(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)

PCT

(10) International Publication Number
WO 2011/017065 A1

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

(21) International Application Number:
PCT/US2010/043292

(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
12/833,860	9 July 2010 (09.07.2010)	US

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

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **KNIGHT, William C.** [US/US]; 1851 Edna Place, Bainbridge Island, WA 98110 (US). **NUSSBAUM, Nicholas I.** [US/US]; 8114 41st Ave. SW, Seattle, WA 98136 (US).

(74) Agents: **WITTMAN, Krista A.** et al.; 500 Union Sreet, Suite 1005, Seattle, WA 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).

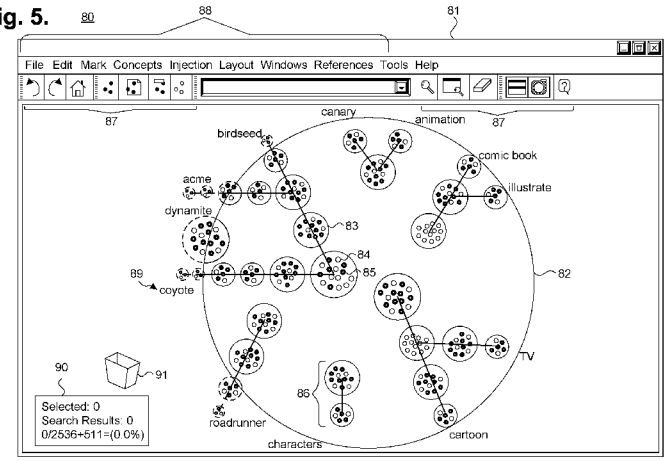
Published:

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

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

(54) **Title**: DISPLAYING RELATIONSHIPS BETWEEN ELECTRONICALLY STORED INFORMATION TO PROVIDE CLASSIFICATION SUGGESTIONS VIA INCLUSION

Fig. 5.



(57) **Abstract**: A system (11) and method (40) for providing reference documents (14b) as a suggestion for classifying uncoded documents (14a) is provided. A set of reference electronically stored information items (14b), each associated with a classification code, is designated. One or more of the reference electronically stored information items (14b) is combined with a set of uncoded electronically stored information items (14a). Clusters (83) of the uncoded electronically stored information items (14a) and the one or more reference electronically stored information items (14b) are generated. Relationships between the uncoded electronically stored information items (14a) and the one or more reference electronically stored information items (14b) in at least one cluster (83) are visually depicted as suggestions for classifying the uncoded electronically stored information items (14a) in that cluster (83).

WO 2011/017065 A1

PCT Patent Application
Docket No. 013.0855.PC.UTL

DISPLAYING RELATIONSHIPS BETWEEN ELECTRONICALLY STORED INFORMATION TO PROVIDE CLASSIFICATION SUGGESTIONS VIA INCLUSION

TECHNICAL FIELD

This application relates in general to using electronically stored information as a reference point and, in particular, to a system and method for displaying relationships between electronically stored information to provide classification suggestions via inclusion.

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 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 can affect the disposition of each document, including admissibility into evidence.

During 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 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, classification 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 growth of ESI exceeds the bounds reasonable for conventional manual human document 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, custom programs ESI review tools, which conduct semi-automated document review through multiple passes over a 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 review requires *a priori* project-specific knowledge engineering, which is only useful for 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 ESI and uncoded ESI and providing a suggestion for classification based on the relationships. The reference ESI and uncoded ESI are clustered based on a similarity of the ESI. The clusters and the relationship between the uncoded ESI and reference ESI within the clusters are visually depicted. The visual relationship of the uncoded ESI and reference ESI provide a suggestion regarding classification for the uncoded ESI.

An embodiment provides a system and method for identifying relationships between electronically stored information to provide a classification suggestion via inclusion. A set of reference electronically stored information items, each associated with a classification code, is designated. One or more of the reference electronically stored information items is combined with a set of uncoded electronically stored information items. Clusters of the uncoded electronically stored information items and the one or more reference electronically stored information items are generated. Relationships between the uncoded electronically stored information items and the one or more reference electronically stored information items in at least one cluster are visually depicted as suggestions for classifying the uncoded electronically stored information items in that cluster.

A further embodiment provides a system and method for clustering reference documents to generate suggestions for classification of uncoded documents. A set of reference documents, each associated with a classification, is designated. One or more of the reference documents are selected and combined with uncoded documents as a set of documents. Clusters of the documents in the document set are generated. A similarity between each document is determined. The documents are grouped into the clusters based on the similarity. At least one cluster having reference documents is identified. Relationships between the uncoded documents and the one or more reference documents in the at least one cluster are visually depicted as suggestions for classifying the uncoded electronically stored information items in that cluster.

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, 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 THE DRAWINGS

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

FIGURE 2 is a process flow diagram showing a method for displaying relationships between electronically stored information to provide classification suggestions via inclusion, in accordance with one embodiment.

FIGURE 3 is a block diagram showing, by way of example, measures for selecting reference document subsets for use in the method of FIGURE 2.

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 screenshot showing, by way of example, a visual display of reference documents in relation to uncoded documents.

FIGURE 6A is a block diagram showing, by way of example, a cluster with "privileged" reference documents and uncoded documents.

FIGURE 6B is a block diagram showing, by way of example, a cluster with "non-responsive" reference documents and uncoded documents.

FIGURE 6C is a block diagram showing, by way of example, a cluster with uncoded documents and a combination of differently classified reference documents.

FIGURE 7 is a process flow diagram showing, by way of example, a method for classifying uncoded documents for use in the method of FIGURE 2.

FIGURE 8 is a screenshot showing, by way of example, a reference options dialogue box for entering user preferences for clustering documents.

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. Previously coded ESI, known as reference ESI, offer knowledge gleaned from earlier work in similar legal projects, as well as a reference point for classifying uncoded ESI.

Reference ESI is previously classified by content and can be used to influence classification of uncoded, that is unclassified, ESI. Specifically, relationships between the uncoded ESI and the reference ESI can be visually depicted to provide suggestions, for instance to a human reviewer, for classifying the visually-proximal uncoded ESI.

Complete ESI review requires a support environment within which classification can be performed. FIGURE 1 is a block diagram showing a system 10 for displaying relationships between electronically stored information to provide classification suggestions via inclusion, 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, and a lookup database 38 for storing many-to-many mappings 39 between documents and document features, such as concepts. The storage device 13 also stores reference documents 14b, which can provide a training set of trusted and known results for use in guiding ESI classification. The reference documents 14b are each associated with an assigned classification code and considered as classified or coded. Hereinafter, the terms "classified" and "coded" are used interchangeably with the same intended meaning, unless otherwise indicated. A set of reference documents can be hand-selected or automatically selected through guided review, which is further discussed below. Additionally, the set of reference documents can be predetermined or can be generated dynamically, as uncoded documents are classified and subsequently added to the set of reference documents.

The backend server 11 is coupled to an intranetwork 21 and executes a workbench 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 document scoring and clustering of uncoded documents and reference documents, such as described in commonly-assigned U.S. Patent No. 7,610,313, the disclosure of which is incorporated by reference. Clusters of uncoded documents 14a and reference documents 14b are formed and organized along vectors, known as spines, based on a similarity of the clusters. The similarity can be expressed in terms of distance.

Document clustering is further discussed below with reference to FIGURE 4. The classifier 35 provides a machine-generated suggestion and confidence level for classification of selected uncoded documents 14b, clusters, or spines, as further described below with reference to FIGURE 7.

The display generator 36 arranges the clusters and spines in thematic relationships in a two-dimensional visual display space, 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 and classify one or more uncoded documents by designating a code. Hereinafter, the terms "reviewer" and "custodian" are used interchangeably with the same intended meaning, unless otherwise indicated. Other types of reviewers are possible, including machine-implemented reviewers.

The document mapper 32 operates on uncoded documents 14a, which can be retrieved from the storage 13, as well as from a plurality of local and remote sources. As well, the local and remote sources can also store the reference documents 14b. 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 backend server 11 and the work client 12 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. Other document sources, either local or remote, are 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 stored in a spreadsheet or database.

In one embodiment, the individual documents 14a, 14b, 17, 20, 26, 29 include electronic message folders storing email and attachments, such as maintained by the Outlook and Outlook Express products, licensed by Microsoft Corporation, Redmond, WA. The database can be an SQL-based relational database, such as the Oracle database management system, Release 8, licensed by Oracle Corporation, Redwood Shores, CA.

The individual documents 17, 20, 26, 29 can be designated and stored as uncoded documents or reference documents. One or more of the uncoded documents can be selected for a document review project and stored as a document corpus, as described *infra*. The reference documents are initially uncoded documents that can be selected from the corpus or other source of uncoded documents, and subsequently classified. The reference documents can assist in providing suggestions for classification of the remaining uncoded documents in the corpus based on visual relationships between the uncoded documents and reference documents. In a further embodiment, the reference documents can provide suggestions for classifying uncoded documents in a different corpus. In yet a further embodiment, the reference documents can be used as a training set to form machine-generated suggestions for classifying uncoded documents, as further described below with reference to FIGURE 8.

The document corpus for a document review project can be divided into subsets of uncoded documents, which are each provided to a particular reviewer as an assignment. To maintain consistency, the same classification codes can be used across all assignments in the document review project. Alternatively, the classification codes can be different for each assignment. The classification codes can be determined using taxonomy generation, during which a list of classification codes can be provided by a reviewer or determined automatically. For purposes of legal discovery, the list of classification codes can include "privileged," "responsive," or "non-responsive;" however, other classification codes 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 unintentional waivers of the subject matter disclosed. A "responsive" document contains information that is related to a legal matter on which the document review project is based and a "non-responsive" document includes information that is not related to the legal matter.

