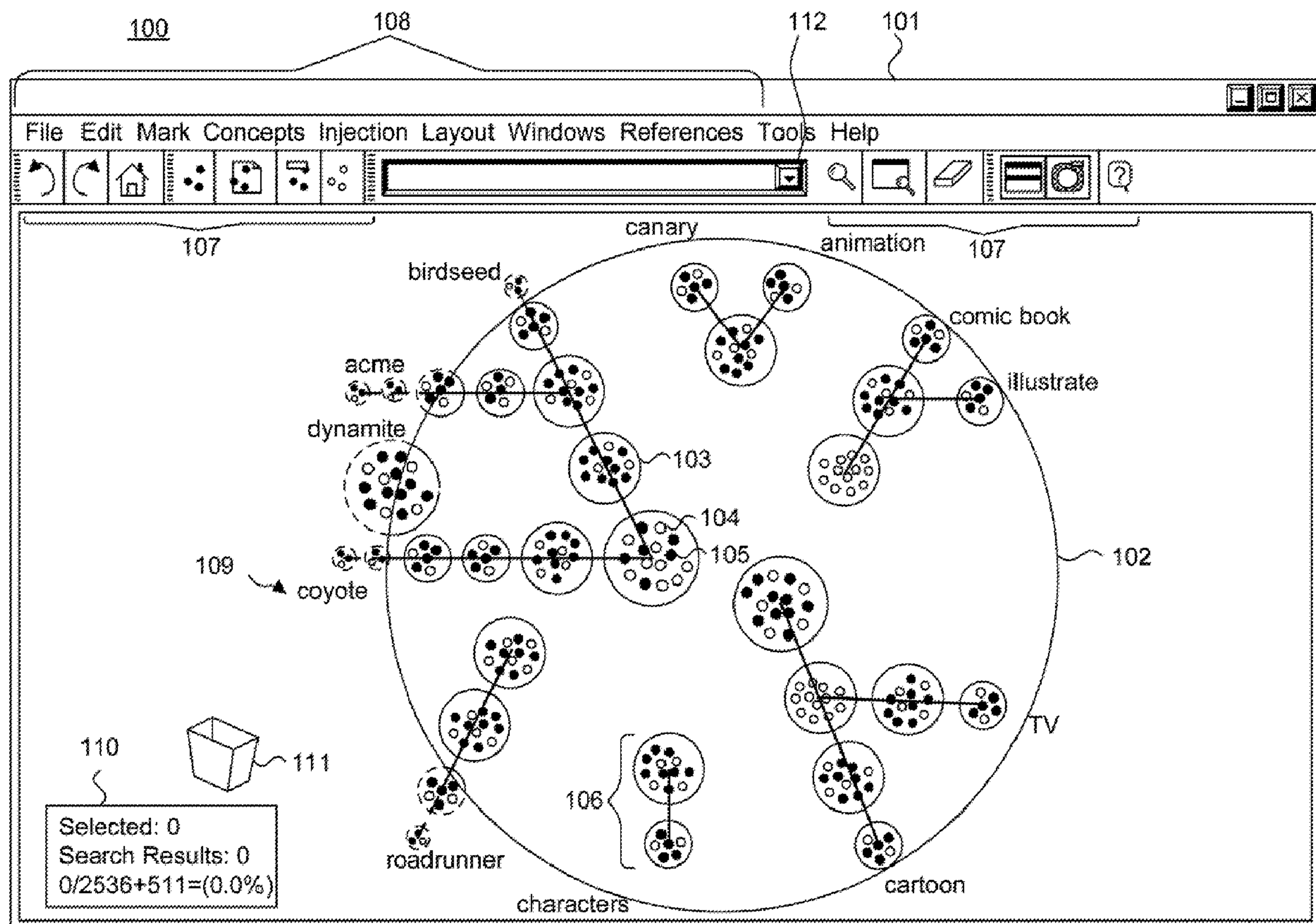




(86) **Date de dépôt PCT/PCT Filing Date:** 2010/07/27
(87) **Date publication PCT/PCT Publication Date:** 2011/02/10
(45) **Date de délivrance/Issue Date:** 2015/08/25
(85) **Entrée phase nationale/National Entry:** 2012/01/26
(86) **N° demande PCT/PCT Application No.:** US 2010/043453
(87) **N° publication PCT/PCT Publication No.:** 2011/017134
(30) **Priorités/Priorities:** 2009/07/28 (US61/229,216);
2009/08/24 (US61/236,490); 2010/07/27 (US12/844,792)

(51) **Cl.Int./Int.Cl. G06F 17/30** (2006.01)
(72) **Inventeurs/Inventors:**
KNIGHT, WILLIAM C., US;
NUSSBAUM, NICHOLAS I., US;
CONWELL, JOHN W., US
(73) **Propriétaire/Owner:**
FTI CONSULTING, INC., US
(74) **Agent:** INTEGRAL IP

(54) **Titre : AFFICHAGE DE RELATIONS ENTRE DES CONCEPTS DANS LE BUT DE PROPOSER DES SUGGESTIONS DE
CLASSEMENT PAR INJECTION**
(54) **Title: DISPLAYING RELATIONSHIPS BETWEEN CONCEPTS TO PROVIDE CLASSIFICATION SUGGESTIONS VIA INJECTION**



(57) **Abrégé/Abstract:**

A system (10) and method (50) for displaying relationships between concepts to provide classification suggestions via injection is provided. A reference set (14d) of concepts each associated with a classification code is designated. Clusters (130) of uncoded concepts (14c) are designated. One or more of the uncoded concepts (14c) from at least one cluster (130) are compared to the

(57) **Abrégé(suite)/Abstract(continued):**

reference set (14d). At least one of the concepts in the reference set (14d) that is similar to the one or more uncoded concepts (14c) is identified. The similar concepts (14d) are injected into the at least one cluster (130). Relationships between the uncoded concepts (14c) and the similar concepts (14d) in the at least one cluster (130) are visually depicted as suggestions for classifying the uncoded concepts (14c).

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

(19) World Intellectual Property Organization
International Bureau(43) International Publication Date
10 February 2011 (10.02.2011)(10) International Publication Number
WO 2011/017134 A4(51) International Patent Classification:
G06F 17/30 (2006.01)

(21) International Application Number:

PCT/US2010/043453

(22) International Filing Date:

27 July 2010 (27.07.2010)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

61/229,216	28 July 2009 (28.07.2009)	US
61/236,490	24 August 2009 (24.08.2009)	US
12/844,792	27 July 2010 (27.07.2010)	US

(71) Applicant (for all designated States except US): **FTI TECHNOLOGY LLC** [US/US]; 500 East Pratt Street, Suite 1400, Baltimore, Maryland 21202 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **KNIGHT, William, C.** [US/US]; 1851 Edna Place, Bainbridge Island, Washington 98110 (US). **NUSSBAUM, Nicholas, I.** [US/US]; 8114 41st Ave. SW, Seattle, Washington 98136 (US). **CONWELL, John, W.** [US/US]; 16052 48th Ave South, Tukwila, Washington 98188 (US).(74) Agents: **WITTMAN, Krista, A.** et al.; Cascadia Intellectual Property, 500 Union Street, Ste 1005, Seattle, Washington 98101 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, 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, RO, RS, RU, 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.

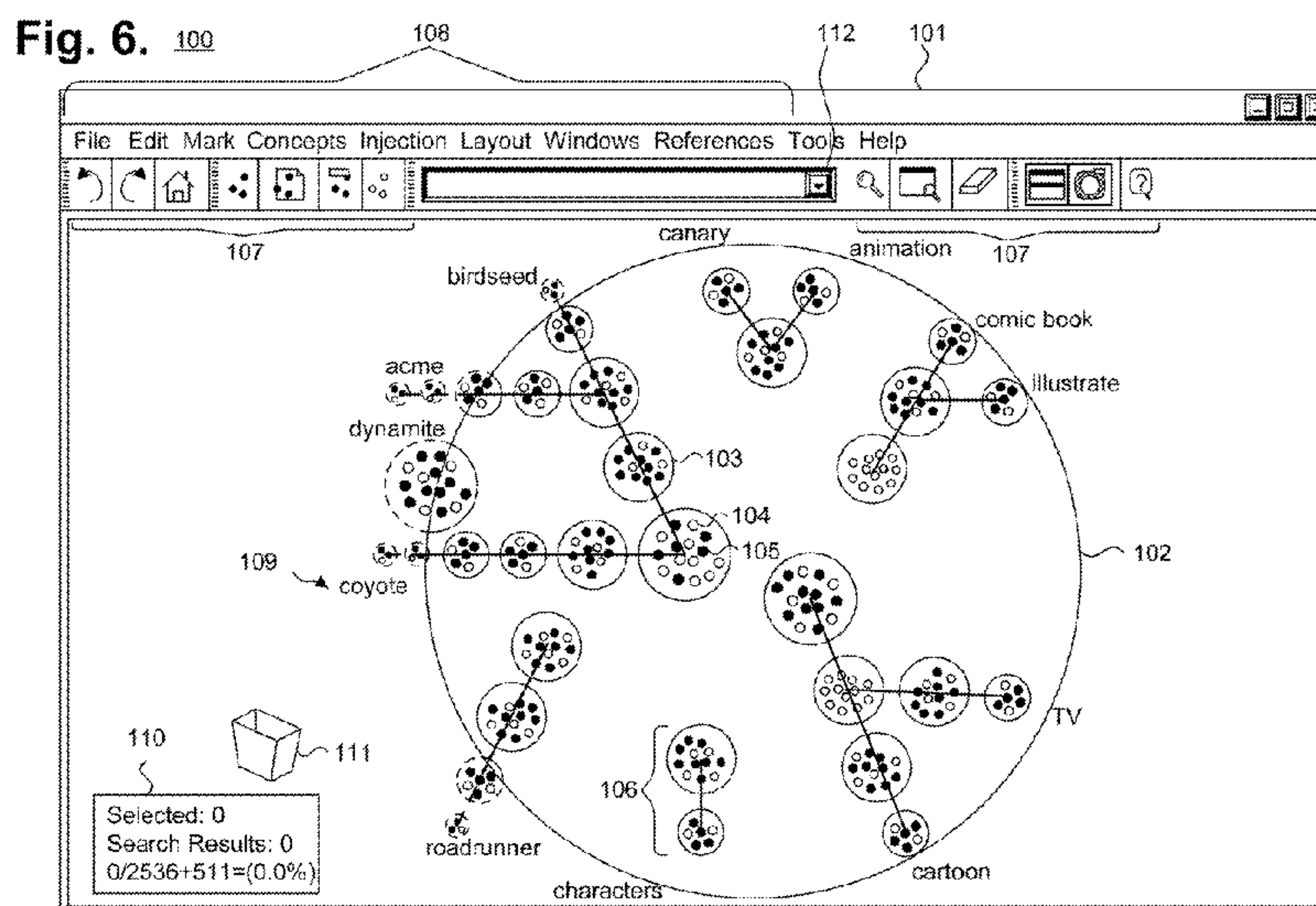
(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, 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))
- with amended claims (Art. 19(1))

[Continued on next page]

(54) Title: DISPLAYING RELATIONSHIPS BETWEEN CONCEPTS TO PROVIDE CLASSIFICATION SUGGESTIONS VIA INJECTION



(57) **Abstract:** A system (10) and method (50) for displaying relationships between concepts to provide classification suggestions via injection is provided. A reference set (14d) of concepts each associated with a classification code is designated. Clusters (130) of uncoded concepts (14c) are designated. One or more of the uncoded concepts (14c) from at least one cluster (130) are compared to the reference set (14d). At least one of the concepts in the reference set (14d) that is similar to the one or more uncoded concepts (14c) is identified. The similar concepts (14d) are injected into the at least one cluster (130). Relationships between the uncoded concepts (14c) and the similar concepts (14d) in the at least one cluster (130) are visually depicted as suggestions for classifying the uncoded concepts (14c).

WO 2011/017134 A4



Date of publication of the amended claims: 31 March 2011

CSCD026-1CA

1

DISPLAYING RELATIONSHIPS BETWEEN CONCEPTS TO PROVIDE CLASSIFICATION SUGGESTIONS VIA INJECTION

TECHNICAL FIELD

This application relates in general to document concepts and, in particular, to a system
5 and method for displaying relationships between concepts to provide classification
suggestions via injection.

BACKGROUND ART

Historically, document review during the discovery phase of litigation and for other
types of legal matters, such as due diligence and regulatory compliance, have been conducted
10 manually. During document review, individual reviewers, generally licensed attorneys, are
assigned sets of documents for coding. A reviewer must carefully study each document and
categorize the document by assigning a code or other marker from a set of descriptive
classifications, such as “privileged,” “responsive,” and “non-responsive.” The classifications
affect the disposition of each document, including admissibility into evidence. During
15 discovery, document review can potentially affect the outcome of the underlying legal matter,
so consistent and accurate results are crucial.

Manual document review is tedious and time-consuming. Marking documents is
solely at the discretion of each reviewer and inconsistent results may occur due to
misunderstanding, time pressures, fatigue, or other factors. A large volume of documents
20 reviewed, often with only limited time, can create a loss of mental focus and a loss of purpose
for the resultant classification. Each new reviewer also faces a steep learning curve to
become familiar with the legal matter, coding categories, and review techniques.

Currently, with the increasingly widespread movement to electronically stored
information (ESI), manual document review is no longer practicable. The often exponential
25 growth of ESI exceeds the bounds reasonable for conventional manual human review and
underscores the need for computer-assisted ESI review tools.

Conventional ESI review tools have proven inadequate to providing efficient,
accurate, and consistent results. For example, DiscoverReady LLC, a Delaware limited
liability company, conducts semi-automated document review through multiple passes over a
30 document set in ESI form. During the first pass, documents are grouped by category and
basic codes are assigned. Subsequent passes refine and further assign codings. Multiple pass

CSCD026-1CA

2

review also requires *a priori* project-specific knowledge engineering, which is useful for only the single project, thereby losing the benefit of any inferred knowledge or know-how for use in other review projects.

Thus, there remains a need for a system and method for increasing the efficiency of document review that bootstraps knowledge gained from other reviews while ultimately ensuring independent reviewer discretion.

