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

Canadian Intellectual Property Office
An agency of Industry Canada

CA 2772082 A1 2011/03/10

(21) 2 772 082

(12) DEMANDE DE BREVET CANADIEN
CANADIAN PATENT APPLICATION

(13) A1

(86) Date de dépôt PCT/PCT Filing Date: 2010/08/24
(87) Date publication PCT/PCT Publication Date: 2011/03/10
(85) Entrée phase nationale/National Entry: 2012/02/23
(86) N° demande PCT/PCT Application No.: US 2010/046557
(87) N° publication PCT/PCT Publication No.: 2011/028553
(30) Priorités/Priorities: 2009/08/24 (US61/236,490);
2010/08/24 (US12/862,682)

(51) Cl.Int./Int.Cl. *G06F 17/30* (2006.01)

(71) Demandeur/Applicant:
FTI CONSULTING, INC., US

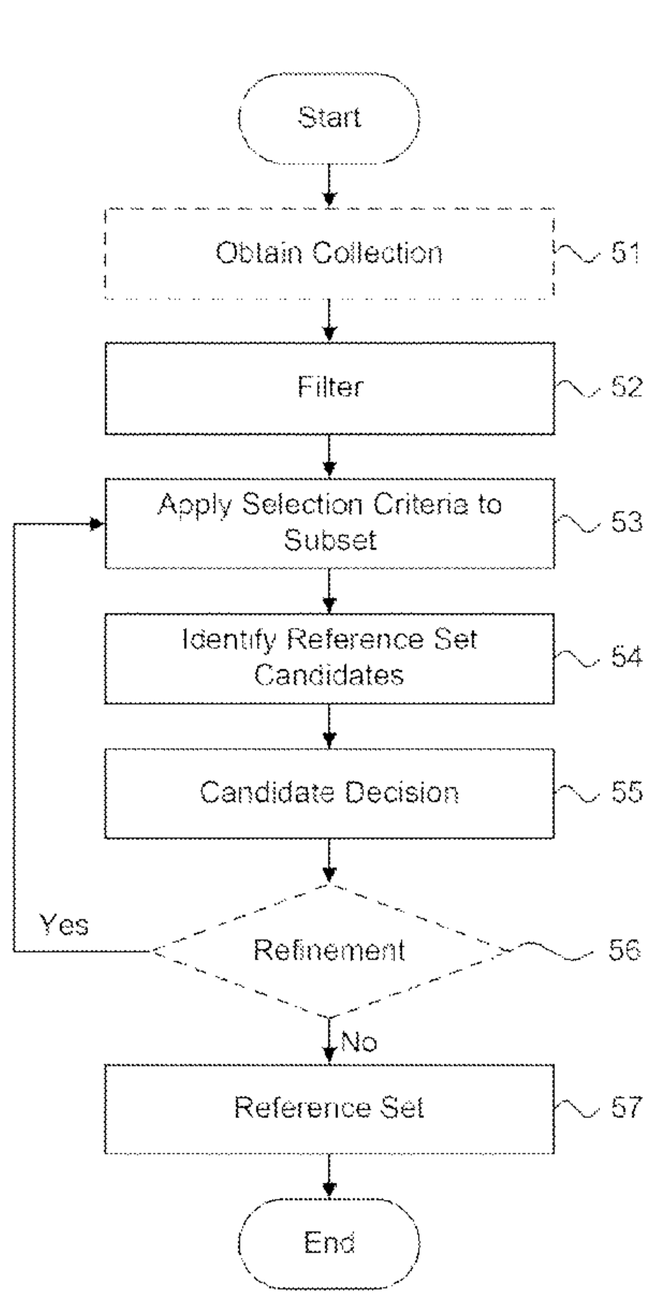
(72) Inventeurs/Inventors:
KNIGHT, WILLIAM C., US;
MCNEE, SEAN M., US;
CONWELL, JOHN, US

(74) Agent: INTEGRAL IP

(54) Titre : GENERATION D'UN ENSEMBLE DE REFERENCE POUR UTILISATION LORS DE LA REVISION D'UN DOCUMENT

(54) Title: GENERATING A REFERENCE SET FOR USE DURING DOCUMENT REVIEW

Fig. 2.



(57) Abrégé/Abstract:
A system (10) and method (50) for providing generating reference sets (14b) for use during document review is provided. A collection of unclassified documents (14a) is obtained. Selection criteria (61) are applied (53) to the document collection and those

CA 2772082 A1 2011/03/10

(21) 2 772 082

(13) A1

(57) Abrégé(suite)/Abstract(continued):
unclassified documents (14a) that satisfy the selection criteria (61) are selected as reference set candidates. A classification code is assigned (55) to each reference set candidate. A reference set (14b) is formed (57) from the classified reference set candidates. The reference set (14b) is quality controlled and shared between one or more users.

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
10 March 2011 (10.03.2011)

PCT

(10) International Publication Number
WO 2011/028553 A1

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

(21) **International Application Number**:
PCT/US2010/046557

(22) **International Filing Date**:
24 August 2010 (24.08.2010)

(25) **Filing Language**: English

(26) **Publication Language**: English

(30) **Priority Data**:

61/236,490	24 August 2009 (24.08.2009)	US
12/862,682	24 August 2010 (24.08.2010)	US

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

(72) **Inventors; and**

(75) **Inventors/Applicants** *(for US only)*: **KNIGHT, William C.** [US/US]; 1851 Edna Place, Bainbridge Island, Washington 98110 (US). **MCNEE, Sean M.** [US/US]; 1618 Bellevue Avenue, Apt. No. 308, Seattle, Washington 98122 (US).

(74) **Agents**: **WITTMAN, Krista A.** et al.; Cascadia Intellectual Property, 500 Union Street, Suite 1005, Seattle, Washington 98101 (US).

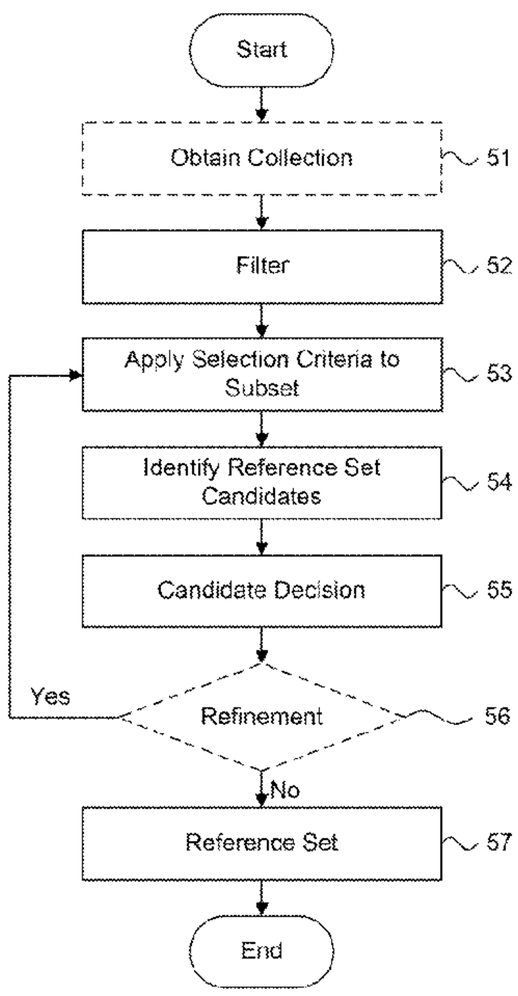
(81) **Designated States** *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) **Title**: GENERATING A REFERENCE SET FOR USE DURING DOCUMENT REVIEW

Fig. 2.



(57) **Abstract**: A system (10) and method (50) for providing generating reference sets (14b) for use during document review is provided. A collection of unclassified documents (14a) is obtained. Selection criteria (61) are applied (53) to the document collection and those unclassified documents (14a) that satisfy the selection criteria (61) are selected as reference set candidates. A classification code is assigned (55) to each reference set candidate. A reference set (14b) is formed (57) from the classified reference set candidates. The reference set (14b) is quality controlled and shared between one or more users.

WO 2011/028553 A1

WO 2011/028553 A1

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))*

GENERATING A REFERENCE SET FOR USE DURING DOCUMENT REVIEW

TECHNICAL FIELD

The invention relates in general to information retrieval and, specifically, to a system and method for generating a reference set for use during document review.

BACKGROUND ART

Document review is an activity frequently undertaken in the legal field during the discovery phase of litigation. Typically, document classification requires reviewers to assess the relevance of documents to a particular topic as an initial step. Document reviews can be conducted manually by human reviewers, automatically by a machine, or by a combination of human reviewers and a machine.

Generally, trained reviewers analyze documents and provide a recommendation for classifying each document in regards to the particular legal issue being litigated. A set of exemplar documents is provided to the reviewer as a guide for classifying the documents. The exemplar documents are each previously classified with a particular code relevant to the legal issue, such as "responsive," "non-responsive," and "privileged." Based on the exemplar documents, the human reviewers or machine can identify documents that are similar to one or more of the exemplar documents and assign the code of the exemplar document to the uncoded documents.

The set of exemplar documents selected for document review can dictate results of the review. A cohesive representative exemplar set can produce accurately coded documents, while effects of inaccurately coded documents can be detrimental to a legal proceeding. For example, a "privileged" document contains information that is protected by a privilege, meaning that the document should not be disclosed to an opposing party. Disclosing a "privileged" document can result in an unintentional waiver of privilege to the subject matter.

The prior art focuses on document classification and generally assumes that exemplar documents are already defined and exist as a reference set for use in classifying document. Such classification can benefit from having better reference sets generated to increase the accuracy of classified documents.

Thus, there remains a need for a system and method for generating a set of exemplar documents that are cohesive and which can serve as an accurate and efficient example for use in classifying documents.

DISCLOSURE OF THE INVENTION

A system and method for providing generating reference sets for use during document review is provided. A collection of unclassified documents is obtained. Selection criteria are applied to the document collection and those unclassified documents that satisfy the selection criteria are selected as reference set candidates. A classification code is assigned to each reference set candidate. A reference set is formed from the classified reference set candidates. The reference set is quality controlled and shared between one or more users.

A further embodiment provides a method for generating a reference set via clustering. A collection of documents is obtained. The documents are grouped into clusters of documents. One or more of the documents are selected from at least one cluster as reference set candidates. A classification code is assigned to each of the reference set candidates. The classified reference set candidates are grouped as the reference set.

A still further embodiment provides a method for generating a reference set via seed documents. A collection of documents is obtained. One or more seed documents are identified. The seed documents are compared to the document collection. Those documents that are similar to the seed documents are identified as reference set candidates. A size threshold is applied to the reference set candidates and the reference set candidates are grouped as the reference set when the size threshold is satisfied.

An even further embodiment provides a method for generating a training set for use during document review. Classification codes are assigned to a set of documents. Further classification codes assigned to the same set of documents are received. The classification code for at least one document is compared with the further classification code for that document. A determination as to whether a disagreement exists between the assigned classification code and the further classification code is made for at least one document. Those documents with disagreeing classification codes are identified as training set candidates. A stop threshold is applied to the training set candidates and the training set candidates are grouped as a training set when the stop threshold is satisfied.

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 generating a reference set for use during document review, in accordance with one embodiment.

FIGURE 2 is a flow diagram showing a method for generating a reference set for use during document review, in accordance with one embodiment.

FIGURE 3 is a data flow diagram showing examples of the selection criteria of FIGURE 2.

FIGURE 4 is a flow diagram showing, by way of example, a method for generating a reference set via hierarchical clustering.