The system 10 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 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 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 relationships between the reference documents and uncoded documents includes clustering. FIGURE 2 is a process flow diagram showing a method 40 for displaying relationships between electronically stored information to provide classification suggestions via inclusion, in accordance with one embodiment. A subset of reference documents is identified and selected (block 41) from a representative set of reference documents. The subset of reference documents can be predefined, arbitrary, or specifically selected, as discussed further below with reference to FIGURE 3. Upon identification, the reference document subset is grouped with uncoded documents (block 42). The uncoded documents can include all uncoded documents in an assignment or in a corpus. The grouped documents, including uncoded and reference documents are organized into clusters (block 43). Clustering of the documents is discussed further below with reference to FIGURE 4.

Once formed, the clusters can be displayed to visually depict relationships (block 44) between the uncoded documents and the reference documents. The relationships can provide a suggestion, which can be used by an individual reviewer for classifying one or more of the uncoded documents, clusters, or spines. Based on the relationships, the reviewer can classify the uncoded documents, clusters, or spines by assigning a classification code, which can represent a relevancy of the uncoded document to the document review project. Further, machine classification can provide a suggestion for classification, including a classification code, based on a calculated confidence level (block 45). Classifying uncoded documents is further discussed below with reference to FIGURE 7.

Prior to clustering, the uncoded documents and reference documents are obtained. The reference documents used for clustering can include a particular subset of reference documents, which are selected from a general set of reference documents. Alternatively, the entire set of reference documents can be clustered with the uncoded documents. The set of reference documents is representative of the document corpus for a document review project in which data organization or classification is desired. The reference document set can be previously defined and maintained for related document review projects or can be specifically generated for each review project. A predefined reference set provides knowledge previously obtained during the related document review project to increase efficiency, accuracy, and consistency. Reference sets newly generated for each review project can include arbitrary or customized reference sets that are determined by a reviewer or a machine.

The set of reference documents can be generated during guided review, which assists a reviewer in building a reference document set. During guided review, the uncoded documents that are dissimilar to the other uncoded documents are identified based on a similarity threshold. Other methods for determining dissimilarity are possible. Identifying a set of dissimilar documents provides a group of uncoded documents that is representative of the corpus for the document review project. Each identified dissimilar document is then classified by assigning a particular classification code based on the content of the document to collectively generate a set of reference documents. Guided review can be performed by a reviewer, a machine, or a combination of the reviewer and machine.

Other methods for generating a reference document set for a document review project using guided review are possible, including clustering. For example, a set of uncoded documents to be classified is clustered, as described in commonly-assigned U.S. Patent No. 7,610,313, the disclosure of which is incorporated by reference. A plurality of the clustered uncoded documents are selected based on selection criteria, such as cluster centers or sample clusters. The cluster centers can be used to identify uncoded documents in a cluster that are most similar or dissimilar to the cluster center. The identified uncoded documents are then selected for classification by assigning classification codes. After classification, the documents represent a reference set. In a further embodiment, sample clusters can be used to generate a reference document set by selecting one or more sample clusters based on cluster relation criteria, such as size, content, similarity, or dissimilarity. The uncoded documents in the selected sample clusters are then assigned classification codes. The classified documents represent a document reference set for the document review project. Other methods for selecting documents for use as a reference set are possible.

Once generated, a subset of reference documents is selected from the reference document set for clustering with uncoded documents. FIGURE 3 is a block diagram showing, by way of example, measures 50 for selecting reference document subsets 51 for use in the method of FIGURE 2. A reference document subset 51 includes one or more reference documents selected from a set of reference documents associated with a document review project for use in clustering with uncoded documents. The reference document subset can be predefined 52, customized 54, selected arbitrarily 53, or based on similarity 55.

A subset of predefined reference documents 52 can be selected from a reference set, which is associated with another document review project that is related to the current document review project. An arbitrary reference subset 53 includes reference documents randomly selected from a reference set, which can be predefined or newly generated for the current

document review project or a related document review project. A customized reference subset 54 includes reference documents specifically selected from a current or related reference set based on criteria, such as reviewer preference, classification category, document source, content, and review project. Other criteria are possible. The number of reference documents in a subset can be determined automatically or by a reviewer based on reference factors, such as a size of the document review project, an average size of the assignments, types of classification codes, and a number of reference documents associated with each classification code. Other reference factors are possible. In a further embodiment, the reference document subset can include more than one occurrence of a reference document. Other types of reference document subsets and methods for selecting the reference document subsets are possible.

Once identified, the reference document subset can be used for clustering with uncoded documents from a corpus associated with a particular document review project. The corpus of uncoded documents for a review project can be divided into assignments using assignment criteria, such as custodian or source of the uncoded document, content, document type, and date. Other criteria are possible. In one embodiment, each assignment is assigned to an individual reviewer for analysis. The assignments can be separately clustered with the reference document subset or alternatively, all of the uncoded documents in the corpus can be clustered with the reference document subset. The content of each uncoded document within the corpus can be converted into a set of tokens, which are word-level or character-level *n*-grams, raw terms, concepts, or entities. Other tokens are 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. Concepts typically include nouns and noun phrases obtained through part-of-speech tagging that have a common semantic meaning. Entities further refine nouns and noun phrases into people, places, and things, such as meetings, animals, relationships, and various other objects. Entities can be extracted using entity extraction techniques known in the field. Clustering of the uncoded documents can be based on cluster criteria, such as the similarity of tokens, including *n*-grams, raw terms, concepts, entities, email addresses, or other metadata.

Clustering provides groupings of related uncoded documents and reference documents. FIGURE 4 is a flow diagram showing a routine 60 for forming clusters for use in the method 40 of FIGURE 2. The purpose of this routine is to use score vectors associated with the documents, including uncoded and reference documents, to form clusters based on relative similarity. Hereinafter, the term "document" is intended to include uncoded documents and reference

documents selected for clustering, unless otherwise indicated. The score vector associated with each document includes a set of paired values for tokens identified in that document and weights, which are based on scores. The score vector is generated by scoring the tokens extracted from each uncoded document and reference document, as described in commonly-assigned U.S. Patent No. 7,610,313, the disclosure of which is incorporated by reference.

As an initial step for generating score vectors, each token within a document is individually scored. Next, a normalized score vector is created for the document by identifying paired values, consisting of a token occurring in that document and the scores for that token. The paired values are ordered along a vector to generate the score vector. The paired values can be ordered based on the tokens, including concept or frequency, as well as other factors. For example, assume a normalized score vector for a first document *A* is $\vec{S}_A = \{(5, 0.5), (120, 0.75)\}$ and a normalized score vector for another document *B* is $\vec{S}_B = \{(3, 0.4), (5, 0.75), (47, 0.15)\}$. Document *A* has scores corresponding to tokens '5' and '120' and Document *B* has scores corresponding to tokens '3,' '5' and '47.' Thus, these documents only have token '5' in common. Once generated, the score vectors can be compared to determine similarity or dissimilarity between the corresponding documents during clustering.

The routine for forming clusters of documents, including uncoded documents and reference documents, proceeds in two phases. During the first phase (blocks 63-68), the documents are evaluated to identify a set of seed documents, which can be used to form new clusters. During the second phase (blocks 70-76), any documents not previously placed are evaluated and grouped into the existing clusters based on a best-fit criterion.

Initially, a single cluster is generated with one or more documents as seed documents and additional clusters of documents are added, if necessary. Each cluster is represented by a cluster center that is associated with a score vector, which is representative of the tokens in all the documents for that cluster. In the following discussion relating to FIGURE 4, the tokens include concepts. However, other tokens are possible, as described *supra*. The cluster center score vector can be generated by comparing the score vectors for the individual documents in the cluster and identifying the most common concepts shared by the documents. The most common concepts and associated weights are ordered along the cluster center score vector. Cluster centers and thus, cluster center score vectors may continually change due to the addition and removal of documents during clustering.

During clustering, the documents are identified (block 61) and ordered by length (block 62). The documents can include all reference documents in a subset and one or more assignments of uncoded documents. Each document is then processed in an iterative processing

loop (blocks 63-68) as follows. The similarity between each document and a center of each cluster is determined (block 64) as the cosine (cos) σ of the score vectors for the document and cluster being compared. The cos σ provides a measure of relative similarity or dissimilarity between tokens, including the concepts, in the documents and is equivalent to the inner products between the score vectors for the document and cluster center.

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

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

where $\cos\sigma_{AB}$ comprises the similarity metric between document A and cluster center B, $\vec{S}_A$ comprises a score vector for the document 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.

Only those documents that are sufficiently distinct from all cluster centers (block 65) are selected as seed documents for forming new clusters (block 66). If the document being compared is not sufficiently distinct (block 65), the document is then grouped into a cluster with the most similar cluster center (block 67). Processing continues with the next document (block 68).

In the second phase, each document not previously placed is iteratively processed in an iterative processing loop (blocks 70-76) as follows. Again, the similarity between each remaining document and each of the cluster centers is determined based on a distance (block 71), such as the cos σ of the normalized score vectors for each of the remaining documents and the cluster centers. A best fit between a remaining document and a cluster center can be found subject to a minimum fit criterion (block 72). 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 73), the remaining document is grouped into the cluster having the best fit (block 75). Otherwise, the remaining document is grouped into a miscellaneous cluster (block 74). Processing continues with the next remaining document (block 76). Finally, a dynamic threshold can be applied to each cluster (block 77) 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 No. 7,610,313, the disclosure of which is incorporated by reference. The routine then returns. Other methods and processes for forming clusters are possible.

