database. To better understand the invention, a text processing system will now be described.

Figure 1 is a block diagram of a text processing system 10. The text processing system 10 may include a parser 12, a clusterizer 14, a map generator 16, a database (DB) 18, a taxonomy generator 20 and a central processing unit (CPU) 22. The parser, the clusterizer, the map generator, and the taxonomy generator may all be separate software applications stored in the memory (not shown) of the text processing system 10 which are executed by the CPU 22. The text processing system may receive one or more pieces of text , such as stories, press releases or documents, and may generate a searchable taxonomy which may be searched by one or more client computers (Client #1, Client #2 and Client #N). To generate the searchable taxonomy, each piece of text may be received by the parser 12 which processes each piece of incoming text and generates one or more key phrases for each piece of text which characterizes the piece of text. The key phrases may be stored in the database 18. Once the key phrases are extracted from each piece of text, the clusterizer 14 may generate one or more clusters of the key phrases based on the relationships between the phrases. The clusters generated may also be stored in the database 18. The map generator 16 may use the generated clusters in the database in order to generate a graphical map showing the relationships of the key phrases within the various pieces of text to each other so that a user of the system may easily search through the database of pieces of text by viewing

the key phrases of the pieces of text. The clusters may also be used to generate the taxonomy in accordance with the invention as described below with reference to Figures 2 - 9. The database 18 may also store a topical library, as described below with reference to Figure 5, which also may be used to generate the taxonomy.

Additional details about the clusterizer and the map generator are disclosed in co-pending U.S. patent application serial No. 08/801,970 which is owned by the assignee of the present invention and which is incorporated herein by reference. The text processing system may be implemented in a variety of manners including a client/server type computer system in which the client computers access the server via a public computer network, such as the Internet.

In operation, a client computer may access the text processing system 10 via a computer network, such as the Internet, using a browser application. The text processing system may generate a user interface showing the top level of the taxonomy previously generated. The user may then select a particular subject matter from within the taxonomy and another level of the taxonomy will be provided to the user. As some point, the user interface will also provide the user with the ability to look at documents which satisfy a given taxonomy classification. The interaction with the user will be described in more detail with reference to Figures 8 and 9. Now, the taxonomy generator 20 in accordance with the invention will be described.

Figure 2 is a block diagram of the taxonomy generator 20 in accordance with the invention. The taxonomy generator 20 may include a taxonomy creator 30, a

combiner 32 and the database 18 from the text processing system. The taxonomy generator 20 may receive the extracted phrases from the parser, the clusters from the clusterizer (both of which are stored in the database 18) and a topical library/index which may also be stored in the database 18. The topical library may be a high level classification of subject matter within a particular area such as might be available in a typical text processing system for searching through the document in the database. The extracted phrases and the clusters may be input to the taxonomy creator 30 which generates a taxonomy from the extracted phrases and the clusters. An example of the taxonomy generated will be described below with reference to Figure 6. The taxonomy from the taxonomy creator is output to the combiner 32 which then combines the taxonomy with the topical library to generate a final taxonomy. The final taxonomy may be searched by a user accessing the text processing system. Since the final taxonomy has been created from the extracted phrases and the clusters, the final taxonomy reflects the documents in the database more accurately than a typical classification system which has no association with the documents in the database. Now, a method for generating a taxonomy in accordance with the invention will be described.

Figure 3 is a flowchart illustrating a method 40 for generating a taxonomy in accordance with the invention. The method begins at step 42 in which the clusters received by the taxonomy system are analyzed to identify the leader clusters. The leader clusters are phrases in a map, as described below, which have associations (as shown by the lines in the map) with a large number of other phrases. The leader

clusters tend to be broader phrases than the less connected other clusters so that the leader clusters form the best broad subject matter classifiers for the documents in the database. The leader clusters may then be used to form the first level of the taxonomy in step 44. In particular, the leader phrases are arranged in a hierarchical outline format as shown in Figure 6. Next, the phrases associated with the leader phrases (those phrases with links to the leader phrases) are identified in step 46 and the next level of the taxonomy is generated from these clusters in step 48. In particular, the phrases related to each leader phrase are arranged underneath the leader phrase in a hierarchical relationship to form more specific sub-headings. Next, the taxonomy generator determines if there are any more phrases which are going to be added into the taxonomy based on the complexity and depth of the map. If there are more clusters, the method returns to step 48 and generates the next level of the taxonomy. If there are no more clusters to process, the taxonomy generated is combined with a topical library in step 52 to generate a final taxonomy in accordance with the invention. Now, an example of generating a final taxonomy from a map showing one or more clusters of phrases and a topical library will be described.

Figure 4 is a diagram illustrating a map 60 containing a plurality of clustered phrases. The map 60 may be generated as described in the above incorporated co-pending patent application. As shown, the map contains one or more phrases 62 whose relationships to other phrases are shown by a line 64 connecting the two phrases. As described above, the leader phrase clusters may be used to generate a first level of the taxonomy. In this example, the phrase "fracture intervention trial" may be a leader

phrase since it has a relationship with a plurality of other phrases. In contrast, the phrase "mission bay" is not a leader phrase. In this example, a taxonomy is generated using only the phrases shown in Figure 4 will be described. It will be appreciated, however, that the taxonomy normally includes a plurality of phrases. To generate the taxonomy, a topical library 70 as shown in Figure 5 may be used to define the broad subject matter classifications of the topics. In this example, the sub-classifications underneath the fracture heading will be generated in accordance with the invention based on the phrases shown in Figure 4.

