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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
1 September 2005 (01.09.2005)

PCT

(10) International Publication Number
WO 2005/081139 A1

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

(21) International Application Number:
PCT/US2005/004241

(22) International Filing Date: 11 February 2005 (11.02.2005)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
10/778,416 13 February 2004 (13.02.2004) US

(71) Applicant *(for all designated States except US)*: **ATTENEX CORPORATION** [US/US]; 925 Fourth Avenue, Ste 1700, Seattle, WA 98104-1125 (US).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: **EVANS, Lynne, Marie** [US/US]; 10382 172nd Ave SE, Renton, WA 98059 (US).

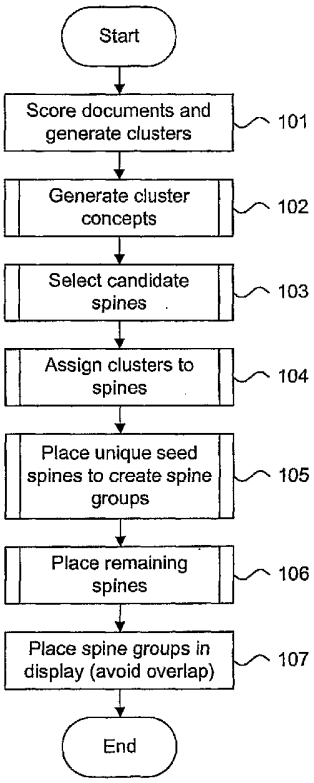
(74) Agent: **INOUYE, Patrick, Joseph, Sus**; 810 Third Avenue, Ste 258, Seattle, WA 98104 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, 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, MZ, NA, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, MC, NL, PL, PT, RO,

[Continued on next page]

(54) Title: ARRANGING CONCEPT CLUSTERS IN THEMATIC NEIGHBORHOOD RELATIONSHIPS IN A TWO-DIMENSIONAL DISPLAY



(57) Abstract: A set of clusters (50) is selected from a concept space. The concept space includes clusters (50) with concepts (53) visualizing document content (49) based on extracted concepts (47). A theme in each of a plurality of the clusters (50) is identified. Each theme includes at least one such concept (53) ranked within the cluster (50). Unique candidate spines (55) is logically formed. Each candidate spine (55) includes clusters (50) commonly sharing at least one such concept (54). The clusters (50) are assigned to one such candidate spine (55) having a substantially best fit. Each such sufficiently unique best fit candidate spine (56) is identified and placed in a visual display space (43). Each non-identified best fit candidate spine (56) is placed in the visual display space (43) relative to an anchor cluster (60) on one such identified best fit candidate spine (56).

WO 2005/081139 A1

SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

ARRANGING CONCEPT CLUSTERS IN THEMATIC NEIGHBORHOOD RELATIONSHIPS IN A TWO-DIMENSIONAL DISPLAY

TECHNICAL FIELD

The present invention relates in general to data visualization and, in particular, to a system and method for arranging concept clusters in thematic neighborhood relationships in a two-dimensional visual display space.

BACKGROUND ART

In general, data visualization transforms numeric or textual information into a graphical display format to assist users in understanding underlying trends and principles in the data. Effective data visualization complements and, in some instances, supplants numbers and text as a more intuitive visual presentation format than raw numbers or text alone. However, graphical data visualization is constrained by the physical limits of computer display systems. Two-dimensional and three-dimensional visualized information can be readily displayed. However, visualized information in excess of three dimensions must be artificially compressed if displayed on conventional display devices. Careful use of color, shape and temporal attributes can simulate multiple dimensions, but comprehension and usability become difficult as additional layers of modeling are artificially grafted into a two- or three-dimensional display space.

Mapping multi-dimensional information into a two- or three-dimensional display space potentially presents several problems. For instance, a viewer could misinterpret dependent relationships between discrete objects displayed adjacently in a two or three dimensional display. Similarly, a viewer could erroneously interpret dependent variables as independent and independent variables as dependent. This type of problem occurs, for example, when visualizing clustered data, which presents discrete groupings of related data. Other factors further complicate the interpretation and perception of visualized data, based on the Gestalt principles of proximity, similarity, closed region, connectedness, good continuation, and closure, such as described in R.E. Horn, "Visual Language: Global Communication for the 21st Century," Ch. 3, MacroVU Press (1998), the disclosure of which is incorporated by reference.

Conventionally, objects, such as clusters, modeled in multi-dimensional concept space are generally displayed in two- or three-dimensional display space as geometric objects. Independent variables are modeled through object attributes, such as radius, volume, angle, distance and so forth. Dependent variables are modeled within the two or three dimensions.

However, poor cluster placement within the two or three dimensions can mislead a viewer into misinterpreting dependent relationships between discrete objects.

Consider, for example, a group of clusters, which each contain a group of points corresponding to objects sharing a common set of traits. Each cluster is located at some distance from a common origin along a vector measured at a fixed angle from a common axis. The radius of each cluster reflects the number of objects contained. Clusters located along the same vector are similar in traits to those clusters located on vectors separated by a small cosine rotation. However, the radius and distance of each cluster from the common origin are independent variables relative to other clusters. When displayed in two dimensions, the overlaying or overlapping of clusters could mislead the viewer into perceiving data dependencies between the clusters where no such data dependencies exist.

Conversely, multi-dimensional information can be advantageously mapped into a two- or three-dimensional display space to assist with comprehension based on spatial appearances. Consider, as a further example, a group of clusters, which again each contain a group of points corresponding to objects sharing a common set of traits and in which one or more "popular" concepts or traits frequently appear in some of the clusters. Since the distance of each cluster from the common origin is an independent variable relative to other clusters, those clusters that contain popular concepts or traits may be placed in widely separated regions of the display space and could similarly mislead the viewer into perceiving no data dependencies between the clusters where such data dependencies exist.

One approach to depicting thematic relationships between individual clusters applies a force-directed or "spring" algorithm. Clusters are treated as bodies in a virtual physical system. Each body has physics-based forces acting on or between them, such as magnetic repulsion or gravitational attraction. The forces on each body are computed in discrete time steps and the positions of the bodies are updated. However, the methodology exhibits a computational complexity of order $O(n^2)$ per discrete time step and scales poorly to cluster formations having a few hundred nodes. Moreover, large groupings of clusters tend to pack densely within the display space, thereby losing any meaning assigned to the proximity of related clusters.

Therefore, there is a need for an approach to efficiently placing clusters based on popular concepts or traits into thematic neighborhoods that map multiple cluster relationships in a visual display space.

There is a further need for an approach to orienting data clusters to properly visualize independent and dependent variables while compressing thematic relationships to emphasize thematically stronger relationships.

DISCLOSURE OF INVENTION

Relationships between concept clusters are shown in a two-dimensional display space by combining connectedness and proximity. Clusters sharing "popular" concepts are identified by evaluating thematically-closest neighboring clusters, which are assigned into linear cluster spines arranged to avoid object overlap. The cluster arrangement methodology exhibits a highly-scalable computational complexity of order $O(n)$.

An embodiment provides a system and method for arranging concept clusters in thematic neighborhood relationships in a two-dimensional visual display space. A set of clusters is selected from a concept space. The concept space includes a multiplicity of clusters with concepts visualizing document content based on extracted concepts. A theme in each of a plurality of the clusters is identified. Each theme includes at least one such concept ranked within the cluster. A plurality of unique candidate spines is logically formed. Each candidate spine includes clusters commonly sharing at least one such concept. One or more of the clusters are assigned to one such candidate spine having a substantially best fit. Each such best fit candidate spine sufficiently unique from each other such best fit candidate spine is identified. The identified best fit candidate spine is placed in a visual display space. Each non-identified best fit candidate spine is placed in the visual display space relative to an anchor cluster on one such identified best fit candidate spine.

Still other embodiments of the present invention will become readily apparent to those skilled in the art from the following detailed description, wherein are one embodiments of the invention 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, all without departing from the spirit and the scope of the present invention. Accordingly, the drawings and detailed description are to be regarded as illustrative in nature and not as restrictive.

DESCRIPTION OF DRAWINGS

FIGURE 1 is a block diagram showing a system for arranging concept clusters in thematic neighborhood relationships in a two-dimensional visual display space, in accordance with the present invention.

FIGURE 2 is a block diagram showing the system modules implementing the display generator of FIGURE 1.

FIGURE 3 is a flow diagram showing a method for arranging concept clusters in thematic neighborhood relationships in a two-dimensional visual display space, in accordance with the present invention.

FIGURE 4 is a flow diagram showing the routine for generating cluster concepts for use in the method of FIGURE 3.

FIGURE 5 is a flow diagram showing the routine for selecting candidate spines for use in the method of FIGURE 3.

FIGURE 6 is a flow diagram showing the routine for assigning clusters to candidate spines for use in the method of FIGURE 3.

FIGURE 7 is a flow diagram showing the routine for placing unique seed spines for use in the method of FIGURE 3.

FIGURE 8 is a flow diagram showing the routine for placing remaining best fit spines for use in the method of FIGURE 3.

FIGURE 9 is a flow diagram showing the function for selecting an anchor cluster for use in the routine of FIGURE 8.

FIGURE 10 is a data representation diagram showing, by way of example, a view of a cluster spine.

FIGURES 11A-C are data representation diagrams showing anchor points within cluster spines.

FIGURE 12 is a flow diagram showing the function for grafting a spine cluster onto a spine for use in the routine of FIGURE 8.

FIGURE 13 is a data representation diagram showing cluster placement relative to an anchor point.

FIGURE 14 is a data representation diagram showing a completed cluster placement.

BEST MODE FOR CARRYING OUT THE INVENTION

Concept: One or more preferably root stem normalized words defining a specific meaning.

Theme: One or more concepts defining a semantic meaning.

Cluster: Grouping of documents containing one or more common themes.

Spine: Grouping of clusters sharing a single concept preferably arranged linearly along a vector. Also referred to as a *cluster spine*.

Spine Group: Set of connected and semantically-related spines.

The foregoing terms are used throughout this document and, unless indicated otherwise, are assigned the meanings presented above.

FIGURE 1 is a block diagram showing a system 10 for arranging concept clusters in thematic neighborhood relationships in a two-dimensional visual display space, in accordance with the present invention. By way of illustration, the system 10 operates in a distributed computing environment, which includes a plurality of heterogeneous systems and document

sources. A backend server 11 executes a workbench suite 31 for providing a user interface framework for automated document management, processing and analysis. The backend server 11 is coupled to a storage device 13, which stores documents 14, in the form of structured or unstructured data, and a database 30 for maintaining document information. A production server 12 includes a document mapper 32, that includes a clustering engine 33 and display generator 34. The clustering engine 33 performs efficient document scoring and clustering, such as described in commonly-assigned U.S. Patent application Serial No. 10/626,984, filed July 25, 2003, pending, the disclosure of which is incorporated by reference. The display generator 34 arranges concept clusters in thematic neighborhood relationships in a two-dimensional visual display space, as further described below beginning with reference to FIGURE 2.

The document mapper 32 operates on documents retrieved from a plurality of local sources. The local sources include documents 17 maintained in a storage device 16 coupled to a local server 15 and documents 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 production system 11 over an intranetwork 21. In addition, the document mapper 32 can identify and retrieve documents from remote sources over an 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.

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

In one embodiment, the individual documents 17, 20, 26, 29 include electronic message folders, such as maintained by the Outlook and Outlook Express products, licensed by Microsoft Corporation, Redmond, Washington. The database is an SQL-based relational database, such as the Oracle database management system, release 8, licensed by Oracle Corporation, Redwood Shores, California.

The individual computer systems, including backend server 11, production server 32, server 15, client 18, remote server 24 and remote client 27, are general purpose, programmed digital 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. 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.

FIGURE 2 is a block diagram showing the system modules implementing the display generator 34 of FIGURE 1. The display generator 34 includes clustering 44, theme generator 41 and spine placement 42 components and maintains attached storage 44 and database 46. Individual documents 14 are analyzed by the clustering component 44 to form clusters 50 of semantically scored documents, such as described in commonly-assigned U.S. Patent application Serial No. 10/626,984, filed July 25, 2003, pending, the disclosure of which is incorporated by reference. In one embodiment, document concepts 47 are formed from concepts and terms extracted from the documents 14 and the frequencies of occurrences and reference counts of the concepts and terms are determined. Each concept and term is then scored based on frequency, concept weight, structural weight, and corpus weight. The document concept scores 48 are compressed and assigned to normalized score vectors for each of the documents 14. The similarities between each of the normalized score vectors are determined, preferably as cosine values. A set of candidate seed documents is evaluated to select a set of seed documents 49 as initial cluster centers based on relative similarity between the assigned normalized score vectors for each of the candidate seed documents or using a dynamic threshold based on an analysis of the similarities of the documents 14 from a center of each cluster 15, such as described in commonly-assigned U.S. Patent application Serial No. 10/626,984, filed July 25, 2003, pending, the disclosure of which is incorporated by reference. The remaining non-seed documents are evaluated against the cluster centers also based on relative similarity and are grouped into the clusters 50 based on best-fit, subject to a minimum fit criterion.