Once formed, the clusters of documents can be can be organized to generate spines of thematically related clusters, as described in commonly-assigned U.S. Patent No. 7,271,804, the disclosure of which is incorporated by reference. Each spine includes those clusters that share one or more tokens, such as concepts, which are placed along a vector. Also, the cluster spines can be positioned in relation to other cluster spines based on a theme shared by those cluster spines, as described in commonly-assigned U.S. Patent No. 7,610,313, the disclosure of which is incorporated by reference. Each theme can include one or more concepts defining a semantic meaning. Organizing the clusters into spines and groups of cluster spines provides an individual reviewer with a display that presents the documents according to a theme while maximizing the number of relationships depicted between the documents.

FIGURE 5 is a screenshot 80 showing, by way of example, a visual display 81 of reference documents 85 in relation to uncoded documents 84. Clusters 83 can be located along a spine, which is a straight vector, based on a similarity of the documents 84, 85 in the clusters 83. Each cluster 83 is represented by a circle; however, other shapes, such as squares, rectangles, and triangles are possible, as described in U.s. Patent No. 6,888,548, the disclosure of which is incorporated by reference. The uncoded documents 84 are each represented by a smaller circle within the clusters 83, while the reference documents 85 are each represented by a circle having a diamond shape within the boundaries of the circle. The reference documents 85 can be further represented by their assigned classification code. The classification codes can include "privileged," "responsive," and "non-responsive" codes, as well as other codes. Each group of reference documents associated with a particular classification code can be identified by a different color. For instance, "privileged" reference documents can be colored blue, while "non-responsive" reference documents are red and "responsive" reference documents are green. In a further embodiment, the reference documents for different classification codes can include different symbols. For example, "privileged" reference documents can be represented by a circle with an "X" in the center , while "non-responsive" reference documents can include a circle with striped lines and "responsive" reference documents can include a circle with dashed lines. Other classification representations for the reference documents are possible. Each cluster spine 86 is represented as a straight vector along which the clusters are placed.

The display 81 can be manipulated by an individual reviewer via a compass 82, which enables the reviewer to navigate, explore, and search the clusters 83 and spines 86 appearing within the compass 82, as further described in commonly-assigned U.S. Patent No. 7,356,777, the disclosure of which is incorporated by reference. Visually, the compass 82 emphasizes

clusters 83 located within the compass 82, while deemphasizing clusters 83 appearing outside of the compass 82.

Spine labels 89 appear outside of the compass 82 at an end of each cluster spine 86 to connect the outermost cluster of a cluster spine 86 to the closest point along the periphery of the compass 82. In one embodiment, the spine labels 89 are placed without overlap and circumferentially around the compass 82. Each spine label 89 corresponds to one or more concepts that most closely describe the cluster spines 86 appearing within the compass 82. Additionally, the cluster concepts for each of the spine labels 89 can appear in a concepts list (not shown) also provided in the display. Toolbar buttons 87 located at the top of the display 81 enable a user to execute specific commands for the composition of the spine groups displayed. A set of pull down menus 88 provide further control over the placement and manipulation of clusters 83 and cluster spines 86 within the display 81. Other types of controls and functions are possible.

A document guide 90 can be placed within the display 81. The document guide 90 can include a "Selected" field, a "Search Results" field, and details regarding the numbers of uncoded documents and reference documents provided in the display. The number of uncoded documents includes all uncoded documents selected for clustering, such as within a corpus of uncoded documents for a review project or within an assignment. The number of reference documents includes the reference document subset selected for clustering. The "Selected" field in the document guide 90 provides a number of documents 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. The "Search Results" field provides a number of uncoded documents and reference documents that include a particular search term identified by the reviewer in a search query box 92.

In one embodiment, a garbage can 91 is provided to remove tokens, such as cluster concepts, from consideration in the current set of clusters 83. Removed cluster concepts prevent those concepts from affecting future clustering, as may occur when a reviewer considers a concept irrelevant to the clusters 83.

The display 81 provides a visual representation of the relationships between thematically-related documents, including the uncoded documents and reference documents. The uncoded documents and reference documents located within a cluster or spine can be compared based on characteristics, such as the assigned classification codes of the reference documents, a number of reference documents associated with each classification code, and a number of different classification codes to identify relationships between the uncoded documents and reference

documents. The reviewer can use the displayed relationships as suggestions for classifying the uncoded documents. For example, FIGURE 6A is a block diagram showing, by way of example, a cluster 93 with "privileged" reference documents 95 and uncoded documents 94. The cluster 93 includes nine uncoded documents 94 and three reference documents 95. Each reference document 95 is classified as "privileged." Accordingly, based on the number of "privileged" reference documents 95 present in the cluster 93, the absence of other classifications of reference documents, and the thematic relationship between the uncoded documents 94 and the "privileged" reference documents 95, the reviewer may be more inclined to review the uncoded documents 94 in that cluster 93 or to classify one or more of the uncoded documents 94 as "privileged" without review.

Alternatively, the three reference documents can be classified as "non-responsive," instead of "privileged" as in the previous example. FIGURE 6B is a block diagram showing, by way of example, a cluster 96 with "non-responsive" reference documents 97 and uncoded documents 94. The cluster 96 includes nine uncoded documents 94 and three "non-responsive" documents 97. Since the uncoded documents 94 in the cluster are thematically related to the "non-responsive" reference documents 97, the reviewer may wish to assign a "non-responsive" code to one or more of the uncoded documents 94 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 documents in the cluster, the presence or absence of other reference document classification codes, and the thematic relationship between the "non-responsive" reference documents and the uncoded documents. Thus, the presence of the three "non-responsive" reference documents 97 in the cluster provides a suggestion that the uncoded documents 94 may also be "non-responsive." Further, the label 89 associated with the spine 86 upon which the cluster is located can also be used to influence a suggestion.

A further example can include a cluster with combination of "privileged" and "non-responsive" reference documents. For example, FIGURE 6C is a block diagram showing, by way of example, a cluster 98 with uncoded documents 94 and a combination of differently classified reference documents 95, 97. The cluster 98 can include one "privileged" reference document 95, two "non-responsive" reference documents 97, and nine uncoded documents 94. The "privileged" 95 and "non-responsive" 97 reference documents can be distinguished by different colors or shape, as well as other identifiers. The combination of "privileged" 95 and "non-responsive" 97 reference documents within the cluster 98 can suggest to a reviewer that the uncoded reference documents 94 should be reviewed before classification or that one or more

uncoded reference documents 94 should be classified as "non-responsive" based on the higher number of "non-responsive" reference documents 97 in the cluster 98. In making a classification decision, the reviewer may consider the number of "privileged" reference documents 95 versus the number of "non-responsive" reference documents 97, as well as the thematic relationships between the uncoded documents 94 and the "privileged" 95 and "non-responsive" 97 reference documents. Additionally, the reviewer can identify the closest reference document to an uncoded document and assign the classification code of the closest reference document to the uncoded document. Other examples, classification codes, and combinations of classification codes are possible.

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

The display of relationships between the uncoded documents and reference documents provides suggestion to an individual reviewer. The suggestions can indicate a need for manual review of the uncoded documents, when review may be unnecessary, and hints for classifying the uncoded documents. Additional information can be generated to assist the reviewer in making classification decisions for the uncoded documents, such as a machine-generated confidence level associated with a suggested classification code, as described in common-assigned U.S. Patent Application Serial No. 12/833,769, entitled "System and Method for Providing a Classification Suggestion for Electronically Stored Information," filed on July 9, 2010, pending, the disclosure of which is incorporated by reference.

The machine-generated suggestion for classification and associated confidence level can be determined by a classifier. FIGURE 7 is a process flow diagram 100 showing, by way of example, a method for classifying uncoded documents by a classifier for use in the method of FIGURE 2. An uncoded document is selected from a cluster within a cluster set (block 101) and compared to a neighborhood of x-reference documents (block 102), also located within the cluster, to identify those reference documents that are most relevant to the selected uncoded

document. 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 documents determined for the selected cluster or spine.

The neighborhood of *x*-reference documents is determined separately for each selected uncoded document and can include one or more reference documents within that cluster. During neighborhood generation, an *x* number of reference documents is first determined automatically or by an individual reviewer. Next, the *x*-number of reference documents nearest in distance to the selected uncoded document are identified. Finally, the identified *x*-number of reference documents are provided as the neighborhood for the selected uncoded document. In a further embodiment, the x-number of reference documents are defined for each classification code, rather than across all classification codes. Once generated, the *x*-number of reference documents in the neighborhood and the selected uncoded document are analyzed by the classifier to provide a machine-generated classification suggestion (block 103). A confidence level for the suggested classification is also provided (block 104).

The analysis of the selected uncoded document and *x*-number of reference documents 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 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 document and assigning the classification code of the closest neighbor as the suggested classification code for the selected uncoded document. The closest neighbor is determined by comparing the score vectors for the selected uncoded document with each of the *x*-number of reference documents in the neighborhood as the cos σ to determine a distance metric. The distance metrics for the *x*-number of reference documents are compared to identify the reference document closest to the selected uncoded document as the closest neighbor.