Figure 6 illustrates an example of a taxonomy 80 generated in accordance with the invention based on the leader phrases shown in Figure 4 and in particular the "fracture intervention trial" phrase. In particular, the phrase is located within the taxonomy and includes three sub-categories 82 in this example. In particular, the sub-categories may include the phrase "called alendranate", the phrase "hip fracture" and the phrase "participant received alendronate". As shown in Figure 4, these phrases are related to the leader phrase and therefore appear as sub-headings underneath the leader phrase. Similarly, the leader phrase "hip fracture" is identified from Figure 4 as a leader phrase and there are one or more phrases underneath it in the taxonomy. The symbols prior to the phrase in the diagram indicate that the phrase is contained elsewhere in the taxonomy. Referring back to Figure 4, each of the phrases underneath a leader phrase has a relationship with the leader phrase so that those phrases become sub-categories underneath the leader phrase.

Figure 7 illustrating a portion of a final taxonomy 90 generated when the generated taxonomy shown in Figure 6 is combined with the topical library shown in Figure 5. As shown, the final taxonomy generated in accordance with the invention contains more detail than the typical topical library. In addition, the headings (leader phrases) and the sub-headings of the taxonomy are generated from the pieces of text in the database so that the headings and sub-headings more accurately reflect the pieces of text in the database. Thus, it is easier to locate documents within the database using the automatically generated taxonomy in accordance with the invention. Now, several examples of the user interfaces for searching through the taxonomy in accordance with the invention will be described.

Figure 8 illustrates an example of a user interface screen 100 for searching a top level of the taxonomy in accordance with the invention. The user interface may include a keyword search portion 102 and a classification search section 104. The keyword search portion permits the user to enter keywords which will be compared against the taxonomy to generate a list of documents which match the keywords. The classification search portion 104 permits the user to browse through the high level classifications and click on any classification of interest. In this example, assume that the user clicks on the "Accidents & Injuries" sub-category underneath the "Disease Areas" heading. The user is then shown another user interface page which will now be described.

Figure 9 illustrates an example of a user interface screen 110 for searching a lower level of the taxonomy in accordance with the invention. As shown, this screen may include a keyword search section 112 for keyword searching as above, a current search status line 114 which may list, for example, the current category within the taxonomy being searched and the path to get to that category, and a sub-classification section 116 for listing the sub-classifications and selecting them. In this example, the taxonomy does not have more specific classifications. Therefore, the selection of one or more sub-classification headings (a phrase associated with a leader phrase in this example) and the selection of a document button 118 by the user causes those documents within the selected one or more sub-classifications to be shown to the user of the system. In this manner, the user may rapidly narrow down the number of documents to be reviewed by using the more comprehensive taxonomy in accordance with the invention.

While the foregoing has been with reference to a particular embodiment of the invention, it will be appreciated by those skilled in the art that changes in this embodiment may be made without departing from the principles and spirit of the invention, the scope of which is defined by the appended claims.

Claims:

1. A system for generating a taxonomy for a database, the system comprising:

a database including a plurality of pieces of text and a plurality of phrases extracted from the pieces of text, the phrases being associated with one or more other phrases;

means for identifying a leader phrase from the phrases in the database, the leader phrase being associated with a predetermined number of other phrases in the database;

means for generating a first level of a taxonomy based on the identified leader phrases, the leader phrases forming a first level of headings in a hierarchical topical outline; and

means for generating a second level of the taxonomy based on phrases in the database associated with the leader phrases, the phrases being sub-headings underneath the leader phrases with which they are associated, the taxonomy reflecting the phrases extracted from the pieces of text in the database so that a user searches through the database using the final taxonomy.

2. The system of Claim 1 further comprising means for combining the taxonomy with a topical library to form a final taxonomy so that the final taxonomy reflects the phrases extracted from the pieces of text in the database so that a user searches through the database using the final taxonomy.

3. A method for generating a taxonomy for a database, the method comprising:

processing a plurality of pieces of text and a plurality of phrases extracted from the pieces of text stored in a database , the phrases being associated with one or more other phrases;

identifying a leader phrase from the phrases in the database, the leader phrase being associated with a predetermined number of other phrases in the database;

generating a first level of a taxonomy based on the identified leader phrases, the leader phrases forming a first level of headings in a hierarchical topical outline; and

generating a second level of the taxonomy based on phrases in the database associated with the leader phrases, the phrases being sub-headings underneath the leader phrases with which they are associated, the taxonomy reflecting the phrases extracted from the pieces of text in the database so that a user searches through the database using the final taxonomy.

4. The method of Claim 3 further comprising combining the taxonomy with a topical library to form a final taxonomy so that the final taxonomy reflects the phrases extracted from the pieces of text in the database so that a user searches through the database using the final taxonomy.

