US 20160253408A1

(19) **United States**

(12) **Patent Application Publication**

Knight et al.

(10) **Pub. No.: US 2016/0253408 A1**

(43) **Pub. Date: Sep. 1, 2016**

(54) **COMPUTER-IMPLEMENTED SYSTEM AND METHOD FOR PROVIDING CLASSIFICATION SUGGESTIONS**

(71) Applicant: **FTI Consulting, Inc.**, Annapolis, MD (US)

(72) Inventors: **William C. Knight**, Bainbridge Island, WA (US); **Nicholas I. Nussbaum**, Seattle, WA (US)

(21) Appl. No.: **15/150,382**

(22) Filed: **May 9, 2016**

Related U.S. Application Data

(63) Continuation of application No. 14/065,364, filed on Oct. 28, 2013, now Pat. No. 9,336,303, which is a continuation of application No. 12/833,880, filed on Jul. 9, 2010, now Pat. No. 8,572,084.

(60) Provisional application No. 61/229,216, filed on Jul. 28, 2009, provisional application No. 61/236,490, filed on Aug. 24, 2009.

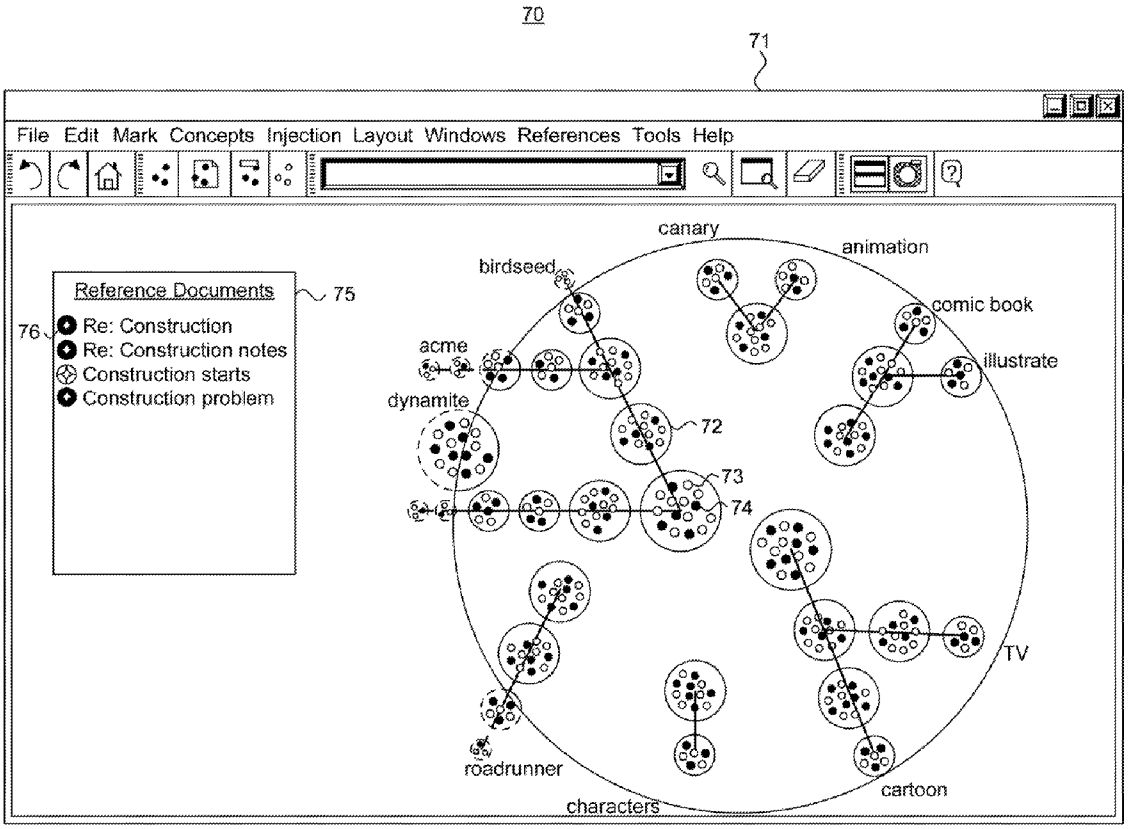
Publication Classification

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

(52) **U.S. Cl.**
CPC ***G06F 17/30601*** (2013.01); ***G06F 17/30675*** (2013.01); ***G06F 17/3064*** (2013.01)

(57) **ABSTRACT**

A computer-implemented system and method for providing classification suggestions is provided. A set of uncoded documents is maintained. One of the uncoded documents is selected and compared with a set of reference documents, each associated with a classification. Those reference documents that are similar to the uncoded document are identified. Relationships between the uncoded document and each reference document are identified by counting a number of similar reference documents associated with each different classification. The classification having a highest count of similar reference documents is selected for the selected uncoded document as a suggestion.