The theme generator 41 evaluates the document concepts 47 assigned to each of the clusters 50 and identifies cluster concepts 53 for each cluster 50, as further described below with reference to FIGURE 4. Briefly, the document concepts 47 for each cluster 50 are ranked into ranked cluster concepts 52 based on cumulative document concept scores 51. The top-ranked document concepts 47 are designated as cluster concepts 53. In the described embodiment, each cluster concept 53 must also be a document concept 47 appearing in the initial cluster center, be contained in a minimum of two documents 14 or at least 30% of the documents 14 in the cluster 50. Other cluster concept membership criteria are possible.

The cluster placement component 42 places spines and certain clusters 50 into a two-dimensional display space as a visualization 43. The cluster placement component 42 performs four principal functions. First, the cluster placement component 42 selects candidate spines 55,

as further described below with reference to FIGURE 5. Briefly, the candidate spines 55 are selected by surveying the cluster concepts 53 for each cluster 50. Each cluster concept 53 shared by two or more clusters 50 can potentially form a spine of clusters 50. However, those cluster concepts 53 referenced by just a single cluster 50 or by more than 10% of the clusters 50 are discarded. The remaining clusters 50 are identified as candidate spine concepts 54, which each logically form a candidate spine 55.

Second, the cluster placement component 42 assigns each of the clusters 50 to a best fit spine 56, as further described below with reference to FIGURE 6. Briefly, the fit of each candidate spine 55 to a cluster 50 is determined by evaluating the candidate spine concept 54 to the cluster concept 53. The candidate spine 545 exhibiting a maximum fit is selected as the best fit spine 56 for the cluster 50.

Third, the cluster placement component 42 selects and places unique seed spines 58, as further described below with reference to FIGURE 7. Briefly, spine concept score vectors 57 are generated for each best fit spine 56 and evaluated. Those best fit spines 56 having an adequate number of assigned clusters 50 and which are sufficiently dissimilar to any previously selected best fit spines 56 are designated and placed as seed spines 58.

The cluster placement component 42 places any remaining unplaced best fit spines 56 and clusters 50 that lack best fit spines 56 into spine groups, as further described below with reference to FIGURE 8. Briefly, anchor clusters 60 are selected based on similarities between unplaced candidate spines 55 and candidate anchor clusters. Cluster spines are grown by placing the clusters 50 in similarity precedence to previously placed spine clusters or anchor clusters along vectors originating at each anchor cluster 60. As necessary, clusters 50 are placed outward or in a new vector at a different angle from new anchor clusters 55. Finally, the spine groups are placed within the visualization 43 by translating the spine groups until there is no overlap, such as described in commonly-assigned U.S. Patent application Serial No. 10/084,401, filed February 25, 2002, pending, the disclosure of which is incorporated by reference.

Each module or component is a computer program, procedure or module written as source code in a conventional programming language, such as the C++ programming language, and is presented for execution by the CPU as object or byte code, as is known in the art. The various implementations of the source code and object and byte codes can be held on a computer-readable storage medium or embodied on a transmission medium in a carrier wave. The display generator 32 operates in accordance with a sequence of process steps, as further described below with reference to FIGURE 3.

FIGURE 3 is a flow diagram showing a method 100 for arranging concept clusters 50 in thematic neighborhood relationships in a two-dimensional visual display space, in accordance with the present invention. The method 80 is described as a sequence of process operations or steps, which can be executed, for instance, by a display generator 32 (shown in FIGURE 1).

As an initial step, documents 14 are scored and clusters 50 are generated (block 101), such as described in commonly-assigned U.S. Patent application Serial No. 10/626,984, filed July 25, 2003, pending, the disclosure of which is incorporated by reference. Next, one or more cluster concepts 53 are generated for each cluster 50 based on cumulative cluster concept scores 51 (block 102), as further described below with reference to FIGURE 4. The cluster concepts 53 are used to select candidate spines 55 (block 103), as further described below with reference to FIGURE 5, and the clusters 50 are then assigned to the candidate spines 55 as best fit spines 56 (block 104), as further described below with reference to FIGURE 6. Unique seed spines are identified from the best fit spines 56 and placed to create spine groups (block 105), along with any remaining unplaced best fit spines 56 and clusters 50 that lack best fit spines 56 (block 106), as further described below with reference to FIGURE 8. Finally, the spine groups are placed within the visualization 43 in the display space. In the described embodiment, each of the spine groups is placed so as to avoid overlap with other spine groups. In a further embodiment, the spine groups can be placed by similarity to other spine groups. Other cluster, spine, and spine group placement methodologies could also be applied based on similarity, dissimilarity, attraction, repulsion, and other properties in various combinations, as would be appreciated by one skilled in the art. The method then terminates.

FIGURE 4 is a flow diagram showing the routine 110 for generating cluster concepts 53 for use in the method 100 of FIGURE 3. One purpose of this routine is to identify the top ranked cluster concepts 53 that best summarizes the commonality of the documents in any given cluster 50 based on cumulative document concept scores 51.

A cluster concept 53 is identified by iteratively processing through each of the clusters 50 (blocks 111-118). During each iteration, the cumulative score 51 of each of the document concepts 47 for all of the documents 14 appearing in a cluster 50 are determined (block 112). The cumulative score 51 can be calculated by summing over the document concept scores 48 for each cluster 50. The document concepts 47 are then ranked by cumulative score 51 as ranked cluster concepts 52 (block 113). In the described embodiment, the ranked cluster concepts 52 appear in descending order, but could alternatively be in ascending order. Next, a cluster concept 53 is determined. The cluster concept 53 can be user provided (block 114). Alternatively, each ranked cluster concept 52 can be evaluated against an acceptance criteria

(blocks 115 and 116) to select a cluster concept 53. In the described embodiment, cluster concepts 53 must meet the following criteria:

(1) be contained in the initial cluster center (block 115); and

(2) be contained in a minimum of two documents 14 or 30% of the documents 14 in the cluster 50, whichever is greater (block 116).

The first criteria restricts acceptable ranked cluster concepts 52 to only those document concepts 47 that appear in a seed cluster center theme of a cluster 50 and, by implication, are sufficiently relevant based on their score vectors. Generally, a cluster seed theme corresponds to the set of concepts appearing in a seed document 49, but a cluster seed theme can also be specified by a user or by using a dynamic threshold based on an analysis of the similarities of the documents 14 from a center of each cluster 50, such as described in commonly-assigned U.S. Patent application Serial No. 10/626,984, filed July 25, 2003, pending, the disclosure of which is incorporated by reference The second criteria filters out those document concepts 47 that are highly scored, yet not popular. Other criteria and thresholds for determining acceptable ranked cluster concepts 52 are possible.

If acceptable (blocks 115 and 116), the ranked cluster concept 52 is selected as a cluster concept 53 (block 117) and processing continues with the next cluster (block 118), after which the routine returns.

FIGURE 5 is a flow diagram showing the routine 120 for selecting candidate spines 55 for use in the method 100 of FIGURE 3. One purpose of this routine is to identify candidate spines 55 from the set of all potential spines 55.

Each cluster concept 53 shared by two or more clusters 50 can potentially form a spine of clusters 50. Thus, each cluster concept 53 is iteratively processed (blocks 121-126). During each iteration, each potential spine is evaluated against an acceptance criteria (blocks 122-123). In the described embodiment, a potential spine cannot be referenced by only a single cluster 50 (block 122) or by more than 10% of the clusters 50 in the potential spine (block 123). Other criteria and thresholds for determining acceptable cluster concepts 53 are possible. If acceptable (blocks 122, 123), the cluster concept 53 is selected as a candidate spine concept 54 (block 124) and a candidate spine 55 is logically formed (block 125). Processing continues with the next cluster (block 126), after which the routine returns.

FIGURE 6 is a flow diagram showing the routine 130 for assigning clusters 50 to candidate spines 55 for use in the method 100 of FIGURE 3. One purpose of this routine is to match each cluster 50 to a candidate spine 55 as a best fit spine 56.

The best fit spines 56 are evaluated by iteratively processing through each cluster 50 and candidate spine 55 (blocks 131-136 and 132-134, respectively). During each iteration for a given cluster 50 (block 131), the spine fit of a cluster concept 53 to a candidate spine concept 54 is determined (block 133) for a given candidate spine 55 (block 132). In the described embodiment, the spine fit F is calculated according to the following equation:

$$F = \log\left(\frac{popularity}{rank^2}\right) \times scale$$

where *popularity* is defined as the number of clusters 50 containing the candidate spine concept 54 as a cluster concept 53, *rank* is defined as the rank of the candidate spine concept 54 for the cluster 50, and *scale* is defined as a bias factor for favoring a user specified concept or other predefined or dynamically specified characteristic. In the described embodiment, a scale of 1.0 is used for candidate spine concept 54 while a scale of 5.0 is used for user specified concepts. Processing continues with the next candidate spine 55 (block 134). Next, the cluster 50 is assigned to the candidate spine 55 having a maximum spine fit as a best fit spine 56 (block 135). Processing continues with the next cluster 50 (block 136). Finally, any best fit spine 56 that attracts only a single cluster 50 is discarded (block 137) by assigning the cluster 50 to a next best fit spine 56 (block 138). The routine returns.

FIGURE 7 is a flow diagram showing the routine 140 for placing unique seed spines for use in the method 100 of FIGURE 3. One purpose of this routine identify and place best fit spines 56 into the visualization 43 as unique seed spines 58 for use as anchors for subsequent candidate spines 55.

Candidate unique seed spines are selected by first iteratively processing through each best fit spine 56 (blocks 141-144). During each iteration, a spine concept score vector 57 is generated for only those spine concepts corresponding to each best fit spine 56 (block 142). The spine concept score vector 57 aggregates the cumulative cluster concept scores 51 for each of the clusters 50 in the best fit spine 56. Each spine concept score in the spine concept score vector 57 is normalized, such as by dividing the spine concept score by the length of the spine concept score vector 57 (block 143). Processing continues for each remaining best fit spine 56 (block 144), after which the best fit spines 56 are ordered by number of clusters 50. Each best fit spine 56 is again iteratively processed (blocks 146-151). During each iteration, best fit spines 56 that are not sufficiently large are discarded (block 147). In the described embodiment, a sufficiently large best fit spine 56 contains at least five clusters 50. Next, the similarities of the best fit spine 56 to each previously-selected unique seed spine 58 is calculated and compared (block 148). In the described embodiment, best fit spine similarity is calculated as the cosine of the cluster

concept score vectors 59, which contains the cumulative cluster concept scores 51 for the cluster concepts 53 of each cluster 50 in the best fit spine 56 or previously-selected unique seed spine 58. Best fit spines 56 that are not sufficiently dissimilar are discarded (block 14). Otherwise, the best fit spine 56 is identified as a unique seed spine 58 and is placed in the visualization 43 (block 150). Processing continues with the next best fit spine 56 (block 151), after which the routine returns.

FIGURE 8 is a flow diagram showing the routine 160 for placing remaining candidate spines 55 for use in the method 100 of FIGURE 3. One purpose of this routine identify and place any remaining unplaced best fit spines 56 and clusters 50 that lack best fit spines 56 into the visualization 43.

First, any remaining unplaced best fit spines 56 are ordered by number of clusters 50 assigned (block 161). The unplaced best fit spine 56 are iteratively processed (blocks 162-175) against each of the previously-placed spines (blocks 163-174). During each iteration, an anchor cluster 60 is selected from the previously placed spine 58 (block 164), as further described below with reference to FIGURE 9. The cluster 50 contained in the best fit spine 56 that is most similar to the selected anchor cluster 60 is then selected (block 165). In the described embodiment, cluster similarity is calculated as cosine value of the cumulative cluster concept vectors 51, although other determinations of cluster similarity are possible, including minimum, maximum, and median similarity bounds. The spine clusters 50 are grafted onto the previously placed spine along a vector defined from the center of the anchor cluster 55 (block 166), as further described below with reference to FIGURE 12. If any of the spine clusters are not placed (block 167), another anchor cluster 60 is selected (block 168), as further described below with reference to FIGURE 9. Assuming another anchor cluster 60 is selected (block 169), the spine clusters are again placed (block 166), as further described below with reference to FIGURE 12. Otherwise, if another anchor cluster 60 is not selected (block 169), the cluster 50 is placed in a related area (block 170). In one embodiment, unanchored best fit spines 56 become additional spine group seeds. In a further embodiment, unanchored best fit spines 56 can be placed adjacent to the best fit anchor cluster 60 or in a display area of the visualization 43 separately from the placed best fit spines 56.

If the cluster 50 is placed (block 167), the best fit spine 56 is labeled as containing candidate anchor clusters 60 (block 171). If the current vector forms a maximum line segment (block 172), the angle of the vector is changed (block 173). In the described embodiment, a maximum line segment contains more than 25 clusters 50, although any other limit could also be applied. Processing continues with each seed spine (block 174) and remaining unplaced best fit

spine 56 (block 175). Finally, any remaining unplaced clusters 50 are placed (block 176). In one embodiment, unplaced clusters 50 can be placed adjacent to a best fit anchor cluster 60 or in a display area of the visualization 43 separately from the placed best fit spines 56. The routine then returns.