The minimum average distance classification measure includes calculating an average distance of the reference documents in a cluster for each classification code. The classification code with the reference documents having the closest average distance to the selected uncoded document 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 documents within the cluster for each classification code and assigning a count or

"vote" to the reference documents based on the assigned classification code. The classification code with the highest number of reference documents or "votes" is assigned to the selected uncoded document as the suggested classification. The distance weighted maximum count classification measure includes identifying a count of all reference documents within the cluster for each classification code and determining a distance between the selected uncoded document and each of the reference documents. Each count assigned to the reference documents is weighted based on the distance of the reference document from the selected uncoded document. The classification code with the highest count, after consideration of the weight, is assigned to the selected uncoded document as the suggested classification.

The machine-generated classification code is provided for the selected uncoded document with a confidence level, which can be presented as an absolute value or a percentage. Other confidence level measures are possible. The reviewer can use the suggested classification code and confidence level to assign a classification to the selected uncoded document. Alternatively, the *x*-NN classifier can automatically assign the suggested classification. In one embodiment, the *x*-NN classifier only assigns an uncoded document with the suggested classification code if the confidence level is above a threshold value, which can be set by the reviewer or the *x*-NN classifier.

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 documents that are closest to the cluster center can be selected for inclusion in the neighborhood, as described above. Each reference document 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 document to determine an *x*-number of reference documents that are closest to the cluster center. However, other methods for 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.

During classification, either by an individual reviewer or a machine, the reviewer can retain control over many aspects, such as a source of the reference documents and a number of reference documents to be selected. FIGURE 8 is a screenshot 110 showing, by way of example, an options dialogue box 111 for entering user preferences for clustering and display of the uncoded documents and reference documents. The dialogue box 111 can be accessed via a pull-down menu as described above with respect to FIGURE 5. Within the dialogue box 111, the

reviewer can utilize user-selectable parameters to define a reference source 112, category filter 113, command details 114, advanced options 115, classifier parameters 116, and commands 117. 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 the reviewer. Other user-selectable options and displays are possible.

The reference source parameter 112 allows the reviewer to identify one or more sources of the reference documents. The sources can include all reference documents for which the associated classification has been verified, all reference documents that have been analyzed, and all reference documents in a particular binder. The binder can include reference documents particular to a current document review project or that are related to a prior document review project. The category filter parameter 113 allows the reviewer to generate and display the subset of reference documents using only those reference documents associated with a particular classification code. Other options for generating the reference set are possible, including custodian, source, and content. The command parameters 114 allow the reviewer to enter instructions regarding actions for the uncoded and reference documents, such as indicating counts of the documents, and display of the documents. The advanced option parameters 115 allow the reviewer to specify clustering thresholds and classifier parameters. The parameters entered by the user can be compiled as command parameters 116 and provided in a drop-down menu on a display of the clusters. Other user selectable parameters, options, and actions are possible.

Providing suggestions for classification has been described in relation to uncoded documents and reference documents; however, in a further embodiment, suggestions can be provided for tokens extracted from the uncoded documents using reference tokens. For example, the uncoded tokens and reference tokens are clustered and displayed to provide classification suggestions based on relationships between the uncoded tokens and similar reference tokens. The uncoded documents can then be classified based on the classified tokens. In one embodiment, the tokens include concepts, *n*-grams, raw terms, and entities.

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.

CLAIMS:

1. A system (11) for providing reference items (14b) as a suggestion for classifying uncoded electronically stored information items (14a), comprising:

a set (13) of reference electronically stored information items (14b) each associated with a classification code;

a clustering module (33) to combine one or more of the reference electronically stored information items (14b) with a set of uncoded electronically stored information items (14a) and to generate clusters (83) of the uncoded electronically stored information items (14a) and the one or more reference electronically stored information items (14b); and

a display (37) to visually depict relationships between the uncoded electronically stored information items (14a) and the one or more reference electronically stored information items (14b) in at least one cluster (83) as suggestions for classifying the uncoded electronically stored information items (14a) in that cluster (83).

2. A system (11) according to Claim 1, further comprising:

a reference module (32) to generate the set of reference electronically stored information items (14b), comprising at least one of:

a similarity module to identify dissimilar electronically stored information items (14a) for a document review project and to assign a classification code to each of the dissimilar electronically stored information items (14a); and

a reference clustering module to cluster (83) electronically stored information items (14a) for a document review project, to select one or more of the electronically stored information items (14a) in at least one cluster (83), and to assign a classification code to each of the selected electronically stored information items (14a).

3. A system (11) according to Claim 1, wherein the clusters (83) are generated based on a similarity metric comprising forming a score vector for each uncoded electronically stored information (14a) in the portion and each electronically stored information in the reference set (14b) and

calculating the similarity metric by comparing the score vectors for one of the uncoded electronically stored information (14a) and one of the electronically stored information in the reference set (14b) as an inner product.

4. A system (11) according to Claim 3, wherein the inner product is determined according to the following equation:

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

where $\cos\sigma_{AB}$ comprises a similarity between uncoded electronically stored information item A and reference electronically stored information item B, $\vec{S}_A$ comprises a score vector for uncoded electronically stored information item A, and $\vec{S}_B$ comprises a score vector for reference electronically stored information item B.

5. A system (11) according to Claim 1, further comprising:

a classification module to assign a classification code to one or more of the uncoded electronically stored information items (14a) in the at least one cluster (83).

6. A system (11) according to Claim 1, wherein each uncoded electronically stored information item (14a) in the at least one cluster (83) is represented by a symbol in the display (37) and each of the one or more reference electronically stored information items (14b) is represented by an additional symbol in the display (37), and further wherein the reference electronically stored information items (14b) associated with different classification codes are distinguished by assigning a different color to the different symbols.

7. A method (40) for providing reference items (14b) as a suggestion for classifying uncoded electronically stored information items (14a), comprising:

designating a set of reference electronically stored information items (14b) each associated with a classification code;

combining one or more of the reference electronically stored information items (14b) with a set of uncoded electronically stored information items (14a);

generating clusters (83) of the uncoded electronically stored information items (14a) and the one or more reference electronically stored information items (14b); and

visually depicting relationships between the uncoded electronically stored information items (14a) and one or more reference electronically stored information items (14b) in at least one cluster (83) as suggestions for classifying the uncoded electronically stored information items (14a) in that cluster (83).

8. A method (40) according to Claim 7, further comprising:

generating the set of reference electronically stored information items (14b), comprising at least one of:

identifying (41) dissimilar electronically stored information items (14a) for a document review project and assigning a classification code to each of the dissimilar electronically stored information items (14a); and

clustering (43) electronically stored information items (14a) for a document review project, selecting one or more of the electronically stored information items (14a) in at least one cluster (83) and assigning a classification code to each of the selected electronically stored information items (14a).

9. A method (40) according to Claim 7, wherein the clusters (83) are generated based on a similarity metric, comprising:

forming a score vector for each uncoded electronically stored information (14a) in the portion and each electronically stored information in the reference set (14b); and

calculating the similarity metric by comparing the score vectors for one of the uncoded electronically stored information (14a) and one of the electronically stored information in the reference set (14b) as an inner product.

10. A method (40) according to Claim 9, wherein the inner product is determined according to the following equation:

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

where $\cos\sigma_{AB}$ comprises a similarity between uncoded electronically stored information item A and reference electronically stored information item B, $\vec{S}_A$ comprises a score vector for uncoded electronically stored information item A, and $\vec{S}_B$ comprises a score vector for reference electronically stored information item B.

11. A method (40) according to Claim 7, further comprising:

assigning a classification code to one or more of the uncoded electronically stored information items (14a) in the at least one cluster (83).

12. A method (40) according to Claim 7, further comprising:

representing each uncoded electronically stored information item (14a) in the at least one cluster (83) with a symbol; and

representing each of the one or more reference electronically stored information items (14b) with a different symbol; and

distinguishing the reference electronically stored information items (14b) associated with different classification codes by assigning a different color to the different symbols.

13. A system (11) for clustering reference documents (14b) to generate suggestions for classification of uncoded documents (14a), comprising:

a set of reference documents (14b) each associated with a classification;

a clustering module to selecting one or more of the reference documents (14b), to combine the one or more reference documents (14b) selected with uncoded documents (14a) as a set of documents, and to generate clusters (83) of the documents in the document set, further comprising:

a cluster similarity module to determine a similarity between each document; and

a grouping module to group the documents into the clusters (83) based on the similarity;

an identification module to identify at least one cluster (83) with the reference documents (14b); and

a display (37) to visually depict relationships between the uncoded documents (14a) and the one or more reference documents (14b) in the at least one cluster (83) as suggestions for classifying the uncoded electronically stored information items (14a) in that cluster (83).

14. A system (11) according to Claim 13, further comprising:

a reference module to generate the set of reference documents (14b), comprising at least one of:

a reference similarity module to identify dissimilar documents for a document review project and assigning a classification code to each of the dissimilar documents; and

a reference cluster module to generate clusters (83) of documents for a document review project, selecting one or more of the documents in at least one of the clusters (83) and assigning a classification code to each of the documents.

15. A system (11) according to Claim 13, wherein the one or more reference documents (14b) are selected from at least one of a predefined, customized, or arbitrary reference document set.

16. A system (11) according to Claim 13, wherein the similarity is determined by forming a score vector for each uncoded document and each reference document and calculating a similarity metric between the score vectors for the uncoded documents (14a) and reference documents (14b) as an inner product.

17. A system (11) according to Claim 16, wherein the inner product is determined according to the following equation:

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

where $\cos\sigma_{AB}$ comprises a similarity between uncoded document A and reference document B, $\vec{S}_A$ comprises a score vector for uncoded document A, and $\vec{S}_B$ comprises a score vector for reference document B.