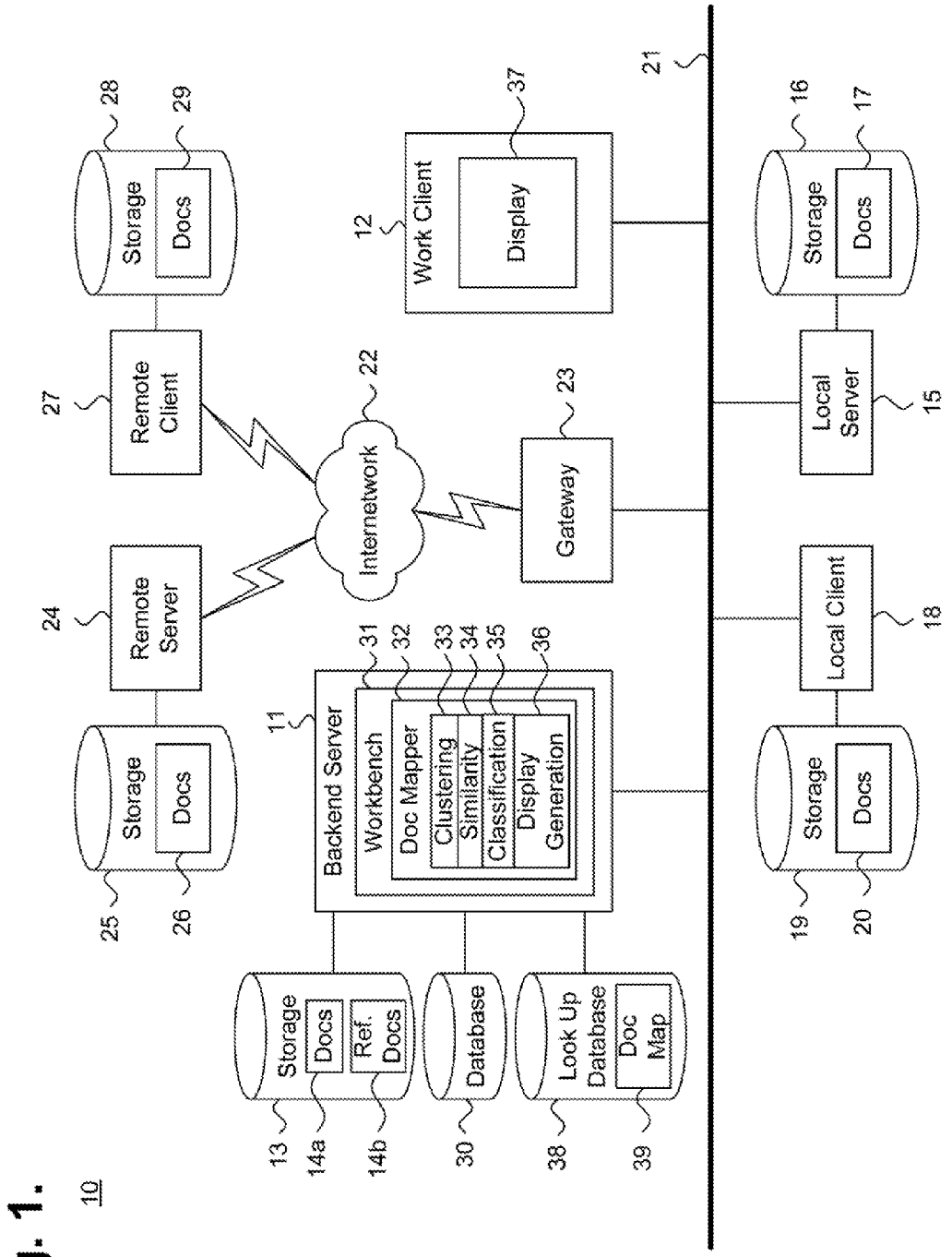

10
21
Storage
Docs
28
29
Work Client
Display
12
37
Storage
Docs
16
17
Remote Client
27
Internetwork
22
Gateway
23
Local Server
15
Remote Server
24
Backend Server
11
Workbench
Doc Mapper
Clustering
Similarity
Classification
Display Generation
31
32
33
34
35
36
Local Client
18
Storage
Docs
25
26
Storage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

Fig. 1.