FIGURE 9 is a flow diagram showing the function 180 for selecting an anchor cluster 60 for use in the routine 160 of FIGURE 8. One purpose of this routine is to return a set of anchor clusters 60, which contain the spine concept and which are ordered by similarity to the largest cluster 50 in the spine.

Each candidate anchor cluster 60 is iteratively processed (blocks 181-183) to determine the similarity between a given cluster 50 and each candidate anchor cluster 60 (block 182). In one embodiment, each cluster similarity is calculated as cosine value concept vectors, although other determinations of cluster similarity are possible, including minimum, maximum, and median similarity bounds. The most similar candidate anchor cluster 60 is identified (block 184) and, if found, chosen as the anchor cluster 60 (block 187), such as described in commonly-assigned U.S. Patent application Serial No. 10/084,401, filed February 25, 2002, pending, the disclosure of which is incorporated by reference. Otherwise, if not found (block 185), the largest cluster 50 assigned to the unique seed spine 58 is chosen as the anchor cluster 60 (block 186). The function then returns set of the anchor clusters 60 and the unique seed spine 58 becomes a seed for a new spine group (block 188).

FIGURE 10 is a data representation diagram 200 showing, by way of example, a view of a cluster spine 202. Clusters are placed in a cluster spine 202 along a vector 203, preferably defined from center of an anchor cluster. Each cluster in the cluster spine 202, such as endpoint clusters 204 and 206 and midpoint clusters 205, group documents 207 sharing a popular concept, that is, assigned to a best-fit concept 53. The cluster spine 202 is placed into a visual display area 201 to generate a two-dimensional spatial arrangement. To represent data inter-relatedness, the clusters 204-206 in each cluster spine 202 are placed along a vector 203 arranged in order of cluster similarity, although other line shapes and cluster orderings can be used.

The cluster spine 202 visually associates those clusters 204-206 sharing a common popular concept. A theme combines two or more concepts. During cluster spine creation, those clusters 204-206 having available anchor points are identified for use in grafting other cluster spines sharing popular thematically-related concepts, as further described below with reference to FIGURES 11A-C.

FIGURES 11A-C are data representation diagrams 210, 220, 230 showing anchor points within cluster spines. A placed cluster having at least one open edge constitutes a candidate

anchor point 54. Referring first to FIGURE 11A, a starting endpoint cluster 212 of a cluster spine 211 functions as an anchor point along each open edge 215a-e at primary and secondary angles.

An open edge is a point along the edge of a cluster at which another cluster can be adjacently placed. In the described embodiment, clusters are placed with a slight gap between each cluster to avoid overlapping clusters. Otherwise, a slight overlap within 10% with other clusters is allowed. An open edge is formed by projecting vectors 214a-e outward from the center 213 of the endpoint cluster 212, preferably at normalized angles. The clusters in the cluster spine 211 are arranged in order of cluster similarity.

In one embodiment, given $0 \leq \sigma < \Pi$, where σ is the angle of the current cluster spine 211, the normalized angles for largest endpoint clusters are at one third Π to minimize interference with other spines while maximizing the degree of interrelatedness between spines. If the cluster ordinal spine position is even, the primary angle is $\sigma + \frac{\Pi}{3}$ and the secondary angle is $\sigma - \frac{\Pi}{3}$. Otherwise, the primary angle is $\sigma - \frac{\Pi}{3}$ and the secondary angle is $\sigma + \frac{\Pi}{3}$. Other evenly divisible angles could be also used.

Referring next to FIGURE 11B, the last endpoint cluster 222 of a cluster spine 221 also functions as an anchor point along each open edge. The endpoint cluster 222 contains the fewest number of concepts. The clusters in the cluster spine 221 are arranged in order of similarity to the last placed cluster. An open edge is formed by projecting vectors 224a-c outward from the center 223 of the endpoint cluster 222, preferably at normalized angles.

In one embodiment, given $0 \leq \sigma < \Pi$, where σ is the angle of the current cluster spine 221, the normalized angles for smallest endpoint clusters are at one third Π, but only three open edges are available to graft other thematically-related cluster spines. If the cluster ordinal spine position is even, the primary angle is $\sigma + \frac{\Pi}{3}$ and the secondary angle is $\sigma - \frac{\Pi}{3}$. Otherwise, the primary angle is $\sigma - \frac{\Pi}{3}$ and the secondary angle is $\sigma + \frac{\Pi}{3}$. Other evenly divisible angles could be also used.

Referring finally to FIGURE 11C, a midpoint cluster 237 of a cluster spine 231 functions as an anchor point for a cluster spine 236 along each open edge. The midpoint cluster 237 is located intermediate to the clusters in the cluster spine 236 and defines an anchor point along each open edge. An open edge is formed by projecting vectors 239a-b outward from the center 238 of the midpoint cluster 237, preferably at normalized angles. Unlike endpoint clusters 52,

232 the midpoint cluster 237 can only serve as an anchor point along tangential vectors non-coincident to the vector forming the cluster spine 236. Accordingly, endpoint clusters 212, 222 include one additional open edge serving as a coincident anchor point.

In one embodiment, given $0 \leq \sigma < \Pi$, where σ is the angle of the current cluster spine 231, the normalized angles for midpoint clusters are at one third Π, but only two open edges are available to graft other thematically-related cluster spines. Empirically, limiting the number of available open edges to those facing the direction of cluster similarity helps to maximize the interrelatedness of the overall display space.

FIGURE 12 is a flow diagram showing the function 240 for grafting a spine cluster 50 onto a spine for use in the routine 160 of FIGURE 8. One purpose of this routine is to attempt to place a cluster 50 at an anchor point in a cluster spine either along or near an existing vector, if possible, as further described below with reference to FIGURE 13.

An angle for placing the cluster 50 is determined (block 241), dependent upon whether the cluster against which the current cluster 50 is being placed is a starting endpoint, midpoint, or last endpoint cluster, as described above with reference to FIGURES 11A-C. If the cluster ordinal spine position is even, the primary angle is $\sigma + \frac{\Pi}{3}$ and the secondary angle is $\sigma - \frac{\Pi}{3}$. Otherwise, the primary angle is $\sigma - \frac{\Pi}{3}$ and the secondary angle is $\sigma + \frac{\Pi}{3}$. Other evenly divisible angles could be also used. The cluster 50 is then placed using the primary angle (block 242). If the cluster 50 is the first cluster in a cluster spine but cannot be placed using the primary angle (block 243), the secondary angle is used and the cluster 50 is placed (block 244). Otherwise, if the cluster 50 is placed but overlaps more than 10% with existing clusters (block 245), the cluster 50 is moved outward (block 246) by the diameter of the cluster 50. Finally, if the cluster 50 is satisfactorily placed (block 247), the function returns an indication that the cluster 50 was placed (block 248). Otherwise, the function returns an indication that the cluster was not placed (block 249).

FIGURE 13 is a data representation diagram showing cluster placement relative to an anchor point. Anchor points 266, 267 are formed along an open edge at the intersection of a vector 263a, 263b, respectively, drawn from the center 262 of the cluster 261. The vectors are preferably drawn at a normalized angle, such as $\frac{\Pi}{3}$ in one embodiment, relative to the vector 268 forming the cluster spine 268.

FIGURE 14 is a data representation diagram 270 showing a completed cluster placement. The clusters 272, 274, 276, 278 placed in each of the cluster spines 271, 273, 275, 277 are

respectively matched to popular concepts, that is, best-fit concepts 53. Slight overlap 279 between grafted clusters is allowed. In one embodiment, no more than 10% of a cluster can be covered by overlap. The singleton clusters 280, however, do not thematically relate to the placed clusters 272, 274, 276, 278 in cluster spines 271, 273, 275, 277 and are therefore grouped as individual clusters in non-relational placements.

While the invention has been particularly shown and described as referenced to the embodiments thereof, those skilled in the art will understand that the foregoing and other changes in form and detail may be made therein without departing from the spirit and scope of the invention.

CLAIMS:

1. A system (34) for arranging concept clusters (53) in thematic neighborhood relationships in a two-dimensional visual display space (43), comprising:

a set of clusters (50) selected from a concept space comprising a multiplicity of clusters (50) with concepts (53) visualizing document content (49) based on extracted concepts (47);

a theme generator (41) to identify a theme in each of a plurality of the clusters (50), each theme comprising at least one such concept (53) ranked within the cluster (50); and

a spine placer (42), comprising:

a candidate spine selector to logically form a plurality of unique candidate spines (55) comprising clusters (50) commonly sharing at least one such concept (54);

a candidate spine assigner to assign one or more of the clusters (50) to one such candidate spine (55) having a substantially best fit;

a best fit candidate spine placer to identify each such best fit candidate spine (56) sufficiently unique from each other such best fit candidate spine (56) and to place the identified best fit candidate spine (56) in a visual display space (43); and

a remaining candidate spine placer to place each non-identified best fit candidate spine (56) in the visual display space (43) relative to an anchor cluster (60) on one such identified best fit candidate spine (56).

2. A system (34) according to Claim 1, further comprising:

a concept scorer to determine a cumulative score (51) for one or more of the concepts (53) for each of the plurality of clusters (50); and

a concept ranker to rank the concepts (53) by the cumulative score (51) in at least one of descending and ascending order.

3. A system (34) according to Claim 1, further comprising:

a concept evaluator to evaluate each of the plurality of concepts (53) against an acceptance criteria to qualify as the theme of the cluster (50).

4. A system (34) according to Claim 3, wherein the acceptance criteria comprises at least one of being contained in a seed theme of a cluster (50) and being contained in a predetermined minimum of the documents (49).

5. A system (34) according to Claim 1, further comprising:

a candidate spine evaluator to evaluate such candidate spine (55) against an acceptance criteria.

6. A system (34) according to Claim 5, wherein the acceptance criteria comprises the at least one such concept (54) being contained in at least one of a plurality of the plurality of clusters (50) and within a predetermined maximum of the plurality of clusters (50).

7. A system (34) according to Claim 1, further comprising:

a spine fit evaluator to determine a spine fit between the concept (54) in each such cluster (50) and the at least one theme commonly shared by the clusters (50) in each of the candidate spines (55); and

a spine fit selector to select the spine fit comprising a maximum spine fit as the substantially best fit.

8. A system (34) according to Claim 7, wherein the spine fit is calculated in accordance to an equation:

$$F = \log\left(\frac{popularity}{rank^2}\right) \times scale$$

where *popularity* is defined as a number of clusters (50) containing each such concept (54) in the candidate spine (55), *rank* is defined as a rank of the candidate spine concept (54), and *scale* is defined as a bias factor.

9. A system (34) according to Claim 1, wherein each such best fit candidate spine (56) containing only one such cluster (50) is discarded.

10. A system (34) according to Claim 1, further comprising:

a spine concept score vector (57) generated for each such best fit candidate spine (56); and

a similarity evaluator to evaluate a similarity between the best fit candidate spine (56) and each other such other such best fit candidate spine (56).

11. A system (34) according to Claim 10, further comprising:

a concept score aggregator to aggregate a concept score (51) for each such concept (54) contained in each cluster (50) in the best fit candidate spine (56); and

a concept score normalizer to normalize each aggregated concept score (51).

12. A system (34) according to Claim 10, wherein the similarity is calculated as a cosine over the spine concept score vectors (57).

13. A system (34) according to Claim 1, further comprising:

a similarity identifier to determine a similarity between at least one anchor cluster candidate (50) and at least one such cluster (50) in a non-identified best fit candidate spine (56), and to identify the at least one such anchor cluster (60) candidate with acceptable similarity as the anchor cluster (60).

14. A system (34) according to Claim 13, wherein the similarity is calculated as a cosine over the anchor cluster (60) candidate and one such cluster (50) in the spine (56).

15. A system (34) according to Claim 13, wherein the non-identified best fit candidate spine (56) is placed in the visual display space (43) along a vector originating from the anchor cluster (60).

16. A system (34) according to Claim 15, wherein the placement of the non-identified best fit candidate spine (56) is adjusted if overlapping with at least one other cluster (50) already placed.

17. A system (34) according to Claim 15, wherein the non-identified best fit candidate spine (56) is labeled as containing at least one anchor cluster candidate (50) following placement.

18. A system (34) according to Claim 15, wherein the non-identified best fit candidate spine (56) is placed along a vector originating from the anchor cluster (60) with an angle comprising at least one of $\sigma \pm \frac{\Pi}{3}$, where $0 \leq \sigma < \Pi$.