18. A system (11) according to Claim 13, wherein each uncoded document in the at least one cluster (83) is represented by a symbol and each reference document is represented by a different symbol, and further wherein the reference electronically stored information items (14b) associated with different classification codes are distinguished by different color assigned to the different symbols.

19. A method (40) for clustering reference documents (14b) to generate suggestions for classification of uncoded documents (14a), comprising:

designating a set of reference documents (14b) each associated with a classification;

selecting one or more of the reference documents (14b) and combining the one or more reference documents (14b) selected with uncoded documents (14a) as a set of documents;

generating clusters (83) of the documents in the document set, comprising:

determining a similarity between each document; and

grouping the documents into the clusters (83) based on the similarity;

identifying at least one cluster (83) with the reference documents (14b); and

visually depicting relationships between the uncoded documents (14a) and the one or more reference documents (14b) in the at least one cluster (83) as suggestions for classifying the uncoded electronically stored information items (14a) in that cluster (83).

20. A method (40) according to Claim 19, further comprising:

generating the set of reference documents (14b), comprising at least one of:

identifying dissimilar documents for a document review project and assigning a classification code to each of the dissimilar documents; and

generating clusters (83) of documents for a document review project, selecting one or more of the documents in at least one of the clusters (83) and assigning a classification code to each of the documents.

21. A method (40) according to Claim 19, wherein the one or more reference documents (14b) are selected from at least one of a predefined, customized, or arbitrary reference document set.

22. A method (40) according to Claim 19, further comprising:

determining the similarity, comprising:

forming a score vector for each uncoded document and each reference document; and

calculating a similarity metric between the score vectors for the uncoded documents (14a) and reference documents (14b) as an inner product.

23. A method (40) according to Claim 22, wherein the inner product is determined according to the following equation:

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

where $\cos\sigma_{AB}$ comprises a similarity between uncoded document A and reference document B, $\vec{S}_A$ comprises a score vector for uncoded document A, and $\vec{S}_B$ comprises a score vector for reference document B.

24. A method (40) according to Claim 19, further comprising:

representing each uncoded document in the at least one cluster (83) with a symbol; and

representing each reference document with a different symbol; and

distinguishing the reference documents (14b) with different classification codes with different colors of the different symbols.

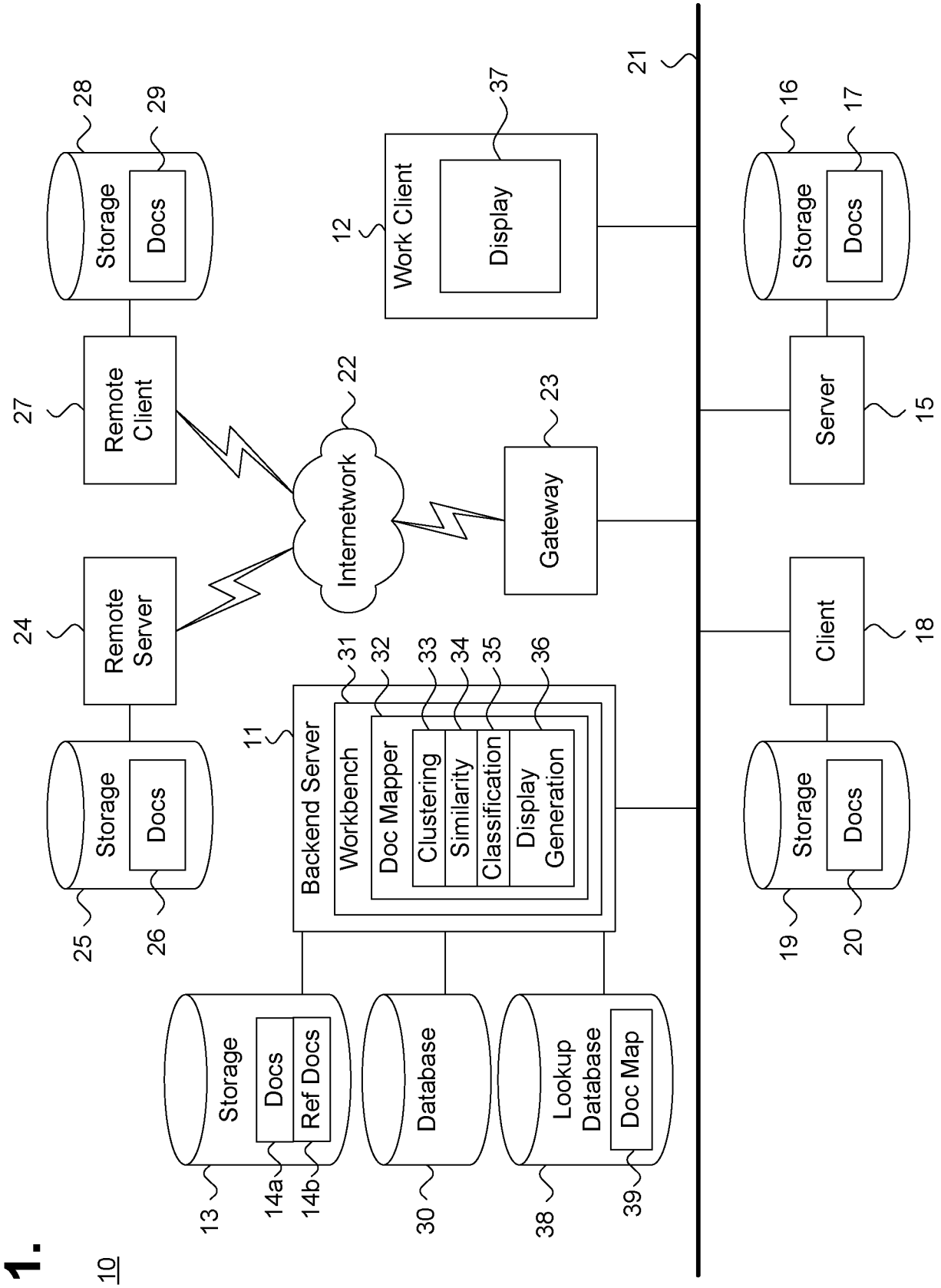

10
21
Storage
Docs
28
29
Remote Client
27
Work Client
12
Display
37
Storage
Docs
16
17
Internetwork
22
Gateway
23
Server
15
Remote Server
24
Backend Server
11
31
32
33
34
35
36
Workbench
Doc Mapper
Clustering
Similarity
Classification
Display Generation
Client
18
Storage
Docs
25
26
Storage
Docs
Ref Docs
13
14a
14b
Database
30
Lookup Database
Doc Map
38
39
Storage
Docs
19
20

Fig. 1.

Fig. 2.

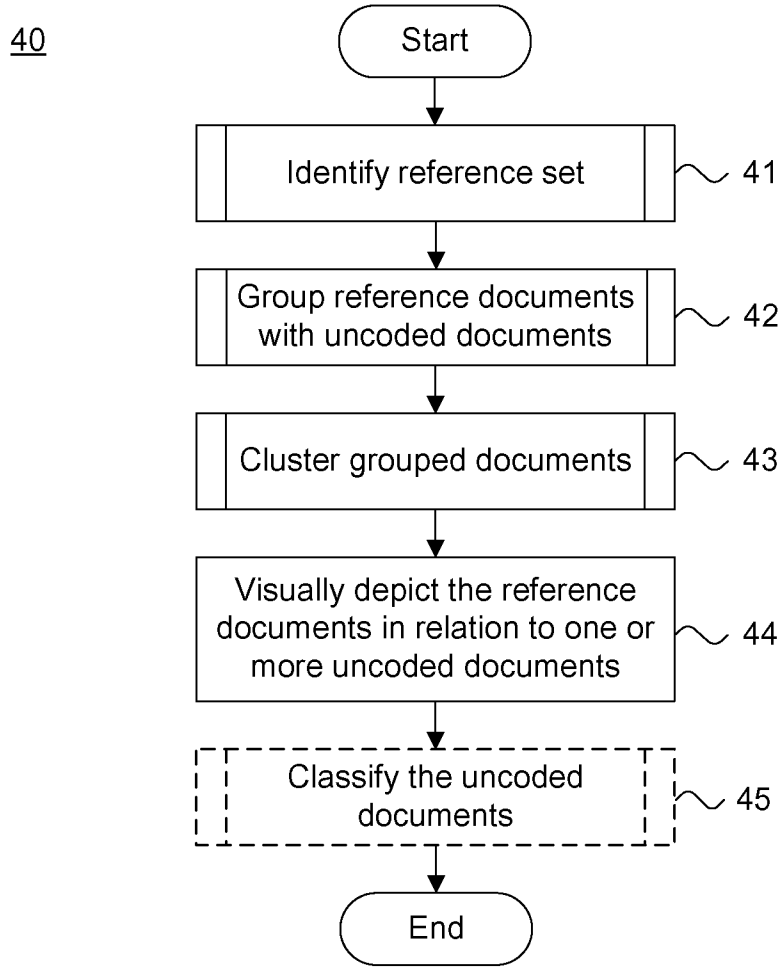

40
Start
Identify reference set
41
Group reference documents with uncoded documents
42
Cluster grouped documents
43
Visually depict the reference documents in relation to one or more uncoded documents
44
Classify the uncoded documents
45
End

Fig. 3.