FIGURE 5 is a flow diagram showing, by way of example, a method for generating a reference set via iterative clustering.

FIGURE 6 is a flow diagram showing, by way of example, a method for generating a reference set via document seeding.

FIGURE 7 is a flow diagram showing, by way of example, a method for generating a reference set via random sampling.

FIGURE 8 is a flow diagram showing, by way of example, a method for generating a reference set via user assisted means.

FIGURE 9 is a flow diagram showing, by way of example, a method for generating a reference set via active learning.

FIGURE 10 is a flow diagram showing, by way of example, a method for generating a training set.

BEST MODE FOR CARRYING OUT THE INVENTION

Reference documents are each associated with a classification code and are selected as exemplar documents or a "reference set" to assist human reviewers or a machine to identify and code unclassified documents. The quality of a reference set can dictate the results of a document review project and an underlying legal proceeding or other activity. Use of a noncohesive or "bad" reference set can provide inaccurately coded documents and could negatively affect a pending legal issue during, for instance, litigation. Generally, reference sets should be cohesive for a particular issue or topic and provide accurate guidance to classifying documents.

Cohesive reference set generation requires a support environment to review, analyze, and select appropriate documents for inclusion in the reference set. FIGURE 1 is a block diagram showing a system for generating a reference set for use in classifying documents, in accordance with one embodiment. By way of illustration, the system 10 operates in a distributed computing environment, including "cloud environments," which include a plurality of systems and sources.

A backend server 11 is coupled to a storage device 13, 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 stores documents 14a and reference sets 14b. The documents 14a can include uncoded or "unclassified" documents and coded or "classified" documents, in the form of structured or unstructured data. Hereinafter, the terms "classified" and "coded" are used interchangeably with the same intended meaning, unless otherwise indicated.

The uncoded and coded documents can be related to one or more topics or legal issues. Uncoded documents are analyzed and assigned a classification code during a document review, while coded documents that have been previously reviewed and associated with a classification code. The storage device 13 also stores reference documents 14b, which together form a reference set of trusted and known results for use in guiding document classification. A set of reference documents can be hand-selected or automatically selected, as discussed *infra*.

Reference sets can be generated for one or more topics or legal issues, as well as for any other data to be organized and classified. For instance, the topic can include data regarding a person, place, or object. In one embodiment, the reference set can be generated for a legal proceeding based on a filed complaint or other court or administrative filing or submission. Documents in the reference set 14b are each associated with an assigned classification code and can highlight important information for the current topic or legal issue. A reference set can include reference documents with different classification codes or the same classification code. Core reference documents most clearly exhibit the particular topic or legal matter, whereas boundary condition reference documents include information similar to the core reference documents, but which are different enough to require assignment of a different classification code.

Once generated, the reference set can be used as a guide for classifying uncoded documents, such as described in commonly-assigned U.S. Patent Application Serial No. 12/833,860, entitled "System and Method for Displaying Relationships Between Electronically Stored Information to Provide Classification Suggestions via Inclusion," filed July 9, 2010, pending; U.S. Patent Application Serial No. 12/833,872, entitled "System and Method for Displaying Relationships Between Electronically Stored Information to Provide Classification Suggestions via Injection," filed July 9, 2010, pending; U.S. Patent Application Serial No. 12/833,880, entitled "System and Method for Displaying Relationships Between Electronically Stored Information to Provide Classification Suggestions via Nearest Neighbor," filed July 9, 2010, pending; and 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 disclosures of which are incorporated by reference.

In a further embodiment, a reference set can also be generated based on features associated with the document. The feature reference set can be used to identify uncoded documents associated with the reference set features and provide classification suggestions, such as described in commonly-assigned U.S. Patent Application Serial No. 12/844,810, entitled "System and Method for Displaying Relationships Between Concepts to Provide Classification Suggestions via Inclusion," filed July 27, 2010, pending; U.S. Patent Application Serial No. 12/844,792, entitled "System and Method for Displaying Relationships Between Concepts to Provide Classification Suggestions via Injection," filed July 27, 2010, pending; U.S. Patent Application Serial No. 12/844,813, entitled "System and Method for Displaying Relationships Between Concepts to Provide Classification Suggestions via Nearest Neighbor," filed July 27, 2010, pending; and U.S. Patent Application Serial No. 12/844,785, entitled "System and Method for Providing a Classification Suggestion for Concepts," filed July 27, 2010, pending, the disclosures of which are incorporated by reference.

The backend server 11 is also 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, selector 34, classifier 35, and display generator 36. Other workbench suite modules are possible. In a further embodiment, the clustering engine, selector, classifier, and display generator can be provided independently of the document mapper.

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, issued on October 27, 2009, the disclosure of which is incorporated by reference. The uncoded documents 14a can be grouped into clusters and one or more documents can be selected from at least one cluster to form reference set candidates, as further discussed below in detail with reference to FIGURES 4 and 5. The clusters can be organized along vectors, known as spines, based on a similarity of the clusters. The selector 34 applies predetermined criteria to a set of documents to identify candidates for inclusion in a reference set, as discussed *infra*. The classifier 35 provides a machine-generated classification code suggestion and confidence level for coding of selected uncoded documents.

The display generator 36 arranges the clusters and spines in thematic neighborhood relationships in a two-dimensional visual display space. 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 human reviewer. The reviewer can include an individual person who is assigned to review and classify one or more uncoded documents by designating a code. 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 data, 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 Windows Live Mail products, licensed by Microsoft Corporation, Redmond, WA. The database can be an SQL-based relational database, such as the Oracle database management system, Release 11, licensed by Oracle Corporation, Redwood Shores, CA. Further, the individual documents 17, 20, 26, 29 can be stored in a "cloud," such as in Windows Live Hotmail, licensed by Microsoft Corporation, Redmond, WA. Additionally, the individual documents 17, 20, 26, 29 include uncoded documents and reference documents.

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 that have 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.

Reference set candidates selected for inclusion in a reference set are identified using selection criteria, which can reduce the number of documents for selection. FIGURE 2 is a flow diagram showing a method for generating a reference set for use in document review, in accordance with one embodiment. A collection of documents is obtained (block 51). The collection of documents can include uncoded documents selected from a current topic or legal matter, previously coded documents selected from a related topic or legal matter, or pseudo documents. Pseudo documents are created using knowledge obtained by a person familiar with the issue or topic that is converted into a document. For example, a reviewer who participated in a verbal conversation with a litigant or other party during which specifics of a lawsuit were discussed could create a pseudo document based on the verbal conversation. A pseudo document can exist electronically or in hardcopy form. In one embodiment, the pseudo document is created specifically for use during the document review. Other types of document collections are possible.

Filter criteria are optionally applied to the document collection to identify a subset of documents (block 52) for generating the reference set. The filter criteria can be based on metadata associated with the documents, including date, file, folder, custodian, or content. Other filter criteria are possible. In one example, a filter criteria could be defined as "all documents created after 1997;" and thus, all documents that satisfy the filter criteria are selected as a subset of the document collection.

The filter criteria can be used to reduce the number of documents in the collection. Subsequently, selection criteria are applied to the document subset (block 53) to identify those documents that satisfy the selection criteria as candidates (block 54) for inclusion in the reference set. The selection criteria can include clustering, feature identification, assignments or random selection, and are discussed in detail below with reference to FIGURE 3. A candidate decision is applied (block 55) to the reference set candidates to identify the reference candidates for potential inclusion in the reference set (block 57). During the candidate decision, the reference set candidates are analyzed and a classification code is assigned to each reference set candidate. A human reviewer or machine can assign the classification codes to the reference set candidates based on features of each candidate. The features include pieces of information that described the document candidate, such as entities, metadata, and summaries, as well as other information. Coding instructions guide the reviewer or machine to assign the correct classification code using the features of the reference set candidates. The coding instructions can

be provided by a reviewer, a supervisor, a law firm, a party to a legal proceeding, or a machine. Other sources of the coding instructions are possible.

Also, a determination as to whether that reference set candidate is a suitable candidate for including in the reference set is made. Once the reference set candidates are coded, each candidate is analyzed to ensure that candidates selected for the reference set cover or "span" the largest area of feature space provided by the document collection. In one embodiment, the candidates that are most dissimilar from all the other candidates are selected as the reference set. A first reference set candidate is selected and placed in a list. The remaining reference set candidates are compared to the first reference set candidate in the list and the candidate most dissimilar to all the listed candidates is also added to the list. The process continues until all the dissimilar candidates have been identified or other stop criteria have been satisfied. The stop criteria can include a predetermined number of dissimilar reference set criteria, all the candidates have been reviewed, or a measure of the most dissimilar document fails to satisfy a dissimilarity threshold. Identifying dissimilar documents is discussed in the paper, Sean M. McNee. "Meeting User Information Needs in Recommender Systems". Ph.D. Dissertation, University of Minnesota-Twin Cities. June 2006, which is hereby incorporated by reference. Other stop criteria are possible.

However, refinement (block 56) of the reference set candidates can optionally occur prior to selection of the reference set. The refinement assists in narrowing the number of reference set candidates used to generate a reference set of a particular size or other criteria. If refinement is to occur, further selection criteria are applied (block 53) to the reference set candidates and a further iteration of the process steps occurs. Each iteration can involve different selection criteria. For example, clustering criteria can be applied during a first pass and random sampling can be applied during a second pass to identify reference set candidates for inclusion in the reference set.

In a further embodiment, features can be used to identify documents for inclusion in a reference set. A collection of documents is obtained and features are identified from the document collection. The features can be optionally filtered to reduce the feature set and subsequently, selection criteria can be applied to the features. The features that satisfy the selection criteria are selected as reference set candidate features. A candidate decision, including assigning classification codes to each of the reference set candidate features, is applied. Refinement of the classified reference set candidate features is optionally applied to broaden or narrow the reference set candidate features for inclusion in the reference set. The refinement can include applying further selection criteria to reference set documents during a second iteration.

Alternatively, the selection criteria can first be applied to documents and in a further iteration; the selection criteria are applied to features from the documents. Subsequently, documents associated with the reference set candidate features are grouped as the reference set.

The candidate criteria can be applied to a document set to identify reference set candidates for potential inclusion in the reference set. FIGURE 3 is a data flow diagram 60 showing examples of the selection criteria of FIGURE 2. The selection criteria 61 include clustering 62, features 63, assignments 64, document seeding 65, and random sampling 66. Other selection criteria are possible. Clustering 62 includes grouping documents by similarity and subsequently selecting documents from one or more of the clusters. A number of documents to be selected can be predetermined by a reviewer or machines, as further described below with reference to FIGURES 4 and 5. Features 63 include metadata about the documents, including nouns, noun phrases, length of document, "To" and "From" fields, date, complexity of sentence structure, and concepts. Assignments 64 include a subset of documents selected from a larger collection of uncoded document to be reviewed. The assignments can be generated based on assignment criteria, such as content, size, or number of reviewers. Other features, assignments, and assignment criteria are possible.

Document seeding 65 includes selecting one or more seed documents and identifying documents similar to the seed documents from a larger collection of documents as reference set candidates. Document seeding is further discussed below in detail with reference to FIGURE 6. Random sampling 66 includes randomly selecting documents from a larger collection of documents as reference set candidates. Random sampling is further discussed below in detail with reference to FIGURE 7.