19. A method (100) for arranging concept clusters (53) in thematic neighborhood relationships in a two-dimensional visual display space (43), comprising:

selecting a set of clusters (50) from a concept space comprising a multiplicity of clusters (50) with concepts (53) visualizing document content (49) based on extracted concepts (47);

identifying (110) a theme in each of a plurality of the clusters (50), each theme comprising at least one such concept (53) ranked within the cluster (50);

logically forming (120) a plurality of unique candidate spines (55) comprising clusters (50) commonly sharing at least one such concept (54) and assigning one or more of the clusters (50) to one such candidate spine (55) having a substantially best fit;

identifying (140) each such best fit candidate spine (56) sufficiently unique from each other such best fit candidate spine (56) and placing the identified best fit candidate spine (56) in a visual display space (43); and

placing (160) each non-identified best fit candidate spine (56) in the visual display space (43) relative to an anchor cluster (60) on one such identified best fit candidate spine (56).

20. A method (100) according to Claim 19, further comprising:

determining a cumulative score (51) for one or more of the concepts (53) for each of the plurality of clusters (50); and

ranking the concepts (53) by the cumulative score (51) in at least one of descending and ascending order.

21. A method (100) according to Claim 19, further comprising:

evaluating each of the plurality of concepts (53) against an acceptance criteria to qualify as the theme of the cluster (50).

22. A method (100) according to Claim 21, wherein the acceptance criteria comprises at least one of being contained in a seed theme of a cluster (50) and being contained in a predetermined minimum of the documents (49).

23. A method (100) according to Claim 19, further comprising:

evaluating such candidate spine (55) against an acceptance criteria.

24. A method (100) according to Claim 23, wherein the acceptance criteria comprises the at least one such concept (54) being contained in at least one of a plurality of the plurality of clusters (50) and within a predetermined maximum of the plurality of clusters (50).

25. A method (100) according to Claim 19, further comprising:

determining a spine fit between the concept (54) in each such cluster (50) and the at least one theme commonly shared by the clusters (50) in each of the candidate spines (55); and

selecting the spine fit comprising a maximum spine fit as the substantially best fit.

26. A method (100) according to Claim 25, wherein the spine fit is calculated in accordance to an equation:

$$F = \log\left(\frac{popularity}{rank^2}\right) \times scale$$

where *popularity* is defined as a number of clusters (50) containing each such concept (54) in the candidate spine (55), *rank* is defined as a rank of the candidate spine concept (54), and *scale* is defined as a bias factor.

27. A method (100) according to Claim 19, further comprising:

discarding each such best fit candidate spine (56) containing only one such cluster (50).

28. A method (100) according to Claim 19, further comprising:

generating a spine concept score vector (57) for each such best fit candidate spine (56); and

evaluating a similarity between the best fit candidate spine (56) and each other such other such best fit candidate spine (56).

29. A method (100) according to Claim 28, further comprising:

aggregating a concept score (51) for each such concept (54) contained in each cluster (50) in the best fit candidate spine (56); and

normalizing each aggregated concept score (51).

30. A method (100) according to Claim 28, further comprising:

calculating the similarity as a cosine over the spine concept score vectors (57).

31. A method (100) according to Claim 19, further comprising:

determining a similarity between at least one anchor cluster candidate (50) and at least one such cluster (50) in a non-identified best fit candidate spine (56); and

identifying the at least one such anchor cluster (60) candidate with acceptable similarity as the anchor cluster (60).

32. A method (100) according to Claim 31, further comprising:

calculating the similarity as a cosine over the anchor cluster (60) candidate and one such cluster (50) in the spine (56).

33. A method (100) according to Claim 31, further comprising:

placing the non-identified best fit candidate spine (56) in the visual display space (43) along a vector originating from the anchor cluster (60).

34. A method (100) according to Claim 33, further comprising:

adjusting placement of the non-identified best fit candidate spine (56) if overlapping with at least one other cluster (50) already placed.

35. A method (100) according to Claim 33, further comprising:

labeling the non-identified best fit candidate spine (56) as containing at least one anchor cluster candidate (50) following placement.

36. A method (100) according to Claim 33, further comprising:

placing the non-identified best fit candidate spine (56) along a vector originating from the anchor cluster (60) with an angle comprising at least one of $\sigma \pm \frac{\Pi}{3}$, where $0 \leq \sigma < \Pi$.

37. A computer-readable storage medium holding code for performing the method (100) according to Claim 19.

38. An apparatus for arranging concept clusters (53) in thematic neighborhood relationships in a two-dimensional visual display space (43), comprising:

means for selecting a set of clusters (50) from a concept space comprising a multiplicity of clusters (50) with concepts (53) visualizing document content (49) based on extracted concepts (47);

means for identifying a theme in each of a plurality of the clusters (50), each theme comprising at least one such concept (53) ranked within the cluster (50);

means for logically forming a plurality of unique candidate spines (55) comprising clusters (50) commonly sharing at least one such concept (54) and means for assigning one or more of the clusters (50) to one such candidate spine (55) having a substantially best fit;

means for identifying each such best fit candidate spine (56) sufficiently unique from each other such best fit candidate spine (56) and means for placing the identified best fit candidate spine (56) in a visual display space (43); and

means for placing each non-identified best fit candidate spine (56) in the visual display space (43) relative to an anchor cluster (60) on one such identified best fit candidate spine (56).

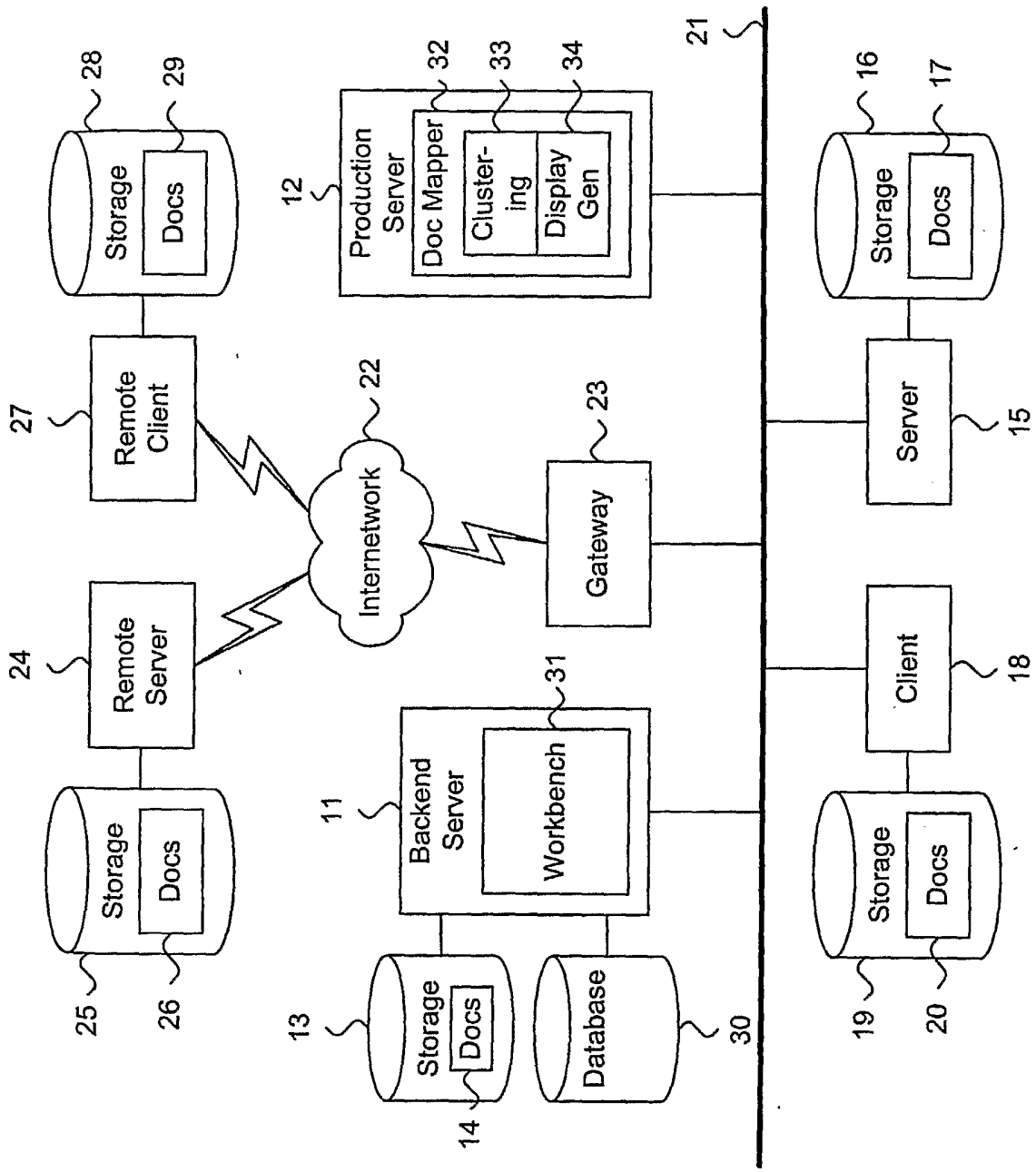

10
Storage
Docs
28
29
Remote Client
27
Remote Server
24
Storage
Docs
25
26
Internetwork
22
Production Server
12
Doc Mapper
Cluster-ing
Display Gen
32
33
34
21
Gateway
23
Backend Server
11
Workbench
31
Storage
Docs
13
14
Database
30
Storage
Docs
16
17
Server
15
Client
18
Storage
Docs
19
20

Fig. 1.

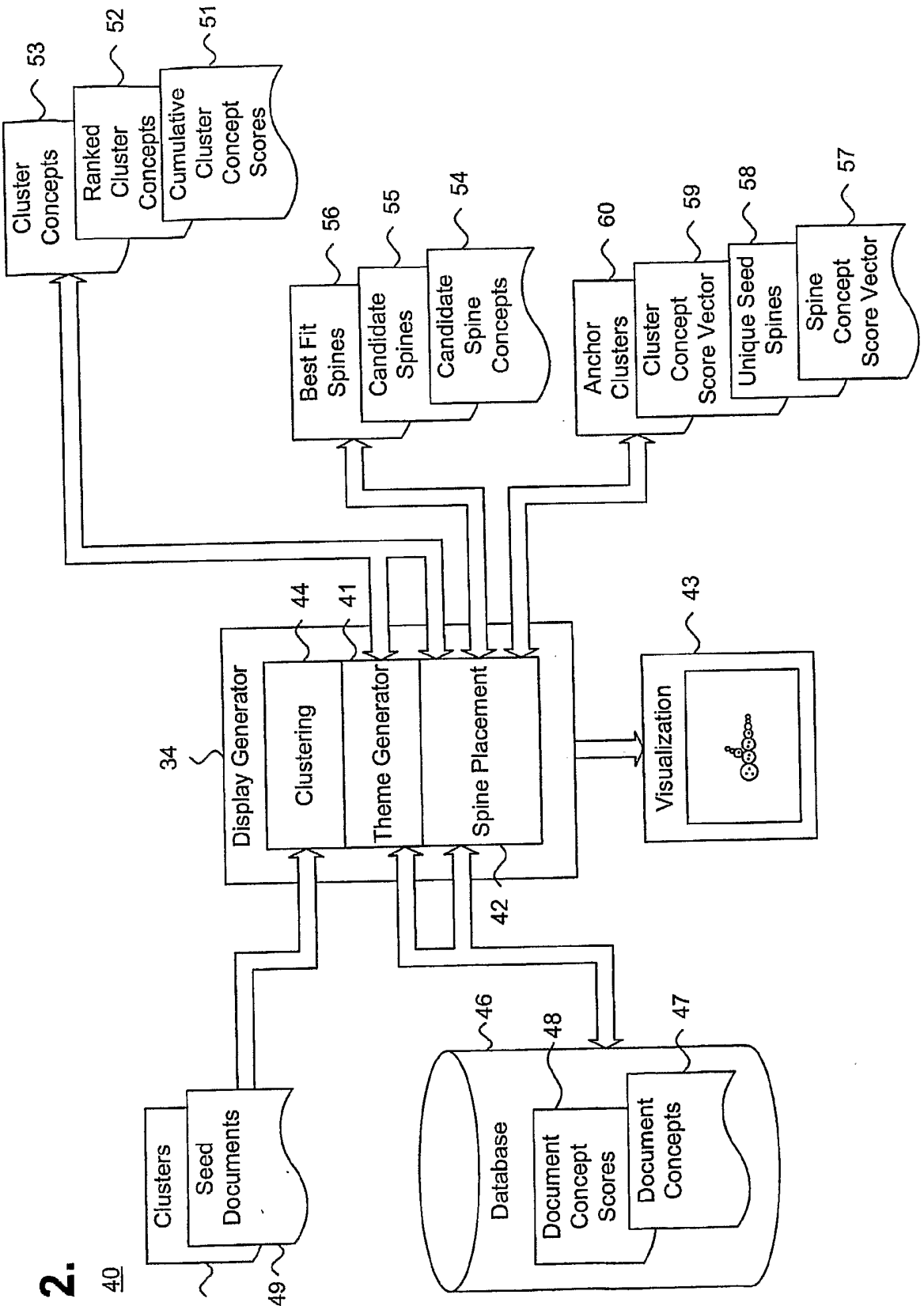

Cluster Concepts
Ranked Cluster Concepts
Cumulative Cluster Concept Scores
Best Fit Spines
Candidate Spines
Candidate Spine Concepts
Anchor Clusters
Cluster Concept Score Vector
Unique Seed Spines
Spine Concept Score Vector
Display Generator
Clustering
Theme Generator
Spine Placement
Visualization
Clusters
Seed Documents
Database
Document Concept Scores
Document Concepts
53
52
51
56
55
54
60
59
58
57
34
44
41
43
42
46
48
47
50
49
40

Fig. 2.