76886-81

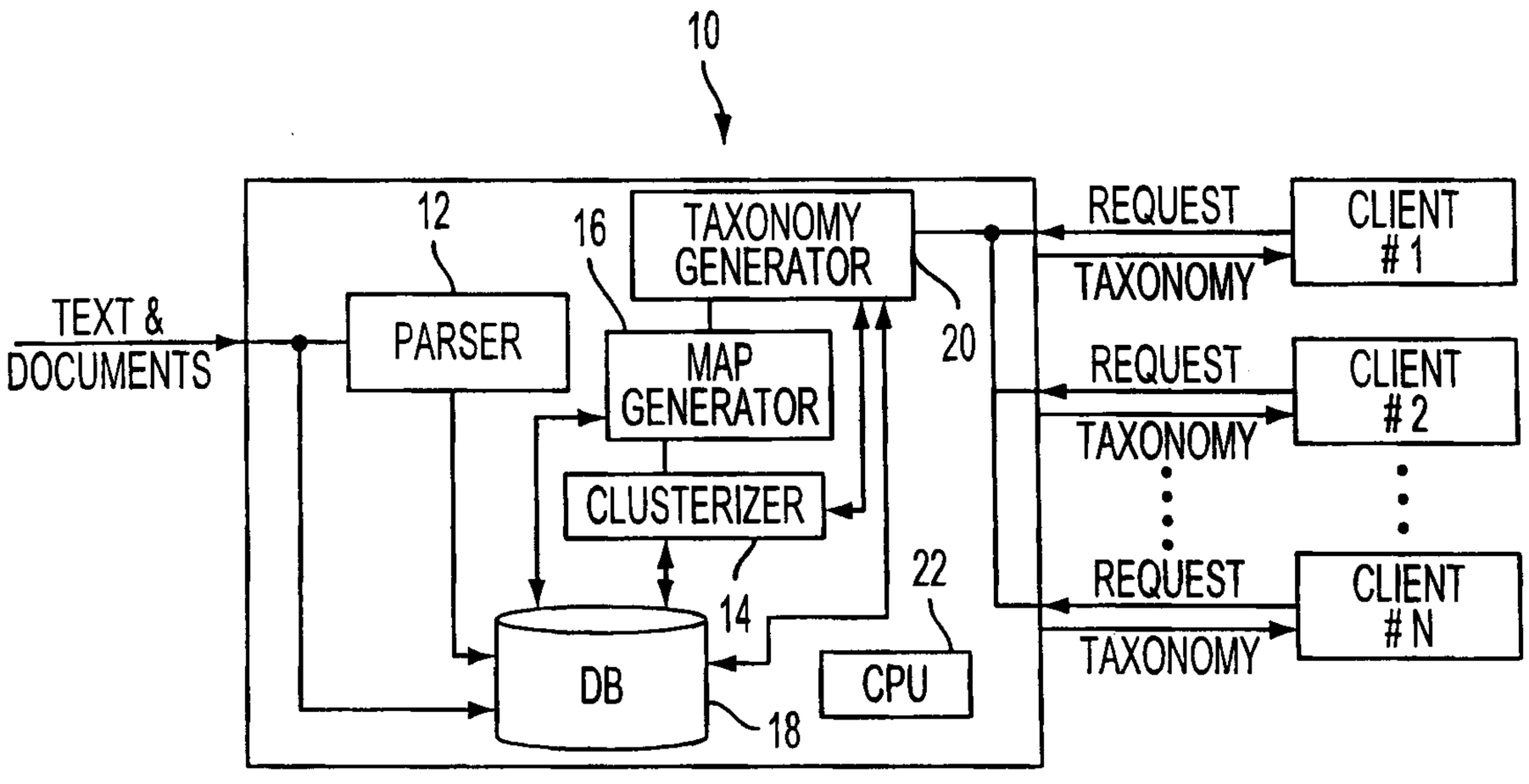


FIG. 1

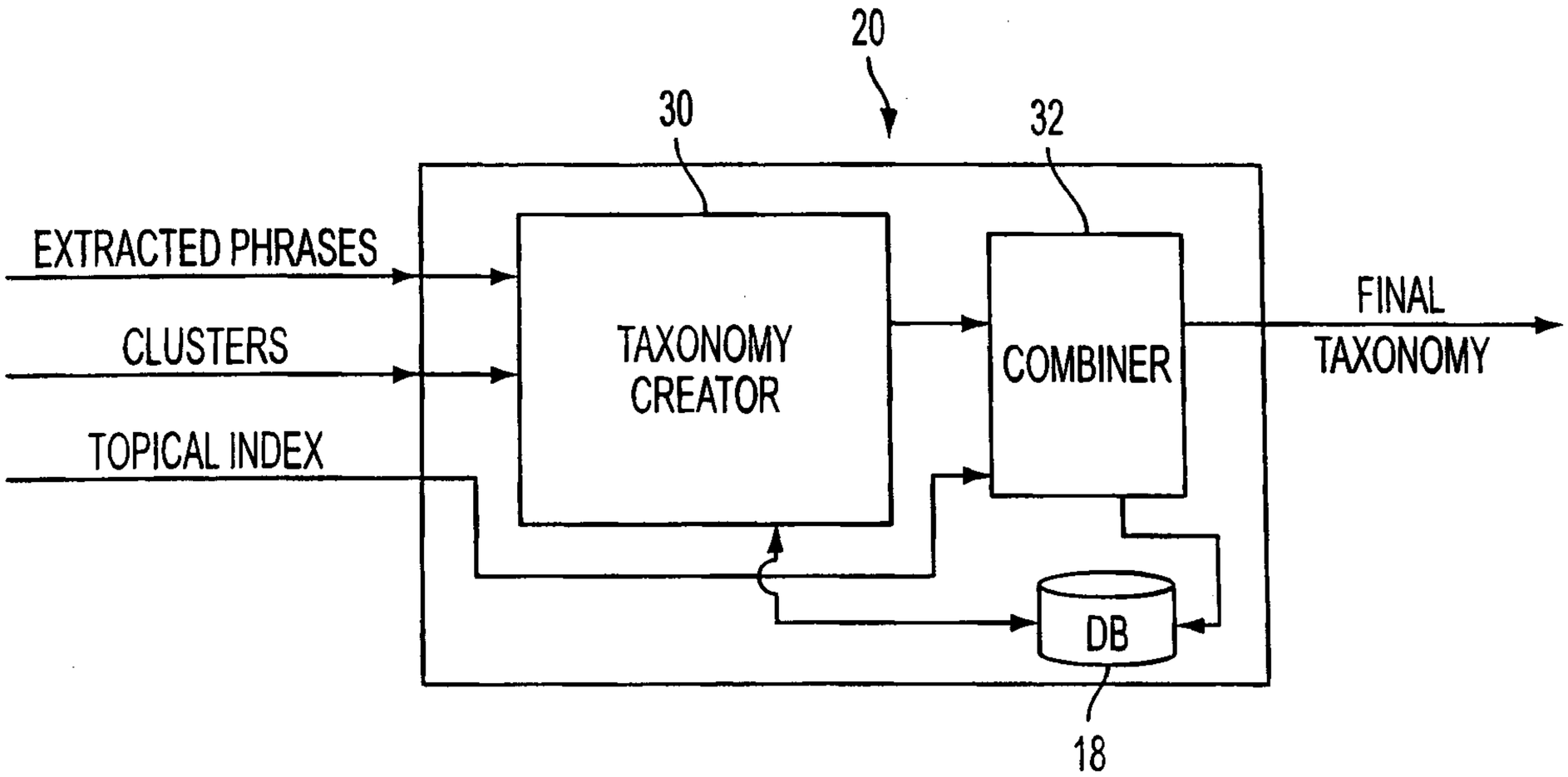


FIG. 2

76886-81

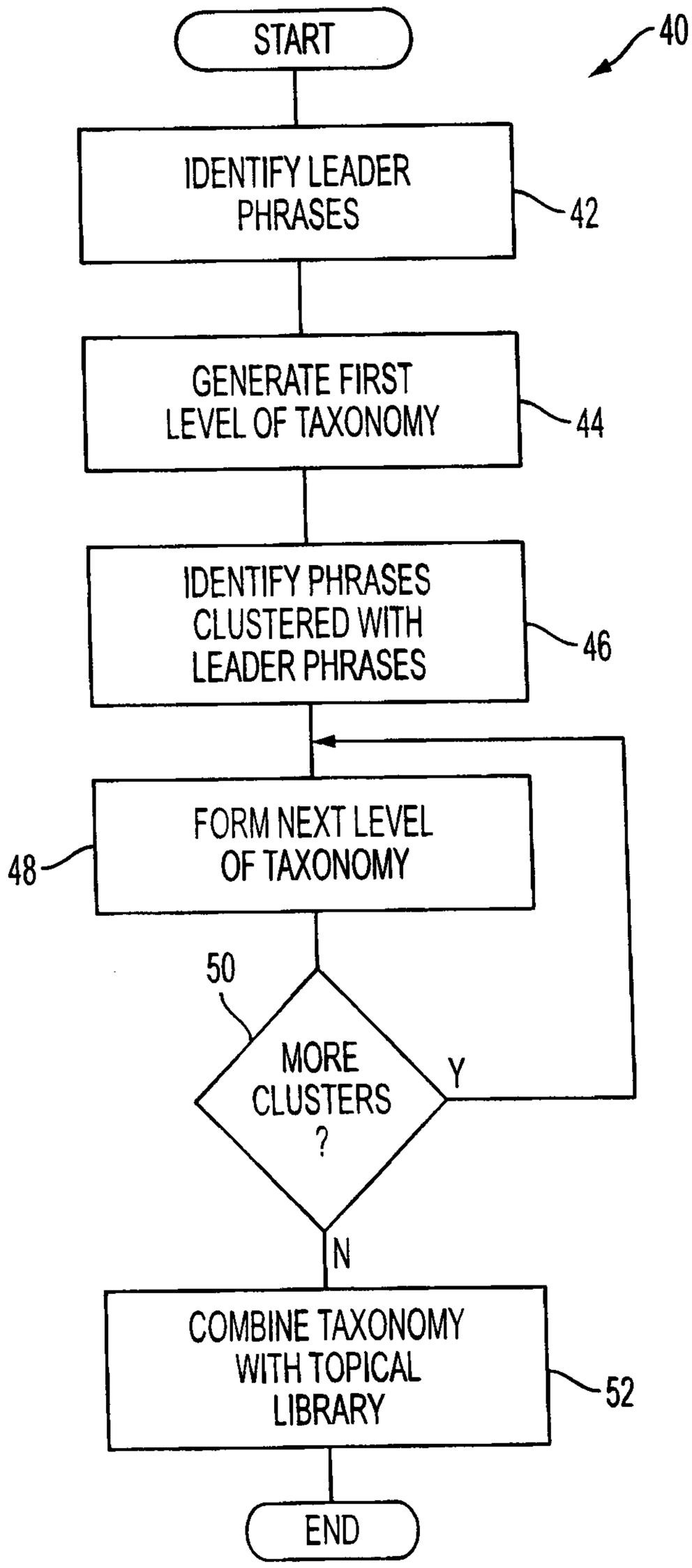
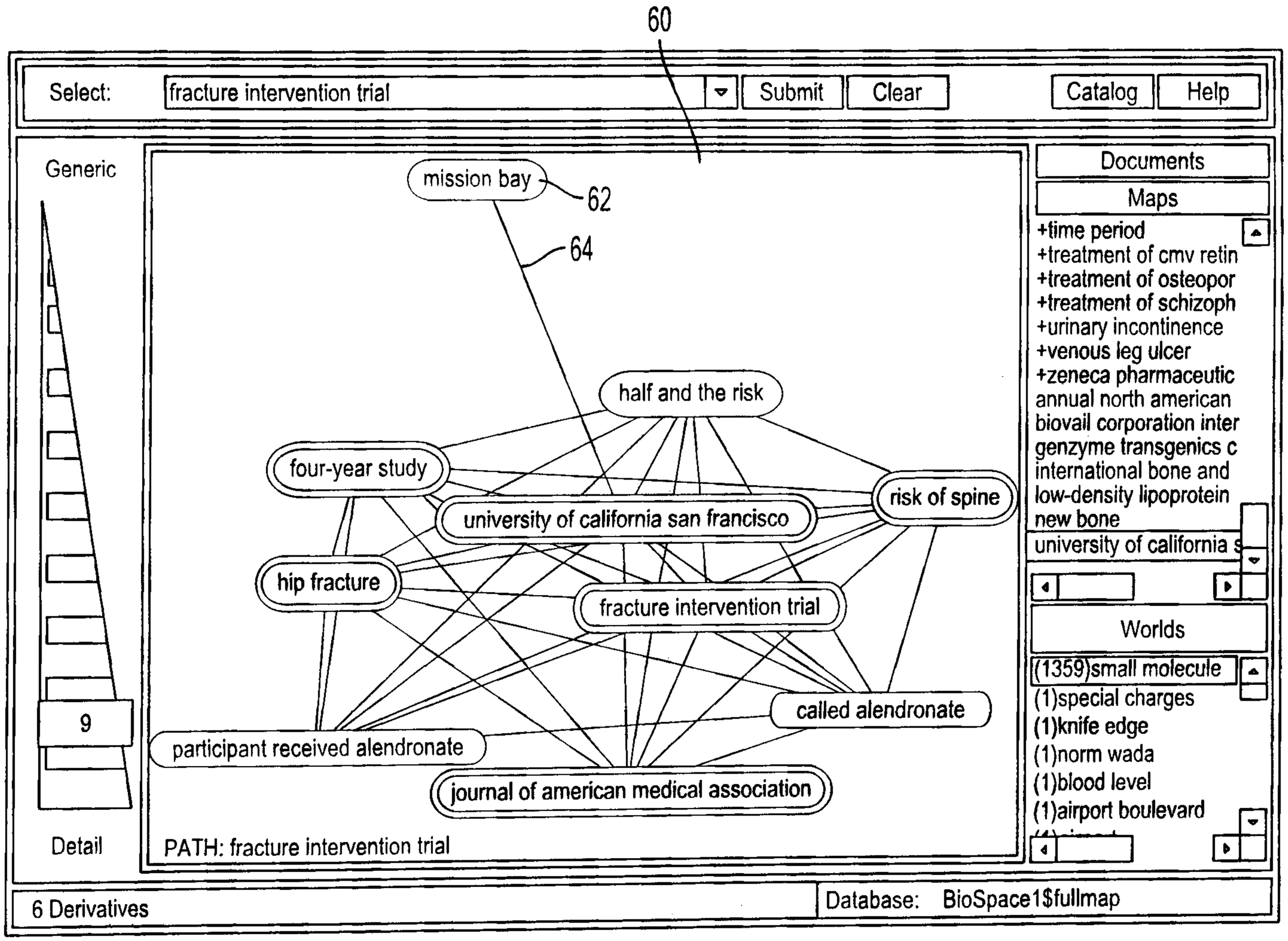


FIG. 3



... A drug for the treatment of osteoporosis dramatically reduced the **risk of spine** and **hip fractures** by almost and the risk of other painful fractures by 30 percent among women with osteoporosis, according to a study by University of California San Francisco researchers. The drug, called alendronate, is a bisphosphonate that binds bones and protects against bone loss. During a four-year study of 4432 postmenopausal women, 2214 participants received alendronate. New results from the study, part of the landmark **Fracture Intervention Trial** (FIT), are published in the December 23 issue of the Journal of American M...