DISCLOSURE OF THE INVENTION

Document review efficiency can be increased by identifying relationships between reference concepts and uncoded concepts and providing a suggestion for classification based on the relationships. A set of clusters including uncoded concepts is obtained. Each of the uncoded concepts represented one or more uncoded documents to be classified. The uncoded concepts for a cluster are compared to a set of reference concepts. Those reference concepts most similar to the uncoded concepts are identified and inserted into the cluster. The relationship between the inserted reference concepts and uncoded concepts for the cluster are visually depicted and provide a suggestion regarding classification of the uncoded concepts. The classified concepts can then be used to classify the documents associated with the concepts.

An embodiment provides a system and method for displaying relationships between concepts to provide classification suggestions via injection. A reference set of concepts each associated with a classification code is designated. Clusters of uncoded concepts are designated. One or more of the uncoded concepts from at least one cluster are compared to the reference set. At least one of the concepts in the reference set that is similar to the one or more uncoded concepts is identified. The similar concepts are injected into the at least one cluster. Relationships between the uncoded concepts and the similar concepts in the at least one cluster are visually depicted as suggestions for classifying the uncoded concepts.

Still other embodiments of the present invention will become readily apparent to those skilled in the art from the following detailed description, wherein are described embodiments by way of illustrating the best mode contemplated for carrying out the invention. As will be realized, the invention is capable of other and different embodiments and its several details are capable of modifications in various obvious respects. Accordingly, the drawings and detailed description are to be regarded as illustrative in nature.

CSCD026-1CA

DESCRIPTION OF THE DRAWINGS

FIGURE 1 is a block diagram showing a system for displaying relationships between concepts to provide classification suggestions via injection, in accordance with one embodiment.

5 FIGURE 2 is a process flow diagram showing a method for displaying relationships between concepts to provide classification suggestions via injection, in accordance with one embodiment.

FIGURE 3 is a table showing, by way of example, a matrix mapping of uncoded concepts and documents.

10 FIGURE 4 is a process flow diagram showing, by way of example, a method for forming clusters for use in the method of FIGURE 2.

FIGURE 5 is a block diagram showing, by way of example, cluster measures for identifying similar reference concepts for use in the method of FIGURE 2.

15 FIGURE 6 is a screenshot showing, by way of example, a visual display of reference concepts in relation to uncoded concepts.

FIGURE 7A is a block diagram showing, by way of example, a cluster with “privileged” reference concepts and uncoded concepts.

FIGURE 7B is a block diagram showing, by way of example, a cluster 96 with “non-responsive” reference concepts and uncoded concepts.

20 FIGURE 7C is a block diagram showing, by way of example, a cluster with a combination of classified reference concepts and uncoded concepts.

FIGURE 8 is a process flow diagram showing, by way of example, a method for classifying uncoded concepts for use in the method of FIGURE 2 using a classifier.

25 FIGURE 9 is a screenshot showing, by way of example, a reference options dialogue box for entering user preferences for reference concept injection.

BEST MODE FOR CARRYING OUT THE INVENTION

The ever-increasing volume of ESI underlies the need for automating document review for improved consistency and throughput. Token clustering via injection utilizes reference, or previously classified tokens, which offer knowledge gleaned from earlier work
30 in similar legal projects, as well as a reference point for classifying uncoded tokens.

The tokens can include word-level, symbol-level, or character-level *n*-grams, raw terms, entities, or concepts. Other tokens, including other atomic parse-level elements, are

CSCD026-1CA

4

possible. An n -gram is a predetermined number of items selected from a source. The items can include syllables, letters, or words, as well as other items. A raw term is a term that has not been processed or manipulated. Entities further refine nouns and noun phrases into people, places, and things, such as meetings, animals, relationships, and various other objects.

5 Additionally, entities can represent other parts of grammar associated with semantic meanings to disambiguate different instances or occurrences of the grammar. Entities can be extracted using entity extraction techniques known in the field.

Concepts are collections of nouns and noun-phrases with common semantic meaning that can be extracted from ESI, including documents, through part-of-speech tagging. Each
10 concept can represent one or more documents to be classified during a review. Clustering of the concepts provides an overall view of the document space, which allows users to easily identify documents sharing a common theme.

The clustering of tokens, for example, concepts, differs from document clustering, which groups related documents individually. In contrast, concept clustering groups related
15 concepts, which are each representative of one or more related documents. Each concept can express an ideas or topic that may not be expressed by individual documents. A concept is analogous to a search query by identifying documents associated with a particular idea or topic.

A user can determine how particular concepts are related based on the concept
20 clustering. Further, users are able to intuitively identify documents by selecting one or more associated concepts in a cluster. For example, a user may wish to identify all documents in a particular corpus that are related to car manufacturing. The user can select the concept “car manufacturing” or “vehicle manufacture” within one of the clusters and subsequently, the associated documents are presented. However, during document clustering, a user is first
25 required to select a specific document from which other documents that are similarly related can then be identified.

Reference tokens are previously classified based on the document content represented by that token and can be injected into clusters of uncoded, that is unclassified, tokens to influence classification of the uncoded tokens. Specifically, relationships between an
30 uncoded token and the reference tokens, in terms of semantic similarity or distinction, can be used as an aid in providing suggestions for classifying uncoded tokens. Once classified, the newly-coded, or reference, tokens can be used to further classify the represented documents.

CSCD026-1CA

5

Although tokens, such as word-level or character-level n -grams, raw terms, entities, or concepts, can be clustered and displayed, the discussion below will focus on a concept as a particular token.

Complete ESI review requires a support environment within which classification can be performed. FIGURE 1 is a block diagram showing a system 10 for providing reference concepts as a suggestion for uncoded concepts, in accordance with one embodiment. By way of illustration, the system 10 operates in a distributed computing environment, which includes a plurality of heterogeneous systems and ESI sources. Henceforth, a single item of ESI will be referenced as a “document,” although ESI can include other forms of non-document data, as described *infra*. A backend server 11 is coupled to a storage device 13, which stores documents 14a, such as uncoded documents in the form of structured or unstructured data, a database 30 for maintaining information about the documents, a lookup database 38 for storing many-to-many mappings between documents and document features, and a concept document index 40, which maps documents to concepts. The storage device 13 also stores reference documents 14b, concepts 14c, and reference concepts 14d. Concepts are collections of nouns and noun-phrases with common semantic meaning. The nouns and noun-phrases can be extracted from one or more documents in the corpus for review. Hereinafter, the terms “classified” and “coded” are used interchangeably with the same intended meaning, unless otherwise indicated.

The backend server 11 is coupled to an intranetwork 21 and executes a workbench software suite 31 for providing a user interface framework for automated document management, processing, analysis, and classification. In a further embodiment, the backend server 11 can be accessed via an internetwork 22. The workbench software suite 31 includes a document mapper 32 that includes a clustering engine 33, similarity searcher 34, classifier 35, and display generator 36. Other workbench suite modules are possible.

The clustering engine 33 performs efficient concept scoring and clustering of uncoded concepts. Efficient concept scoring and clustering is described in commonly-assigned U.S. Patent application Publication No. 2005/0022106, pending. Clusters of uncoded concepts 14c can be organized along vectors, known as spines, based on a similarity of the clusters. Similarity can be expressed in terms of distance. Concept clustering is further discussed below with reference to FIGURE 4. The similarity searcher 34 identifies the reference concepts 14d that are most similar to selected uncoded concepts, clusters, or

CSCD026-1CA

6

spines. The classifier 35 provides a machine-generated suggestion and confidence level for classification of the selected uncoded concepts, cluster, or spine, as further described below with reference to FIGURE 8. The display generator 36 arranges the clusters and spines in thematic relationships in a two-dimensional visual display space and inserts the identified

5 reference concepts into one or more of the clusters, as further described below beginning with reference to FIGURE 2. Once generated, the visual display space is transmitted to a work client 12 by the backend server 11 via the document mapper 32 for presenting to a reviewer on a display 37. The reviewer can include an individual person who is assigned to review a set of documents by classifying one or more uncoded concepts and designating a code.

10 Hereinafter, the terms “reviewer” and “custodian” are used interchangeably with the same intended meaning. Other types of reviewers are possible, including machine-implemented reviewers.

The document mapper 32 operates on uncoded concepts 14c, which can be retrieved from the storage 13, as well as from a plurality of local and remote sources. The local and

15 remote sources can also store the reference concepts 14d, as well as the uncoded documents 14a and reference documents 14b. The local sources include documents and concepts 17 maintained in a storage device 16 coupled to a local server 15, and documents and concepts 20 maintained in a storage device 19 coupled to a local client 18. The local server 15 and local client 18 are interconnected to the backend server 11 and the work client 12 over the

20 intranetwork 21. In addition, the document mapper 32 can identify and retrieve documents from remote sources over the internetwork 22, including the Internet, through a gateway 23 interfaced to the intranetwork 21. The remote sources include documents 26 maintained in a storage device 25 coupled to a remote server 24 and documents 29 maintained in a storage device 28 coupled to a remote client 27. Other document sources, either local or remote, are

25 possible.

The individual documents 14a, 14b, 17, 20, 26, 29 include all forms and types of structured and unstructured ESI, including electronic message stores, word processing documents, electronic mail (email) folders, Web pages, and graphical or multimedia data. Notwithstanding, the documents could be in the form of structurally organized data, such as

30 stored in a spreadsheet or database.

In one embodiment, the individual documents 14a, 14b, 17, 20, 26, 29 can include electronic message folders storing email and attachments, such as maintained by the Outlook

CSCD026-1CA

7

and Outlook Express products, licensed by Microsoft Corporation, Redmond, Washington. The database can be an SQL-based relational database, such as the Oracle database management system, release 8, licensed by Oracle Corporation, Redwood Shores, California.

5 Additionally, the individual concepts 14c, 14d, 17, 20, 26, 29 include uncoded concepts and reference concepts. The uncoded concepts, which are unclassified, represent collections of nouns and noun-phrases that are semantically related and extracted from documents in a document review project.

10 The reference concepts are initially uncoded concepts that can represent documents selected from the corpus or other sources of documents. The reference concepts assist in providing suggestions for classification of the remaining uncoded concepts representative of the document corpus based on visual relationships between the uncoded concepts and reference concepts. The reviewer can classify one or more of the remaining uncoded concepts by assigning a classification code based on the relationships. In a further embodiment, the reference concepts can be used as a training set to form machine-generated suggestions for classifying the remaining uncoded concepts, as further described below with reference to FIGURE 8.

15 The reference concepts are representative of the document corpus for a review project in which data organization or classification is desired. A set of reference concepts can be generated for each document review project or alternatively, the reference concepts can be representative of documents selected from a previously conducted document review project that is related to the current document review project. Guided review assists a reviewer in building a reference concept set representative of the corpus for use in classifying uncoded concepts. During guided review, uncoded concepts that are dissimilar to all other uncoded concepts are identified based on a similarity threshold. Other methods for determining dissimilarity are possible. Identifying the dissimilar concepts provides a group of uncoded concepts that is representative of the document corpus for a document review project. Each identified dissimilar concept is then classified by assigning a particular classification code based on the content of the documents represented by that concept to generate a set of reference concepts for the document review project. Guided review can be performed by a reviewer, a machine, or a combination of the reviewer and machine.

30 Other methods for generating a reference concept set for a document review project using guided review are possible, including clustering. For example, a set of uncoded

CSCD026-1CA

8

concepts to be classified is clustered, as described in commonly-assigned U.S. Patent application Publication No. 2005/0022106, pending. A plurality of the clustered uncoded concepts are selected based on selection criteria, such as cluster centers or sample clusters. The cluster centers can be used to identify uncoded concepts in a cluster that are most similar
5 or dissimilar to the cluster center. The identified uncoded concepts are then selected for classification by assigning codes. After classification, the previously uncoded concepts represent a concept reference set. In a further example, sample clusters can be used to generate a reference set by selecting one or more sample clusters based on cluster relation criteria, such as size, content, similarity, or dissimilarity. The uncoded concepts in the
10 selected sample clusters are then assigned classification codes. The classified concepts represent a reference concept set for the document review project. Other methods for selecting uncoded concepts for use as a reference set are possible.

The document corpus for a document review project can be divided into subsets of documents, which are each provided to a particular reviewer as an assignment. The uncoded
15 documents are analyzed to identify concepts, which are subsequently clustered. A classification code can be assigned to each of the clustered concepts. To maintain consistency, the same codes can be used across all concepts representing assignments in the document review project. The classification codes can be determined using taxonomy generation, during which a list of classification codes can be provided by a reviewer or
20 determined automatically. The classification code of a concept can be assigned to the documents associated with that concept.

For purposes of legal discovery, the classification codes used to classify concepts can include “privileged,” “responsive,” or “non-responsive.” Other codes are possible. The assigned classification codes can be used as suggestions for classification of associated
25 documents. For example, a document associated with three concepts, each assigned a “privileged” classification can also be considered “privileged.” Other types of suggestions are possible. A “privileged” document contains information that is protected by a privilege, meaning that the document should not be disclosed or “produced” to an opposing party. Disclosing a “privileged” document can result in an unintentional waiver of the subject
30 matter disclosed. A “responsive” document contains information that is related to the legal matter, while a “non-responsive” document includes information that is not related to the legal matter.

CSCD026-1CA

Identifying reference concepts that are most similar to an uncoded concept, cluster, or spine can be performed by the system 10, which includes individual computer systems, such as the backend server 11, work server 12, server 15, client 18, remote server 24 and remote client 27. The individual computer systems are general purpose, programmed digital
5 computing devices consisting of a central processing unit (CPU), random access memory (RAM), non-volatile secondary storage, such as a hard drive or CD ROM drive, network interfaces, and peripheral devices, including user interfacing means, such as a keyboard and display. The various implementations of the source code and object and byte codes can be held on a computer-readable storage medium, such as a floppy disk, hard drive, digital video
10 disk (DVD), random access memory (RAM), read-only memory (ROM) and similar storage mediums. For example, program code, including software programs, and data are loaded into the RAM for execution and processing by the CPU and results are generated for display, output, transmittal, or storage.