Fig. 3.

100

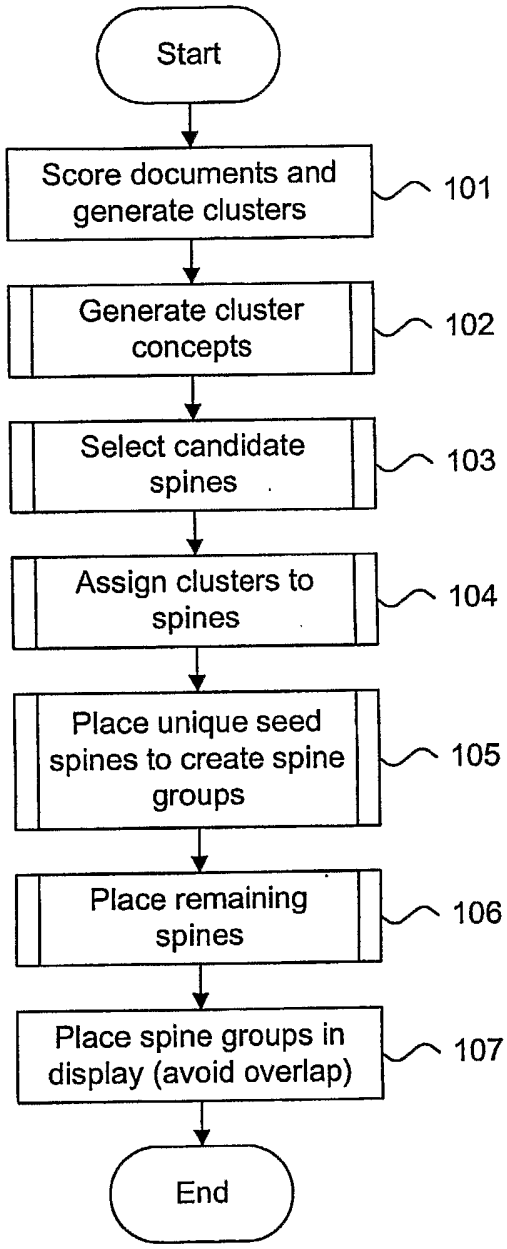

Start
Score documents and generate clusters
101
Generate cluster concepts
102
Select candidate spines
103
Assign clusters to spines
104
Place unique seed spines to create spine groups
105
Place remaining spines
106
Place spine groups in display (avoid overlap)
107
End

Fig. 4.

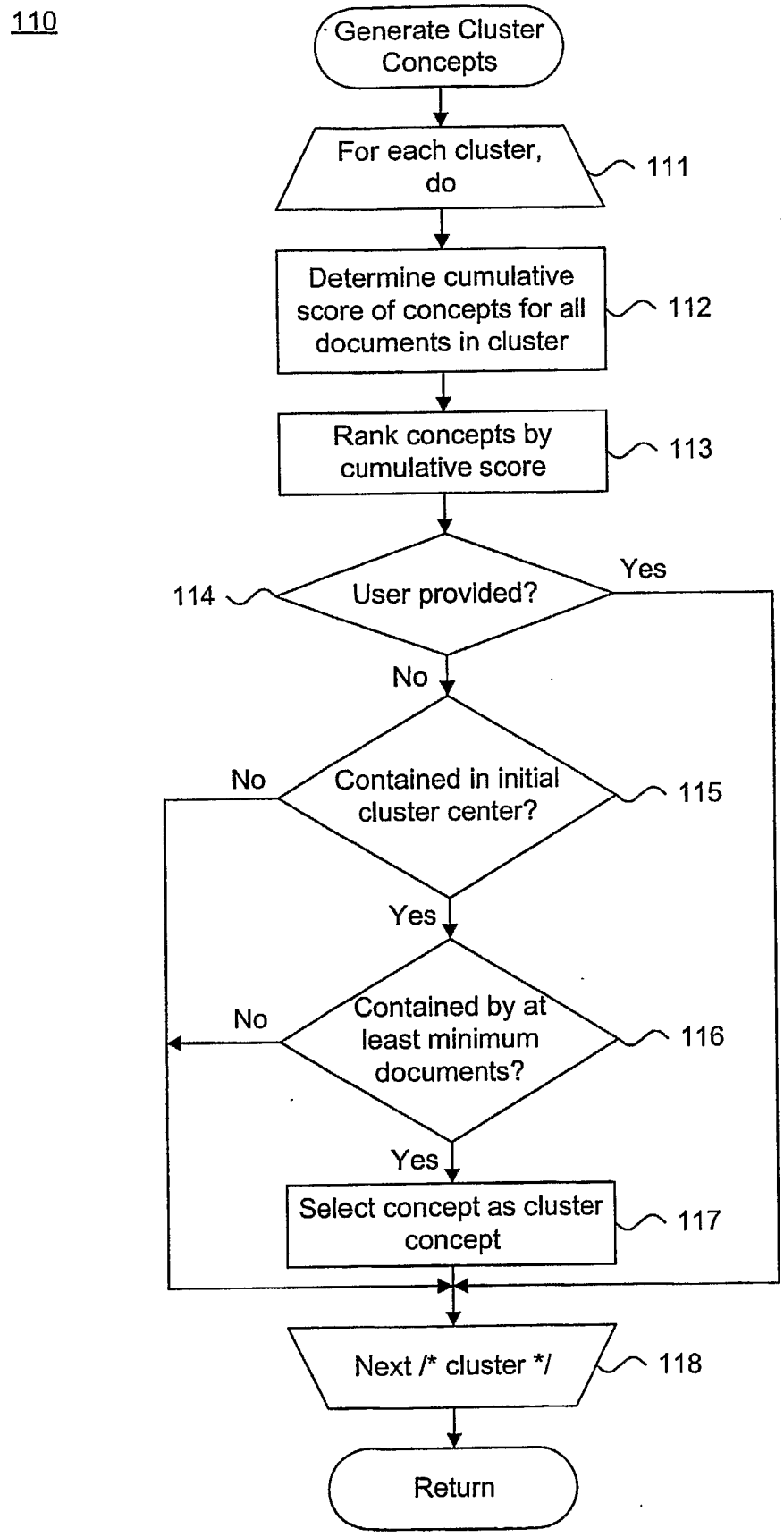

110
Generate Cluster Concepts
For each cluster, do
111
Determine cumulative score of concepts for all documents in cluster
112
Rank concepts by cumulative score
113
114
User provided?
Yes
No
No
Contained in initial cluster center?
115
Yes
No
Contained by at least minimum documents?
116
Yes
Select concept as cluster concept
117
Next /* cluster */
118
Return

Fig. 5.

120

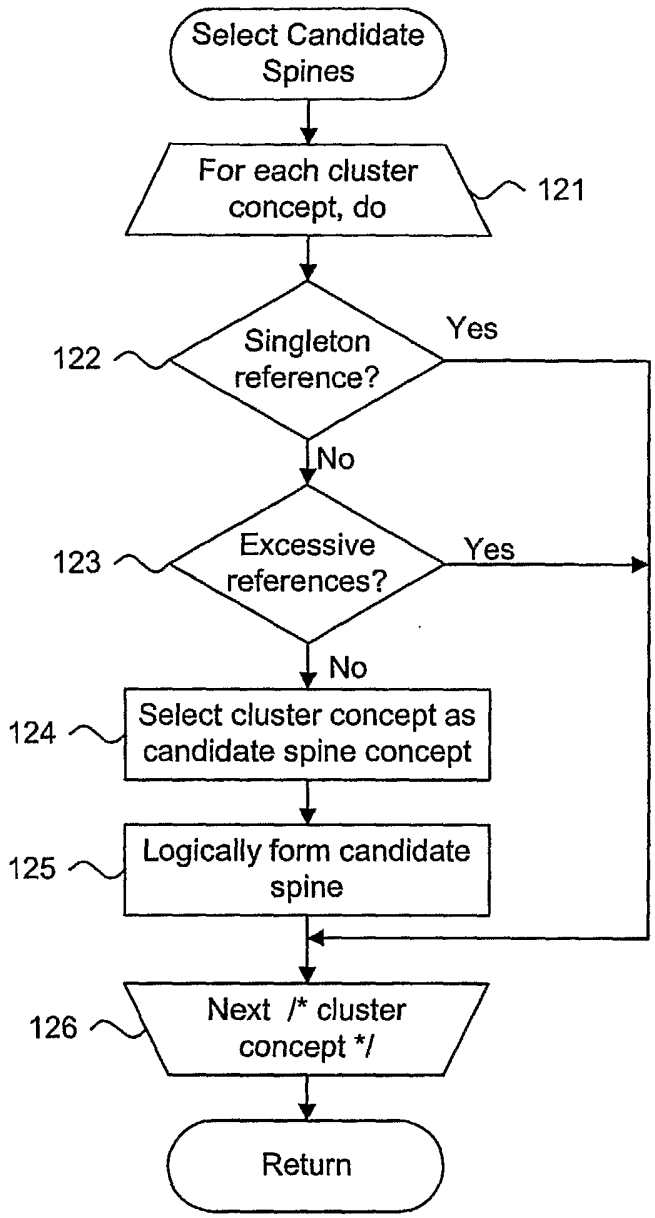

Select Candidate Spines
For each cluster concept, do
121
Singleton reference?
122
Yes
No
Excessive references?
123
Yes
No
Select cluster concept as candidate spine concept
124
Logically form candidate spine
125
Next /* cluster concept */
126
Return

Fig. 6.

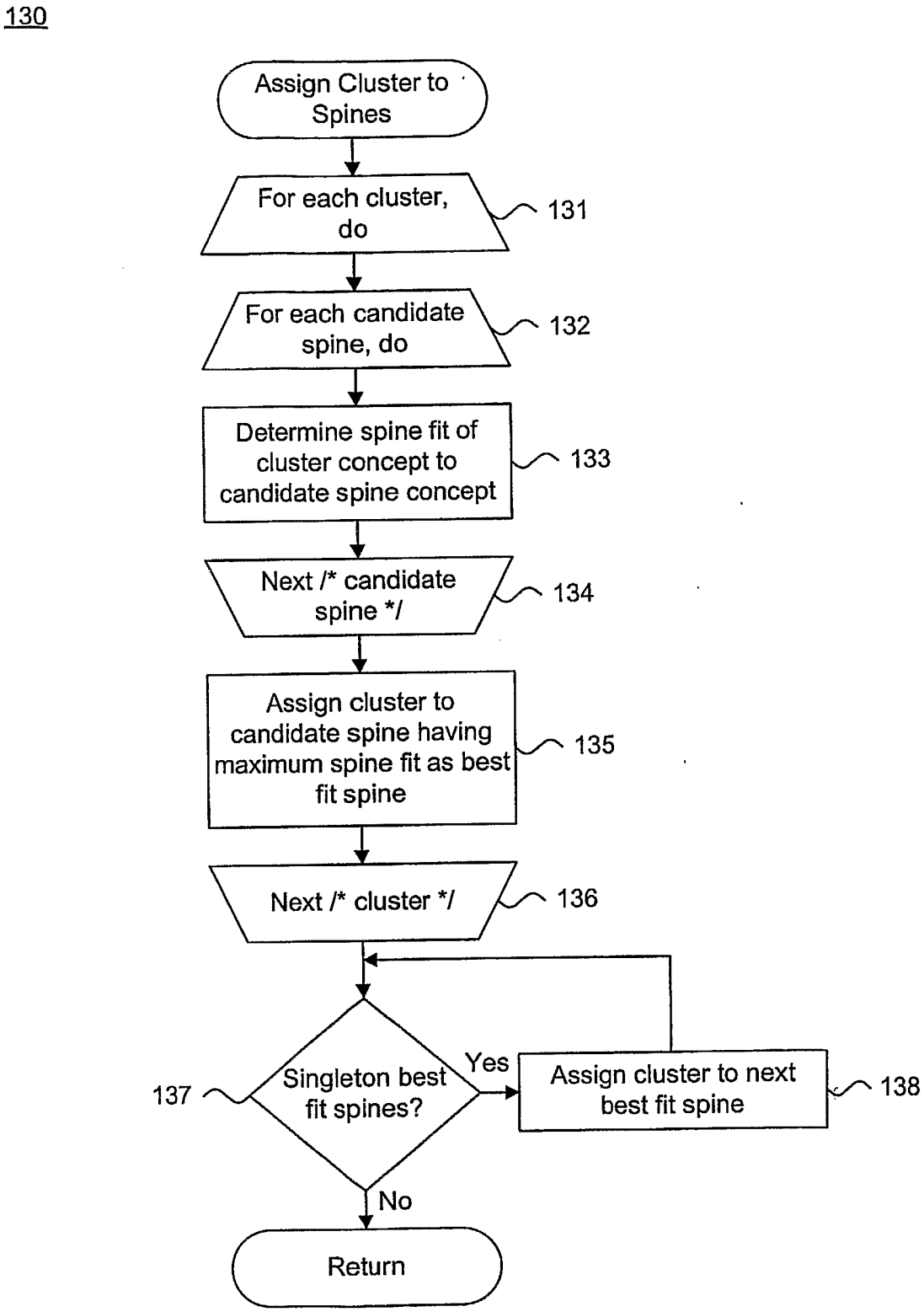

130
Assign Cluster to Spines
For each cluster, do
131
For each candidate spine, do
132
Determine spine fit of cluster concept to candidate spine concept
133
Next /* candidate spine */
134
Assign cluster to candidate spine having maximum spine fit as best fit spine
135
Next /* cluster */
136
Singleton best fit spines?
137
Yes
Assign cluster to next best fit spine
138
No
Return

Fig. 7.