The process for generating a reference set can be iterative and each pass through the process can use different selection criteria, as described above with reference to FIGURE 2. Alternatively, a single pass through the process using only one selection criteria to generate a cohesive reference set is also possible. Use of the clustering selection criteria can identify and group documents by similarity. FIGURE 4 is a flow diagram showing, by way of example, a method for generating a reference set via hierarchical clustering. A collection of documents is obtained (block 71) and filter criteria can optionally be applied to reduce a number of the documents (block 72). The documents are then clustered (block 73) to generate a hierarchical tree via hierarchical clustering. Hierarchical clustering, including agglomerative or devisive clustering, can be used to generate the clusters of documents, which can be used to identify a set of reference documents having a particular predetermined size. During agglomerative clustering, each document is assigned to a cluster and similar clusters are combined to generate the

hierarchical tree. Meanwhile, during devisive clustering, all the documents are grouped into a single cluster and subsequently divided to generate the hierarchical tree.

The clusters of the hierarchical tree can be traversed (block 74) to identify *n*-documents as reference set candidates (block 75). The *n*-documents can be predetermined by a user or a machine. In one embodiment, the *n*-documents are influential documents, meaning that a decision made for the *n*-document, such as the assignment of a classification code, can be propagated to other similar documents. Using influential documents can improve the speed and classification consistency of a document review.

To obtain the *n*-documents, *n*-clusters can be identified during the traversal of the hierarchical tree and one document from each of the identified clusters can be selected. The single document selected from each cluster can be the document closest to the cluster center or another documents. Other values of *n* are possible, such as *n*/2. For example, *n*/2 clusters are identified during the traversal and two documents are selected from each identified cluster. In one embodiment, the selected documents are the document closest to the cluster center and the document furthest from the cluster center. However, other documents can be selected, such as randomly picked documents.

Once identified, the reference set candidates are analyzed and a candidate decision is made (block 76). During the analysis, a classification code is assigned to each reference set candidate and a determination of whether that reference set candidate is appropriate for the reference set is made. If one or more of the reference set candidates are not sufficient for the reference set, refinement of the reference set candidates may optionally occur (block 77) by reclustering the reference set candidates (block 73). Refinement can include changing input parameters of the clustering process and then reclustering the documents, changing the document collection by filtering different documents, or selecting a different subset of *n*-documents from the clusters. Other types of and processes for refinement are possible.. The refinement assists in narrowing the number of reference set candidates to generate a reference set of a particular size during which reference set candidates can be added or removed. One or more of the reference set candidates are grouped to form the reference set (block 78). The size of the reference set can be predetermined by a human reviewer or a machine.

In a further embodiment, features can be used to identify documents for inclusion in a reference set. A collection of documents is obtained and features from the documents are identified. Filter criteria can optionally be applied to the features to reduce the number of potential documents for inclusion in the reference set. The features are then grouped into clusters, which are traversed to identify n-features as reference set candidate features. A

candidate decision, including the assignment of classification codes, is applied to each of the reference set candidate features and refinement of the features is optional. Documents associated with the classified reference set candidate features are then grouped as the reference set.

Iterative clustering is a specific type of hierarchical clustering that provides a reference set of documents having an approximate size. FIGURE 5 is a flow diagram showing, by way of example, a method for generating a reference set via iterative clustering. A collection of documents is obtained (block 81). The documents can be optionally divided into assignments (block 82), or groups of documents, based on document characteristics, including metadata about the document. In general, existing knowledge about the document is used to generate the assignments. Other processes for generating the assignments are possible. In one embodiment, attachments to the document can be included in the same assignment as the document, and in an alternative embodiment, the attachments are identified and set aside for review or assigned to a separate assignment. The documents are then grouped into clusters (block 83). One or more documents can be selected from the clusters as reference set candidates (block 84). In one embodiment, two documents are selected, including the document closest to the cluster center and the document closest to the edge of the cluster. The document closest to the center provides information regarding the center of the cluster, while the outer document provides information regarding the edge of the cluster. Other numbers and types of documents can be selected.

The selected documents are then analyzed to determine whether a sufficient number of documents have been identified as reference set candidates (block 85). The number of documents can be based on a predefined value, threshold, or bounded range selected by a reviewer or a machine. If a sufficient number of reference set candidates are not identified, further clustering (block 83) is performed on the reference set candidates until a sufficient number of reference set candidates exists. However, if a sufficient number of reference set candidates are identified, the candidates are analyzed and a candidate decision is made (block 86). For example, a threshold can define a desired number of documents for inclusion in the reference set. If the number of reference set candidates is equal to or below the threshold, those candidates are further analyzed, whereas if the number of reference set candidates is above the threshold, further clustering is performed until the number of candidates is sufficient. In a further example, a bounded range, having an upper limitation and a lower limitation, is determined and if the number of reference set candidates falls within the bounded range, those reference set candidates are further analyzed.

The candidate decision includes coding of the documents and a determination as to whether each reference set candidate is a good candidate for inclusion in the reference set. The

coded reference set candidates form the reference set (block 87). Once formed, the reference set can be used as a group of exemplar documents to classify uncoded documents.

In a further embodiment, features can be used to identify documents for inclusion in the reference set. A collection of documents is obtained and features are identified within the documents. The features can optionally be divided into one or more assignments. The features are then grouped into clusters and at least one feature is selected from one or more of the clusters. The selected features are compared with a predetermined number of documents for inclusion in the reference set. If the predetermined number is not satisfied, further clustering is performed on the features to increase or reduce the number of features. However, if satisfied, the selected features are assigned classification codes. Refinement of the classified features is optional. Subsequently, documents associated with the classified features are identified and grouped as the reference set.

The selection criteria used to identify reference set candidates can include document seeding, which also groups similar documents. FIGURE 6 is a flow diagram showing, by way of example, a method for generating a reference set via document seeding. A collection of documents is obtained (block 91). The collection of documents includes unmarked documents related to a topic, legal matter, or other theme or purpose. The documents can be optionally grouped into individual assignments (block 92). One or more seed documents are identified (block 93). The seed documents are considered to be important to the topic or legal matter and can include documents identified from the current matter, documents identified from a previous matter, or pseudo documents.

The seed documents from the current case can include the complaint filed in a legal proceeding for which documents are to be classified or other documents, as explained *supra*. Alternatively, the seed documents can be quickly identified using a keyword search or knowledge obtained from a reviewer. In a further embodiment, the seed documents can be identified as reference set candidates identified in a first pass through the process described above with reference to FIGURE 2. The seed documents from a previous related matter can include one or more of the reference documents from the reference set generated for the previous matter. The pseudo documents use knowledge from a reviewer or other user, such as a party to a lawsuit, as described above with reference to FIGURE 2.

The seed documents are then applied to the document collection or at least one of the assignments and documents similar to the seed documents are identified as reference set candidates (block 94). In a further embodiment, dissimilar documents can be identified as reference set candidates. In yet a further embodiment, the similar and dissimilar documents can

be combined to form the seed documents. The similar and dissimilar documents can be identified using criteria, including document injection, linear search, and index look up. However, other reference set selection criteria are possible.

The number of reference set candidates are analyzed to determine whether there are a sufficient number of candidates (block 95). The number of candidates can be predetermined and selected by a reviewer or machine. If a sufficient number of reference set candidates exist, the reference set candidates form the reference set (block 97). However, if the number of reference set candidates is not sufficient, such as too large, refinement of the candidates is performed to remove one or more reference candidates from the set (block 96). Large reference sets can affect the performance and outcome of document classification. The refinement assists in narrowing the number of reference set candidates to generate a reference set of a particular size. If refinement is to occur, further selection criteria are applied to the reference set candidates. For example, if too many reference set candidates are identified, the candidate set can be narrowed to remove common or closely related documents, while leaving the most important or representative document in the candidate set. The common or closely related documents can be identified as described in commonly-assigned U.S. Patent No. 6,745,197, entitled "System and Method for Efficiently Processing Messages Stored in Multiple Message Stores," issued on June 1, 2004, and U.S. Patent No. 6,820,081, entitled "System and Method for Evaluating a Structured Message Store for Message Redundancy," issued on November 16, 2004, the disclosures of which are incorporated by reference. Additionally, the common or closely related documents can be identified based on influential documents, which are described above with reference to FIGURE 4, or other measures of document similarity. After the candidate set has been refined, the remaining reference set candidates form the reference set (block 97).

In a further embodiment, features can be used to identify documents for inclusion in the reference set. A collection of documents is obtained and features from the documents are identified. The features are optionally divided into assignments. Seed features are identified and applied to the identified features. The features similar to the seed features are identified as reference set candidate features and the similar features are analyzed to determine whether a sufficient number of reference set candidate features are identified. If not, refinement can occur to increase or decrease the number of reference set candidate features until a sufficient number exists. If so, documents associated with the reference set candidate features are identified and grouped as the reference set.

Random sampling can also be used as selection criteria to identify reference set candidates. FIGURE 7 is a flow diagram showing, by way of example, a method for generating

a reference set via random sampling. A collection of documents is obtained (block 101), as described above with reference to FIGURE 2. The documents are then grouped into categories (block 102) based on metadata about the documents. The metadata can include date, file, folder, fields, and structure. Other metadata types and groupings are possible. Document identification values are assigned (block 103) to each of the documents in the collection. The identification values can include letters, numbers, symbols or color coding, as well as other values, and can be human readable or machine readable. A random machine generator or a human reviewer can assign the identification values to the documents. Subsequently, the documents are randomly ordered into a list (block 104) and the first *n*-documents are selected from the list as reference candidates (block 105). In a further embodiment, the document identification values are provided to a random number generator, which randomly selects *n* document identification values. The documents associated with the selected identification values are then selected as the reference set candidates. The number of *n*-documents can be determined by a human reviewer, user, or machine. The value of *n* dictates the size of the reference set. The reference candidates are then coded (block 106) and grouped as the reference set (block 107).

In a further embodiment, features or terms selected from the documents in the collection can be sampled. Features can include metadata about the documents, including nouns, noun phrases, length of document, "To" and "From" fields, date, complexity of sentence structure, and concepts. Other features are possible. Identification values are assigned to the features and a subset of the features or terms are selected, as described *supra*. Subsequently, the subset of features is randomly ordered into a list and the first *n*-features are selected as reference candidate features. The documents associated with the selected reference candidate features are then grouped as the reference set. Alternatively, the number of *n*-features can be randomly selected by a random number generator, which provides *n*-feature identification values. The features associated with the selected *n*-feature identification values are selected as reference candidate features.

Reference sets for coding documents by a human reviewer or a machine can be the same set or a different set. Reference sets for human reviewers should be cohesive; but need not be representative of a collection of documents since the reviewer is comparing uncoded documents to the reference documents and identifying the similar uncoded documents to assign a classification code. Meanwhile, a reference or "training" set for classifiers should be representative of the collection of documents, so that the classifier can distinguish between documents having different classification codes. FIGURE 8 is a flow diagram showing, by way of example, a method 110 for generating a reference set with user assistance. A collection of

documents associated with a topic or legal issue is obtained (block 111). A reviewer marks one or more of the documents in the collection by assigning a classification code (block 112). Together, the classified documents can form an initial or candidate reference set, which can be subsequently tested and refined. The reviewer can randomly select the documents, receive review requests for particular documents by a classifier, or receive a predetermined list of documents for marking. In one embodiment, the documents marked by the reviewer can be considered reference documents, which can be used to train a classifier.