Fig. 2.

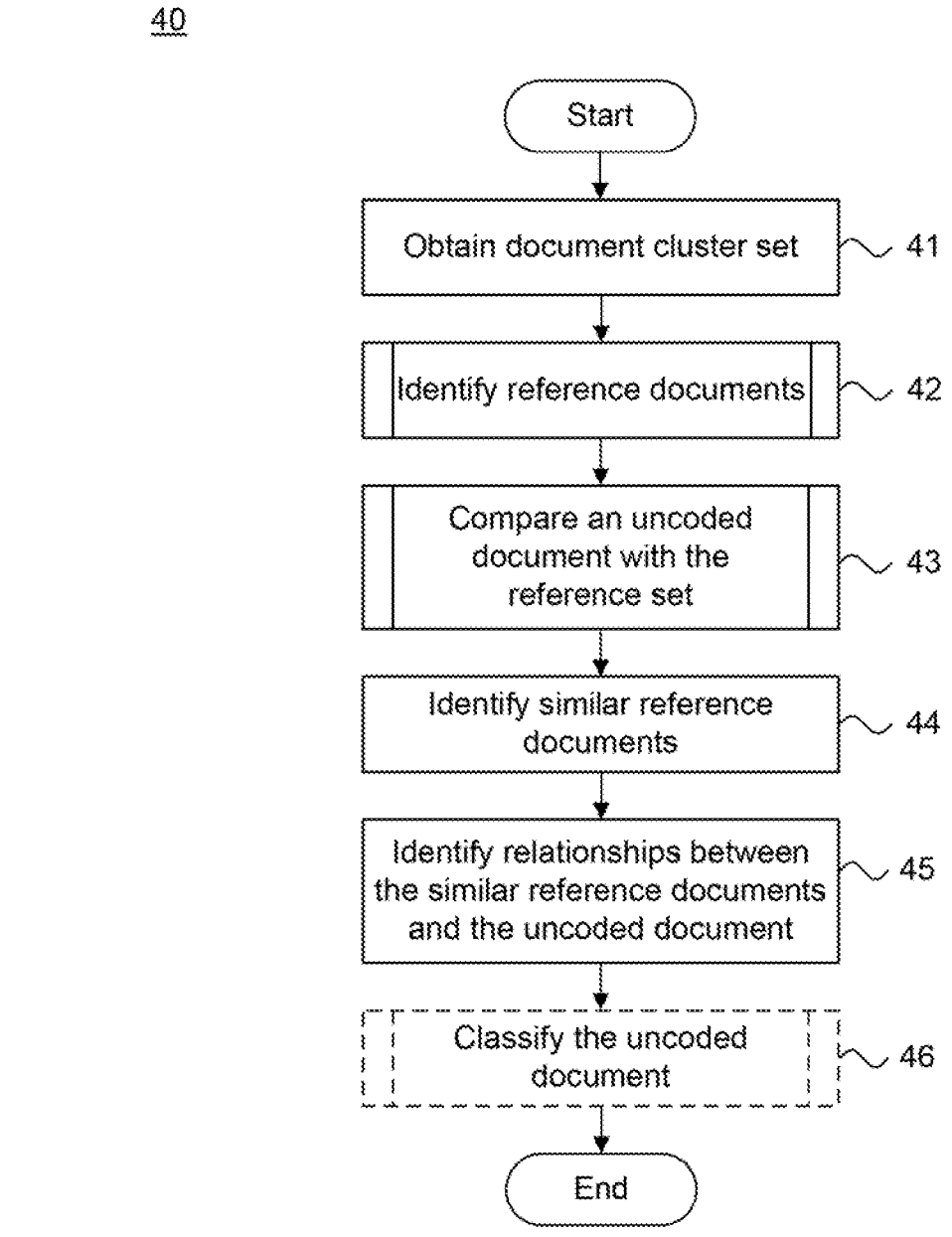

40
Start
Obtain document cluster set
41
Identify reference documents
42
Compare an uncoded document with the reference set
43
Identify similar reference documents
44
Identify relationships between the similar reference documents and the uncoded document
45
Classify the uncoded document
46
End

Fig. 3.