Identifying reference concepts for use as classification suggestions include a
15 comparison of the uncoded concepts and reference concepts. FIGURE 2 is a process flow diagram showing a method 50 for displaying relationships between concepts to provide classification suggestions via injection, in accordance with one embodiment. A cluster set of uncoded concepts is obtained (block 51). For each cluster, a cluster center is determined based on the uncoded concepts included in that cluster. The clusters can be generated upon
20 command or previously generated and stored. Clustering uncoded concepts is further discussed below with reference to FIGURE 3. One or more uncoded concepts can be compared with a reference concept set (block 52) and those reference concepts that satisfy a threshold of similarity are selected (block 53). Determining similar reference concepts is further discussed below with reference to FIGURE 5. The selected reference concepts are
25 then injected into the cluster associated with the one or more uncoded concepts (block 54). The selected reference concepts injected into the cluster can be the same as or different than the selected reference concepts injected into another cluster. The total number of reference concepts and uncoded concepts in the clusters can exceed the sum of the uncoded concepts originally clustered and the reference concept set. In a further embodiment, a single uncoded
30 concept or spine can be compared to the reference concept set to identify similar reference concepts for injecting into the cluster set.

CSCD026-1CA

10

Together, reference concepts injected into the clusters represent a subset of reference concepts specific to that cluster set. The clusters of uncoded concepts and inserted reference concepts can be displayed to visually depict relationships (block 55) between the uncoded concepts in the cluster and the inserted reference concepts. The relationships can provide a suggestion for use by an individual reviewer, for classifying that cluster. Determining relationships between the reference concepts and uncoded concepts to identify classification suggestions is further discussed below with reference to FIGURE 7A-7C. Further, machine classification can optionally provide a classification suggestion based on a calculated confidence level (block 56). Machine-generated classification suggestions and confidence levels are further discussed below with reference to FIGURE 8.

In one embodiment, the classified concepts can be used to classify those documents represented by that concept. For example, in a product liability lawsuit, the plaintiff claims that a wood composite manufactured by the defendant induces and harbors mold growth. During discovery, all documents within the corpus for the lawsuit and relating to mold should be identified for review. The concept for mold is clustered and includes a “responsive” classification code, which indicates that the noun phrase mold is related to the legal matter. Upon selection of the mold concept, all documents that include the noun phrase mold can be identified using the mapping matrix, which is described below with reference to FIGURE 3. The responsive classification code assigned to the concept can be used as a suggestion for the document classification. However, if the document is represented by multiple concepts with different classification codes, each different code can be considered during classification of the document.

In a further embodiment, the concept clusters can be used with document clusters, which are described in commonly-owned in U.S. Patent No. 8,713,018, entitled “System and Method for Displaying Relationships Between Electronically Stored Information to Provide Classification Suggestions via Inclusion,” filed July 9, 2010, and U.S. Patent No. 8,515,957, entitled “System and Method for Displaying Relationships Between Electronically Stored Information to Provide Classification Suggestions via Injection,” filed July 9, 2010. For example, selecting a concept in the concept cluster display can identify one or more documents with a common idea or topic. Further selection of one of the documents represented by the selected cluster in the document concept display can identify documents

CSCD026-1CA

that are similarly related to the content of the selected document. The identified documents can be the same or different as the other documents represented by the concept.

Similar documents can also be identified as described in commonly-assigned U.S. Patent No. 8,572,084, entitled "System and Method for Displaying Relationships Between Electronically Stored Information to Provide Classification Suggestions via Nearest Neighbor," filed July 9, 2010.

In an even further embodiment, the documents identified from one of the concepts can be classified automatically as described in commonly-assigned U.S. Patent No. 8,635,223, entitled "System and Method for Providing a Classification Suggestion for Electronically Stored Information," filed July 9, 2010.

A corpus of documents for a review project can be divided into assignments using assignment criteria, such as custodian or source of the documents, content, document type, and date. Other criteria are possible. Each assignment is assigned to an individual reviewer for analysis. The assignments can be separately analyzed or alternatively, analyzed together to determine concepts for the one or more document assignments. The content of each document within the corpus can be converted into a set of concepts. As described above, concepts typically include nouns and noun phrases obtained through part-of-speech tagging that have a common semantic meaning. The concepts, which are representative of the documents can be clustered to provide an intuitive grouping of the document content.

Clustering of the uncoded concepts provides groupings of related uncoded concepts and is based on a similarity metric using score vectors assigned to each uncoded concept. The score vectors can be generated using a matrix showing the uncoded concepts in relation to documents that contain the concepts. FIGURE 3 is a table showing, by way of example, a matrix mapping 60 of uncoded concepts 64 and documents 63. The uncoded documents 63 are listed along a horizontal dimension 61 of the matrix, while the concepts 64 are listed along a vertical dimension 62. However, the placement of the uncoded documents 63 and concepts 64 can be reversed. Each cell 65 within the matrix 60 includes a cumulative number of occurrences of each concept within a particular uncoded document 63. Score vectors can be generated for each document by identifying the concepts and associated weights within that document and ordering the concepts along a vector with the associated concept weight. In the matrix 60, the score vector 66 for a document 63 can be identified as all the concepts included in that document and the associated weights, which are based on the number of

CSCD026-1CA

12

occurrences of each concept. Score vectors can also be generated for each concept by identifying the documents that contain that concept and determining a weight associated with each document. The documents and associated weights are then ordered along a vector for each concept, as the concept score vector. In the matrix 60, the score vector 67 for a concept
 5 can be identified as all the documents that contain that concept and the associated weights.

As an initial step for generating score vectors, each document associated with an uncoded concept is individually scored. Next, a normalized score vector is created for each uncoded concept by identifying paired values, consisting of a document and an associated score. The associated score for each document can be based on the number of occurrences of
 10 the uncoded concept in that document. Once determined, the paired values can be ordered along a vector based on concept or frequency, as well as other factors. For example, assume a normalized score vector for a first uncoded concept A is $\vec{S}_A = \{(5, 0.5), (120, 0.75)\}$ and a normalized score vector for another uncoded concept B is $\vec{S}_B = \{(3, 0.4), (5, 0.75), (47, 0.15)\}$. Concept A has scores corresponding to documents '5' and '120' and Concept B has
 15 scores corresponding to documents '3,' '5' and '47.' Once generated, the score vectors can be compared to determine similarity or dissimilarity between the uncoded concepts during clustering. Thus, upon comparison, Concepts share document '5' in common.

The uncoded concepts can be clustered using the associated score vectors. The clustering can be based on cluster criteria, such as the similarity of the concepts. Other
 20 clustering criteria are possible, including clustering by entities, email address, source, raw terms, n-grams, and other metadata. FIGURE 4 is a flow diagram showing a routine 70 for forming clusters for use in the method of FIGURE 2. The purpose of this routine is to use the score vectors associated with each uncoded concept to form clusters based on relative similarity. The score vector for each uncoded concept includes a set of paired values of
 25 documents and weights. The score vector for an uncoded concept is generated by scoring the documents associated with that concept, as described in commonly-assigned U.S. Patent Application Publication No. 2005/0022106, pending.

The routine for forming clusters of uncoded concepts proceeds in two phases. During the first phase (blocks 73-78), uncoded concepts are evaluated to identify a set of seed
 30 concepts, which can be used to form new clusters. During the second phase (blocks 80-86),

CSCD026-1CA

13

the uncoded concepts not previously placed are evaluated and grouped into existing clusters based on a best-fit criterion.

Initially, a single cluster is generated with one or more uncoded concepts as seed concepts and additional clusters of uncoded concepts are added. Each cluster is represented
 5 by a cluster center that is associated with a score vector, which is representative of all the uncoded concepts in that cluster. The cluster center score vector can be generated by comparing the score vectors for the individual uncoded concepts in the cluster and identifying the most common documents shared by the uncoded concepts. The most common documents and associated weights are ordered along the cluster center score vector. Cluster centers, and
 10 thus, cluster center score vectors can continually change due to the addition or removal of concepts during clustering.

During clustering, the uncoded concepts are identified (block 71) and ordered by length (block 72). The uncoded concepts can include all uncoded concepts representative of a corpus or can include only those uncoded concepts representative of a single assignment.
 15 Each uncoded concept is then processed in an iterative processing loop (blocks 73-78) as follows. The similarity between each uncoded concept and the cluster centers, based on uncoded concepts already clustered, is determined (block 74) as the cosine (cos) σ of the score vectors for the uncoded concepts and cluster being compared. The cos σ provides a measure of relative similarity or dissimilarity between the concepts associated with the
 20 documents and is equivalent to the inner product between the score vectors for the uncoded concept and cluster center.

In the described embodiment, the cos σ is calculated in accordance with the equation:

$$\cos \sigma_{AB} = \frac{\langle \vec{S}_A \cdot \vec{S}_B \rangle}{|\vec{S}_A| |\vec{S}_B|}$$

where $\cos \sigma_{AB}$ comprises the similarity between uncoded concept A and cluster center B , $\vec{S}_A$
 25 comprises a score vector for the uncoded concept A , and $\vec{S}_B$ comprises a score vector for the cluster center B . Other forms of determining similarity using a distance metric are feasible, as would be recognized by one skilled in the art. An example includes using Euclidean distance.

CSCD026-1CA

14

Only those uncoded concepts that are sufficiently distinct from all cluster centers (block 75) are selected as seed concepts for forming new clusters (block 76). If the uncoded concepts being compared are not sufficiently distinct (block 75), each uncoded concept is grouped into a cluster with the most similar cluster center (block 77). Processing continues
5 with the next uncoded concept (block 78).

In the second phase, each uncoded concept not previously placed is iteratively processed in an iterative processing loop (blocks 80-86) as follows. Again, the similarity between each remaining uncoded concept and each cluster center is determined based on a distance (block 81) as the $\cos \sigma$ of the normalized score vectors for the remaining uncoded
10 concept and the cluster center. A best fit between the remaining uncoded concept and one of the cluster centers can be found subject to a minimum fit criterion (block 82). In the described embodiment, a minimum fit criterion of 0.25 is used, although other minimum fit criteria could be used. If a best fit is found (block 83), the remaining uncoded concept is grouped into the cluster having the best fit (block 85). Otherwise, the remaining uncoded
15 concept is grouped into a miscellaneous cluster (block 84). Processing continues with the next remaining uncoded concept (block 86). Finally, a dynamic threshold can be applied to each cluster (block 87) to evaluate and strengthen document membership in a particular cluster. The dynamic threshold is applied based on a cluster-by-cluster basis, as described in commonly-assigned U.S. Patent application Publication No. 2005/0022106, pending. The
20 routine then returns. Other methods and processes for forming clusters are possible.

Alternatively, clusters can be generated by inclusion as further described in commonly-owned U.S. Patent No. 8,700,627, entitled "System and Method for Displaying Relationships Between Concepts to Provide Classification Suggestions via Inclusion," filed July 27, 2010.

25 Once clustered, similar concepts can be identified as described in commonly-assigned U.S. Patent No. 8,645,378, entitled "System and Method for Displaying Relationships Between Electronically Stored Information to Provide Classification Suggestions via Nearest Neighbor," filed July 27, 2010.

30 Once a cluster set is obtained, one or more uncoded concepts within a cluster are selected for comparing to a set of reference concepts to identify a subset of the reference concepts that are similar. The similarity is determined based on a similarity metric, which can include a distance metric. More specifically, the similarity can be determined as the $\cos$

CSCD026-1CA

15

σ of the score vectors for the reference concepts and clusters. Selection of the one or more uncoded concepts can be determined based a cluster measure. FIGURE 5 is a block diagram showing, by way of example, cluster measures 90 for comparing uncoded concepts with and identifying reference concepts for use in the method of FIGURE 2. One or more uncoded

5 concepts in at least one cluster are compared with the reference concepts to identify a subset of the reference concepts that are similar. More specifically, the cluster of the one or more uncoded concepts can be represented by a cluster measure, which is compared with the reference concepts. The cluster measures 90 can include a cluster center 91, sample 92, cluster center and sample 93, and spine 94. Once compared, a similarity threshold is applied

10 to the reference concepts to identify those reference concepts that are most similar.

Identifying similar reference concepts using the cluster center measure 91 includes determining a cluster center for each cluster, comparing at least one cluster center to a set of reference concepts, and identifying the reference concepts that satisfy a threshold similarity with the particular cluster center. Specifically, the score vector for the cluster center is

15 compared to score vectors associated with each reference concept as $\cos \sigma$ of the score vectors for the reference concept and the cluster center. The cluster center score vector is based on all the uncoded concepts in a cluster.

The sample cluster measure 92 includes generating a sample of one or more uncoded concepts in a single cluster that is representative of that cluster. The number of uncoded

20 concepts in the sample can be defined by the reviewer, set as a default, or determined automatically. Once generated, a score vector is calculated for the sample by comparing the score vectors for the individual uncoded concepts selected for inclusion in the sample and identifying the most common documents shared by the selected uncoded concepts. The most common documents and associated weights for the sample are positioned along a score

25 vector, which is representative of those uncoded concepts in that sample. The cluster center and sample cluster measures 93 includes comparing both the cluster center score vector and the sample score vector for a cluster to identify reference concepts that are most similar to the uncoded concepts in the cluster.

Further, similar reference concepts can be identified based on a spine, which includes

30 those clusters that share common documents and are arranged linearly along a vector. The cluster spines are generated as described in commonly-assigned U.S. Patent No. 7,271,804. Also, the cluster spines can be positioned in relation to other cluster spines, as described in

CSCD026-1CA

16

commonly-assigned U.S. Patent application Publication No. 2005/0022106, pending. Organizing the clusters into spines and groups of cluster spines provides an individual reviewer with a display that presents the uncoded concepts and reference concepts according to shared documents while maximizing the number of relationships depicted between the
5 concepts.

The spine cluster measure 94 involves generating a score vector for the spine by comparing the score vectors for the clusters positioned along that spine and identifying the most common documents shared by the clusters. The spine score vector is compared with the score vectors of the reference concepts in the set. Those reference concepts determined to
10 satisfy a threshold of similarity with the spine score vectors are selected for injection into one or more of the clusters along the spine.

For each measures of similarity discussed above, the similarity can be calculated as $\cos \sigma$ of the score vectors for the reference concepts and the selected uncoded concepts. However, other similarity calculations are possible. The similarity calculations can be
15 applied to a threshold and those references concepts with similarity that satisfy the threshold can be selected as the most similar. The most similar reference concepts selected for a cluster can be the same as or different from the most similar reference concepts for the other clusters. Although four types of similarity metrics are described above, other similarity metrics are possible.