While the reviewer is marking the documents, a machine classifier analyzes the coding decisions provided by the reviewer (block 113). The analysis of the coding decisions by the classifier can include one or more steps, which can occur simultaneously or sequentially. In one embodiment, the analysis process is a training or retraining of the classifier. Retraining of the classifier can occur when new information, such as documents or coding decisions are identified. In a further embodiment, multiple classifiers are utilized. Thereafter, the classifier begins classifying documents (block 114) by automatically assigning classification codes to the documents. The classifier can begin classification based on factors, such as a predetermined number of documents for review by the classifier, after a predetermined time period has passed, or after a predetermined number of documents in each classification category is reviewed. For instance, in one embodiment, the classifier can begin classifying documents after analyzing at least two documents coded by the reviewer. As the number of documents analyzed by the classifier prior to classification increases, a confidence level associated with assigned classification codes by the classifier can increase. . The classification codes provided by the classifier are compared (block 115) with the classification codes for the same documents provided by the reviewer to determine whether there is a disagreement between the assigned codes (block 116). For example, a disagreement exists when the reviewer assigns a classification code of "privileged" to a document and the classifier assigns the same document a classification code of "responsive."

If a disagreement does not exist (block 116), the classifier begins to automatically classify documents (block 118). However, if a disagreement exists (block 116), a degree of the disagreement is analyzed to determine whether the disagreement falls below a predetermined threshold (block 117). The predetermined threshold can be measured using a percentage, bounded range, or value, as well as other measurements. In one embodiment, the disagreement threshold is set as 99% agreement, or alternatively as 1% disagreement. In a further embodiment, the predetermined threshold is based on a number of agreed upon documents. For example, the threshold can require that the last 100 documents coded by the reviewer and the

classifier be in agreement. In yet a further embodiment, zero-defect testing can be used to determine the threshold. A defect can be a disagreement in a coding decision, such as an inconsistency in the classification code assigned. An error rate for classification is determined based on the expected percentages that a particular classification code will be assigned, as well as a confidence level. The error rate can include a percentage, number, or other value. A collection of documents is randomly sampled and marked by the reviewer and classifier. If a value of documents with disagreed upon classification codes exceeds the error rate, further training of the classifier is necessary. However, if the value of documents having a disagreement falls below the error rate, automated classification can begin.

If the disagreement value is below the threshold, the classifier begins to automatically classify documents (block 118). If not, the reviewer continues to mark documents from the collection set (block 112), the classifier analyzes the coding decisions (block 113), the classifier marks documents (block 114), and the classification codes are compared (block 115) until the disagreement of the classification codes assigned by the classifier and the reviewer falls below the predetermined threshold.

In one embodiment, the disagreed upon documents can be selected and grouped as the reference set. Alternatively, all documents marked by the classifier can be included in the reference set, such as the agreed and disagreed upon documents.

In a further embodiment, features can be used to identify documents for inclusion in the reference set. A collection of documents is obtained and features are identified from the collection. A reviewer marks one or more features by assigning classification codes and provides the marked features to a classifier for analysis. After the analysis, the classifier also begins to assign classification codes to the features. The classification codes assigned by the reviewer and the classifier for a common feature are compared to determine whether a disagreement exists. If there is no disagreement, classification of the features becomes automated. However, if there is disagreement, a threshold is applied to determine whether the disagreement falls below threshold. If so, classification of the features becomes automated. However, if not, further marking of the features and analysis occurs.

Reference sets generated using hierarchical clustering, iterative clustering, random sampling, and document seeding rely on the human reviewer for coding of the reference documents. However, a machine, such as a classifier, can also be trained to identify reference sets for use in classifying documents. FIGURE 9 is a flow diagram showing, by way of example, a method for generating a reference set via active learning. A set of coded documents is obtained (block 121). The set of documents can include a document seed set or a reference

set, as well as other types of document sets. The document set can be obtained from a previous related topic or legal matter, as well as from documents in the current matter. The coding of the document set can be performed by a human reviewer or a machine. The document set can be used to train one or more classifiers (block 122) to identify documents for inclusion in a reference set. The classifiers can be the same or different, including nearest neighbor or Support Vector Machine classifiers, as well as other types of classifiers. The classifiers review and mark a set of uncoded documents for a particular topic, legal matter, theme, or purpose by assigning a classification code (block 123) to each of the uncoded documents. The classification codes assigned by each classifier for the same document are compared (block 124) to determine whether there is a disagreement in classification codes provided by the classifiers (block 125). A disagreement exists when one document is assigned different classification codes. If there is no disagreement, the classifiers continue to review and classify the uncoded documents (block 123) until there are no uncoded documents remaining. Otherwise, if there is a disagreement, the document is provided to a human reviewer for review and marking. The human reviewer provides a new classification code or confirms a classification code assigned by one of the classifiers (block 126). The classifiers that incorrectly marked the document and reviewer assigned classification code (block 127) can be analyzed for further training. For the classifiers that correctly marked the document (block 127), no additional training need occur. The documents receiving inconsistent classification codes by the classifiers form the reference set (block 128). The reference set can then be used to train further classifiers for classifying documents.

In a further embodiment, features can be analyzed to identify reference documents for inclusion in a reference set. A collection of coded documents, such as a seed set or reference set, is obtained. The document set can be obtained from a previous related topic, legal matter, theme or purpose, as well as from documents in the current matter. Features within the document set are identified. The features can include metadata about the documents, including nouns, noun phrases, length of document, to and from fields, date, complexity of sentence structure, and concepts. Other features are possible. The identified features are then classified by a human reviewer and used to train one or more classifiers. Once trained, the classifiers review a further set of uncoded documents, identify features within the further set of uncoded documents, and assign classification codes to the features. The classification codes assigned to a common feature by each classifier are compared to determine whether a discrepancy in the assigned classification code exists. If not, the classifiers continue to review and classify the features of the uncoded documents until no uncoded documents remain. If there is a classification

disagreement, the feature is provided to a human reviewer for analysis and coding. The classification code is received from the user and used to retrain the classifiers, which incorrectly coded the feature. Documents associated with the disagreed upon features are identified and grouped to form the reference set.

Feature selection can be used to identify specific areas of two or more documents that are interesting based on the classification disagreement by highlighting or marking the areas of the documents containing the particular disagreed upon features. Documents or sections of documents can be considered interesting based on the classification disagreement because the document data is prompting multiple classifications and should be further reviewed by a human reviewer.

In yet a further embodiment, a combination of the reference documents identified by document and the reference documents identified by features can be combined to create a single reference set of documents.

The reference set can be provided to a reviewer for use in manually coding documents or can be provided to a classifier for automatically coding the documents. In a further embodiment, different reference sets can be used for providing to a reviewer and a classifier. FIGURE 10 is a flow diagram 130 showing, by way of example, a method for generating a training set for a classifier. A set of coded document, such as a reference set, is obtained (block 131). One or more classifiers can be trained (block 132) using the reference set. The classifiers can be the same or different, such as a nearest neighbor classifier or a Support Vector Machine classifier. Other types of classifiers are possible. Once trained, the classifiers are each run over a common sample of assignments to classify documents in that assignment (block 133). The classification codes assigned by each classifier are analyzed for the documents and a determination of whether the classifiers disagree on a particular classification code is made (block 134). If there is no disagreement (block 134), the classifiers are run over further common samples (block 133) of assignments until disagreed upon documents are identified. However, if there is disagreement between the classifiers on a document marking, the classified document in disagreement must then be reviewed (block 135) and identified as training set candidates. A further classification code is assigned to the classified document in disagreement (block 137). The further classification code can be assigned by a human reviewer or a machine, such as one of the classifiers or a different classifier. The classifiers can each be optionally updated (block 132) with the newly assigned code. The review and document coding can occur manually by a reviewer or automatically. The training set candidates are then combined with the reference set (block 137). A stop threshold is applied (block 138) to the combined training set candidates and

reference set to determine whether each of the documents is appropriate for inclusion in the training set. The stop threshold can include a predetermined training set size, a breadth of the training set candidates with respect to the feature space of the reference set, or the zero defect test. Other types of tests and processes for determining the stopping threshold are possible. If the threshold is not satisfied, the classifiers are run over further assignments (block 133) for classifying and comparing. Otherwise, if satisfied, the combined training set candidates and reference set form the training set (block 139). Once generated, the training set can be used for automatic classification of documents, such as described above with reference to FIGURE 8.

In a further embodiment, features can be used to identify documents for inclusion in the reference set. A set of coded documents is obtained and features are identified from the coded documents. Classifiers are trained using the features and then run over a random sample of features to assign classification codes to the features. The classification codes for a common feature are compared to determine whether a disagreement exists. If not, further features can be classified. However, if so, the disagreed upon features are provided to a reviewer for further analysis. The reviewer can assign further classification codes to the features, which are grouped as training set candidate features. The documents associated with the training set candidate features can be identified as training set candidates and combined with the coded documents. A stop threshold is applied to determine whether each of the documents is appropriate for inclusion in the reference set. If so, the training set candidates and coded documents are identified as the training set. However, if not, further coding of features is performed to identify training set candidates appropriate for inclusion in the reference set.

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 method (50) for generating a reference set (14b) for use during document review, comprising:

obtaining a collection of unclassified documents (14a);

applying selection criteria (61) to the collection and selecting those unclassified documents (14a) that satisfy the selection criteria (61) as reference set candidates;

assigning a classification code (55) to each reference set candidate; and

forming a reference set (14b) from the classified reference set candidates, wherein the reference set (14b) comprises coded documents (14a) that are quality controlled and shared between one or more reviewers.

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

refining the reference set (14b) by reducing a number of reference set candidates included in the reference set (14b).

3. A method (50) according to Claim 1, wherein the selection criteria (61) comprises random sampling (100), further comprising:

randomly ordering the unclassified documents (14a) into a list; and

selecting a predetermined number (105) of the unclassified documents (14a) from the list as the reference set candidates.

4. A method (50) according to Claim 1, wherein the selection criteria (61) comprises random sampling (100), further comprising:

assigning document identification values to each of the unclassified documents (14a);

selecting one or more of the document identification values (105); and

identifying the unclassified documents (14a) associated with the selected document identification values as the reference set candidates.

5. A system (10) for generating a reference set (14b) for use during document review, comprising:

a collection of unclassified documents (14a);

a selection module to apply selection criteria (61) to the collection and to select those unclassified documents (14a) that satisfy the selection criteria (61) as reference set candidates;

a classification module to assign a classification code (55) to each reference set candidate; and

a data module to form a reference set (14b) from the classified reference set candidates, wherein the reference set (14b) comprises coded documents (14a) that are quality controlled and shared between one or more reviewers.

6. A system (10) according to Claim 5, further comprising:

refining the reference set (14b) by reducing a number of reference set candidates included in the reference set (14b).

7. A system (10) according to Claim 5, wherein the selection criteria (61) comprises random sampling (100), further comprising:

randomly ordering the unclassified documents (14a) into a list; and

selecting a predetermined number (105) of the unclassified documents (14a) from the list as the reference set candidates.

8. A system (10) according to Claim 5, wherein the selection criteria (61) comprises random sampling (100), further comprising:

assigning document identification values to each of the unclassified documents (14a);

selecting one or more of the document identification values (105); and

identifying the unclassified documents (14a) associated with the selected document identification values as the reference set candidates.

9. A method for generating a reference set (14b) via clustering, comprising:

obtaining (71, 81) a collection of documents (14a);

grouping the documents (14a) into clusters of documents (14a);

selecting (75, 84) one or more documents (14a) from at least one cluster as reference set candidates;

assigning (76, 86) a classification code to each of the reference set candidates; and

grouping the classified reference set candidates as the reference set (14b).