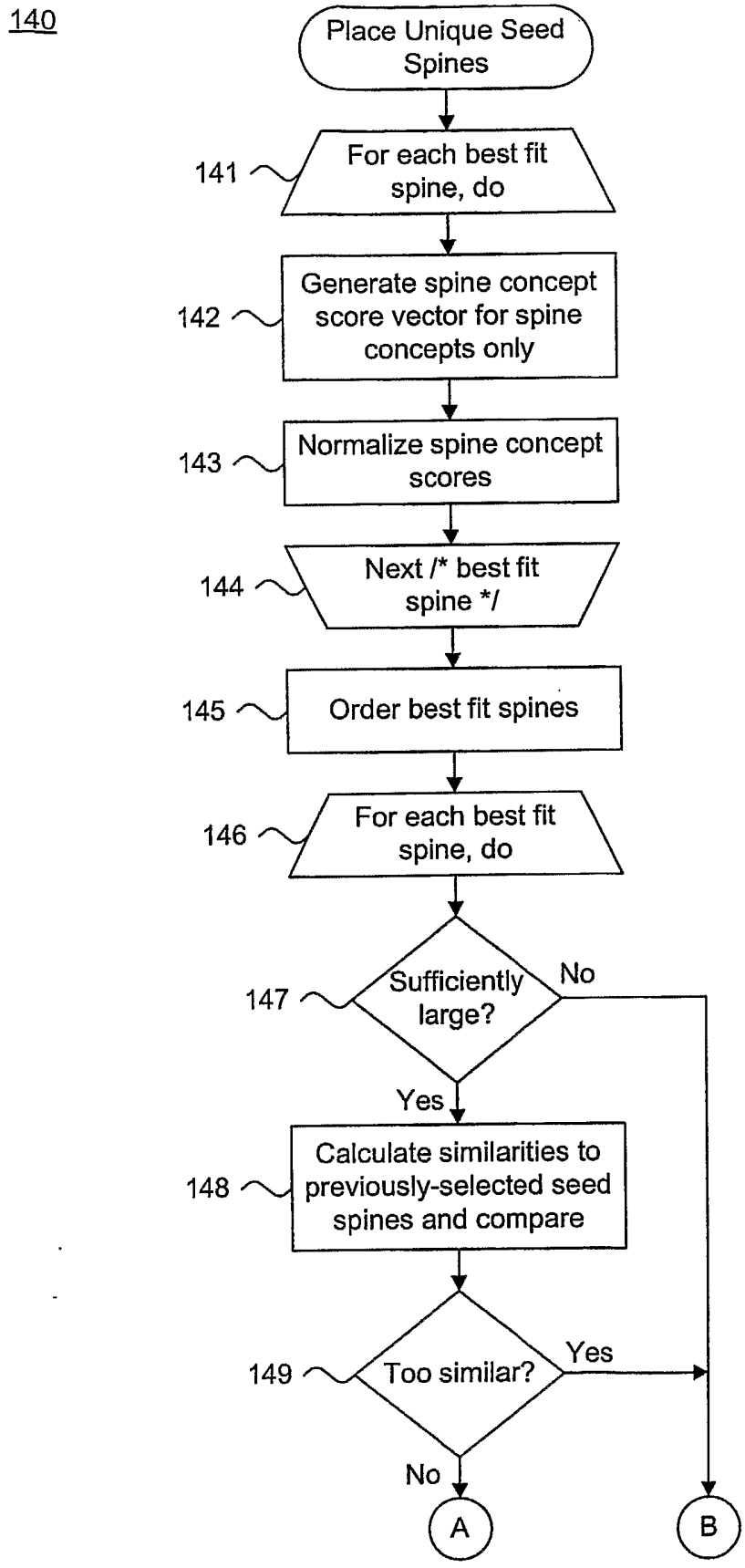

140
Place Unique Seed Spines
141
For each best fit spine, do
142
Generate spine concept score vector for spine concepts only
143
Normalize spine concept scores
144
Next /* best fit spine */
145
Order best fit spines
146
For each best fit spine, do
147
Sufficiently large?
No
Yes
148
Calculate similarities to previously-selected seed spines and compare
149
Too similar?
Yes
No
A
B

Fig. 7 (Cont).

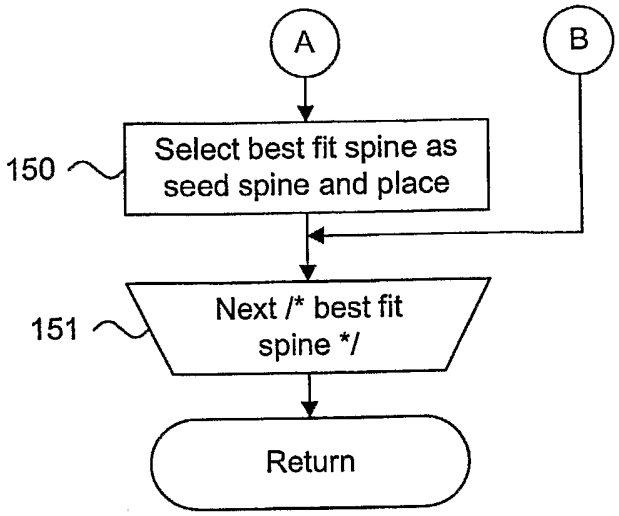


Fig. 8 (Cont).

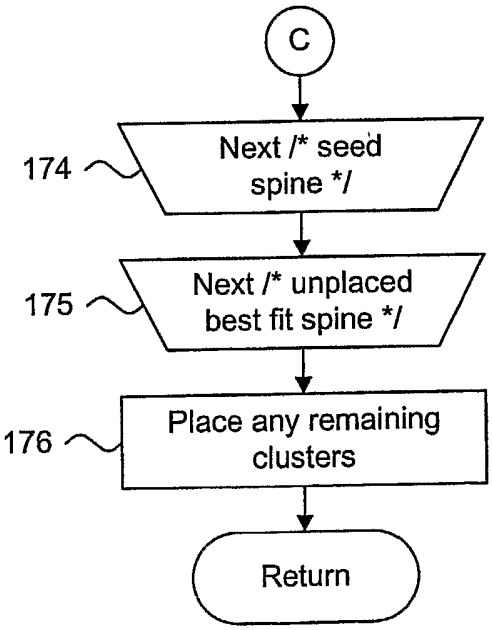


Fig. 8.

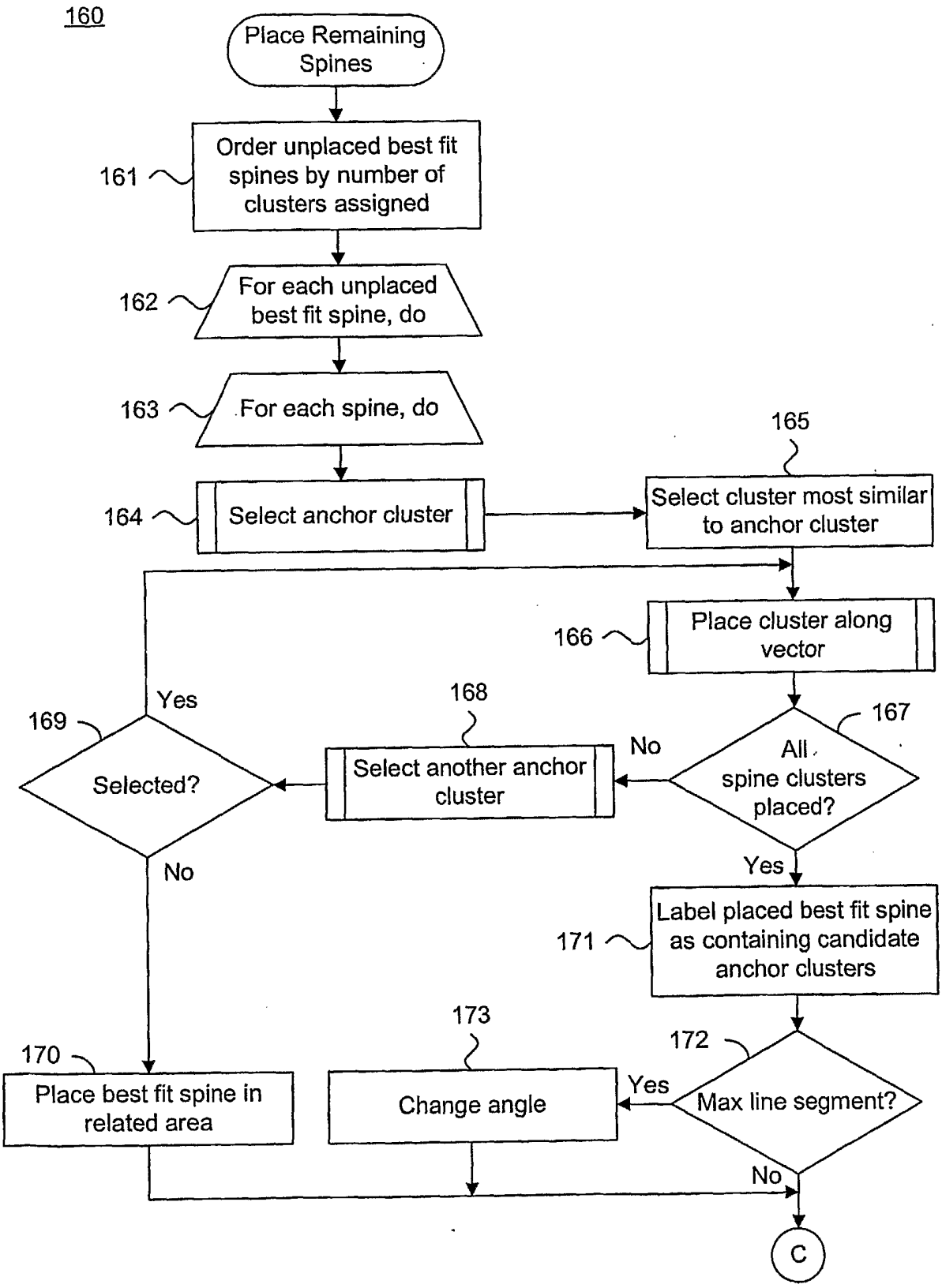
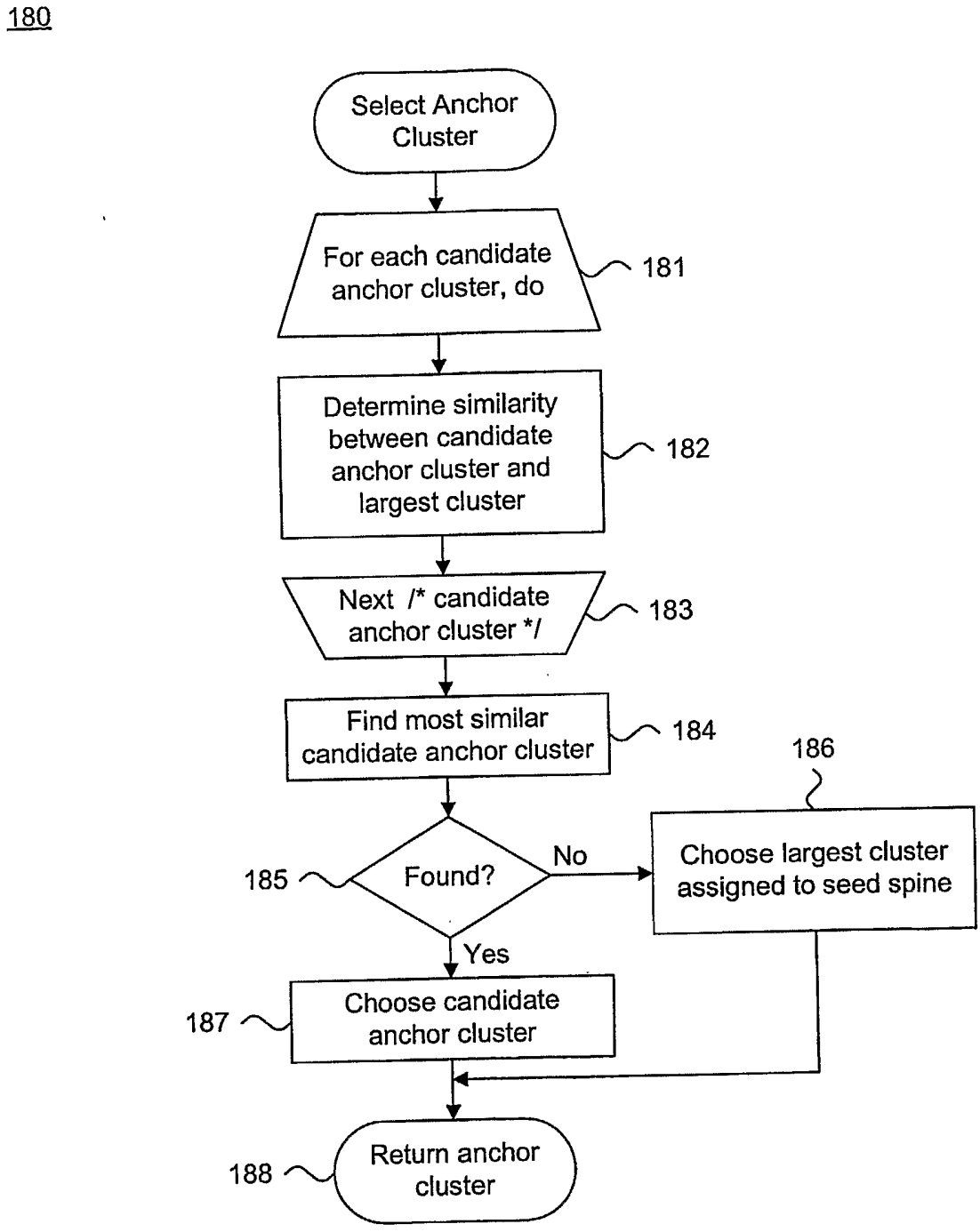