20 Upon identification, the similar reference concepts for a cluster are injected into that cluster to provide relationships between the similar reference concepts and uncoded concepts. Identifying the most similar reference concepts and injecting those concepts can occur cluster-by-cluster or for all the clusters simultaneously. The number of similar reference concepts selected for injection can be defined by the reviewer, set as a default, or determined
25 automatically. Other determinations for the number of similar reference concepts are possible. The similar reference concepts can provide hints or suggestions to a reviewer regarding how to classify the uncoded concepts based on the relationships.

The clusters of uncoded concepts and injected reference concepts can be provided as a display to the reviewer. FIGURE 6 is a screenshot 100 showing, by way of example, a visual
30 display 101 of reference concepts 105 in relation to uncoded concepts 104. Clusters 103 can be located along a spine 106, which is a straight vector, based on a similarity of the uncoded concepts 104 in the clusters 103. Each cluster 103 is represented by a circle; however, other

CSCD026-1CA

17

shapes, such as squares, rectangles, and triangles are possible, as described in U.S. Patent No. 6,888,548. The uncoded concepts 104 are each represented by a smaller circle within the clusters, while the reference concepts 105 are each represented by a circle with a diamond inside the boundaries of the circle. The reference concepts can be further represented by their
 5 assigned classification code. Classification codes can include “privileged,” “responsive,” and “non-responsive,” as well as other codes. Each group of reference concepts associated with a particular classification code can be identified by a different color. For instance, “privileged” reference concepts can be colored blue, while “non-responsive” reference concepts are red and “responsive” reference concepts are green. In further embodiment, the reference
 10 concepts with different classification codes can include different symbols. For example, “privileged” reference concepts can be represented by a circle with an “X” in the center, while “non-responsive” reference concepts can include a circle with striped lines and “responsive” reference concepts include a circle with dashed lines. Other classification representations for the reference concepts are possible.

15 The display 101 can be manipulated by an individual reviewer via a compass 102, which enables the reviewer to navigate, explore, and search the clusters 103 and spines 106 appearing within the compass 102, as further described in commonly-assigned U.S. Patent No. 7,356,777. Visually, the compass 172 emphasizes clusters located 103 within the compass 102, while deemphasizing clusters 103 appearing outside of the compass 102.

20 Spine labels 109 appear outside of the compass 102 at the end of each cluster spine 106 to connect the outermost cluster of the cluster spine 106 to the closest point along the periphery of the compass 102. In one embodiment, the spine labels 109 are placed without overlap and circumferentially around the compass 102. Each spine label 109 corresponds to one or more documents represented by the clustered concepts that most closely describe the
 25 cluster spines 106. Additionally, the documents associated with each of the spine labels 109 can appear in a documents list (not shown) also provided in the display. Toolbar buttons 107 located at the top of the display 101 enable a user to execute specific commands for the composition of the spine groups displayed. A set of pull down menus 108 provide further control over the placement and manipulation of clusters 103 and cluster spines 106 within the
 30 display 101. Other types of controls and functions are possible.

A concept guide 110 can be displayed and include a “Selected” field, a “Search Results” field, and detail the numbers of uncoded concepts and reference concepts provided

CSCD026-1CA

18

in the display. The number of uncoded concepts includes all uncoded concepts within a corpus of documents for a review project or within an assignment for the project. The number of reference concepts includes a total number of reference concepts selected for injection into the cluster set. The “Selected” field in the concept guide 110 provides a
 5 number of concepts within one or more clusters selected by the reviewer. The reviewer can select a cluster by “double clicking” the visual representation of that cluster using a mouse. Other options for selecting a cluster are possible. The “Search Results” field provides a number of uncoded concepts and reference concepts that include or match a particular search term identified by the reviewer in a search query box 112.

10 In one embodiment, a garbage can 111 is provided to remove documents from consideration in the current set of clusters 113. Removed cluster documents prevent those documents from affecting future clustering, as may occur when a reviewer considers a document irrelevant to the clusters 113.

The display 111 provides a visual representation of the relationships between
 15 thematically related concepts, including uncoded concepts and injected reference concepts. The uncoded concepts and injected reference concepts located within a cluster or spine can be compared based on characteristics, such as the assigned classification codes of the reference concepts, a number of reference concepts associated with each classification code, and a number of different classification codes to identify relationships between the uncoded
 20 concepts and injected reference concepts. The reviewer can use the displayed relationships as suggestions for classifying the uncoded concepts. For example, FIGURE 7A is a block diagram showing, by way of example, a cluster with “privileged” reference concepts and uncoded concepts. The cluster 130 includes nine uncoded concepts 131 and three reference
 132 concepts. The three reference concepts 132 are classified as “privileged.” Accordingly,
 25 based on the number of “privileged” reference concepts 132 present in the cluster 130, the absence of other classifications of reference concepts, and the thematic relationship between the uncoded concepts 131 and the “privileged” reference concepts 132, the reviewer may be more inclined to review the uncoded concepts in that cluster 131 or to classify one or more of the uncoded concepts as “privileged” without review.

30 Alternatively, the three reference concepts can be classified as “non-responsive,” instead of “privileged” as in the previous example. FIGURE 7B is a block diagram showing, by way of example, a cluster 135 with “non-responsive” reference concepts 136 and uncoded

CSCD026-1CA

19

concepts 131. The cluster includes nine uncoded concepts 131 and three “non-responsive” concepts 136. Since the uncoded concepts 131 in the cluster are thematically related to the “non-responsive” reference concepts 136, the reviewer may wish to assign a “non-responsive” code to the uncoded concepts 131 without review, as they are most likely not relevant to the legal matter associated with the document review project. In making a decision to assign a code, such as “non-responsive,” the reviewer can consider the number of “non-responsive” reference concepts, the absence of other reference concept classification codes, and the thematic relationship between the “non-responsive” reference concepts and the uncoded concepts. Thus, the presence of three “non-responsive” reference concepts 136 in the cluster of uncoded concepts provides a suggestion that the uncoded concepts 131 may also be “non-responsive.” Further, the label associated with the spine upon which the cluster is located can be used to influence a suggestion.

A further example can include a combination of “privileged” and “non-responsive” reference concepts. For example, FIGURE 7C is a block diagram showing, by way of example, a cluster 140 with uncoded concepts and a combination of classified reference concepts. The cluster 140 can include one “privileged” reference concept 132, two “non-responsive” concepts 136, and nine uncoded concepts 131. The “privileged” and “non-responsive” reference concepts can be distinguished by different colors, shapes, or symbols, as well as by other identifiers. The combination of “privileged” 132 and “non-responsive” 136 reference concepts within the cluster 140 can suggest to a reviewer that the uncoded reference concepts 131 should be reviewed before classification or that one or more of the uncoded reference concepts 131 should be classified as “non-responsive” based on the higher number of “non-responsive” reference concepts 136. In making a classification decision, the reviewer may consider the number of “privileged” reference concepts 132 versus the number of “non-responsive” reference concepts 136, as well as the thematic relationships between the uncoded concepts 131 and the “privileged” 132 and “non-responsive” 136 reference concepts. Additionally, the reviewer can identify the closest reference concept to an uncoded concept and assign the classification code of the closest reference concept to the uncoded concept. Other examples, classification codes, and combinations of classification codes are possible.

Additionally, the reference concepts can also provide suggestions for classifying clusters and spines. The suggestions provided for classifying a cluster can include factors,

CSCD026-1CA

20

such as a presence or absence of classified concepts with different classification codes within the cluster and a quantity of the classified concepts associated with each classification code in the cluster. The classified concepts can include reference concepts and newly classified uncoded concepts. The classification code assigned to the cluster is representative of the concepts in that cluster and can be the same as or different from one or more classified concepts within the cluster. Further, the suggestions provided for classifying a spine include factors, such as a presence or absence of classified concepts with different classification codes within the clusters located along the spine and a quantity of the classified concepts for each classification code. Other suggestions for classifying concepts, clusters, and spines are possible.

The display of relationships between the uncoded concepts and reference concepts provides classification suggestions to an individual reviewer. The suggestions can indicate a need for manual review of the uncoded concepts, when review may be unnecessary, and hints for classifying the uncoded concepts. Additional information can be provided to assist the reviewer in making classification decisions for the uncoded concepts, such as a machine-generated confidence level associated with a suggested classification code, as described in commonly-assigned U.S. Patent Application Serial No. 12/844,785, entitled "System and Method for Providing a Classification Suggestion for Concepts," filed July 27, 2010, pending.

The machine-generated suggestion for classification and associated confidence level can be determined by a classifier. FIGURE 8 is a process flow diagram 150 showing, by way of example, a method for classifying uncoded concepts using a classifier for use in the method of FIGURE 2. An uncoded concept is selected from a cluster within a cluster set (block 151) and compared to a neighborhood of x -reference concepts (block 152), also located within the cluster, to identify those reference concepts in the neighborhood that are most relevant to the selected uncoded concept. Alternatively, a cluster or spine can be selected and compared to a neighborhood of x -reference concepts determined for the selected cluster or spine, as discussed below. In a further embodiment, a machine-generated suggestion for classification and an associated confidence level can be provided for a cluster or spine by selecting and comparing the cluster or spine to a neighborhood of x -reference concepts determined for the selected cluster or spine, as further discussed below.

CSCD026-1CA

The neighborhood of x -reference concepts is determined separately for each selected uncoded concept and can include one or more injected reference concepts within that cluster. During neighborhood generation, the x -number of reference concepts in a neighborhood can first be determined automatically or by an individual reviewer. Next, the x -number of reference concepts nearest in distance to the selected uncoded concept is identified. Finally, the identified x -number of reference concepts are provided as the neighborhood for the selected uncoded concept. In a further embodiment, the x -number of reference concepts are defined for each classification code. Once generated, the x -number of reference concepts in the neighborhood and the selected uncoded concept are analyzed by the classifier to provide a machine-generated classification suggestion (block 153). A confidence level for the suggested classification is also provided (block 154).

The analysis of the selected uncoded concept and x -number of reference concepts can be based on one or more routines performed by the classifier, such as a nearest neighbor (NN) classifier. The routines for determining a suggested classification code for an uncoded concept include a minimum distance classification measure, also known as closest neighbor, minimum average distance classification measure, maximum count classification measure, and distance weighted maximum count classification measure. The minimum distance classification measure includes identifying a neighbor that is the closest distance to the selected uncoded concept and assigning the classification code of the closest neighbor as the suggested classification code for the selected uncoded concept. The closest neighbor is determined by comparing score vectors for the selected uncoded concept with each of the x -number reference concepts in the neighborhood as the $\cos \sigma$ to determine a distance metric. The distance metrics for the x -number of reference concepts are compared to identify the reference document closest to the selected uncoded concept as the closest neighbor.

The minimum average distance classification measure includes calculating an average distance of the reference concepts in a cluster for each classification code. The classification code of the reference concepts having the closest average distance to the selected uncoded concept is assigned as the suggested classification code. The maximum count classification measure, also known as the voting classification measure, includes counting a number of reference concepts within the cluster for each classification code and assigning a count or “vote” to the reference concepts based on the assigned classification code. The classification code with the highest number of reference concepts or “votes” is assigned to the selected

CSCD026-1CA

22

uncoded concept as the suggested classification. The distance weighted maximum count classification measure includes identifying a count of all reference concepts within the cluster for each classification code and determining a distance between the selected uncoded concept and each of the reference concepts. Each count assigned to the reference concepts is

5 weighted based on the distance of the reference concept from the selected uncoded concept. The classification code with the highest count, after consideration of the weight, is assigned to the selected uncoded concept as the suggested classification.

The x -NN classifier provides the machine-generated classification code with a confidence level that can be presented as an absolute value or percentage. Other confidence

10 level measures are possible. The reviewer can use the suggested classification code and confidence level to assign a classification to the selected uncoded concept. Alternatively, the x -NN classifier can automatically assign the suggested classification. In one embodiment, the x -NN classifier only assigns a suggested classification to an uncoded concept if the confidence level is above a threshold value, which can be set by the reviewer or the x -NN

15 classifier.

As briefly described above, classification can also occur on a cluster or spine level. For instance, for cluster classification, a cluster is selected and a score vector for the center of the cluster is determined as described above with reference to FIGURE 4. A neighborhood for the selected cluster is determined based on a distance metric. The x -number of reference

20 concepts that are closest to the cluster center can be selected for inclusion in the neighborhood, as described above. Each reference concept in the selected cluster is associated with a score vector and the distance is determined by comparing the score vector of the cluster center with the score vector of each reference concept to determine an x -number of reference concepts that are closest to the cluster center. However, other methods for

25 generating a neighborhood are possible. Once determined, one of the classification measures is applied to the neighborhood to determine a suggested classification code and confidence level for the selected cluster.

Throughout the process of identifying similar reference concepts and injecting the reference concepts into a cluster to provide a classification suggestion, the reviewer can retain

30 control over many aspects, such as a source of the reference concepts and a number of similar reference concepts to be selected. FIGURE 9 is a screenshot 160 showing, by way of example, a reference options dialogue box 161 for entering user preferences for reference

CSCD026-1CA

23

concept injection. The dialogue box 161 can be accessed via a pull-down menu as described above with respect to FIGURE 6. Within the dialogue box 161, the reviewer can utilize user-selectable parameters to define a source of reference concepts 162, filter the reference concept by category 163, select a target for the reference concepts 164, select an action to be performed upon the reference. concepts 165, define timing of the injection 166, define a count of similar reference concepts to be injected into a cluster 167, select a location for injection within a cluster 168, and compile a list of injection commands 169. Each user-selectable option can include a text box for entry of a user preference or a drop-down menu with predetermined options for selection by a reviewer. Other user-selectable options and displays are possible.

The reference source parameter 162 allows the reviewer to identify one or more sources of the reference concepts. The sources can include all previously classified reference concepts in a document review project, all reference concepts for which the associated classification has been verified, all reference concepts that have been analyzed, or all reference concepts in a particular binder. The binder can include categories of reference concepts, such as reference concepts that are particular to the document review project or that are related to a prior document review project. The category filter parameter 163 allows the reviewer to generate and display the set of reference concepts using only those reference concepts associated with a particular classification code. The target parameter 164 allows the reviewer to select a target for injection of the similar reference concepts. Options available for the target parameter 164 can include an assignment, all clusters, select clusters, all spines, select spines, all concepts, and select concepts. The assignment can be represented as a cluster set; however, other representations are possible, including a file hierarchy and a list of documents, such as an email folder, as described in commonly-assigned U.S. Patent No. 7,404,151.