10. A method according to Claim 9, further comprising:

building a hierarchical tree of the clusters; and

traversing (74) the hierarchical tree to identify the reference set candidates.

11. A method according to Claim 9, further comprising:

applying (85) a size threshold to the reference set candidates; and

selecting the reference set candidates for inclusion in the reference set (14b) when the size threshold is satisfied.

12. A method according to Claim 9, further comprising:

applying (85) a size threshold to the reference set candidates; and

reclustering the reference set candidates when the size threshold is not satisfied until the size threshold is satisfied.

13. A method for generating a reference set (14b) via seed documents, comprising:

obtaining a collection of documents (14a);

identifying (93) one or more seed documents (14a);

comparing (94) the seed documents (14a) to the document collection and identifying those documents (14a) similar to the seed documents (14a) as reference set candidates;

applying (95) a size threshold to the reference set candidates; and

grouping the reference set candidates as the reference set (14b) when the size threshold is satisfied.

14. A method according to Claim 13, further comprising:

refining (96) the reference set (14b) by reducing a number of reference set candidates included in the reference set (14b).

15. A method according to Claim 13, further comprising:

selecting the seed documents (14a) from at least one of a current document set and a previously defined document set.

16. A method according to Claim 13, further comprising:

determining the seed documents (14a) using a keyword search.

17. A method for generating a training set for use during document review, comprising:

assigning classification codes to a set of documents (14a);

receiving further classification codes assigned to the same set of documents (14a);

comparing (124) the classification code for at least one document with the further classification code for that document;

determining (116, 125, 134) whether a disagreement exists between the assigned classification code and the further classification code for at least one document;

identifying those documents (14a) with disagreeing classification codes as training set candidates;

applying a stop threshold (117, 138) to the training set candidates; and

grouping the training set candidates as a training set when the stop threshold is satisfied.

18. A method according to Claim 17, further comprising:

assigning further classification codes to the documents (14a) for which a disagreement exists.

19. A method according to Claim 17, wherein the classification codes are assigned by a machine and the further classification codes are assigned by a reviewer.

20. A method according to Claim 17, wherein the classification codes are assigned by a machine and the further classification codes are assigned by a further machine.

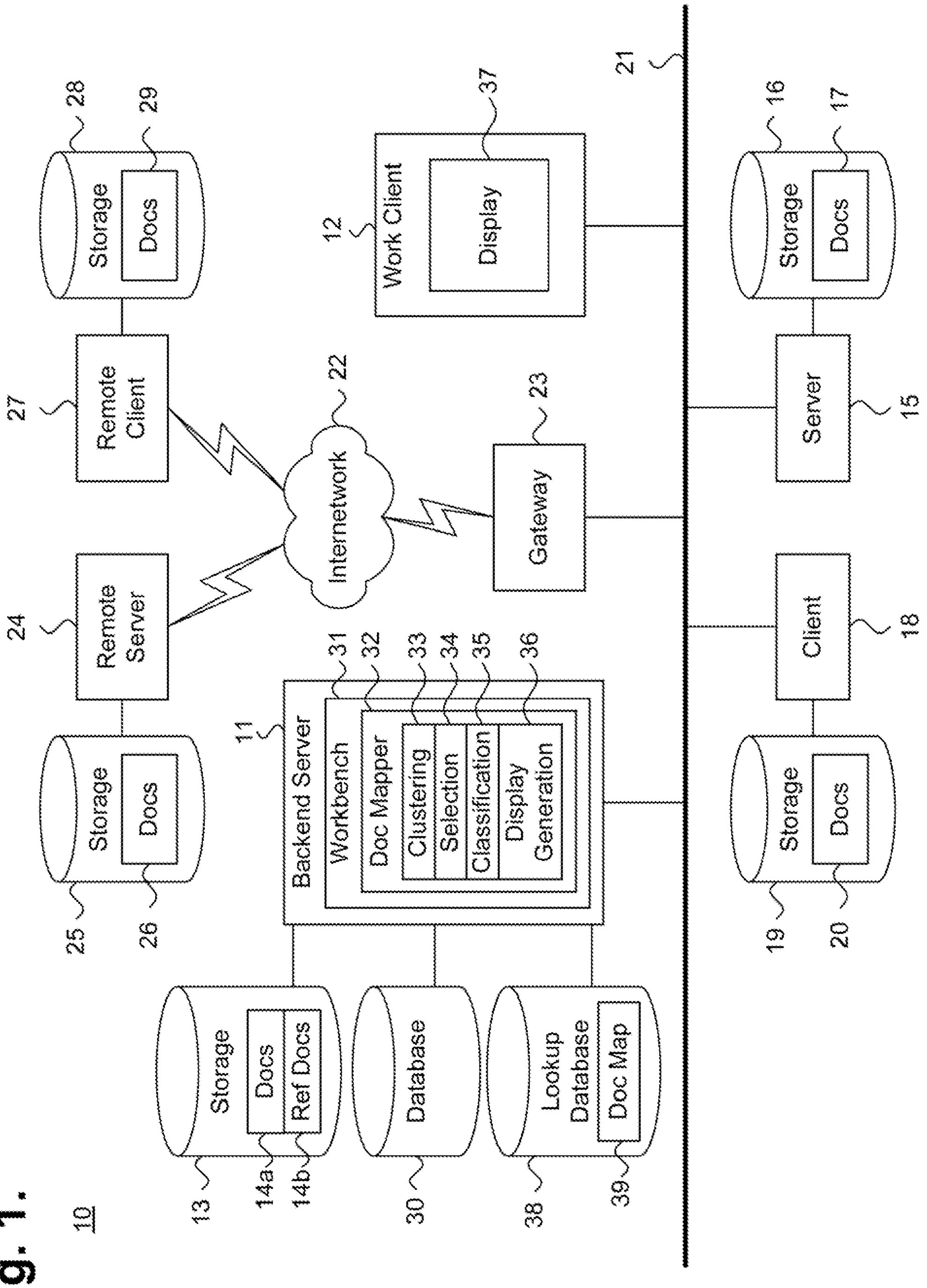

10
11
12
13
14a
14b
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
Storage
Docs
Ref Docs
Database
Lookup Database
Doc Map
Backend Server
Workbench
Doc Mapper
Clustering
Selection
Classification
Display Generation
Work Client
Display
Server
Client
Remote Client
Remote Server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

Fig. 1.

Fig. 2.

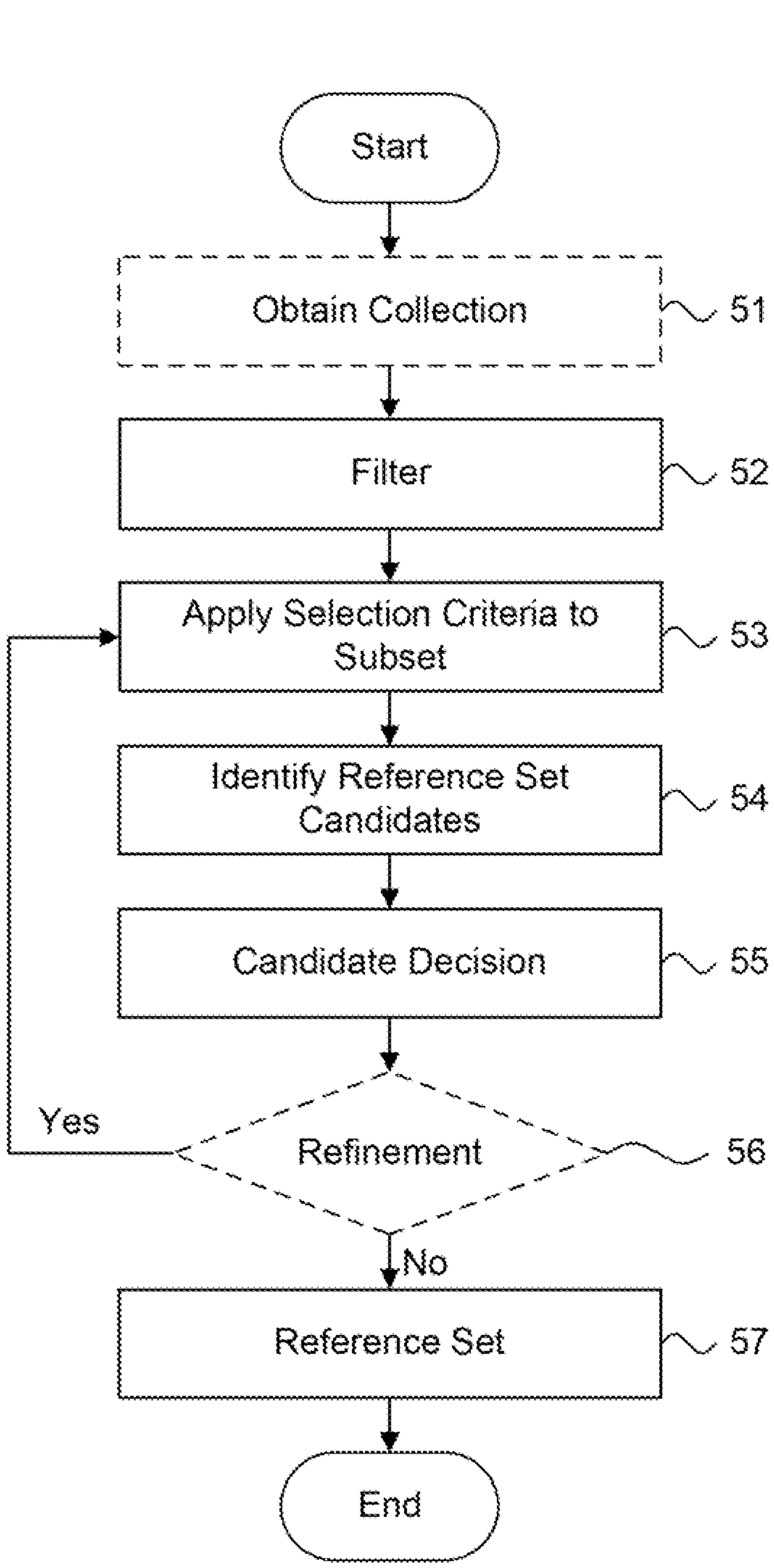

50
Start
Obtain Collection
51
Filter
52
Apply Selection Criteria to Subset
53
Identify Reference Set Candidates
54
Candidate Decision
55
Yes
Refinement
56
No
Reference Set
57
End

Fig. 3.