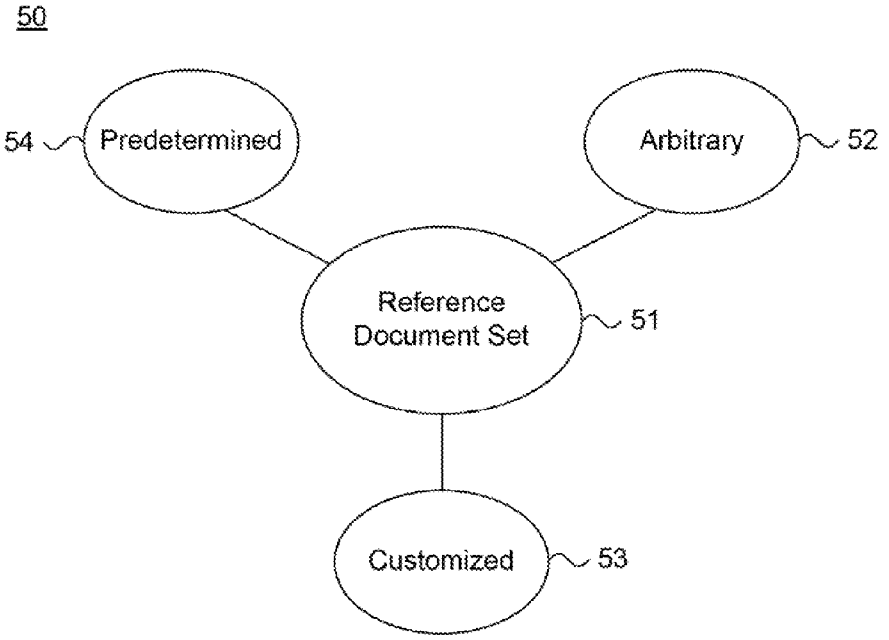

50
54
Predetermined
Arbitrary
52
Reference Document Set
51
Customized
53

Fig. 4.

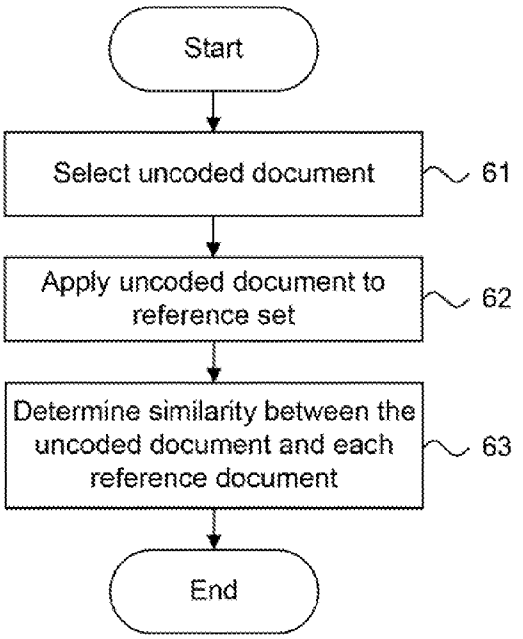

60
Start
Select uncoded document
61
Apply uncoded document to reference set
62
Determine similarity between the uncoded document and each reference document
63
End

Fig. 5.