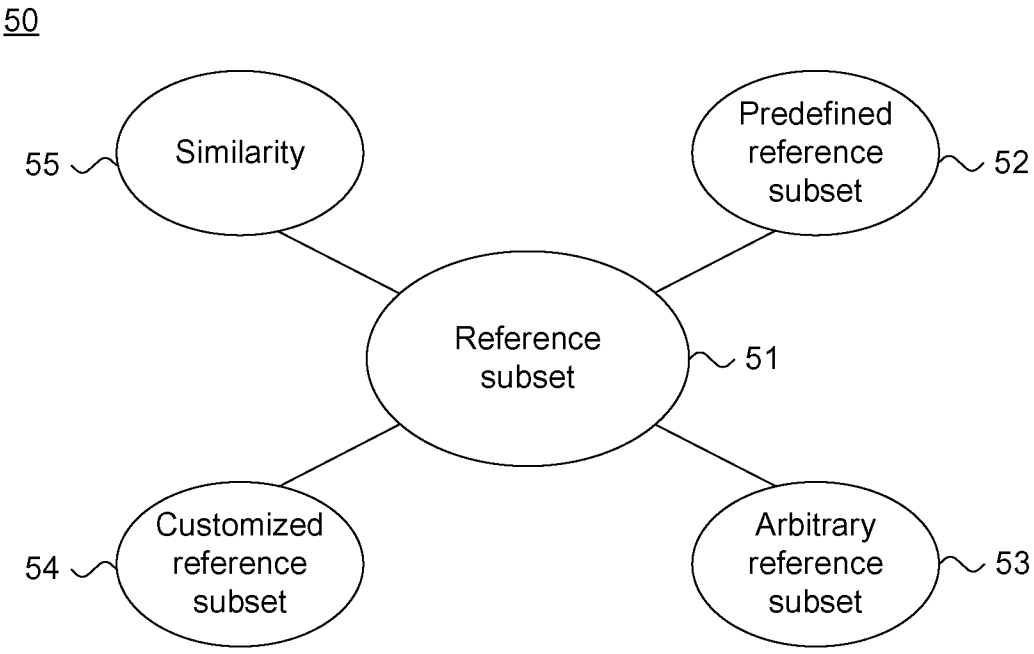

50
55
Similarity
Predefined reference subset
52
Reference subset
51
Customized reference subset
54
Arbitrary reference subset
53

Fig. 4.

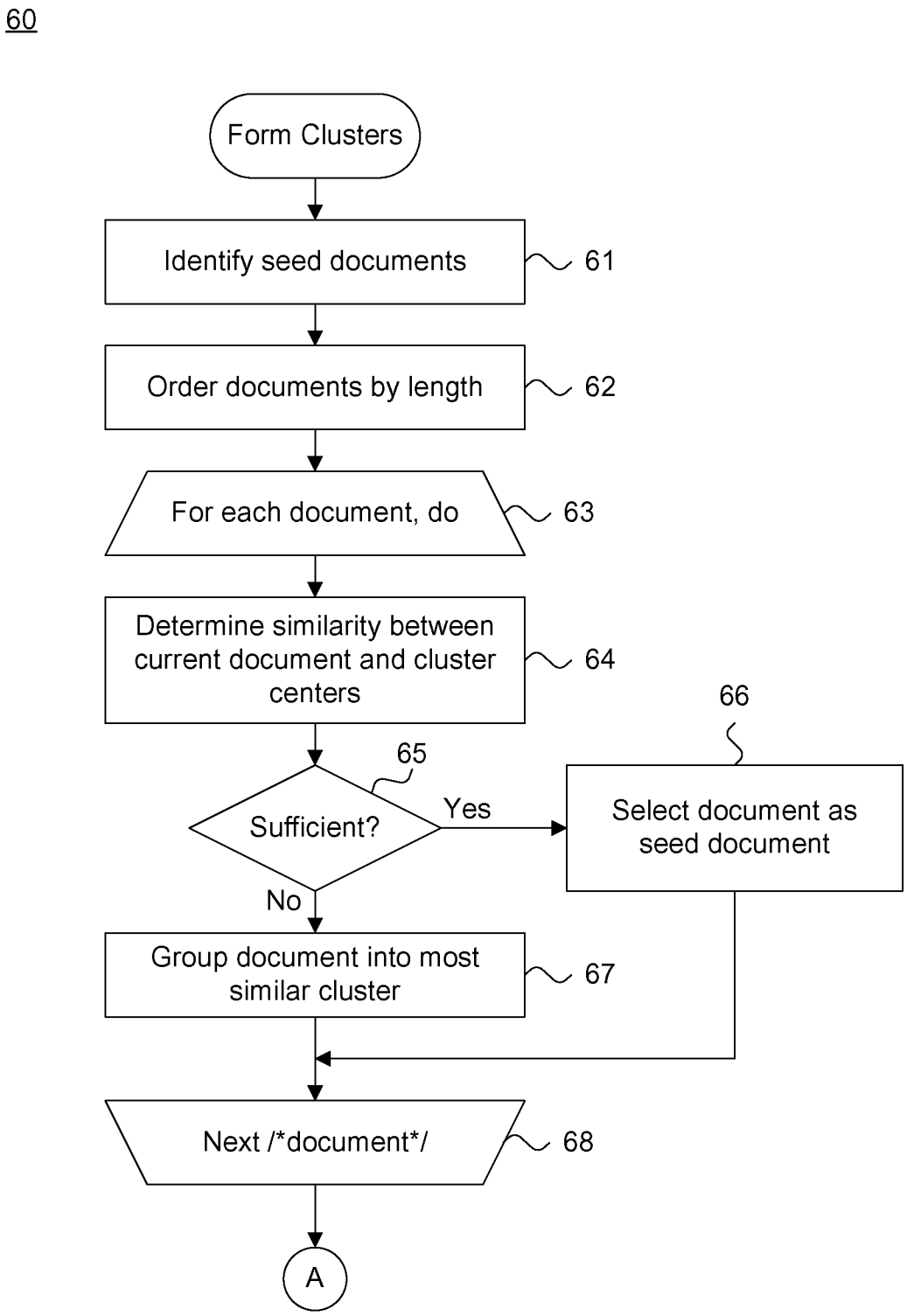

60
Form Clusters
Identify seed documents
61
Order documents by length
62
For each document, do
63
Determine similarity between current document and cluster centers
64
65
Sufficient?
Yes
66
Select document as seed document
No
Group document into most similar cluster
67
Next /*document*/
68
A

Fig. 4 (Cont).

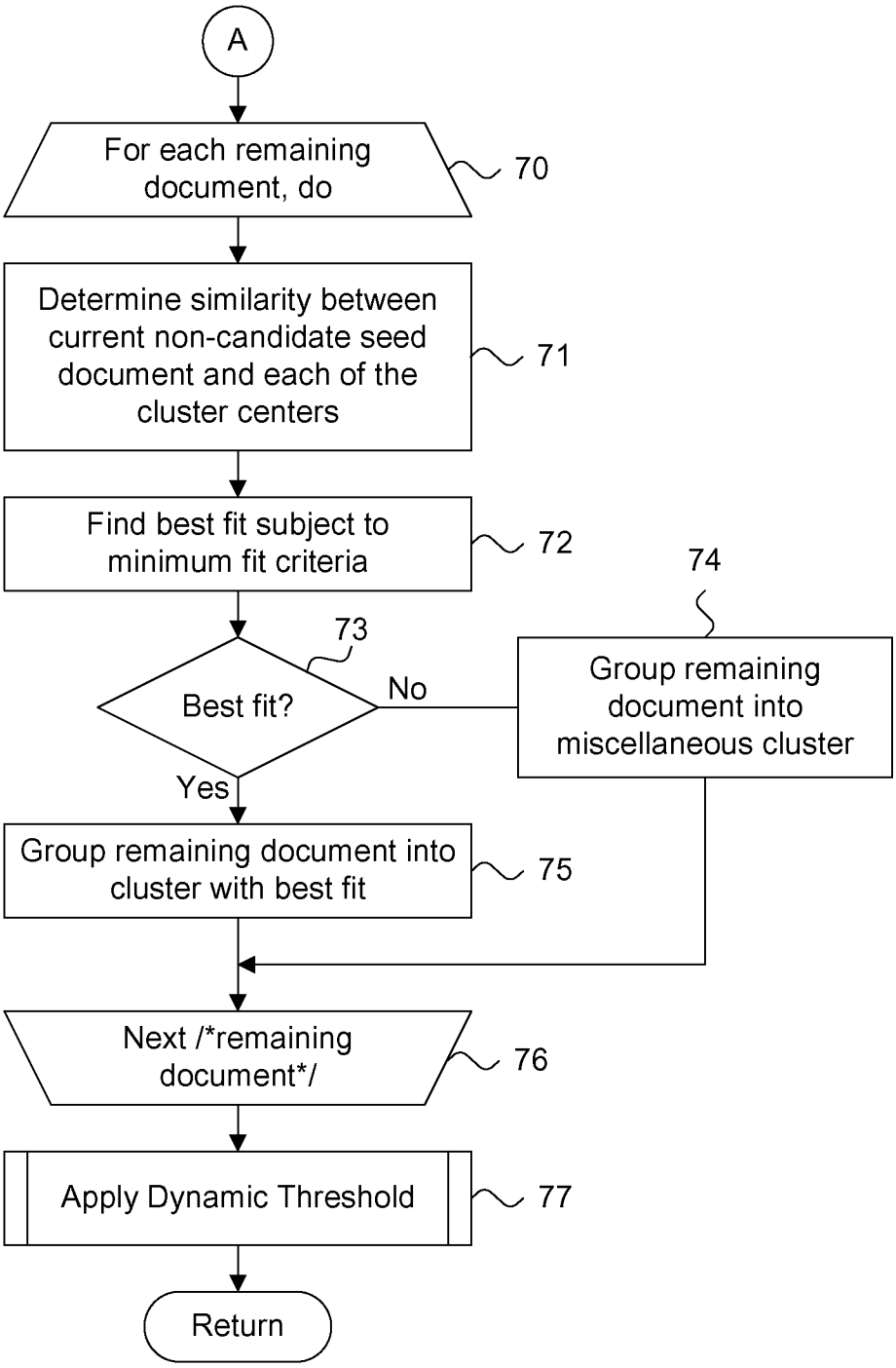

A
For each remaining document, do
70
Determine similarity between current non-candidate seed document and each of the cluster centers
71
Find best fit subject to minimum fit criteria
72
73
Best fit?
No
74
Group remaining document into miscellaneous cluster
Yes
Group remaining document into cluster with best fit
75
Next /*remaining document*/
76
Apply Dynamic Threshold
77
Return

Fig. 5.