The action parameter 165 allows the reviewer to define display options for the injected reference concepts. The display options can include injecting the similar reference concepts into a map display of the clusters, displaying the similar reference concepts in the map until reclustering occurs, displaying the injected reference concepts in the map, and not displaying the injected reference concepts in the map. Using the automatic parameter 166, the reviewer can define a time for injection of the similar reference concepts. The timing options can include injecting the similar reference concepts upon opening of an assignment,

CSCD026-1CA

24

upon reclustering, or upon changing the selection of the target. The reviewer can specify a threshold number of similar reference concepts to be injected in each cluster or spine via the similarity option 167. The number selected by a reviewer is an upper threshold since a lesser number of similar reference concepts may be identified for injecting into a cluster or spine.

- 5 Additionally, the reviewer can use the similarity option to 167 set a value for determining whether a reference document is sufficiently similar to the uncoded concepts.

Further, the reviewer can select a location within the cluster for injection of the similar reference concepts via the cluster site parameter 168. Options for cluster site injection can include the cluster centroid. Other cluster sites are possible. The user-
10 selectable options for each preference can be compiled as a list of injection commands 169 for use in the injection process. Other user selectable parameters, options, and actions are possible.

In a further embodiment, once the uncoded concepts are assigned a classification code, the newly-classified uncoded concepts can be placed into the concept reference set for
15 use in providing classification suggestions for other uncoded concepts.

In yet a further embodiment, each document can be represented by more than one concept. Accordingly, to determine a classification code for the document, the classification codes for each of the associated concepts can be analyzed and compared for consideration in classifying the document. In one example, a classification code can be determined by
20 counting the number of associated concepts for each classification code and then assigned the classification code with the most associated concepts. In a further example, one or more of the associated concepts can be weighted and the classification code associated with the highest weight of concepts is assigned. Other methods for determining a classification code for uncoded documents based on reference concepts are possible.

25 Although clustering and displaying relationships has been described above with reference to concepts, other tokens, such as word-level or character-level n-grams, raw terms, and entities, are possible.

CSCD026-1CA

25

What is claimed is:

1. A method (50) for displaying relationships between concepts to provide classification suggestions via injection, comprising the steps of:
 - designating a reference set (14d) comprising concepts (14d) each associated with a classification code, wherein each concept (14d) comprises nouns and noun phrases with common semantic meaning that are extracted from a set of documents 14a, 17, 20, 26, 29;
 - designating clusters (130) of uncoded concepts (14c);
 - comparing one or more of the uncoded concepts (14c) from at least one cluster to the reference concepts (14d) in the set (14d);
 - identifying at least one of the reference concepts (14d) that is similar to the one or more uncoded concepts (14c);
 - injecting the similar reference concepts (14d) into the at least one cluster (130);
 - visually depicting relationships between the uncoded concepts (14c) and the similar reference concepts (14d) in the at least one cluster (130) as suggestions for classifying the uncoded concepts (14c); and
 - determining a classification code for at least one of the uncoded concepts (14c) in the at least one cluster (130), comprising:
 - comparing the at least one uncoded concept (14c) to a neighborhood of reference concepts (14d) in the at least one cluster (130), wherein the neighborhood comprises a predetermined number of the reference concepts (14d) that have a closest distance to the uncoded concept (14c); and
 - selecting the classification code based on the comparison of the at least one uncoded concept (14c) and the reference concept neighborhood, wherein the steps are performed by a suitably programmed computer.
2. A method (50) according to Claim 1, further comprising:
 - determining the similar reference concepts (14d), comprising:
 - calculating a similarity between each of the one or more uncoded concepts (14c) and each of the reference concepts (14d);

CSCD026-1CA

26

applying a predetermined threshold to the calculated similarities; and

selecting the at least one reference concept (14d) as the similar reference concept when the associated similarity satisfies the similarity threshold.

3. A method (50) according to Claim 1, wherein the comparison between the one or more of the uncoded concepts (14c) and the reference concepts (14d) comprise at least one of:

determining a center of the at least one cluster (130) based on the one or more uncoded concepts (14c) and comparing the cluster (130) center with the reference concepts (14d);

selecting a sample comprising the one or more uncoded concepts (14c) and comparing the sample with the reference concepts (14d), wherein the sample can be selected via at least one of manually and automatically; and

comparing the cluster (130) center and the sample for the at least one cluster (130) with the reference concepts (14d).

4. A method (50) according to Claim 1, further comprising:

automatically classifying the at least one uncoded concept (14c) by assigning the classification code to that uncoded concept (14c).

5. A method (50) according to Claim 4, further comprising:

identifying the documents associated with the classified uncoded concept (14c); and

classifying the associated documents by assigning a classification code to each of the documents.

6. A method (50) according to Claim 5, wherein the documents are identified using a matrix comprising a mapping of concepts (14d) and related documents.

CSCD026-1CA

27

7. A method (50) according to Claim 1, further comprising:
providing the classification code as a suggestion for classifying the at least one uncoded concept (14c) in the at least one cluster (130).
8. A method (50) according to Claim 7, further comprising:
determining a confidence level for the classification code; and
providing the confidence level with the suggested classification code.
9. A method (50) according to Claim 8, wherein the classification code is provided when the confidence level exceeds a predetermined threshold.
10. A system (10) for displaying relationships between concepts (14d) to provide classification suggestions via injection, comprising:
 - a reference set (14d) comprising concepts (14d) each associated with a classification code, wherein each concept comprises nouns and noun phrases with common semantic meaning that are extracted from a set of documents;
 - a clustering module (33) to designate clusters (130) of uncoded concepts (14c);
 - a similarity module (34) to compare one or more of the uncoded concepts (14c) from at least one cluster (130) to the reference concepts (14d) in the set and to identify at least one of the reference concepts (14d) that is similar to the one or more uncoded concepts (14c);
 - an injection module to inject the similar reference concepts (14d) into the at least one cluster (130);
 - a display (37) to visually depict relationships between the uncoded concepts (14c) and the similar reference concepts (14d) in the at least one cluster (130) as suggestions for classifying the uncoded concepts (14c);
 - a classification module (35) to determine a classification code for at least one of the uncoded concepts (14c) in the at least one cluster (130), comprising:
 - a comparison module to compare the at least one uncoded concept (14c) to a neighborhood of reference concepts (14d) in the at least one

CSCD026-1CA

28

cluster (130), wherein the neighborhood comprises a predetermined number of the reference concepts (14d) that have a closest distance to the uncoded concept (14c); and

a selection module to select the classification code based on the comparison of the at least one uncoded concept (14c) and the reference concept neighborhood; and

a processor to execute the modules.

11. A system (10) according to Claim 10, wherein the similarity module (34) comprises:

a determination module to calculate a similarity between each of the one or more uncoded concepts (14c) and each of the reference concepts (14d);

a threshold module to apply a predetermined threshold to the calculated similarities; and

a selection module to select the at least one reference concept as the similar reference concept when the associated similarity satisfies the similarity threshold.

12. A system (10) according to Claim 10, wherein the similarity module (34) comprises at least one of:

a cluster center module to determine a center of the at least one cluster (130) based on the one or more uncoded concepts (14c) and comparing the cluster (130) center with the reference concepts (14d);

a sample module to select a sample comprising the one or more uncoded concepts (14c) and to compare the sample with the reference concepts (14d), wherein the sample can be selected via at least one of manually and automatically; and

a center sample module to compare the cluster (130) center and the sample for the at least one cluster (130) with the reference concepts (14d).

CSCD026-1CA

29

13. A system (10) according to Claim 10, wherein the classification module (34) automatically classifies the at least one uncoded concept (14c) by assigning the classification code to that uncoded concept (14c).
14. A system (10) according to Claim 13, wherein the classification module (34) identifies the documents associated with the classified uncoded concept (14c) and classifies the associated documents by assigning a classification code to each of the documents.
15. A system (10) according to Claim 14, wherein the documents (130) are identified using a matrix (60) comprising a mapping of concepts (14d) and related documents.
16. A system (10) according to Claim 11, wherein the classification module (34) provides the classification code as a suggestion for classifying the at least one uncoded concept (14c) in the at least one cluster (130).
17. A system (10) according to Claim 16, wherein the classification module (34) determines a confidence level for the classification code and provides the confidence level with the suggested classification code.
18. A system (10) according to Claim 17, wherein the classification code is provided when the confidence level exceeds a predetermined threshold.
19. A method (50) for displaying relationships between concepts to provide classification suggestions via injection, comprising the steps of:
 - designating a reference set (14d) comprising concepts (14d) each associated with a classification code, wherein each concept (14d) comprises nouns and noun phrases with common semantic meaning that are extracted from a set of documents 14a, 17, 20, 26, 29;
 - designating clusters (130) of uncoded concepts (14c);

CSCD026-1CA

30

comparing one or more of the uncoded concepts (14c) from at least one cluster to the reference concepts (14d) in the set (14d);

identifying at least one of the reference concepts (14d) that is similar to the one or more uncoded concepts (14c) and determining the similar reference concepts (14d), comprising:

calculating a similarity between each of the one or more uncoded concepts (14c) and each of the reference concepts (14d);

applying a predetermined threshold to the calculated similarities; and

selecting the at least one reference concept (14d) as the similar reference concept when the associated similarity satisfies the similarity threshold;

injecting the similar reference concepts (14d) into the at least one cluster (130); and

visually depicting relationships between the uncoded concepts (14c) and the similar reference concepts (14d) in the at least one cluster (130) as suggestions for classifying the uncoded concepts (14c),

wherein the steps are performed by a suitably programmed computer.

20. A system (10) for displaying relationships between concepts (14d) to provide classification suggestions via injection, comprising:

a reference set (14d) comprising concepts (14d) each associated with a classification code, wherein each concept comprises nouns and noun phrases with common semantic meaning that are extracted from a set of documents;

a clustering module (33) to designate clusters (130) of uncoded concepts (14c);

a similarity module (34) to compare one or more of the uncoded concepts (14c) from at least one cluster (130) to the reference concepts (14d) in the set and to identify at least one of the reference concepts (14d) that is similar to the one or more uncoded concepts (14c), wherein the similarity module (34) comprises:

CSCD026-1CA

31

a determination module to calculate a similarity between each of the one or more uncoded concepts (14c) and each of the reference concepts (14d);

a threshold module to apply a predetermined threshold to the calculated similarities; and

a selection module to select the at least one reference concept as the similar reference concept when the associated similarity satisfies the similarity threshold;

an injection module to inject the similar reference concepts (14d) into the at least one cluster (130);

a display (37) to visually depict relationships between the uncoded concepts (14c) and the similar reference concepts (14d) in the at least one cluster (130) as suggestions for classifying the uncoded concepts (14c); and

a processor to execute the modules.

21. A method (50) for displaying relationships between concepts to provide classification suggestions via injection, comprising the steps of:

designating a reference set (14d) comprising concepts (14d) each associated with a classification code, wherein each concept (14d) comprises nouns and noun phrases with common semantic meaning that are extracted from a set of documents 14a, 17, 20, 26, 29;

designating clusters (130) of uncoded concepts (14c);

comparing one or more of the uncoded concepts (14c) from at least one cluster to the reference concepts (14d) in the set (14d), wherein the comparison between the one or more of the uncoded concepts (14c) and the reference concepts (14d) comprise at least one of:

determining a center of the at least one cluster (130) based on the one or more uncoded concepts (14c) and comparing the cluster (130) center with the reference concepts (14d);

selecting a sample comprising the one or more uncoded concepts (14c) and comparing the sample with the reference concepts (14d),

CSCD026-1CA

32

wherein the sample can be selected via at least one of manually and automatically; and

comparing the cluster (130) center and the sample for the at least one cluster (130) with the reference concepts (14d);

identifying at least one of the reference concepts (14d) that is similar to the one or more uncoded concepts (14c);

injecting the similar reference concepts (14d) into the at least one cluster (130); and

visually depicting relationships between the uncoded concepts (14c) and the similar reference concepts (14d) in the at least one cluster (130) as suggestions for classifying the uncoded concepts (14c),

wherein the steps are performed by a suitably programmed computer.

22. A system (10) for displaying relationships between concepts (14d) to provide classification suggestions via injection, comprising:

a reference set (14d) comprising concepts (14d) each associated with a classification code, wherein each concept comprises nouns and noun phrases with common semantic meaning that are extracted from a set of documents;

a clustering module (33) to designate clusters (130) of uncoded concepts (14c);

a similarity module (34) to compare one or more of the uncoded concepts (14c) from at least one cluster (130) to the reference concepts (14d) in the set and to identify at least one of the reference concepts (14d) that is similar to the one or more uncoded concepts (14c), wherein the similarity module (34) comprises at least one of:

a cluster center module to determine a center of the at least one cluster (130) based on the one or more uncoded concepts (14c) and comparing the cluster (130) center with the reference concepts (14d);

a sample module to select a sample comprising the one or more uncoded concepts (14c) and to compare the sample with the reference concepts (14d), wherein the sample can be selected via at least one of manually and automatically; and

CSCD026-1CA

33

a center sample module to compare the cluster (130) center and the sample for the at least one cluster (130) with the reference concepts (14d);

an injection module to inject the similar reference concepts (14d) into the at least one cluster (130);

a display (37) to visually depict relationships between the uncoded concepts (14c) and the similar reference concepts (14d) in the at least one cluster (130) as suggestions for classifying the uncoded concepts (14c); and

a processor to execute the modules.