160
Place Remaining Spines
161
Order unplaced best fit spines by number of clusters assigned
162
For each unplaced best fit spine, do
163
For each spine, do
164
Select anchor cluster
165
Select cluster most similar to anchor cluster
166
Place cluster along vector
167
All spine clusters placed?
No
168
Select another anchor cluster
169
Selected?
Yes
No
Yes
171
Label placed best fit spine as containing candidate anchor clusters
172
Max line segment?
Yes
173
Change angle
No
170
Place best fit spine in related area
C

Fig. 9.


180
Select Anchor Cluster
For each candidate anchor cluster, do
181
Determine similarity between candidate anchor cluster and largest cluster
182
Next /* candidate anchor cluster */
183
Find most similar candidate anchor cluster
184
186
185
Found?
No
Choose largest cluster assigned to seed spine
Yes
187
Choose candidate anchor cluster
188
Return anchor cluster

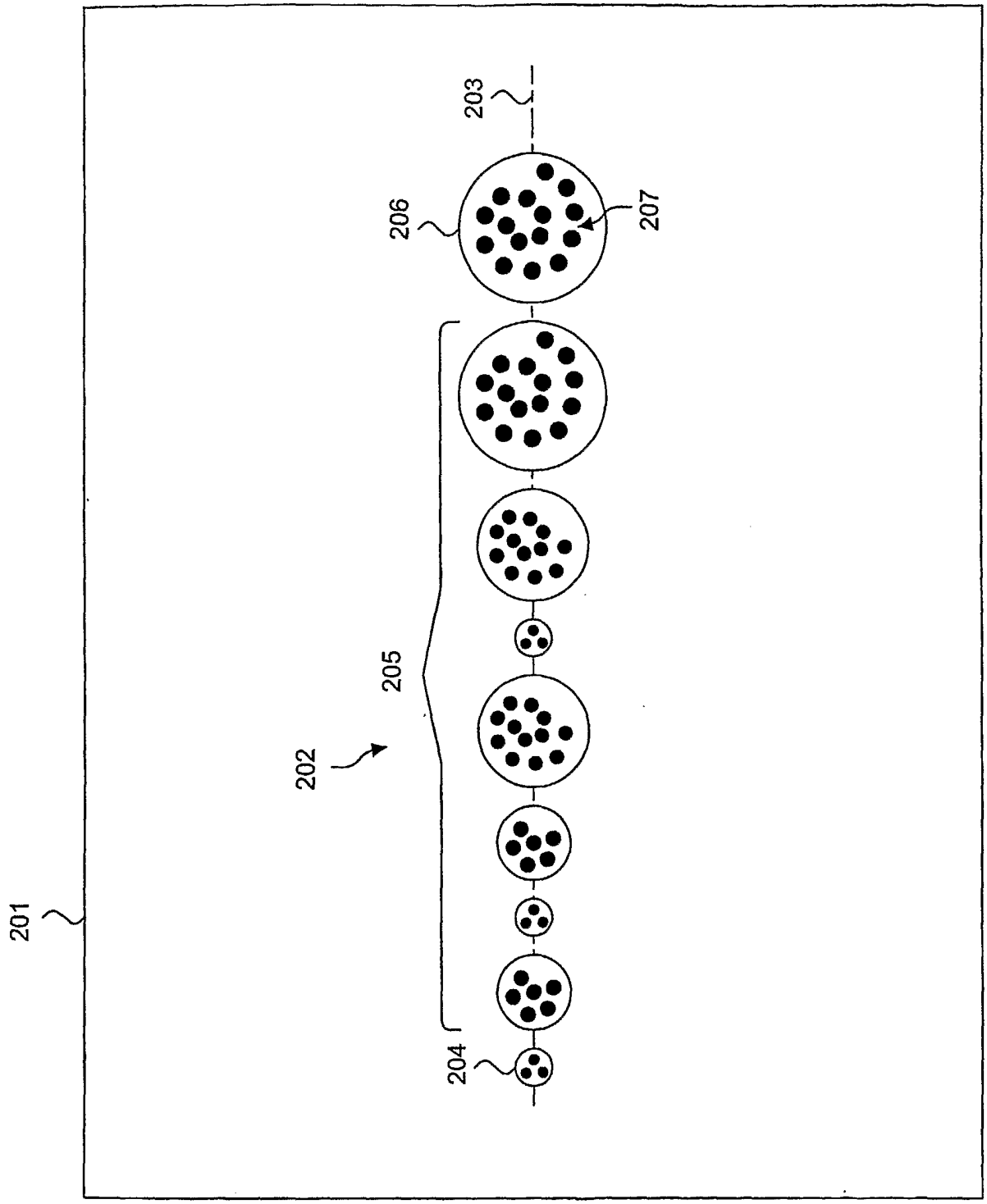


Fig. 10.

Fig. 11A.

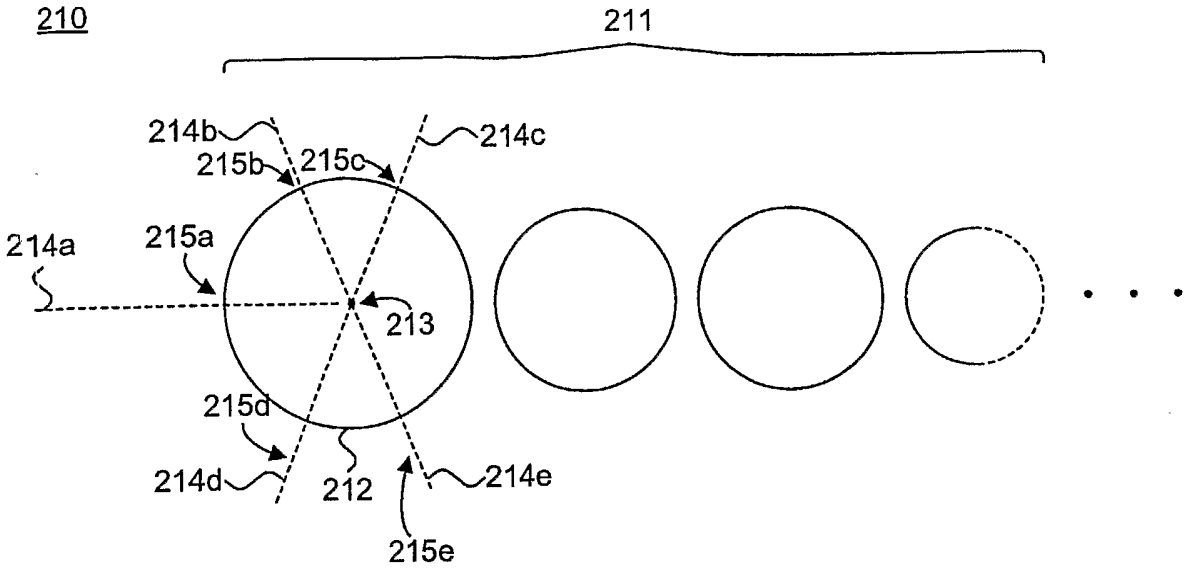


Fig. 11B.

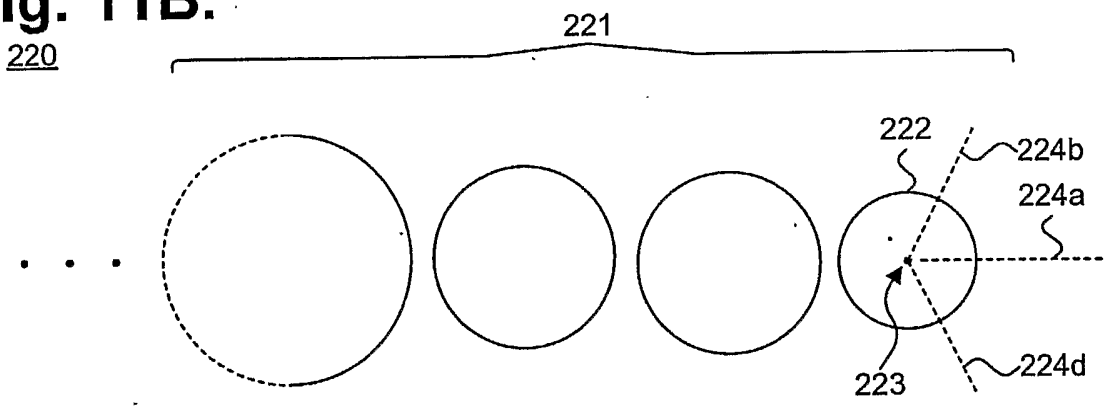


Fig. 11C.