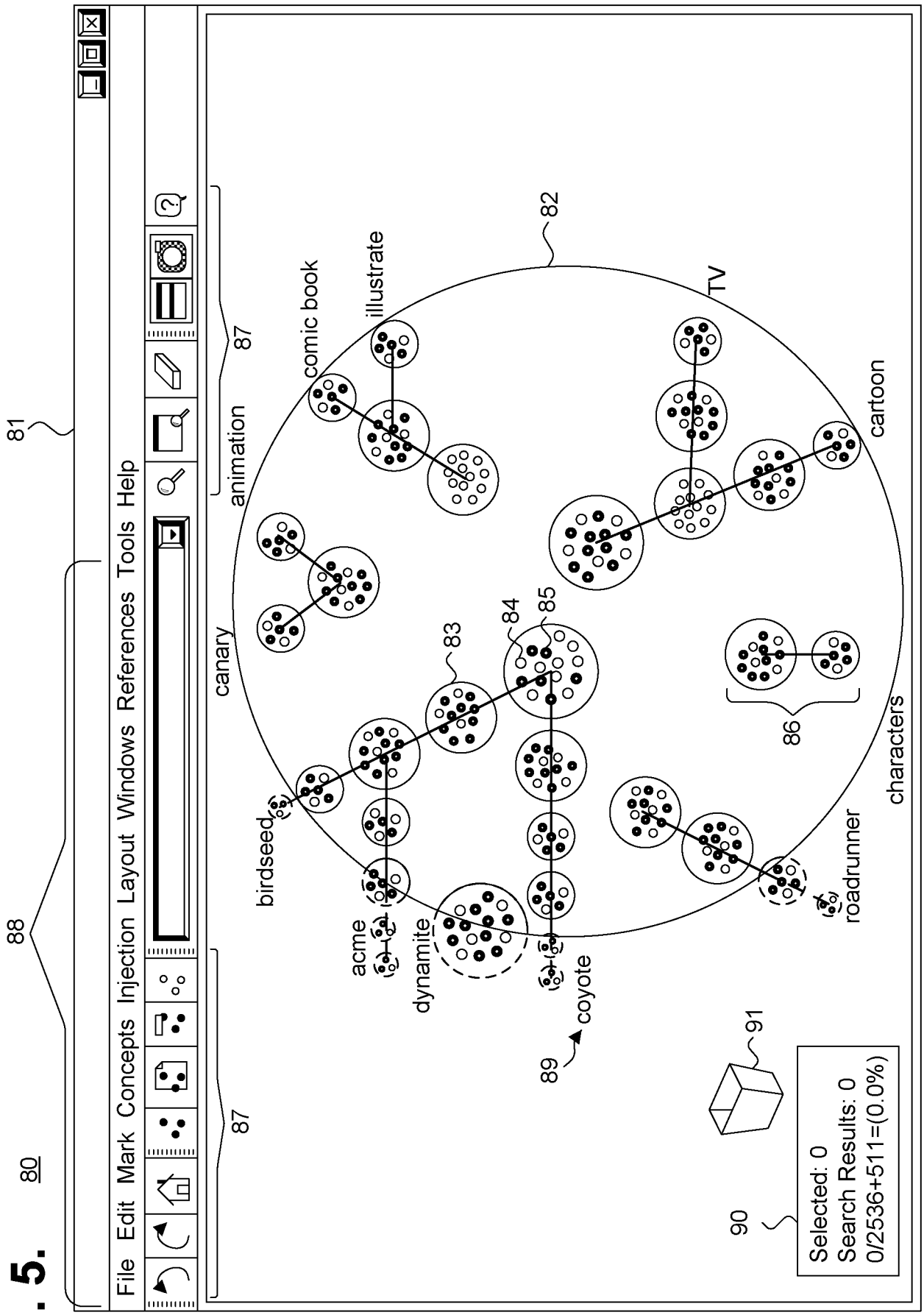

80
81
88
File Edit Mark Concepts Injection Layout Windows References Tools Help
87
87
canary
animation
comic book
illustrate
TV
cartoon
characters
birdseed
acme
dynamite
coyote
roadrunner
82
83
84
85
86
89
90
91
Selected: 0
Search Results: 0
0/2536+511=(0.0%)

Fig. 6A.

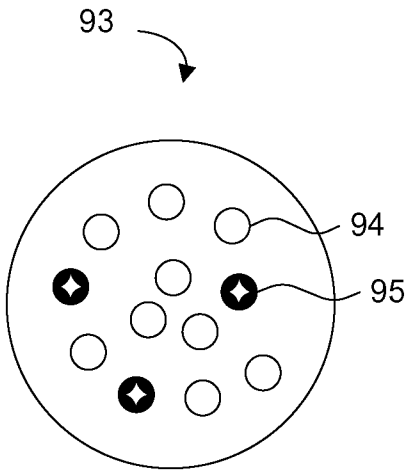


Fig. 6B.

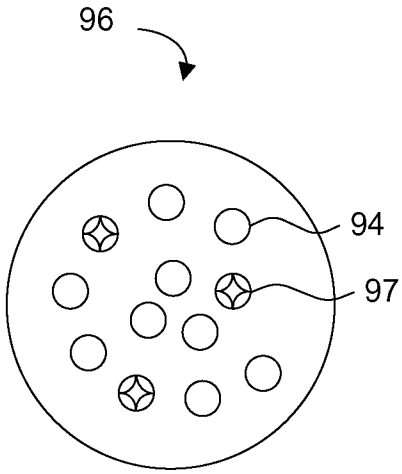


Fig. 6C.

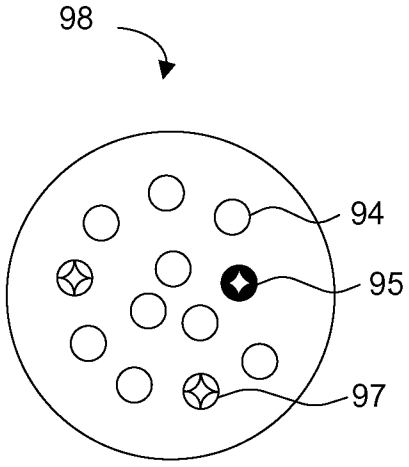


Fig. 7.