1/9

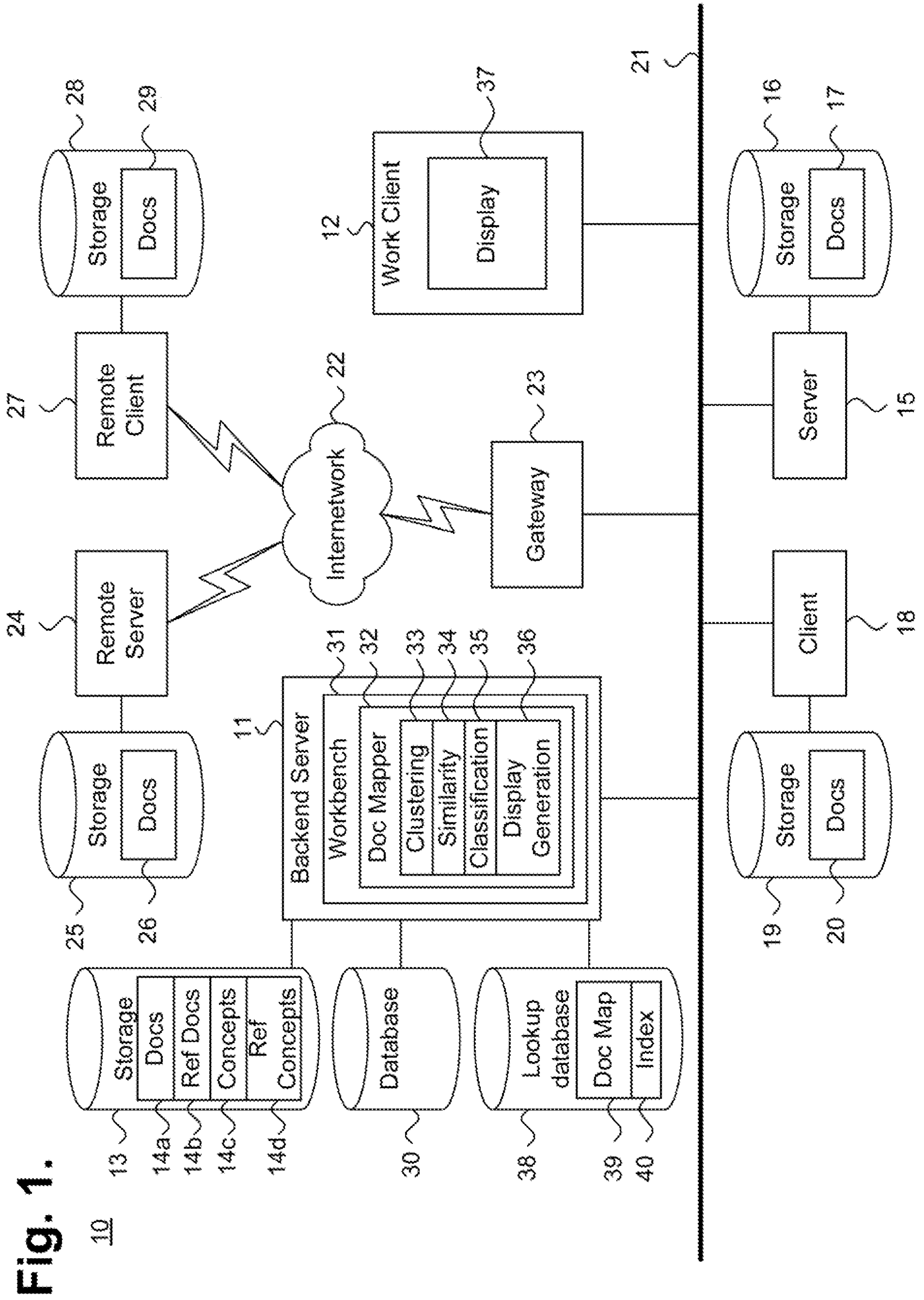


Fig. 2.

2/9

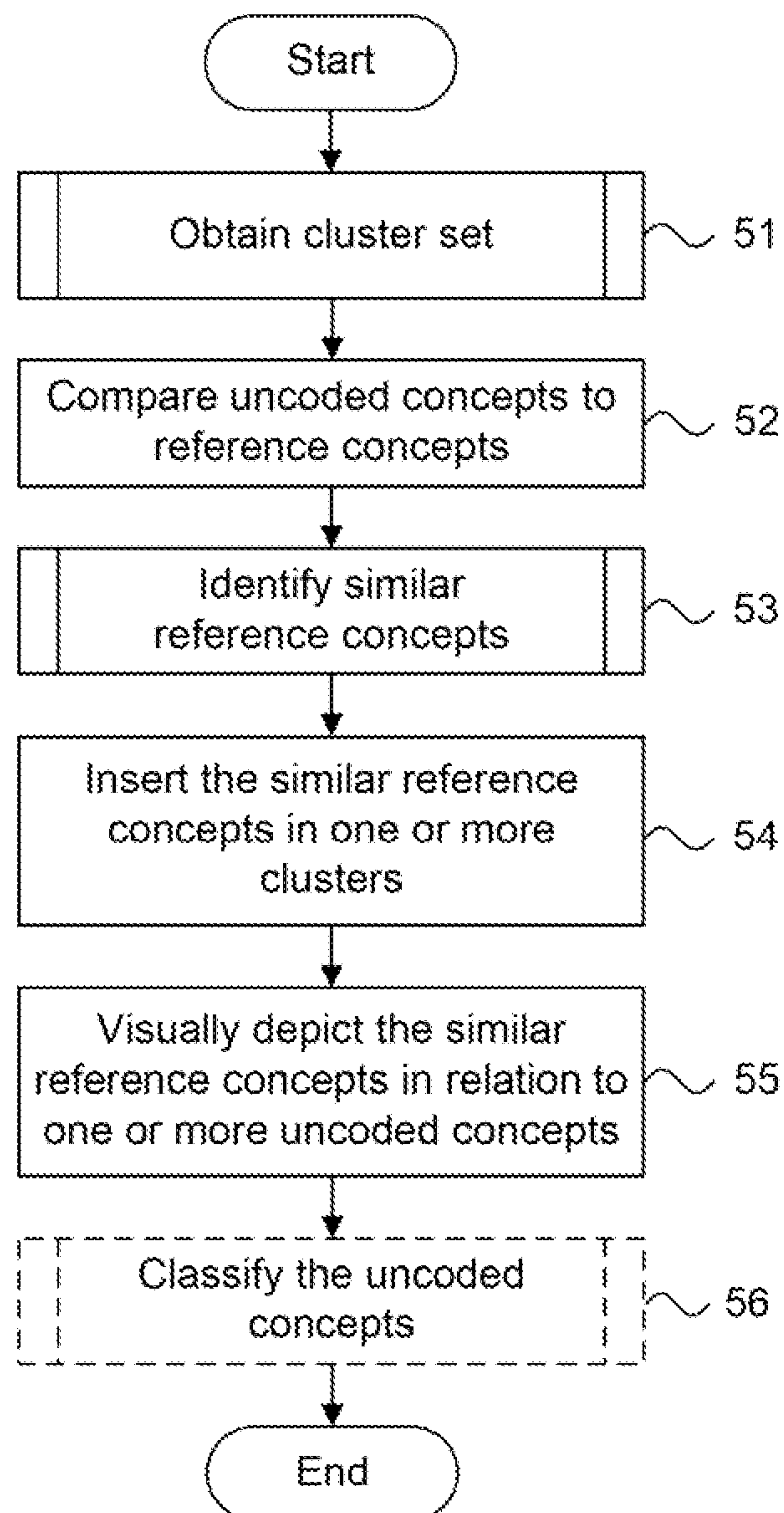
50

Fig. 3.

3/9

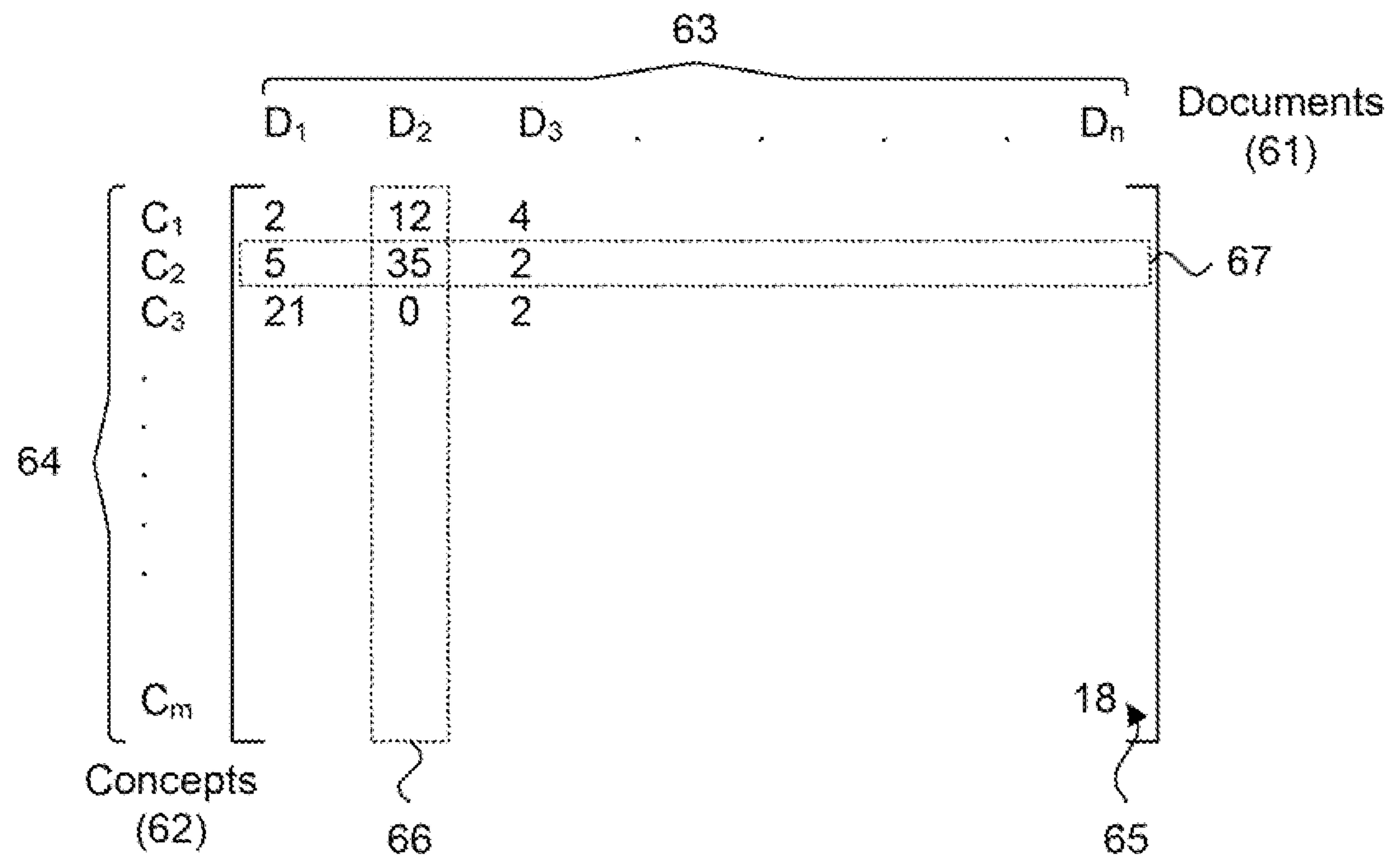
60

Fig. 4.