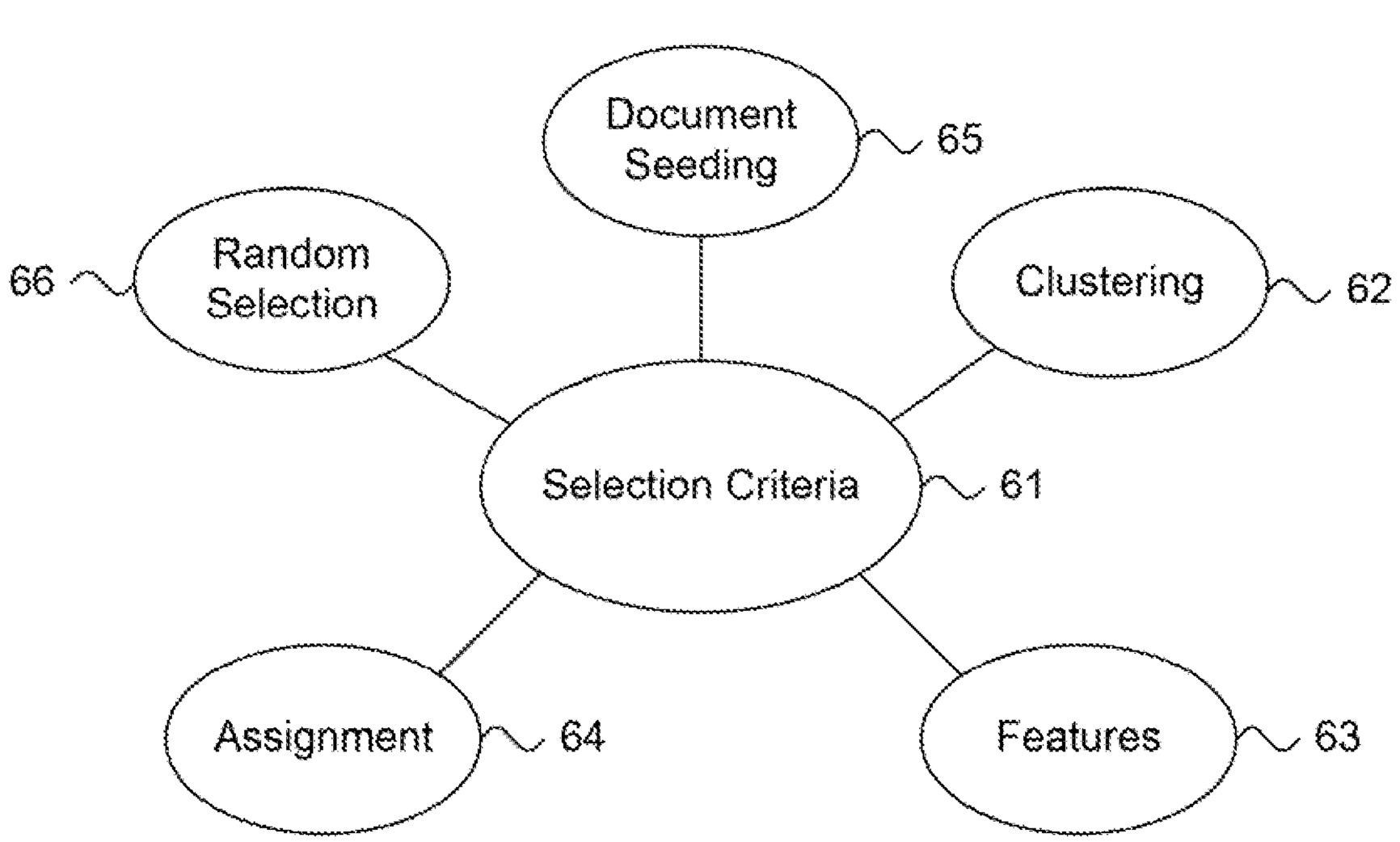

60
Document Seeding
65
Random Selection
66
Clustering
62
Selection Criteria
61
Assignment
64
Features
63

Fig. 4.

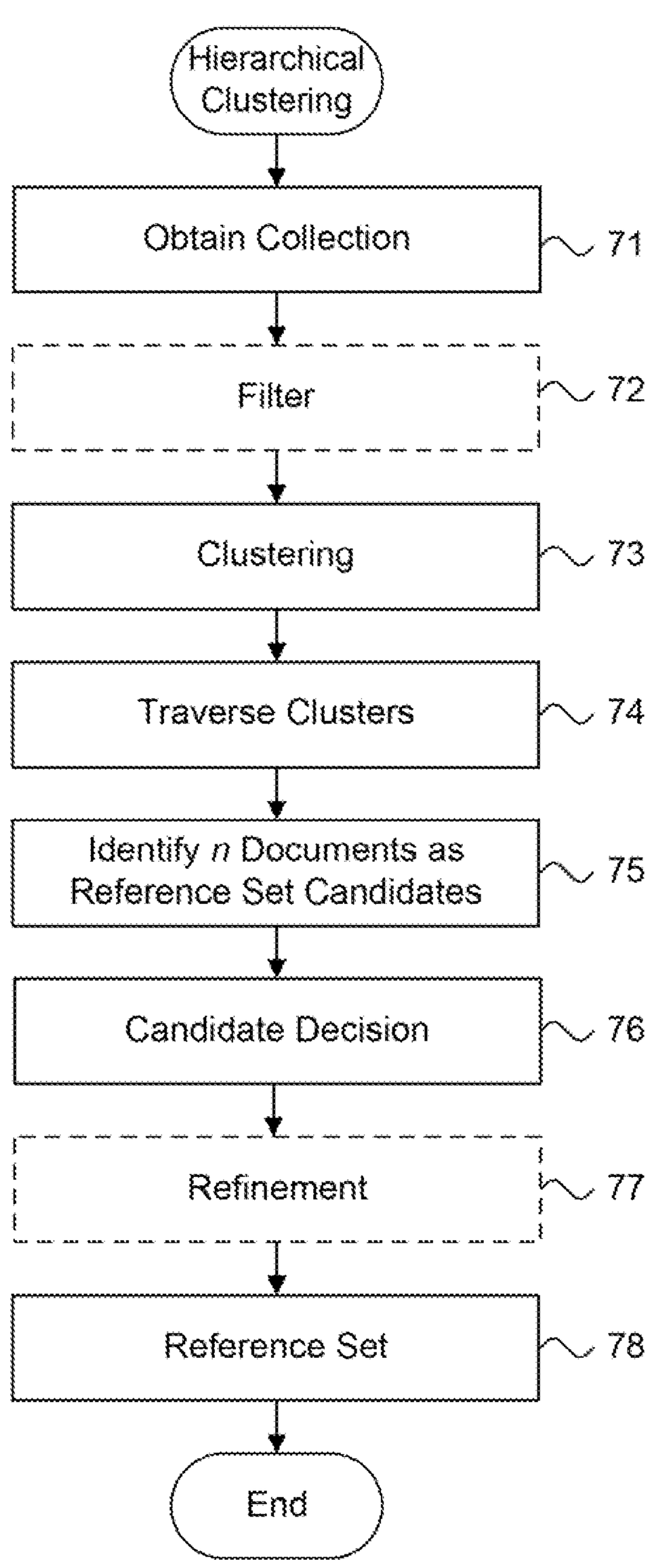

70
Hierarchical Clustering
Obtain Collection
71
Filter
72
Clustering
73
Traverse Clusters
74
Identify n Documents as Reference Set Candidates
75
Candidate Decision
76
Refinement
77
Reference Set
78
End

Fig. 5.

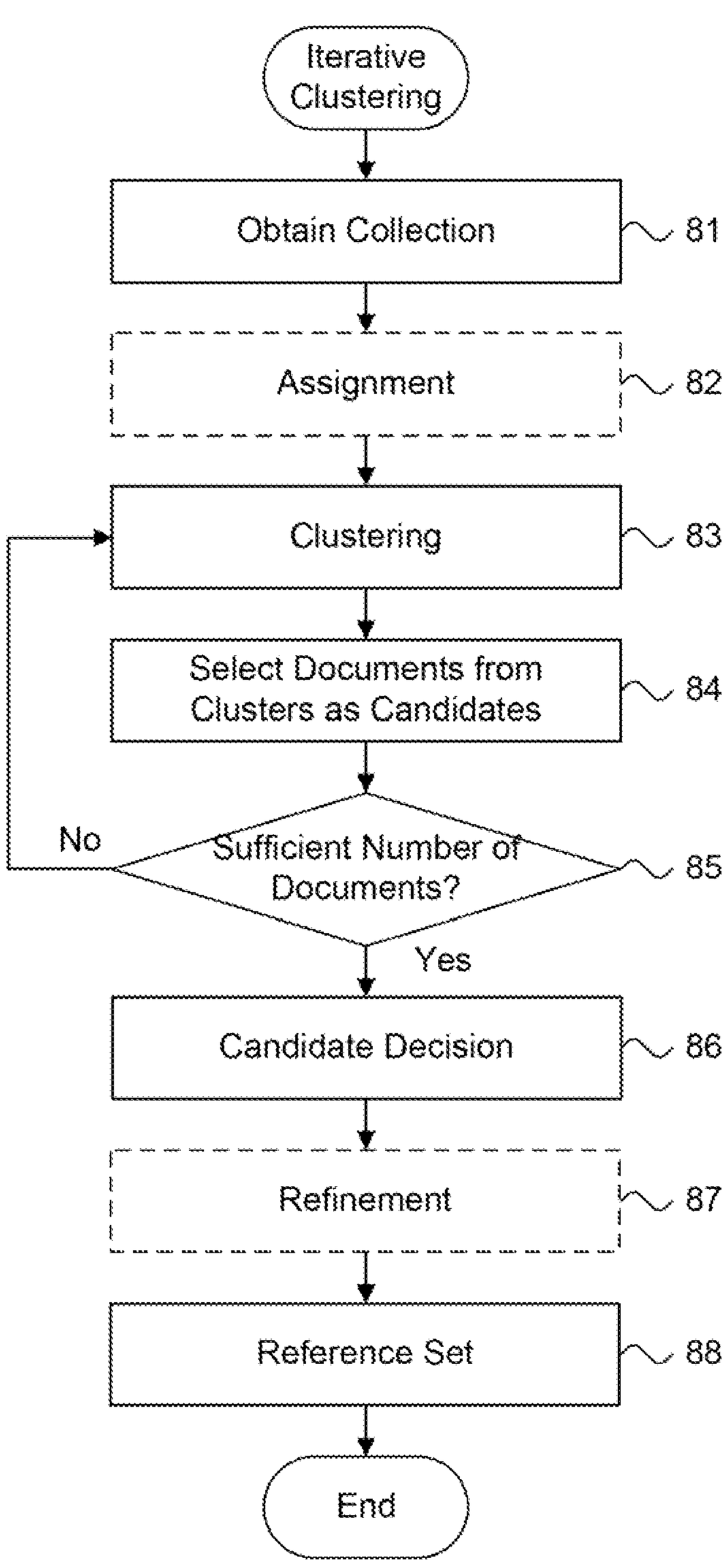

80
Iterative Clustering
Obtain Collection
81
Assignment
82
Clustering
83
Select Documents from Clusters as Candidates
84
No
Sufficient Number of Documents?
85
Yes
Candidate Decision
86
Refinement
87
Reference Set
88
End

Fig. 6.