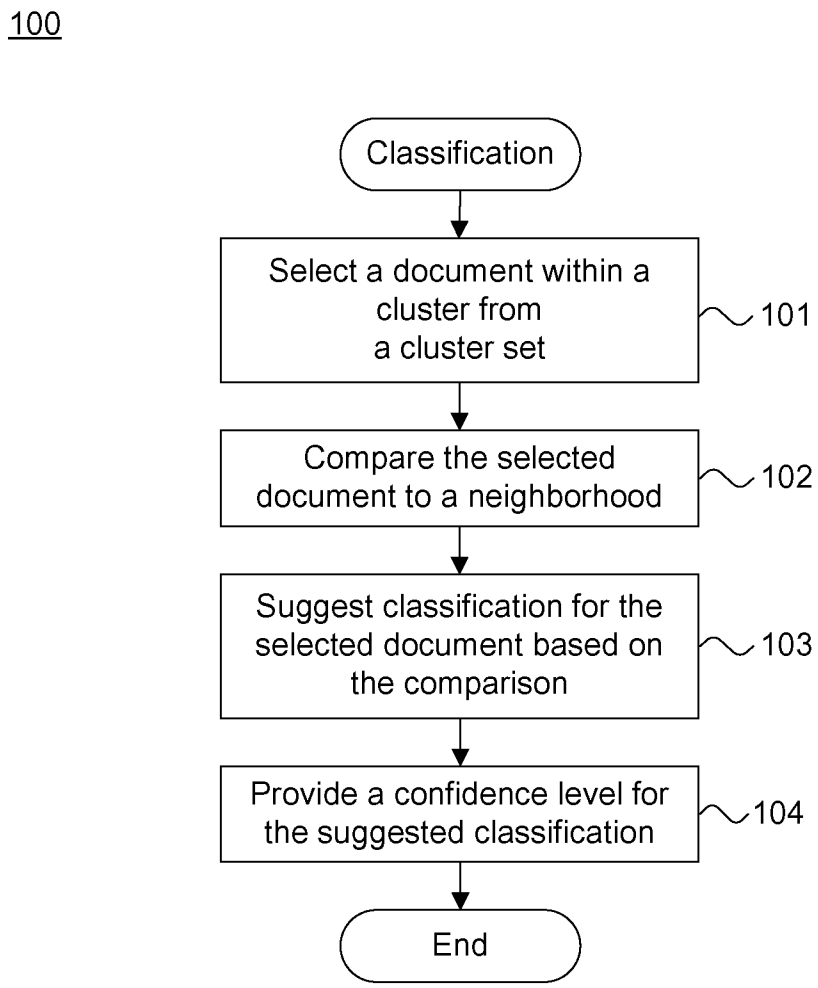
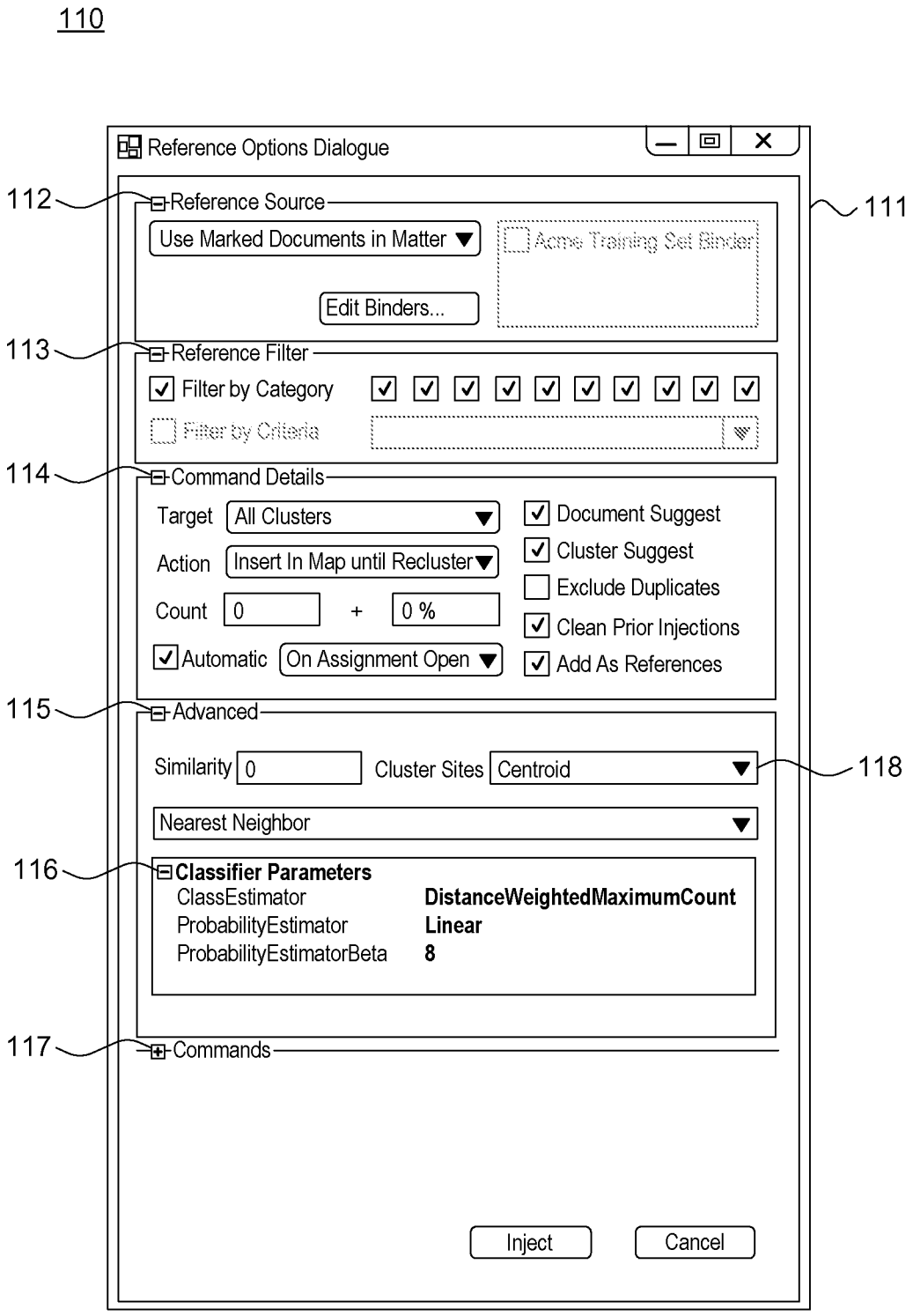

100
Classification
Select a document within a cluster from a cluster set
101
Compare the selected document to a neighborhood
102
Suggest classification for the selected document based on the comparison
103
Provide a confidence level for the suggested classification
104
End

Fig. 8.


110
Reference Options Dialogue
112
Reference Source
111
Use Marked Documents in Matter
Acme Training Set Binder
Edit Binders...
113
Reference Filter
Filter by Category
Filter by Criteria
114
Command Details
Target
All Clusters
Document Suggest
Action
Insert In Map until Recluster
Cluster Suggest
Exclude Duplicates
Count
0
+
0 %
Clean Prior Injections
Automatic
On Assignment Open
Add As References
115
Advanced
Similarity
0
Cluster Sites
Centroid
118
Nearest Neighbor
116
Classifier Parameters
ClassEstimator
DistanceWeightedMaximumCount
ProbabilityEstimator
Linear
ProbabilityEstimatorBeta
8
117
Commands
Inject
Cancel

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2010/043292

A. CLASSIFICATION OF SUBJECT MATTER
INV. 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)
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, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2005/097435 A1 (PRAKASH VIPUL V [US] ET AL PRAKASH VIPUL VED [US] ET AL) 5 May 2005 (2005-05-05) paragraphs [0018], [0023], [0037], [0039], [0040], [0041], [0042], [0043]; figure 1	1-24
X	US 2005/022106 A1 (KAWAI KENJI [US] ET AL) 27 January 2005 (2005-01-27) paragraphs [0010], [0011], [0048], [0054], [0057], [0078], [0102] - [0105]; figures 4,13,14	1-24
X	US 6 502 081 B1 (WILTSHIRE JR JAMES S [US] ET AL) 31 December 2002 (2002-12-31) abstract column 1, lines 19-25 column 4, line 51 - column 6, line 30	1-24

-/--

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

[X] See patent family annex.

* Special categories of cited documents :

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

"E" earlier 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
15 November 2010	08/12/2010

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

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

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	EP 1 049 030 A1 (SER SYSTEME AG PRODUKTE UND AN [DE]) 2 November 2000 (2000-11-02) abstract paragraphs [0008] - [0010], [0020], [0044], [0050], [0062], [0071]; figure 2	1-24
X	WO 2005/073881 A1 (SIFTOLOGY INC [US]) 11 August 2005 (2005-08-11) abstract page 4, line 10 - line 17 page 8, line 15 - line 23 page 9, line 26 - page 10, line 7; figure 3	1-24
X	US 6 697 998 B1 (DAMERAU FREDERICK J [US] ET AL) 24 February 2004 (2004-02-24) abstract the whole document	1-24
X	WO 00/67162 A1 (WEST PUBLISHING CO [US]; YANG STEPHENS BOKYUNG [US]; SWOPE M CHARLES [) 9 November 2000 (2000-11-09) page 5, line 32 - page 6, line 7 page 9, line 24 - page 10, line 31	1-24
X	US 2007/109297 A1 (BORCHARDT JONATHAN M [US] ET AL) 17 May 2007 (2007-05-17) abstract paragraphs [0041], [0058]; figures 2A-B,4A	1-24
X	O'NEILL J et al.: "DISCO: Intelligent Help for Document Review"[Online] 8 June 2009 (2009-06-08), pages 1-10, XP002607216 Workshop DESI at ICAIL 2009, Barcelona, Spain Xerox Retrieved from the Internet: URL:http://www.xrce.xerox.com/content/download/11962/80921/file/2009-036.pdf> [retrieved on 2010-10-28] abstract section "3.1 CategoriX" section "3.2 Using CategoriX for document Review"	1-24

-/--

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2010/043292

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
L	O'NEILL J et al.: "DISCO: Intelligent Help for Document Review"[Online] 8 June 2009 (2009-06-08), XP002607217 Workshop DESI at ICAIL 2009, Barcelona, Spain Xerox Retrieved from the Internet: URL:http://www.xrce.xerox.com/Research-Development/Publications/2009-036> [retrieved on 2010-10-28]	

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

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2010/043292

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2005097435 A1	05-05-2005	US 2009259608 A1	15-10-2009
		WO 2005043417 A2	12-05-2005
US 2005022106 A1	27-01-2005	CA 2534273 A1	10-02-2005
		EP 1652119 A1	03-05-2006
		US 2010049708 A1	25-02-2010
		WO 2005013152 A1	10-02-2005
US 6502081 B1	31-12-2002	NONE	
EP 1049030 A1	02-11-2000	AU 4545000 A	17-11-2000
		WO 0067150 A2	09-11-2000
		EP 1175652 A2	30-01-2002
		US 2009216693 A1	27-08-2009
		US 2006212413 A1	21-09-2006
		US 6976207 B1	13-12-2005
WO 2005073881 A1	11-08-2005	NONE	
US 6697998 B1	24-02-2004	NONE	
WO 0067162 A1	09-11-2000	AU 781157 B2	12-05-2005
		AU 4989800 A	17-11-2000
		CA 2371688 A1	09-11-2000
		EP 1212699 A1	12-06-2002
		JP 2002543528 T	17-12-2002
		NZ 515293 A	30-04-2004
US 2007109297 A1	17-05-2007	CA 2640032 A1	09-08-2007
		EP 1977353 A2	08-10-2008
		US 2008278485 A1	13-11-2008
		WO 2007089588 A2	09-08-2007