4/9

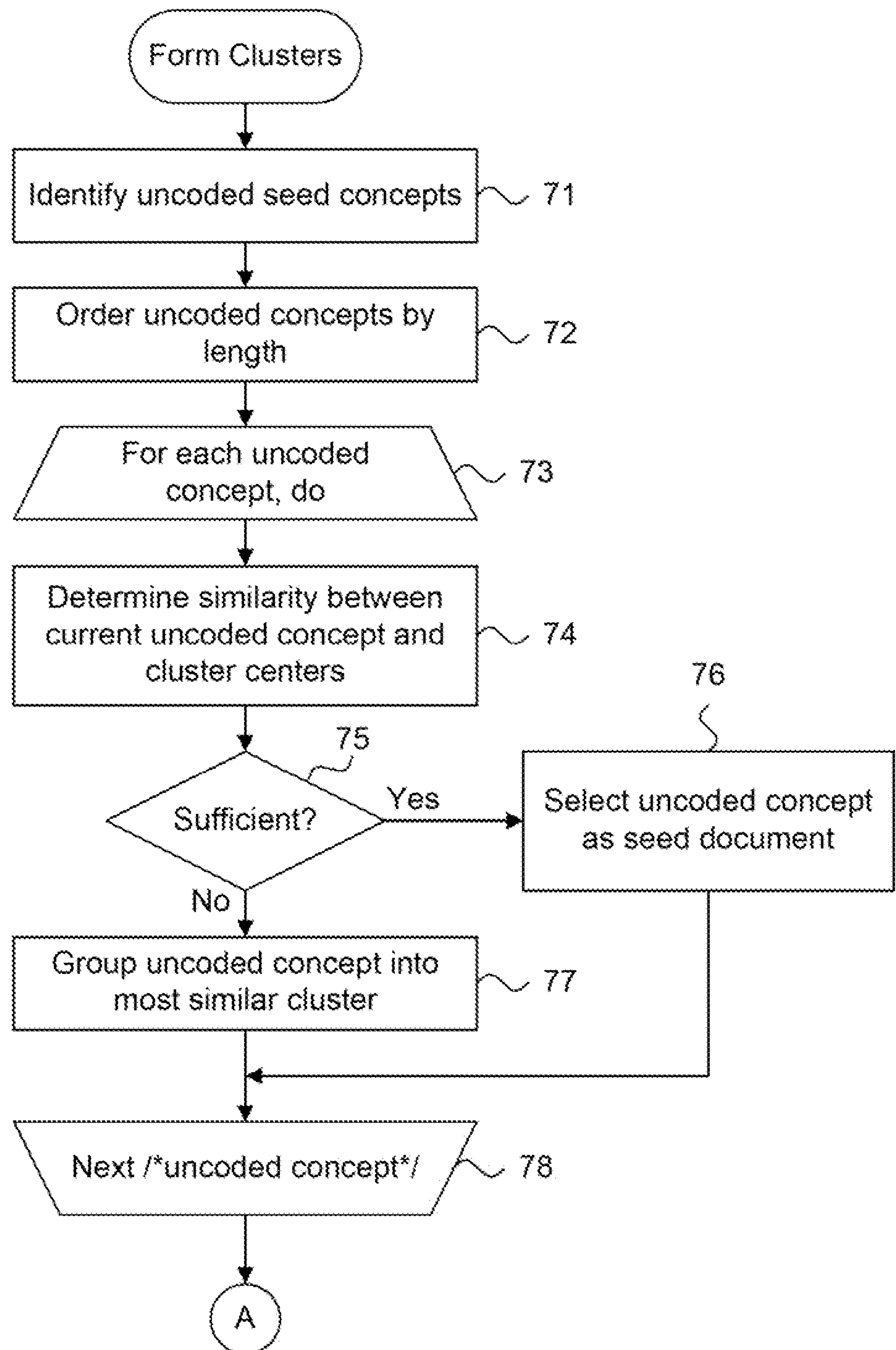
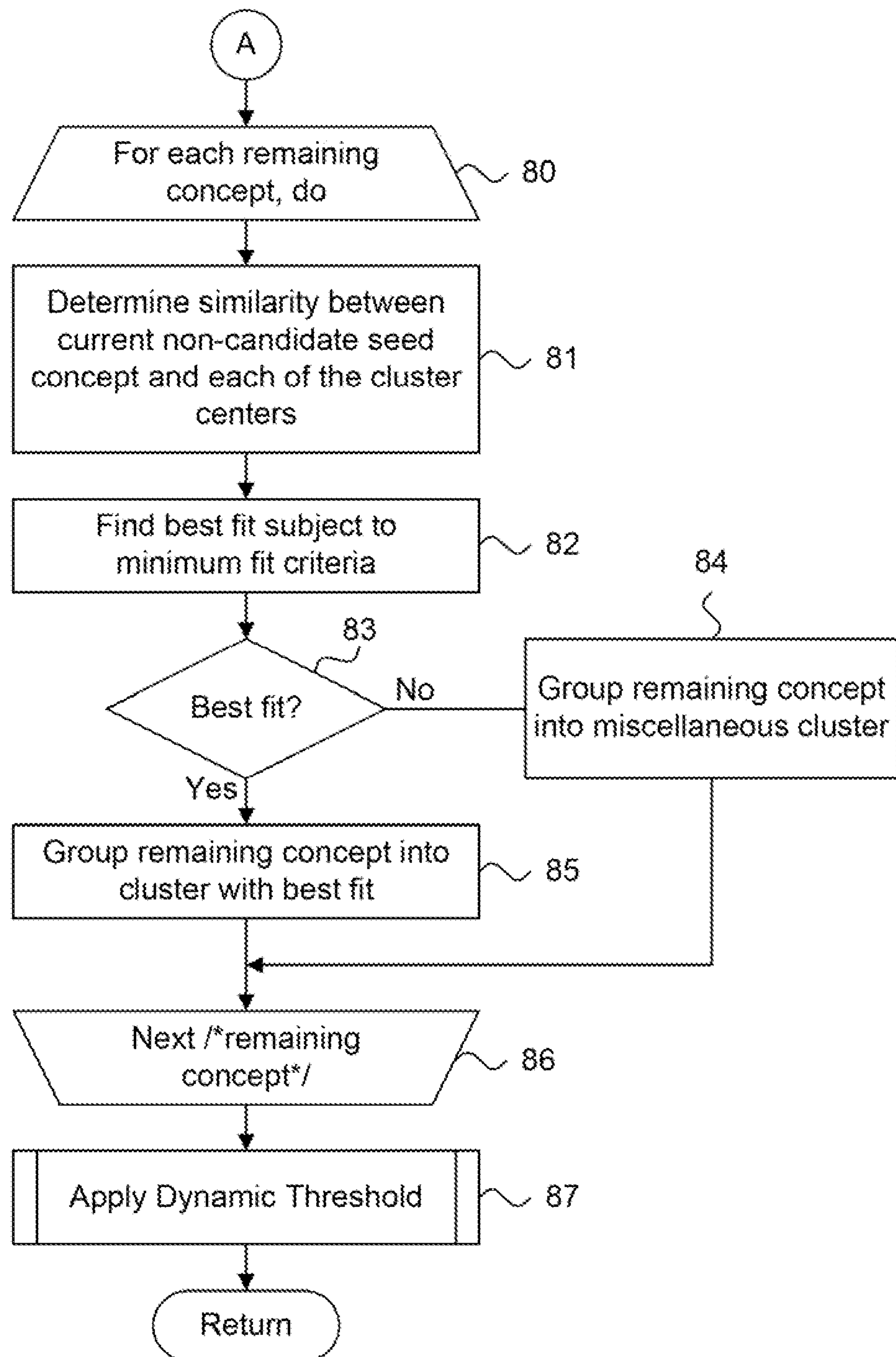
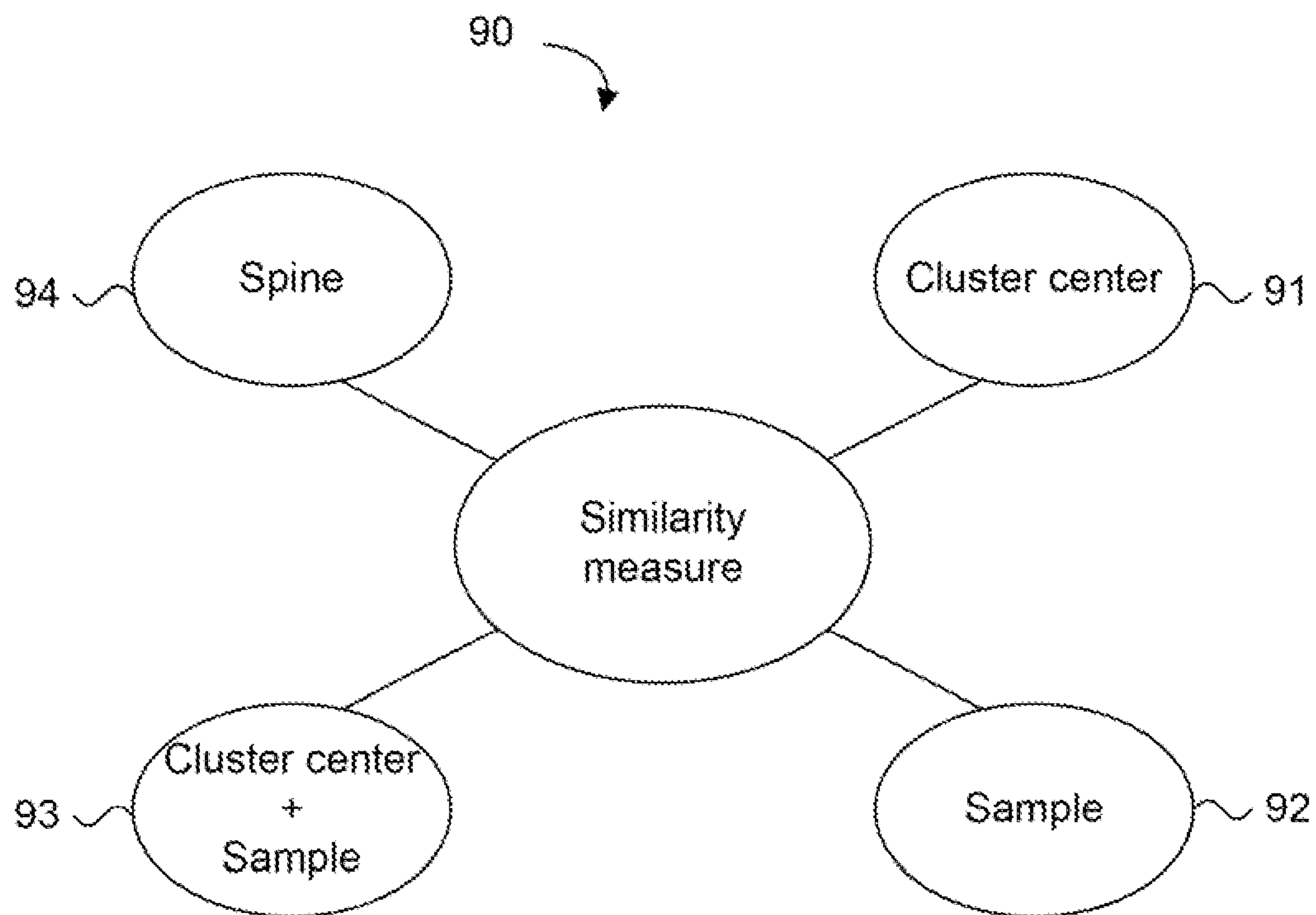
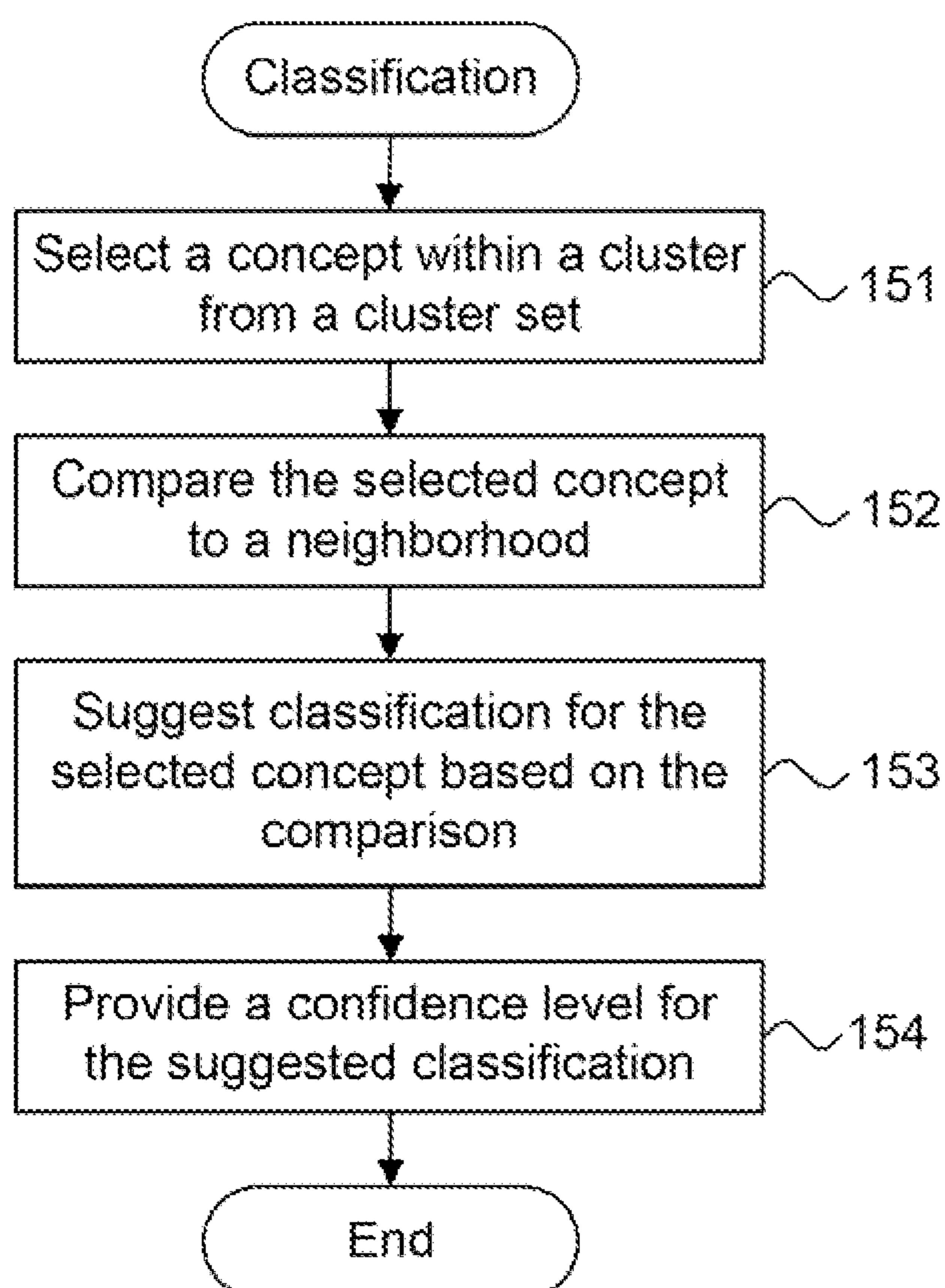
70

Fig. 4 (Cont).