FIG. 4

CA 02366545 2005-09-01

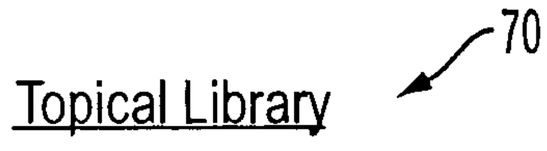

70
Topical Library
Disease Areas
Accidents & Injuries
fracture
burn
trauma
sprain
strain
reperfusion
stroke
wound
ulcer

FIG. 5

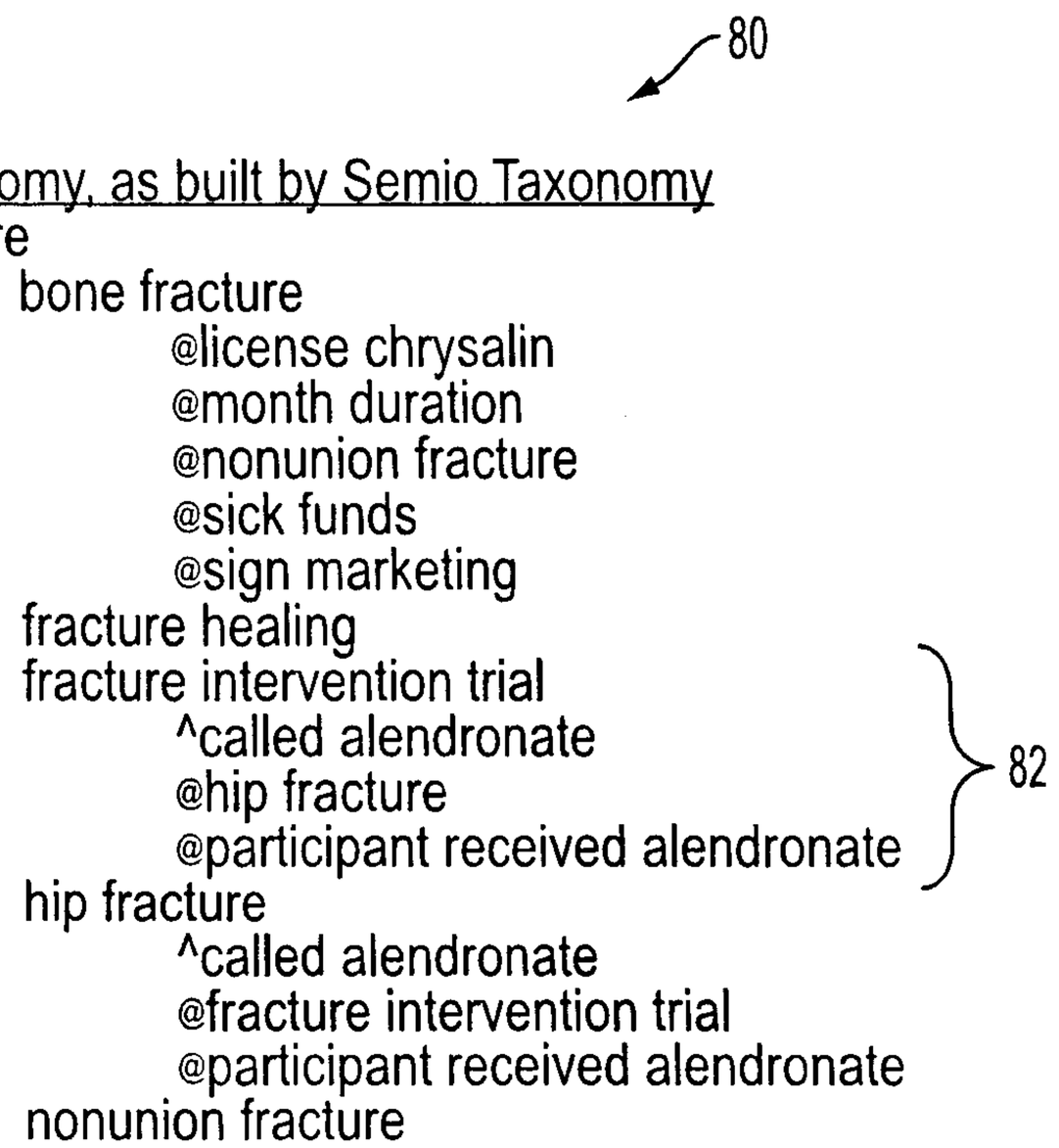

80
Taxonomy, as built by Semio Taxonomy
fracture
bone fracture
@license chrysalin
@month duration
@nonunion fracture
@sick funds
@sign marketing
fracture healing
fracture intervention trial
^called alendronate
@hip fracture
@participant received alendronate
82
hip fracture
^called alendronate
@fracture intervention trial
@participant received alendronate
nonunion fracture
@bone fracture
@month duration
@sick funds

FIG. 6

76886-81

Final taxonomy after integration of extracted taxonomy onto topical library

90

Disease Areas
- Accidents & Injuries
  - fracture
    - bone fracture
      - ^distribution agreement
      - ^license chrysalin
      - ^month duration
      - ^nonunion fracture
      - ^sick funds
      - ^sign marketing
    - fracture healing
    - fracture intervention trial
      - ^called alendronate
      - ^hip fracture
      - ^participant received alendronate
    - hip fracture
      - ^called alendronate
      - ^fracture intervention trial
      - ^participant received alendronate
    - hip fracture surgery
      - ^hip bone
      - ^key measurement