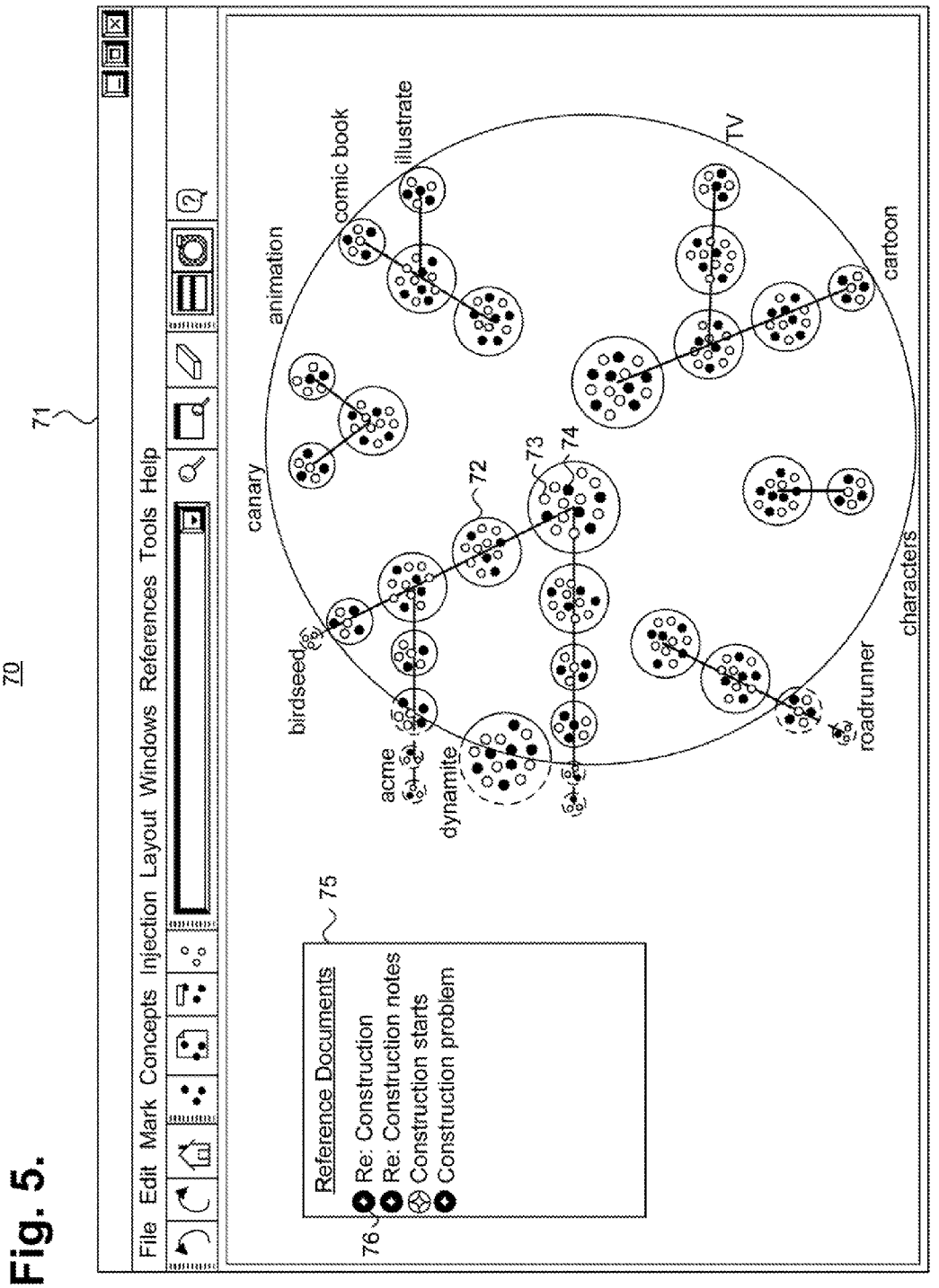

70
71
File Edit Mark Concepts Injection Layout Windows References Tools Help
canary
animation
comic book
illustrate
TV
cartoon
characters
roadrunner
dynamite
acme
birdseed
72
73
74
75
Reference Documents
Re: Construction
Re: Construction notes
Construction starts
Construction problem
76

Fig. 6.