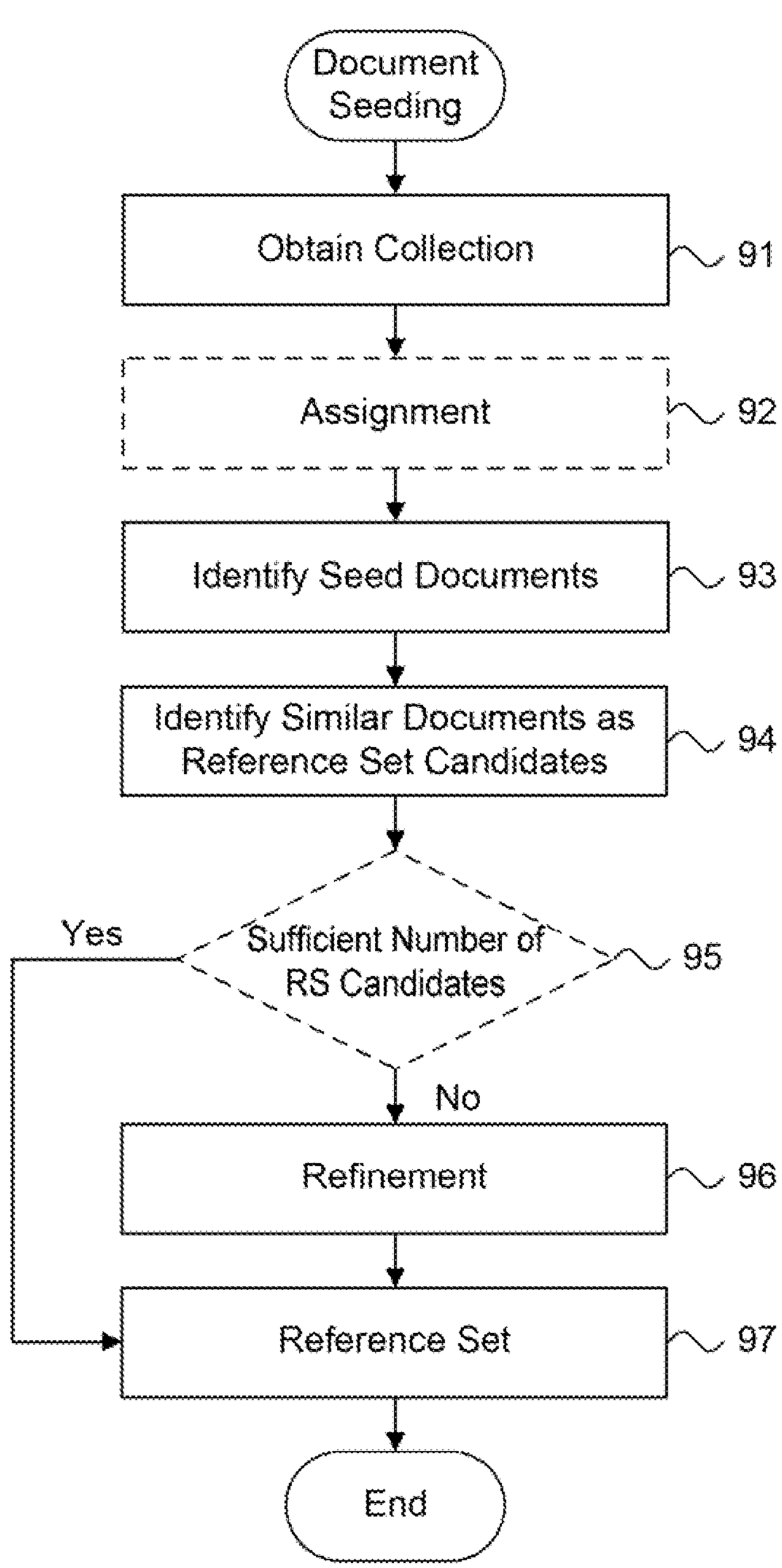

90
Document Seeding
Obtain Collection
91
Assignment
92
Identify Seed Documents
93
Identify Similar Documents as Reference Set Candidates
94
Sufficient Number of RS Candidates
95
Yes
No
Refinement
96
Reference Set
97
End

Fig. 7.

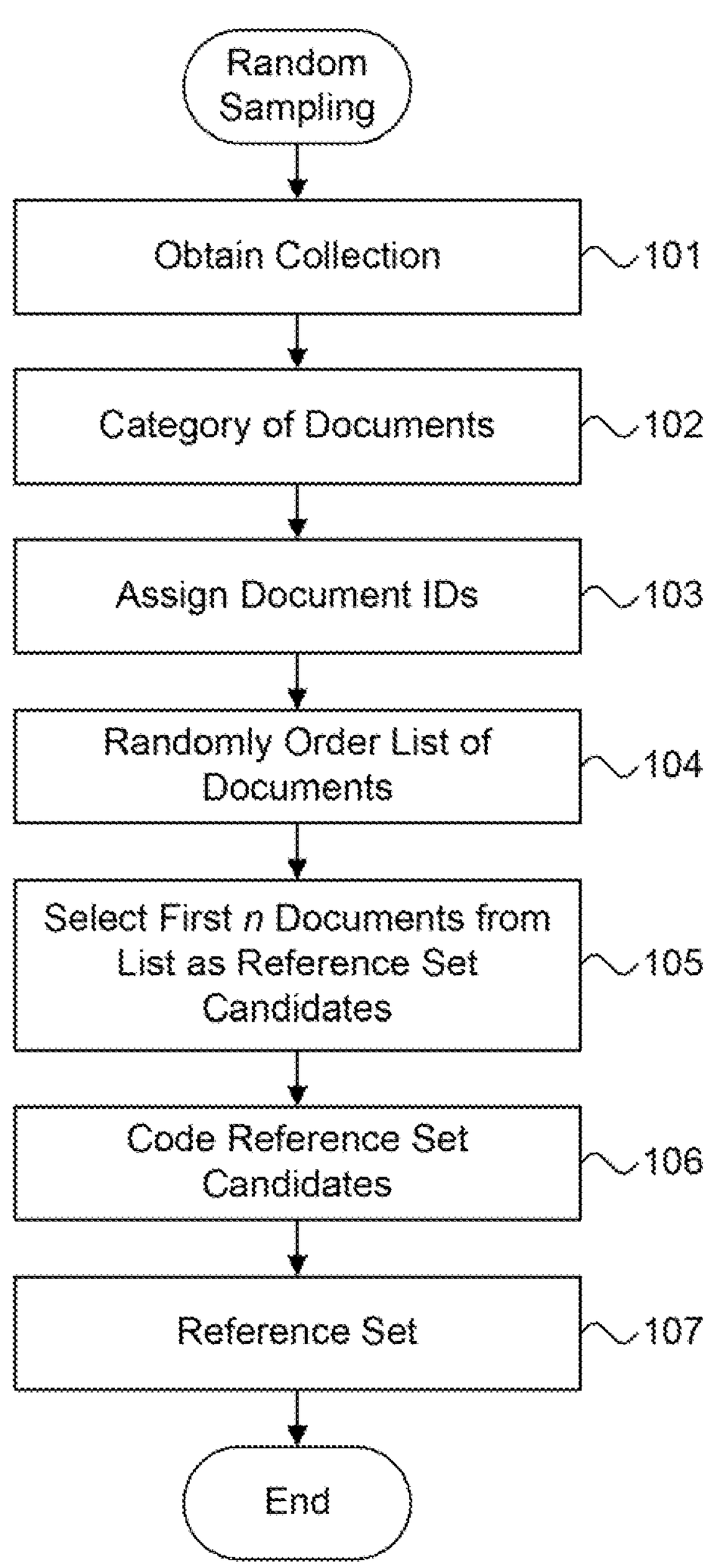

100
Random Sampling
Obtain Collection
101
Category of Documents
102
Assign Document IDs
103
Randomly Order List of Documents
104
Select First n Documents from List as Reference Set Candidates
105
Code Reference Set Candidates
106
Reference Set
107
End

Fig. 8.

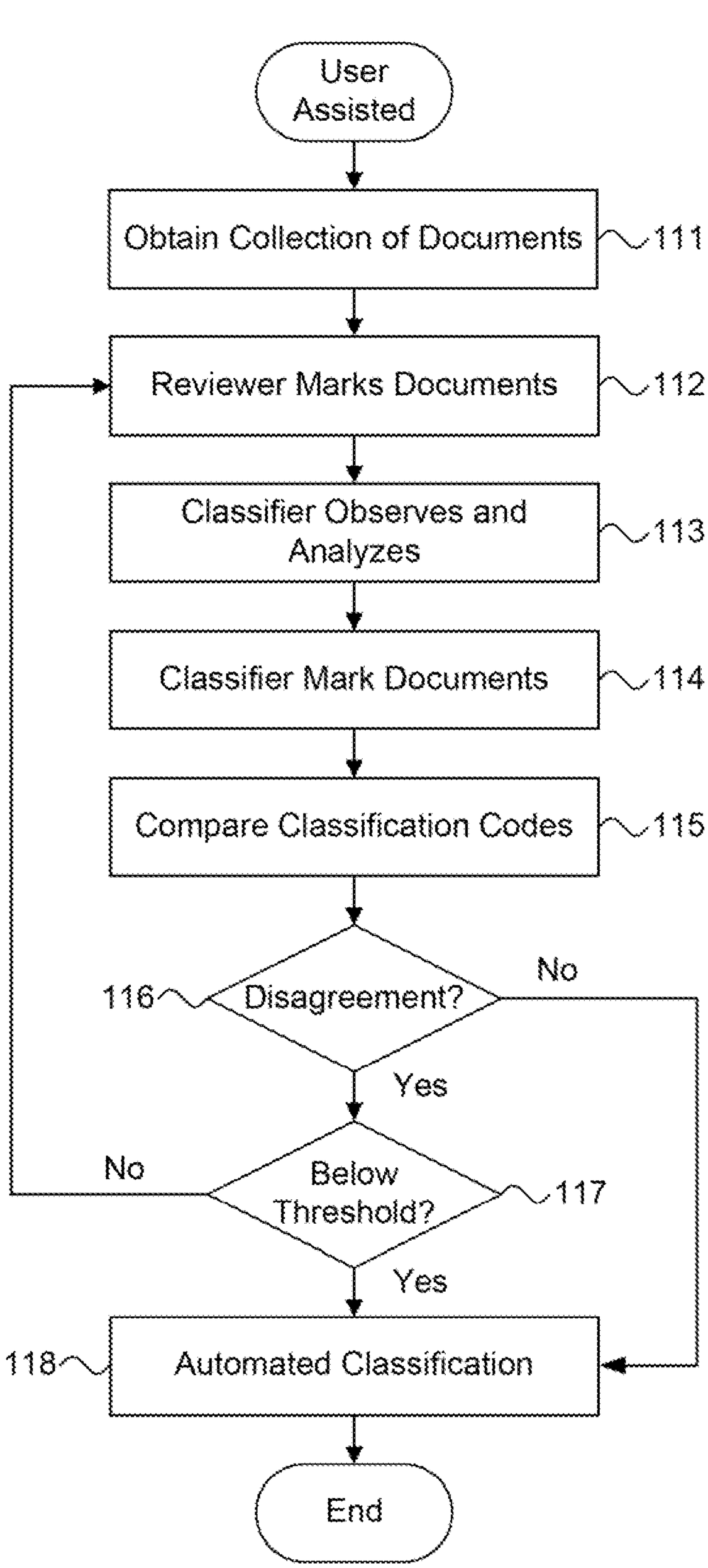

110
User Assisted
Obtain Collection of Documents
111
Reviewer Marks Documents
112
Classifier Observes and Analyzes
113
Classifier Mark Documents
114
Compare Classification Codes
115
116
Disagreement?
No
Yes
No
Below Threshold?
117
Yes
118
Automated Classification
End

Fig. 9.