FIG. 7

76886-81

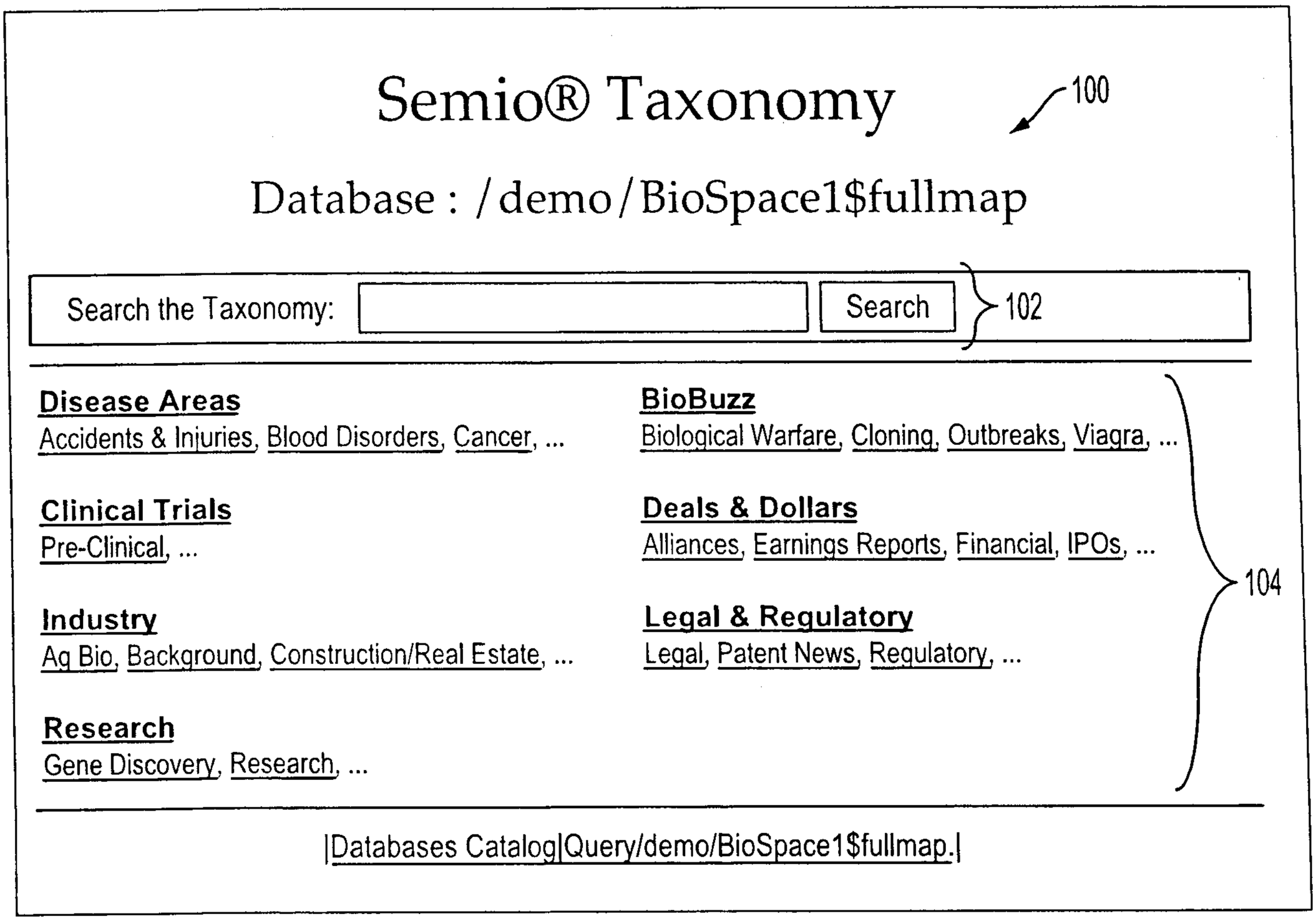


FIG. 8

76886-81

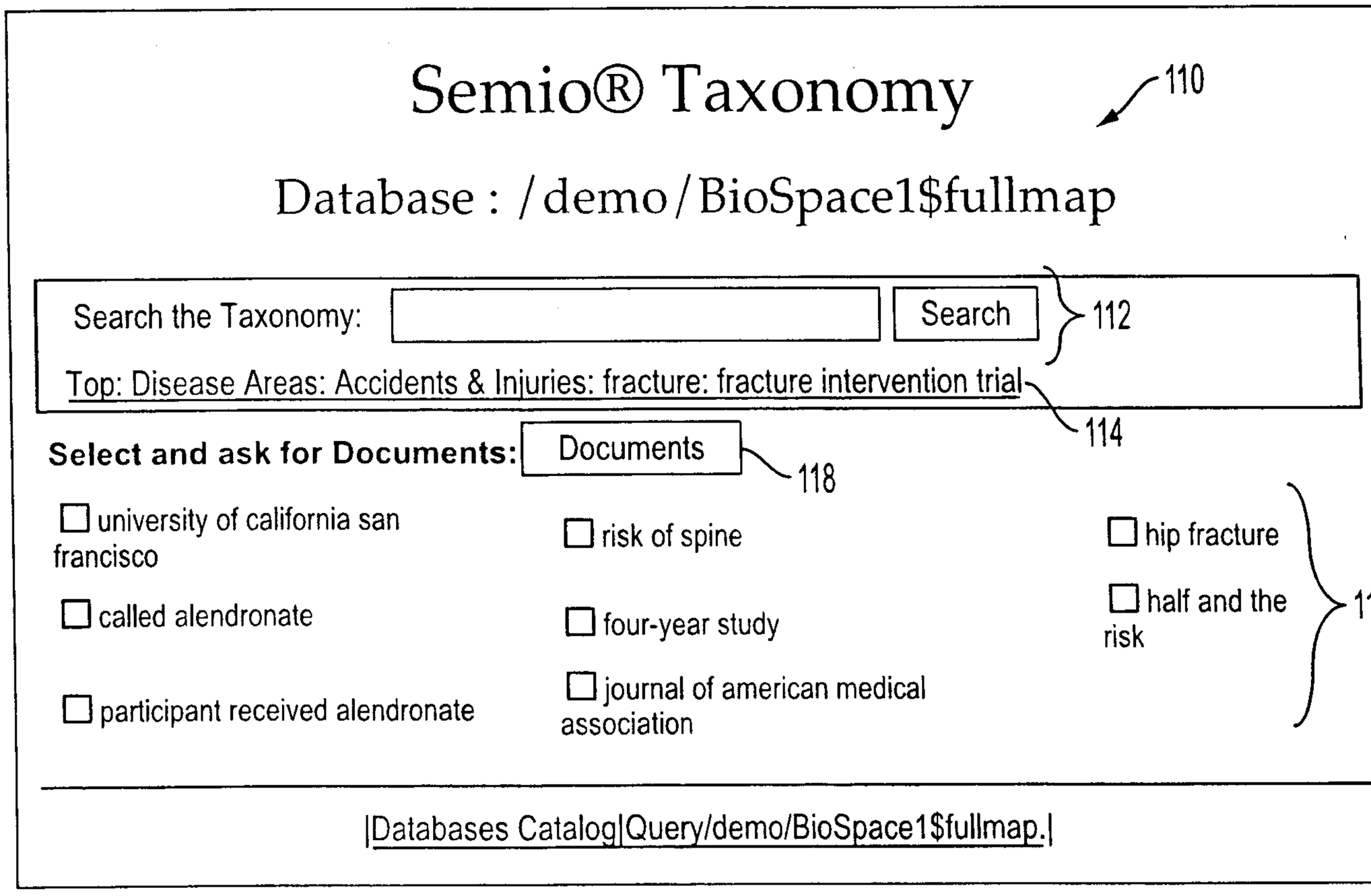