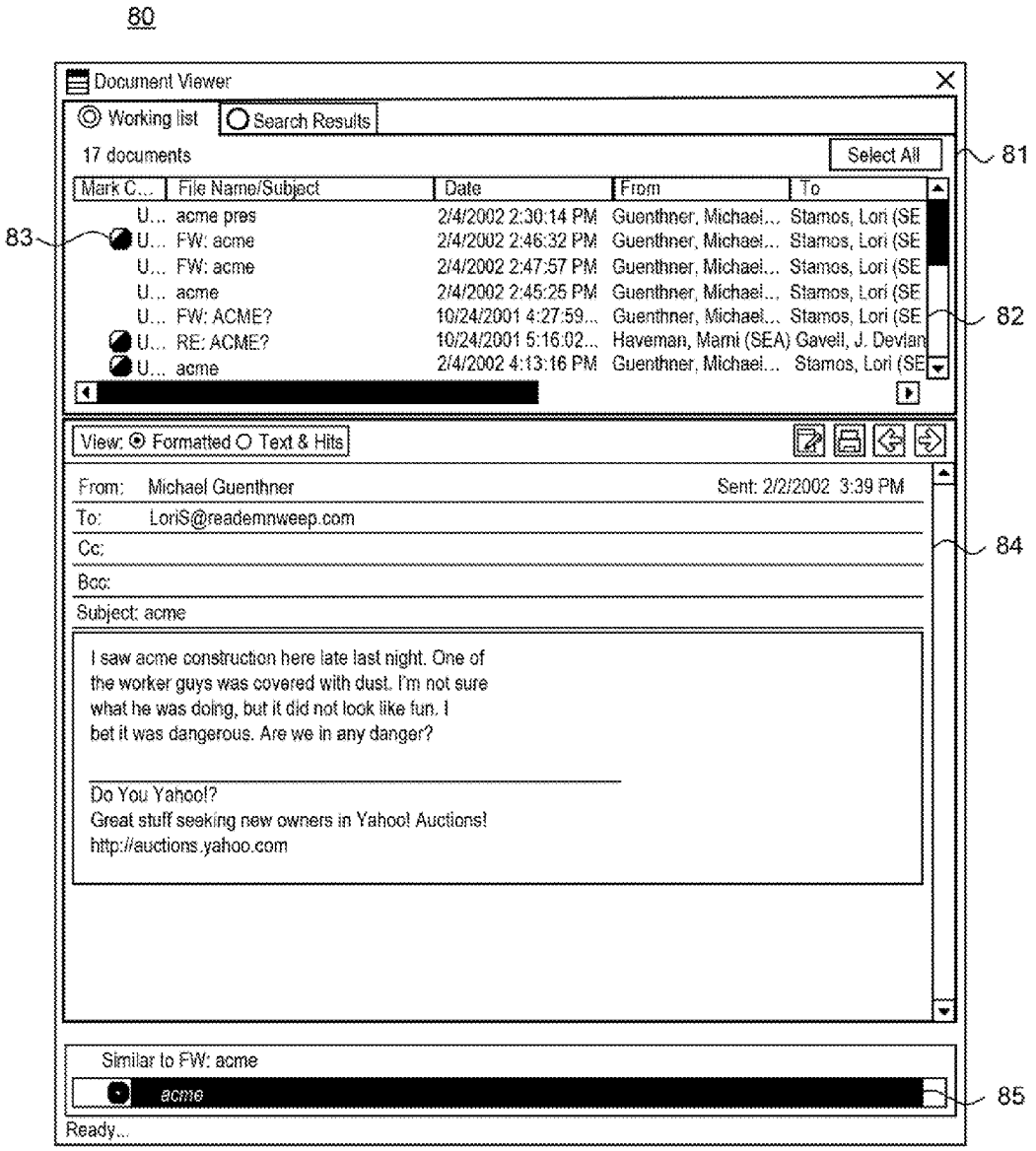
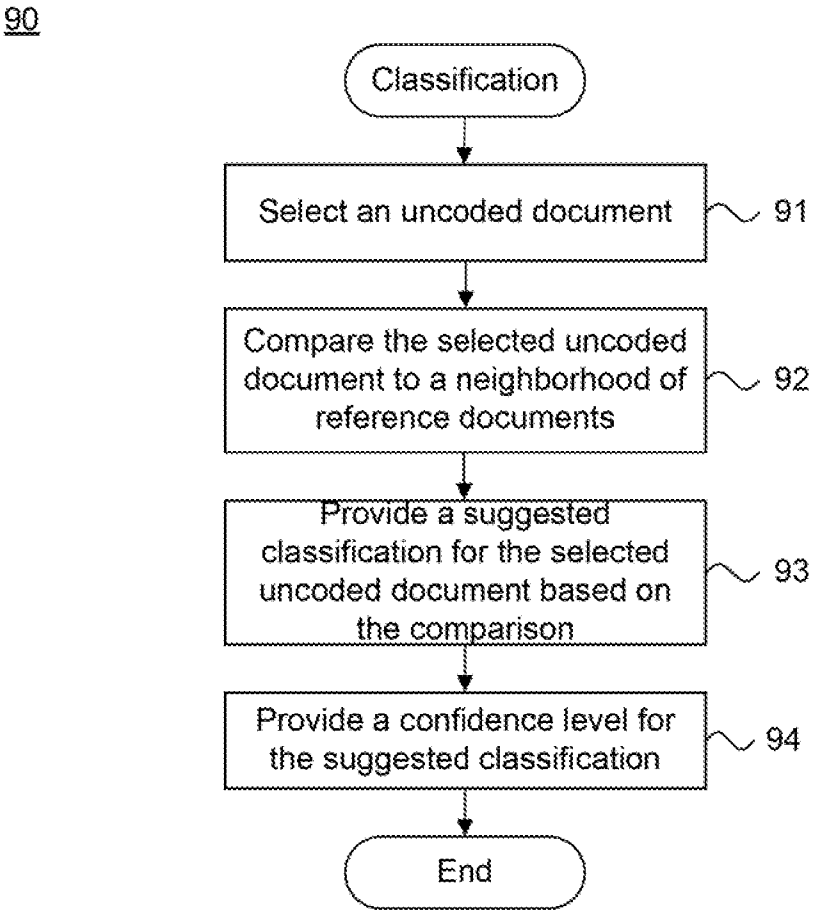