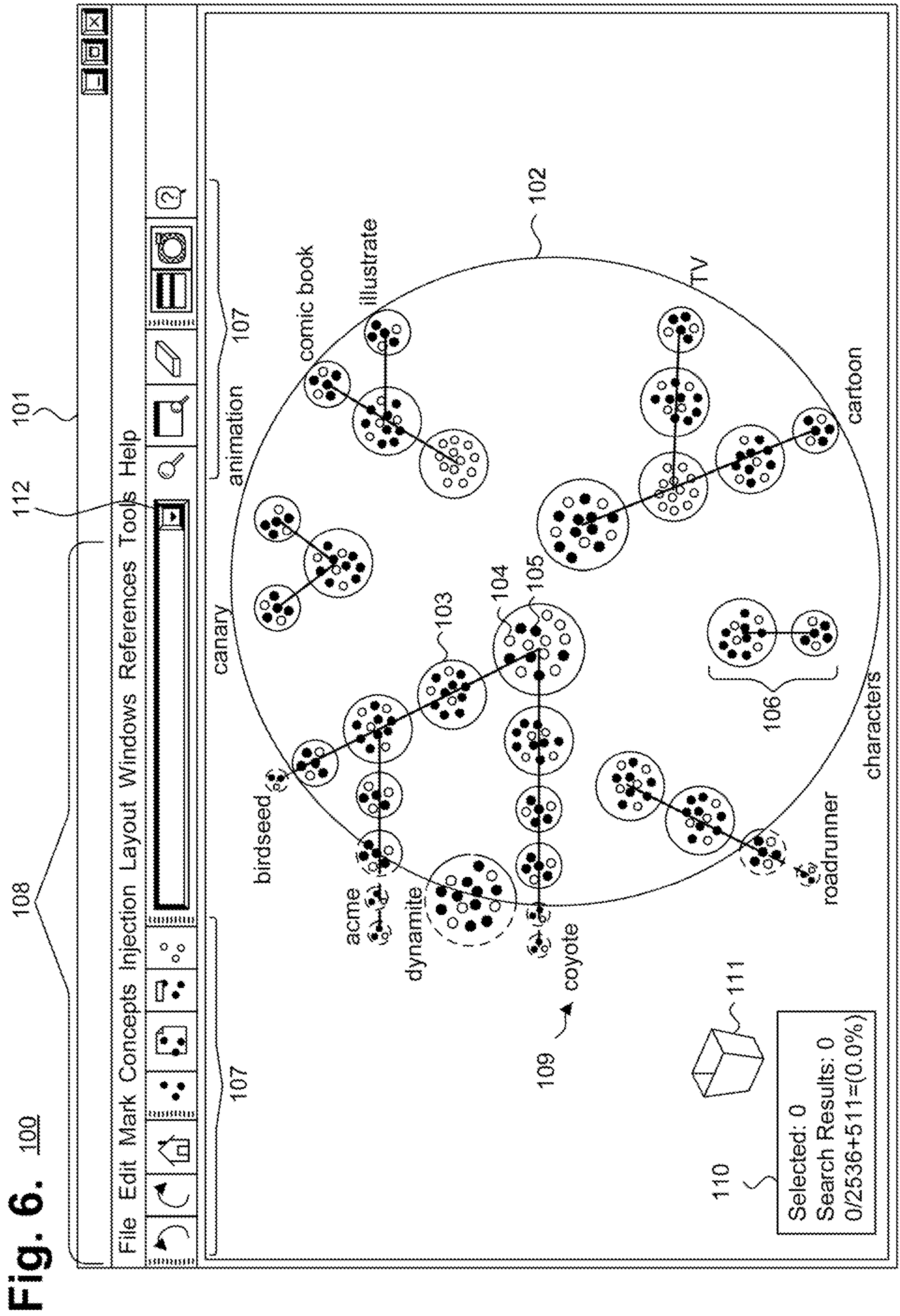
5/9



6/9

Fig. 5.**Fig. 8.**150

7/9



8/9

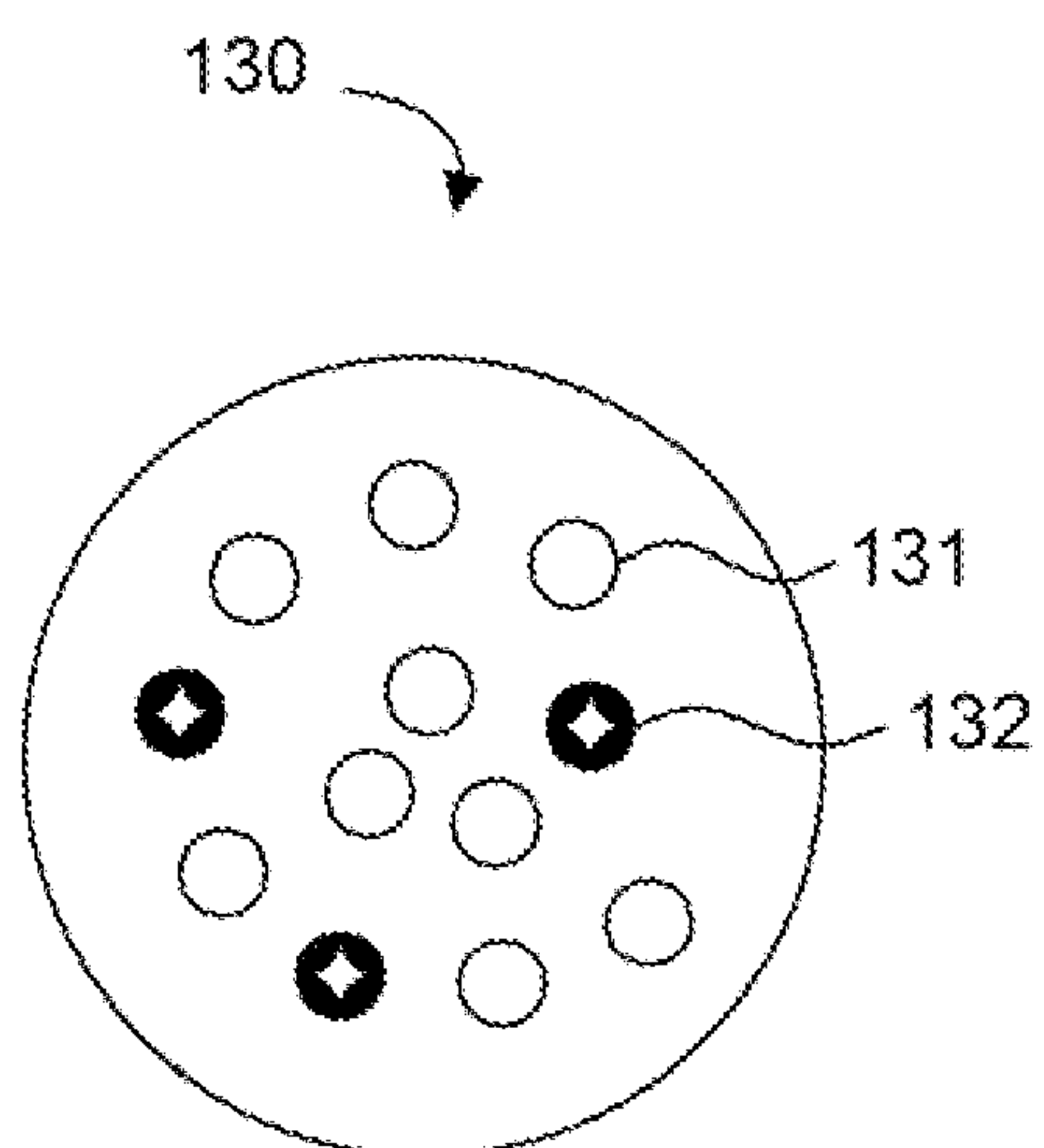
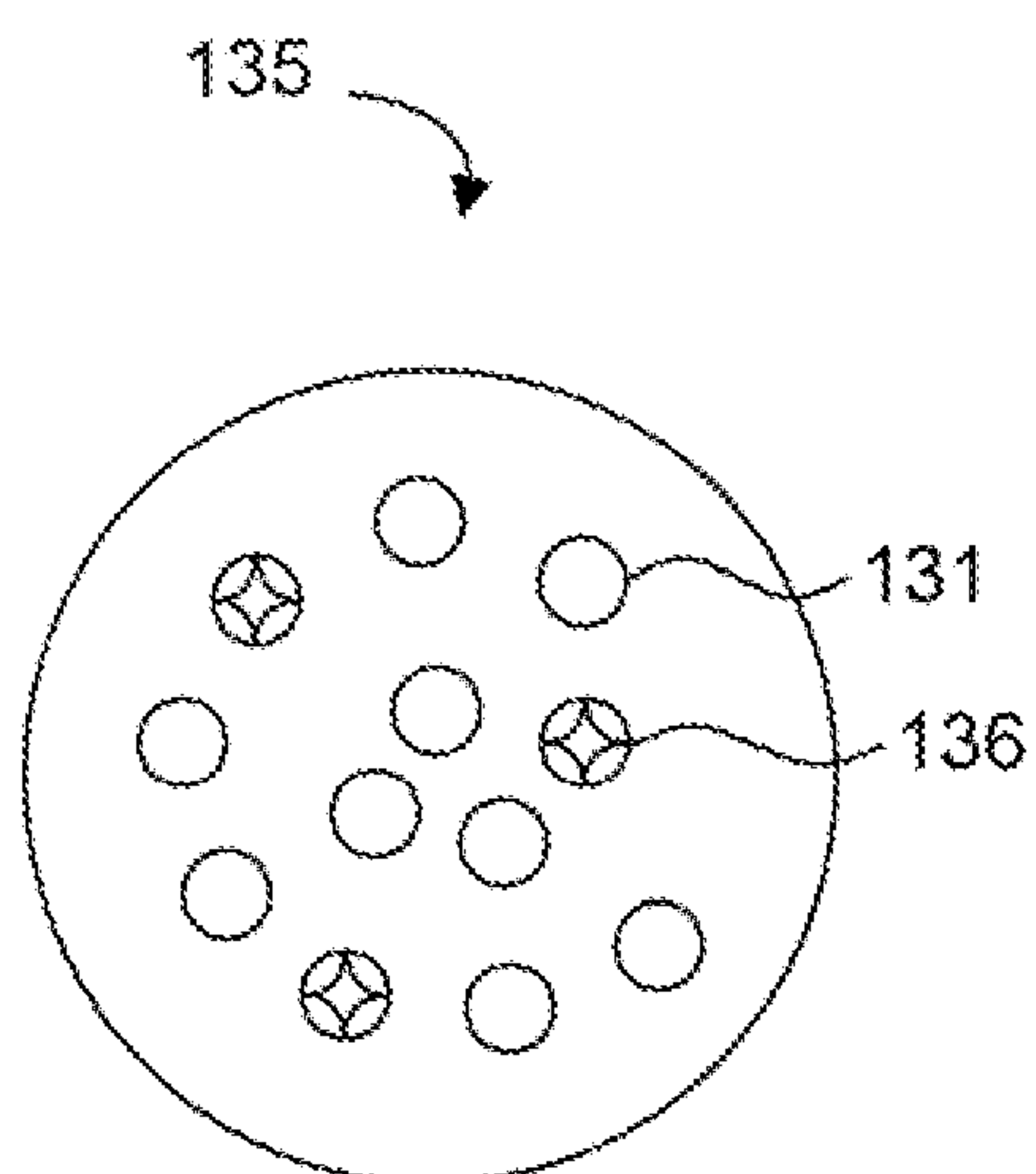
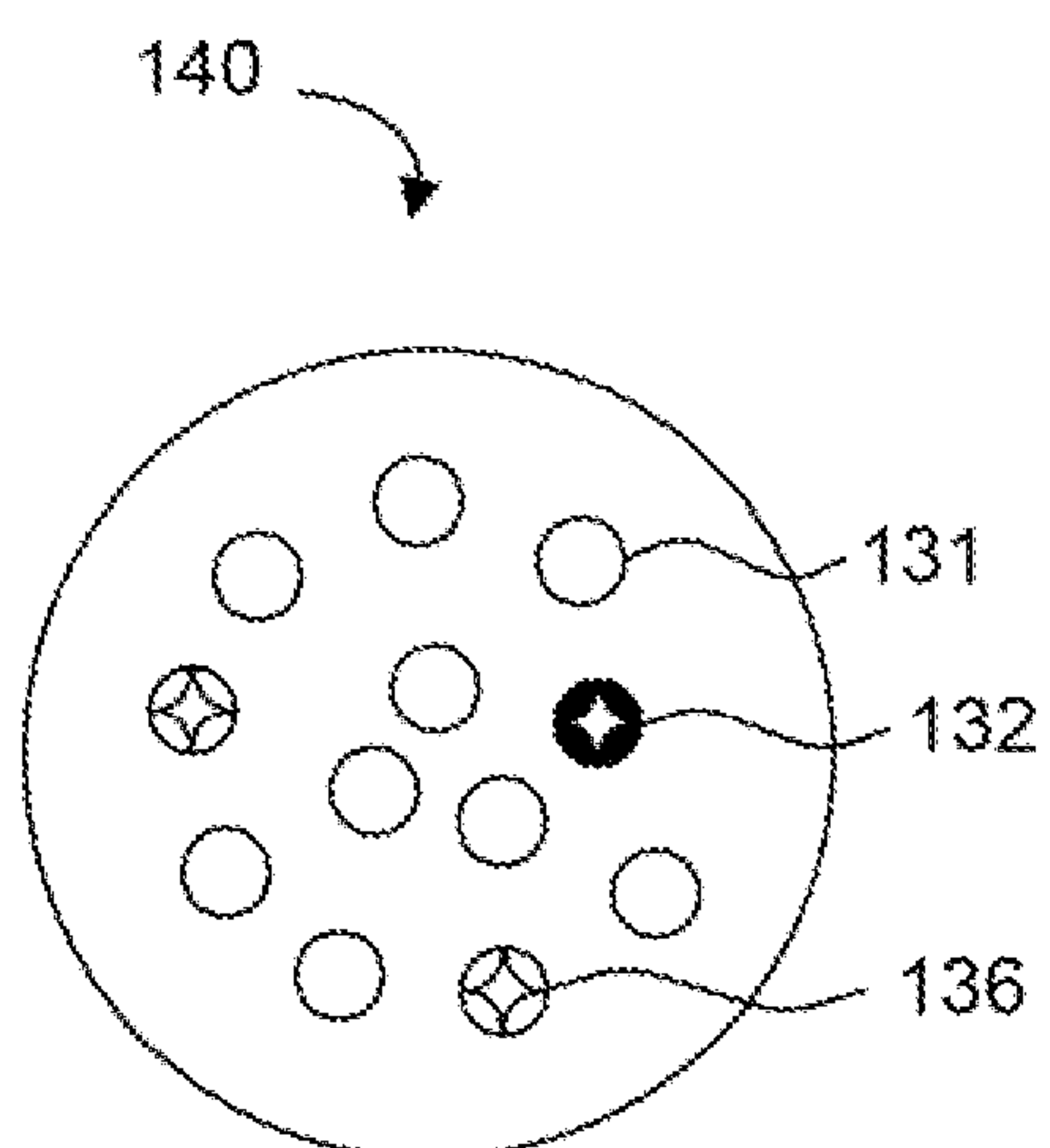
Fig. 7A.**Fig. 7B.****Fig. 7C.**

Fig. 9.

9/9

160

Inject...

162 Reference Source

Use Marked Documents in Matter ▼

☐ Some Training Set Binder

Edit Binders...

163 Reference Filter

☒ Filter by Category ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒ ☒

☐ Filter by Criteria

164 Command Details

Target All Clusters ▼

165 Action Insert In Map until Recluster ▼

Count 0 + 0 %

☒ Automatic On Assignment Open ▼

☒ Document Suggest

☒ Cluster Suggest

☐ Exclude Duplicates

☒ Clean Prior Injections

☒ Add As References

167 Advanced

Similarity 0 Cluster Sites Centroid

Nearest Neighbor

168 Classifier Parameters

ClassEstimator	DistanceWeightedMaximumCount
ProbabilityEstimator	Linear
ProbabilityEstimatorBeta	8

169 Commands

Inject Cancel