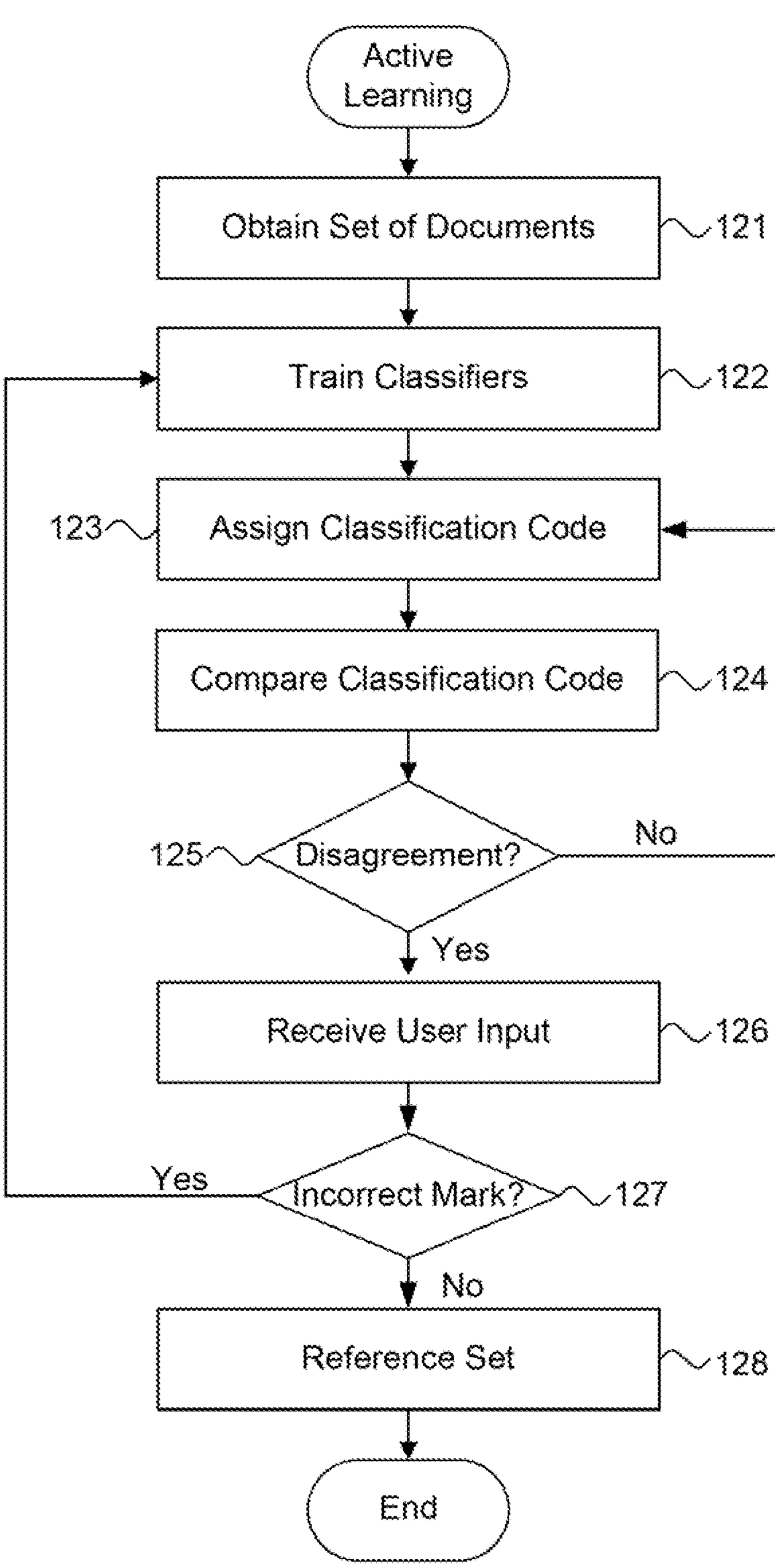

120
Active Learning
Obtain Set of Documents
121
Train Classifiers
122
123
Assign Classification Code
Compare Classification Code
124
125
Disagreement?
No
Yes
Receive User Input
126
Yes
Incorrect Mark?
127
No
Reference Set
128
End

Fig. 10.

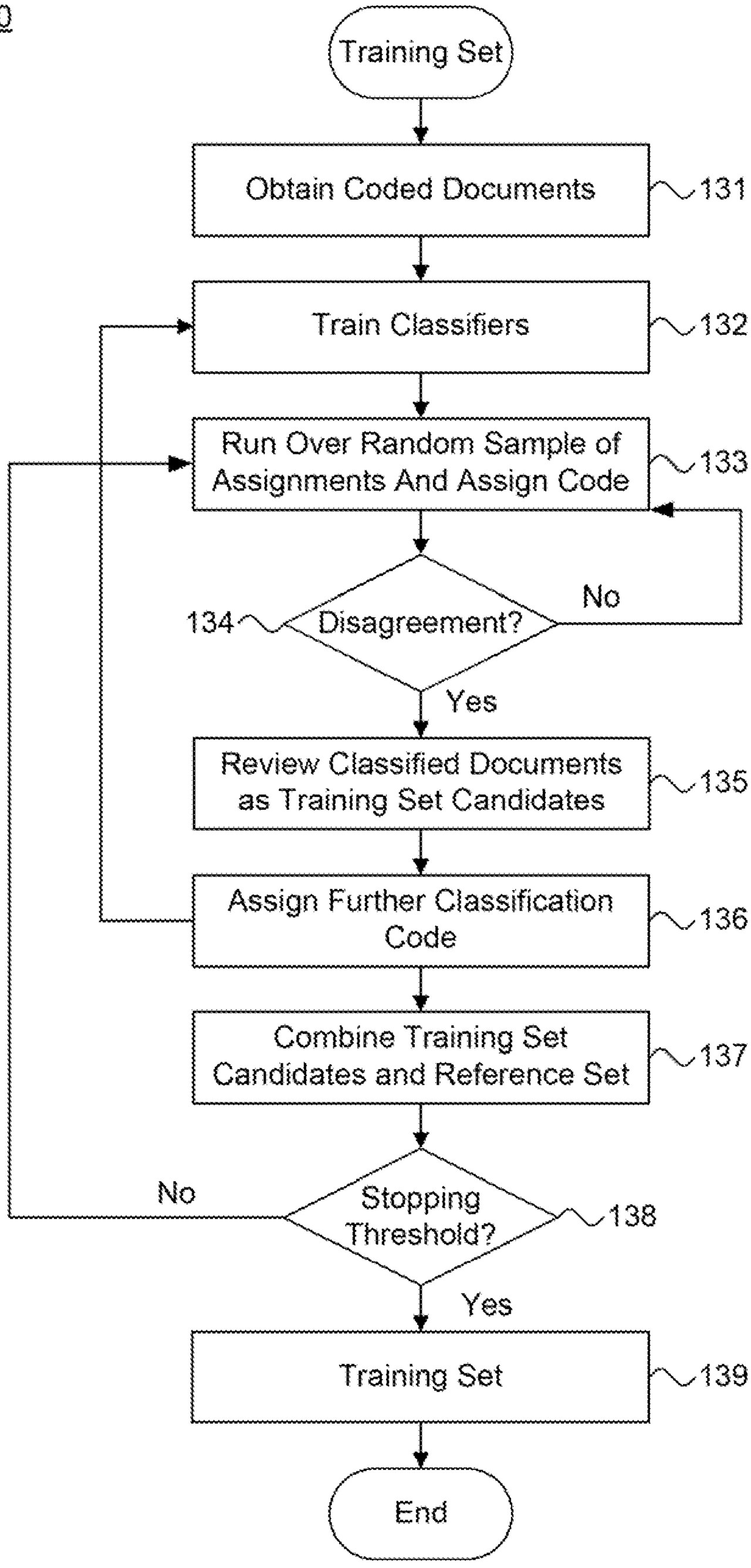
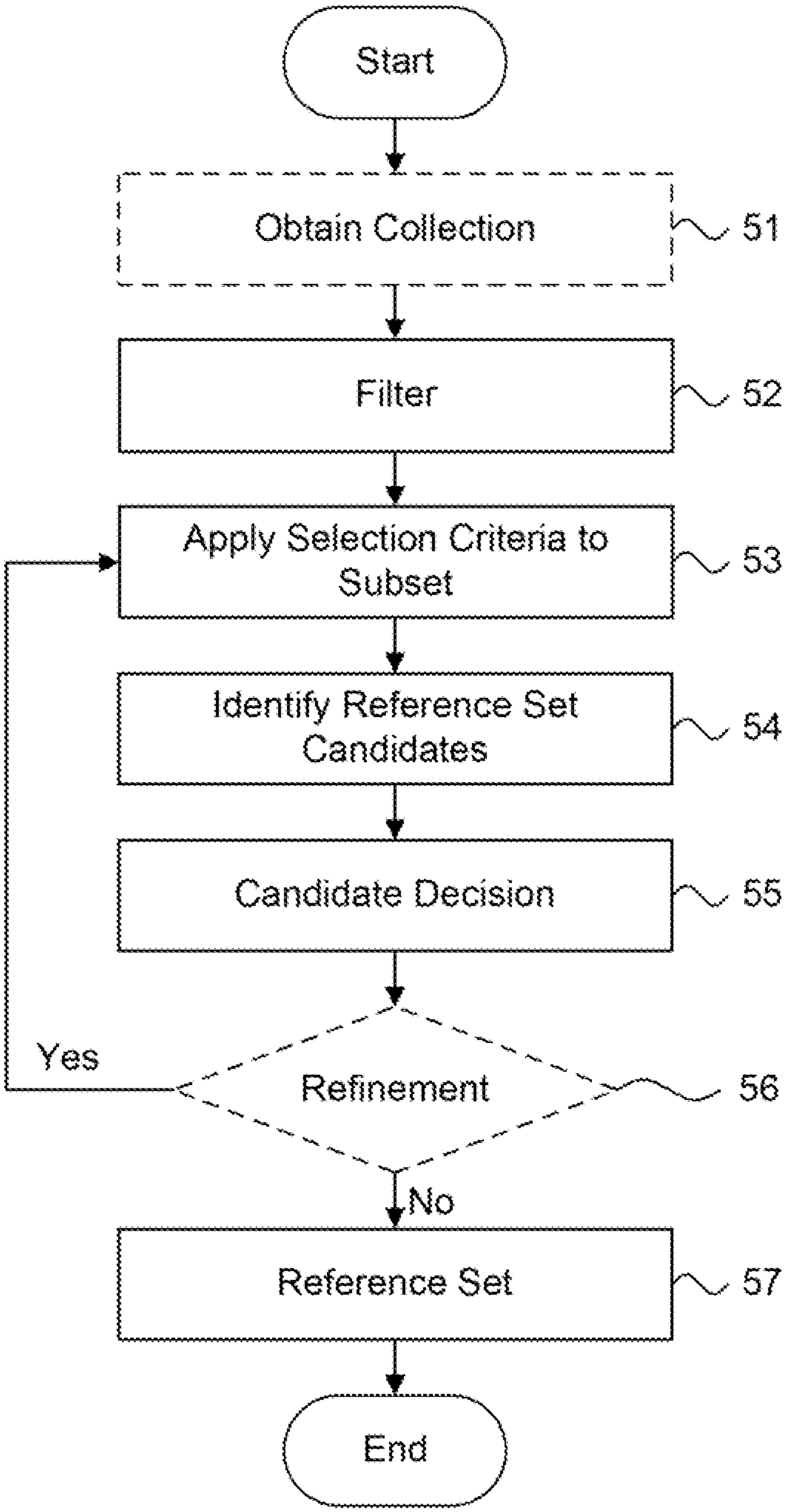

130
Training Set
Obtain Coded Documents
131
Train Classifiers
132
Run Over Random Sample of Assignments And Assign Code
133
134
Disagreement?
No
Yes
Review Classified Documents as Training Set Candidates
135
Assign Further Classification Code
136
Combine Training Set Candidates and Reference Set
137
No
Stopping Threshold?
138
Yes
Training Set
139
End

Fig. 2.


50
Start
Obtain Collection
51
Filter
52
Apply Selection Criteria to Subset
53
Identify Reference Set Candidates
54
Candidate Decision
55
Yes
Refinement
56
No
Reference Set
57
End