80
Document Viewer
Working list
Search Results
17 documents
Select All
81
Mark C...
File Name/Subject
Date
From
To
83
U... acme pres 2/4/2002 2:30:14 PM Guenthner, Michael... Stamos, Lori (SE
U... FW: acme 2/4/2002 2:46:32 PM Guenthner, Michael... Stamos, Lori (SE
U... FW: acme 2/4/2002 2:47:57 PM Guenthner, Michael... Stamos, Lori (SE
U... acme 2/4/2002 2:45:25 PM Guenthner, Michael... Stamos, Lori (SE
U... FW: ACME? 10/24/2001 4:27:59... Guenthner, Michael... Stamos, Lori (SE
82
U... RE: ACME? 10/24/2001 5:16:02... Haveman, Marni (SEA) Gavell, J. Devlan
U... acme 2/4/2002 4:13:16 PM Guenthner, Michael... Stamos, Lori (SE
View: Formatted Text & Hits
From: Michael Guenthner Sent: 2/2/2002 3:39 PM
To: LoriS@rademnweep.com
Cc:
Bcc:
Subject: acme
84
I saw acme construction here late last night. One of the worker guys was covered with dust. I'm not sure what he was doing, but it did not look like fun. I bet it was dangerous. Are we in any danger?
Do You Yahoo!?
Great stuff seeking new owners in Yahoo! Auctions!
http://auctions.yahoo.com
Similar to FW: acme
acme
85
Ready...

Fig. 7.


90
Classification
Select an uncoded document
91
Compare the selected uncoded document to a neighborhood of reference documents
92
Provide a suggested classification for the selected uncoded document based on the comparison
93
Provide a confidence level for the suggested classification
94
End

COMPUTER-IMPLEMENTED SYSTEM AND METHOD FOR PROVIDING CLASSIFICATION SUGGESTIONS

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This patent application is a continuation of commonly-assigned U.S. patent application Ser. No. 14/065,364, filed on Oct. 28, 2013, pending; which is a continuation of U.S. Pat. No. 8,572,084, issued Oct. 29, 2013; which claims priority under 35 U.S.C. §119(e) to U.S. Provisional Patent Application, Serial No. 61/229,216, filed July 28, 2009, and U.S. Provisional Patent Application, Ser. No. 61/236,490, filed Aug. 24, 2009, the priority dates of which are claimed and the disclosures of which are incorporated by reference.

FIELD

[0002] This application relates in general to using documents as a reference point and, in particular, to a system and method for providing classification suggestions.

BACKGROUND

[0003] 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.

[0004] 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.

[0005] 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.

[0006] 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.

[0007] 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.

SUMMARY

[0008] 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 uncoded ESI for a document review project are identified and clustered. At least one of the uncoded ESI is selected from the clusters and compared with the reference ESI based on a similarity metric. The reference ESI most similar to the selected uncoded ESI are identified. Classification codes assigned to the similar reference ESI can be used to provide suggestions for classification of the selected uncoded ESI. Further, a machine-generated suggestion for classification code can be provided with a confidence level.

[0009] An embodiment provides a computer-implemented system and method for providing classification suggestions. A set of uncoded documents is maintained.

[0010] One of the uncoded documents is selected and compared with a set of reference documents, each associated with a classification. Those reference documents that are similar to the uncoded document are identified. Relationships between the uncoded document and each reference document are identified by counting a number of similar reference documents associated with each different classification. The classification having a highest count of similar reference documents is selected for the selected uncoded document as a suggestion.

[0011] 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.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] FIG. **1** is a block diagram showing a system for displaying relationships between electronically stored information to provide classification suggestions via nearest neighbor, in accordance with one embodiment.

[0013] FIG. **2** is a process flow diagram showing a method for displaying relationships between electronically stored information to provide classification suggestions via nearest neighbor, in accordance with one embodiment.