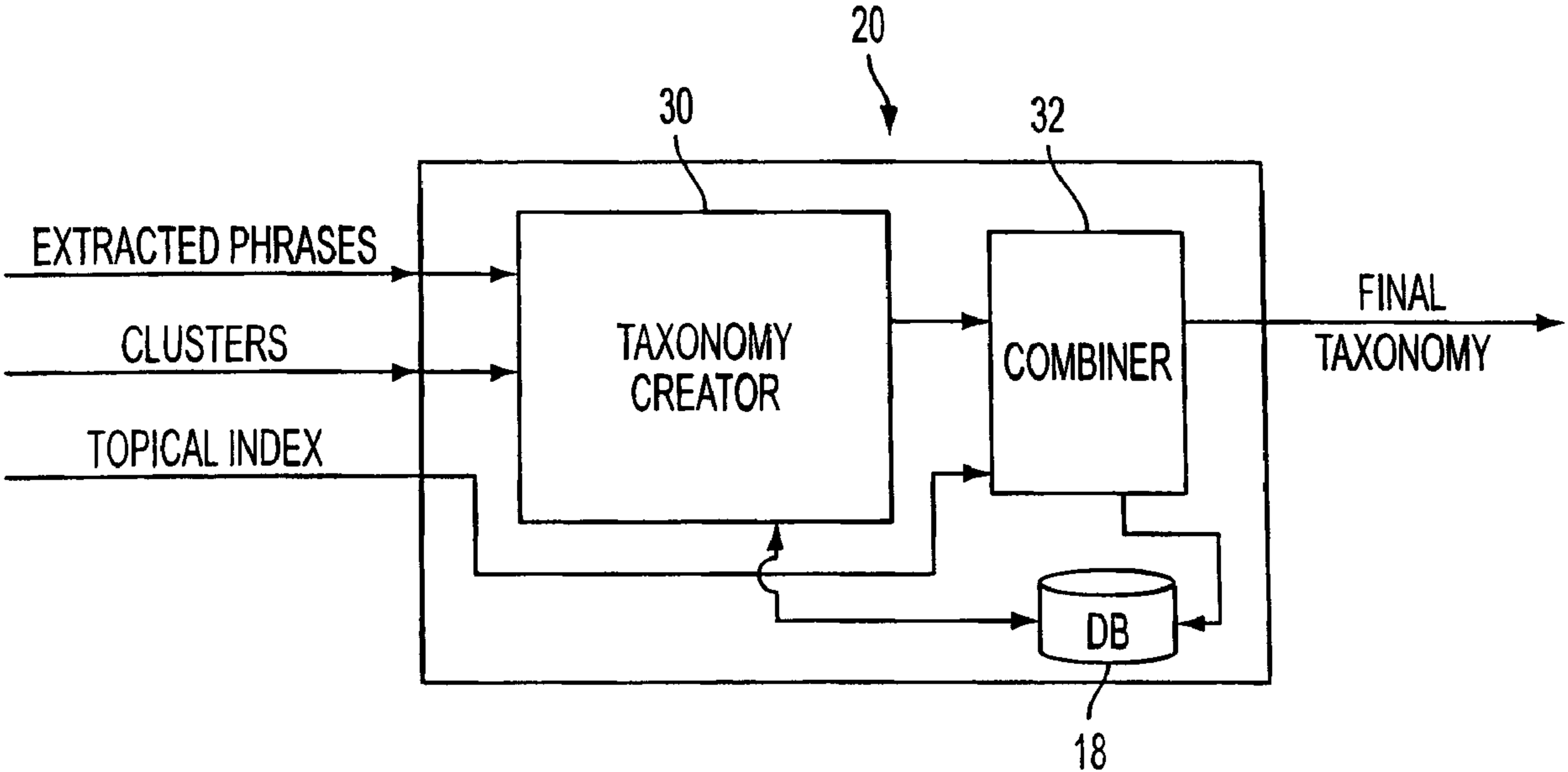


FIG. 9


20
30
32
EXTRACTED PHRASES
CLUSTERS
TOPICAL INDEX
TAXONOMY CREATOR
COMBINER
FINAL TAXONOMY
DB
18