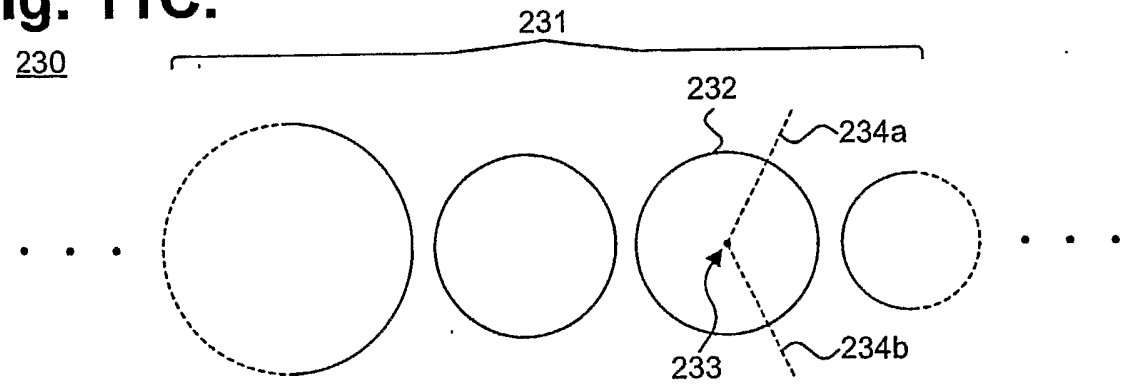
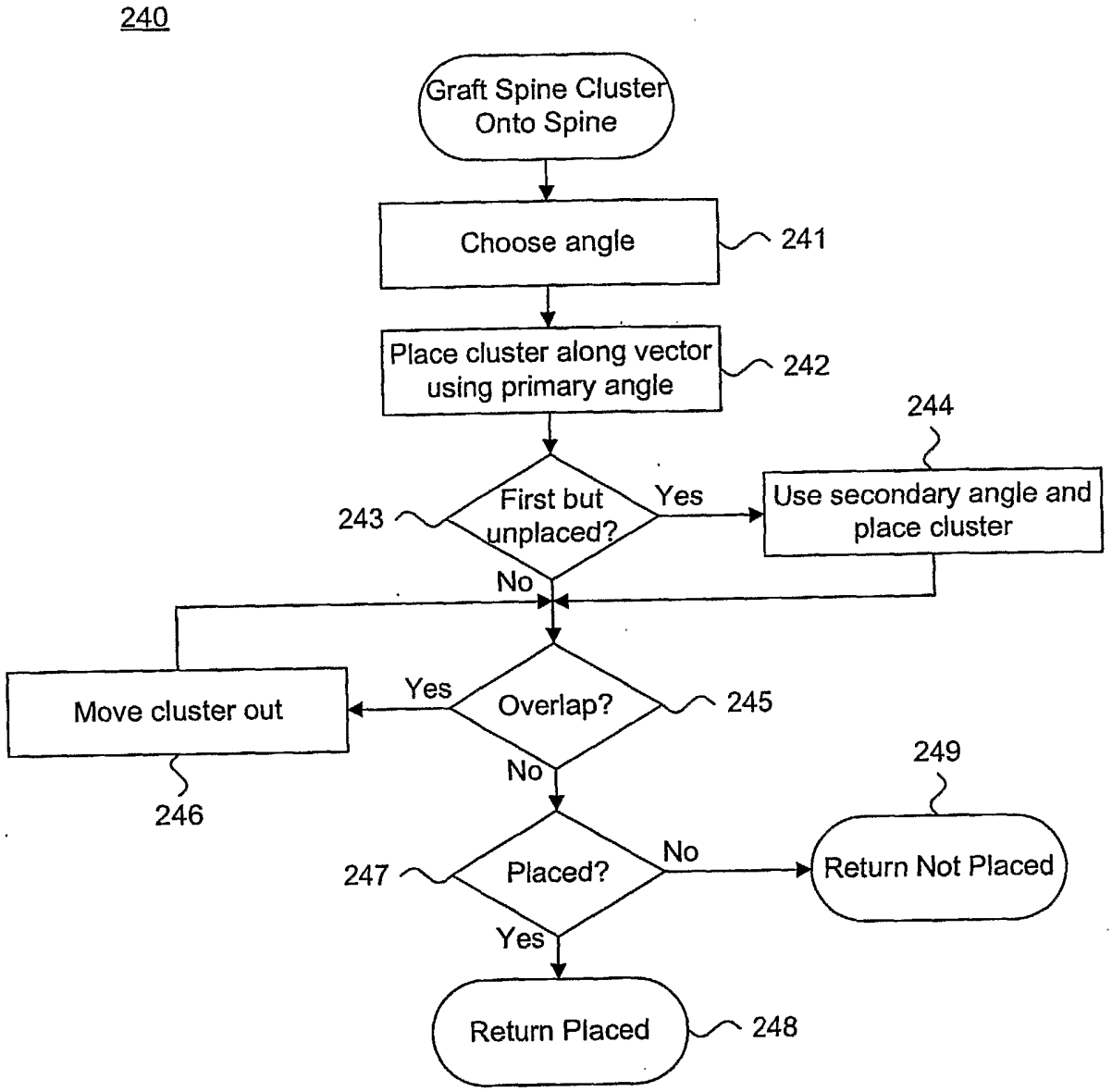


Fig. 12.


240
Graft Spine Cluster Onto Spine
Choose angle
241
Place cluster along vector using primary angle
242
244
243
First but unplaced?
Yes
Use secondary angle and place cluster
No
Move cluster out
Yes
Overlap?
245
No
249
246
247
Placed?
No
Return Not Placed
Yes
Return Placed
248

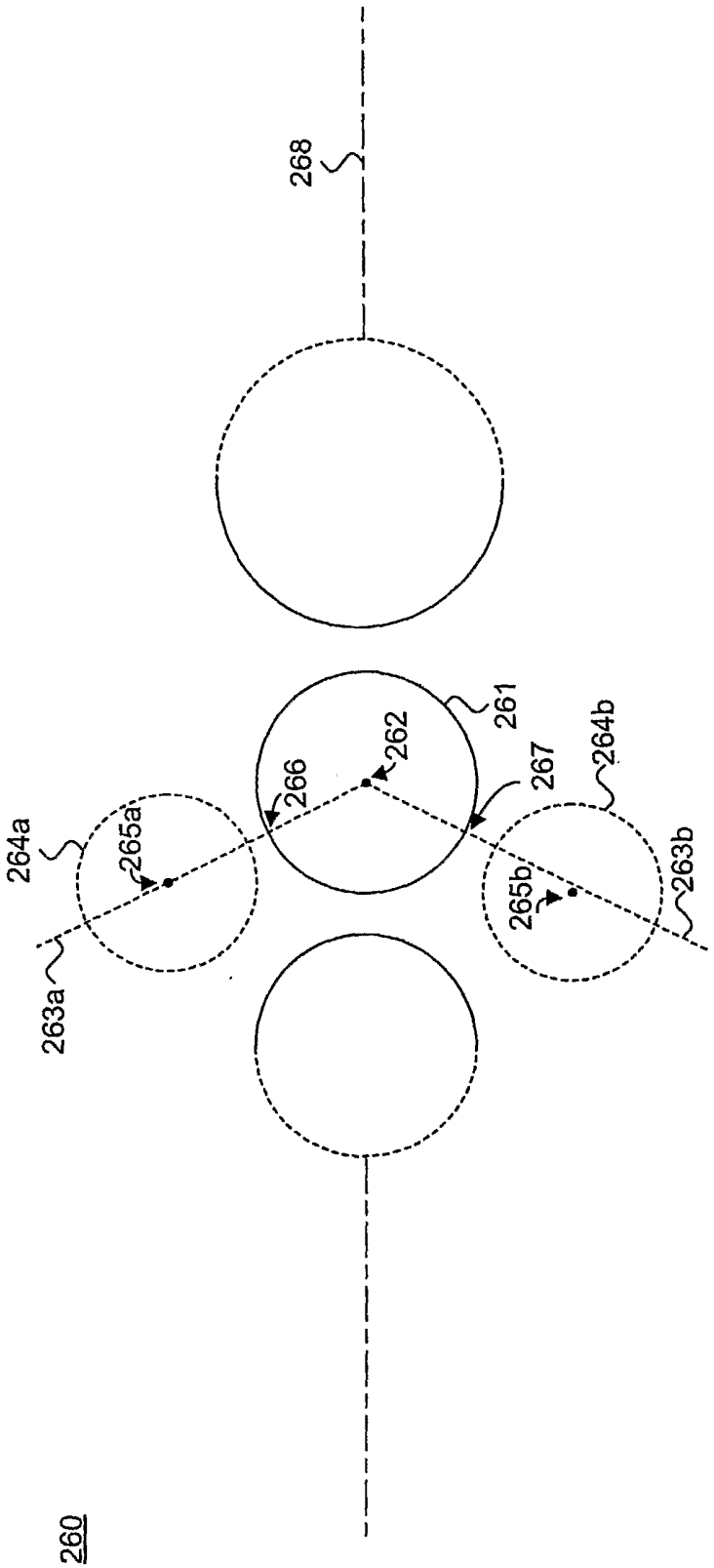
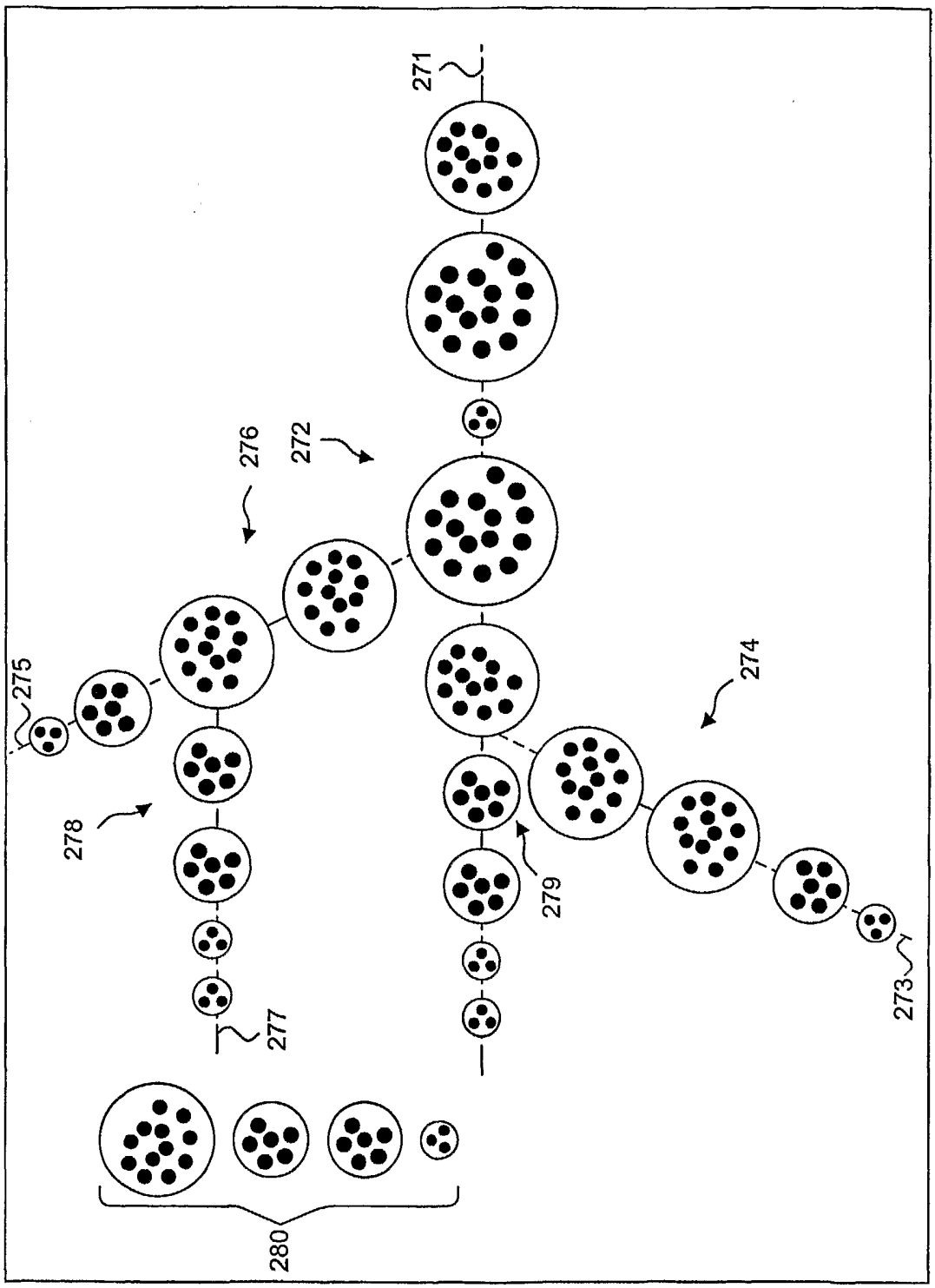


Fig. 13.

Fig. 14.



INTERNATIONAL SEARCH REPORT

PCT/US2005/004241

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 03/073331 A (ATTENEX CORPORATION) 4 September 2003 (2003-09-04) cited in the application page 1, lines 7-9 page 3, lines 15-17 page 3, lines 23-26 page 7, lines 7-17 page 8, lines 24-30 ----- -/--	1-38

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

[X] Patent family members are listed in annex.

° Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
4 July 2005	08/07/2005

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Haffner, R

INTERNATIONAL SEARCH REPORT

PCT/US2005/004241

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CAN F: "Incremental Clustering for Dynamic Information Processing" ACM TRANSACTIONS ON INFORMATION SYSTEMS, ASSOCIATION FOR COMPUTING MACHINERY, NEW YORK, US, vol. 11, no. 2, April 1993 (1993-04), pages 143-164, XP002308022 ISSN: 1046-8188 page 144, paragraph 1-4 page 150, lines 1-5 page 153, paragraph 5 page 157, paragraph 5.4	1-38
A	US 6 449 612 B1 (BRADLEY PAUL S ET AL) 10 September 2002 (2002-09-10) column 3, lines 4-7	1-38
A	US 6 026 397 A (SHEPPARD ET AL) 15 February 2000 (2000-02-15) column 13, lines 14-18	1-38
A	US 2003/130991 A1 (REIJERSE FIDEL ET AL) 10 July 2003 (2003-07-10) column 2, paragraph 17-19	1-38

INTERNATIONAL SEARCH REPORT

PCT/US2005/004241

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 03073331 A	04-09-2003	US 2004172600 A1	02-09-2004
		AU 2003219883 A1	09-09-2003
		CA 2477453 A1	04-09-2003
		EP 1479011 A2	24-11-2004
		WO 03073331 A2	04-09-2003
US 6449612 B1	10-09-2002	US 6263337 B1	17-07-2001
		US 6374251 B1	16-04-2002
		EP 1090362 A1	11-04-2001
		WO 9962007 A1	02-12-1999
		US 6581058 B1	17-06-2003
		EP 1062590 A1	27-12-2000
		WO 9948018 A1	23-09-1999
US 6026397 A	15-02-2000	NONE	
US 2003130991 A1	10-07-2003	WO 02080022 A2	10-10-2002
		WO 02080079 A2	10-10-2002