[0014] FIG. **3** is a block diagram showing, by way of example, measures for selecting a document reference subset.

[0015] FIG. **4** is a process flow diagram showing, by way of example, a method for comparing an uncoded document to reference documents for use in the method of FIG. **2**.

[0016] FIG. **5** is a screenshot showing, by way of example, a visual display of reference documents in relation to uncoded documents.

[0017] FIG. **6** is an alternative visual display of the similar reference documents and uncoded documents.

[0018] FIG. **7** is a process flow diagram showing, by way of example, a method for classifying uncoded documents for use in the method of FIG. **2**.

DETAILED DESCRIPTION

[0019] The ever-increasing volume of ESI underlies the need for automating document review for improved consistency and throughput. Previously coded documents offer knowledge gleaned from earlier work in similar legal projects, as well as a reference point for classifying uncoded ESI.

Providing Suggestions Using Reference Documents

[0020] Reference documents are documents that have been 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.

[0021] Complete ESI review requires a support environment within which classification can be performed. FIG. **1** is a block diagram showing a system **10** for displaying relationships between electronically stored information to provide classification suggestions via nearest neighbor, 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 **14***a*, 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 **14***b*, which can provide a training set of trusted and known results for use in guiding ESI classification. The reference documents **14***b* 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 the selected uncoded documents are classified and subsequently added to the set of reference documents.

[0022] 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.

[0023] The clustering engine **33** performs efficient document scoring and clustering of documents, including uncoded and coded documents, such as described in commonly-assigned U.S. Pat. No. 7,610,313, the disclosure of which is incorporated by reference. Clusters of uncoded documents **14***a* can be formed and organized along vectors, known as spines, based on a similarity of the clusters, which can be expressed in terms of distance. During clustering, groupings of related documents are provided. The content of each document 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 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.

[0024] In a further embodiment, the clusters can include uncoded and coded documents, which are generated based on a similarity with the uncoded documents, as discussed in commonly-owned U.S. Pat. No. 8,713,018, issued on Apr. 29, 2014, and U.S. Pat. No. 8,515,957, issued Aug. 20, 2013, the disclosures of which are incorporated by reference.

[0025] The similarity searcher **34** identifies the reference documents **14***b* that are most similar to selected uncoded documents **14***a*, clusters, or spines, as further described below with reference to FIG. **4**. For example, the uncoded documents, reference documents, clusters, and spines can each be represented by a score vector, which includes paired values consisting of a token, such as a term occurring in that document, cluster or spine, and the associated score for that token. Subsequently, the score vector of the uncoded document, cluster, or spine is then compared with the score vectors of the reference documents to identify similar reference documents.

[0026] The classifier **35** provides a machine-generated suggestion and confidence level for classification of selected uncoded documents **14***a*, clusters, or spines, as further described below with reference to FIG. **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 FIG. **5**. 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.

[0027] The document mapper **32** operates on uncoded **14***a* and coded documents **14***b*, which can be retrieved from the storage **13**, as well as from a plurality of local and remote sources. The local sources include a local server **15**, which is coupled to a storage device **16** with documents **17** and a local client **18**, which is coupled to a storage device **19** with documents **20**. 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 a remote server 24, which is coupled to a storage device 25 with documents 26 and a remote client 27, which is coupled to a storage device 28 with documents 29. Other document sources, either local or remote, are possible.

[0028] The individual documents 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.

[0029] In one embodiment, the individual documents 14*a*, 14*b*, 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, Wash. The database can be an SQL-based relational database, such as the Oracle database management system, Release 8, licensed by Oracle Corporation, Redwood Shores, Calif.

[0030] The individual documents 17, 20, 26, 29 can be designated and stored as uncoded documents or reference documents. The uncoded documents, which are unclassified, are selected for a document review project and stored as a document corpus for classification. 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 based on visual relationships between the uncoded documents and reference documents. In a further embodiment, the reference documents can provide classification suggestions for a document corpus associated with a related document review project. 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 FIG. 7.

[0031] 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 an unintentional waiver 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.

[0032] 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.

[0033] Identifying relationships between the reference documents and uncoded documents includes clustering and similarity measures. FIG. 2 is a process flow diagram showing a method 40 for displaying relationships between electronically stored information to provide classification suggestions via nearest neighbor, in accordance with one embodiment. A set of document clusters is obtained (block 41). In one embodiment, the clusters can include uncoded documents, and in a further embodiment, the clusters can include uncoded and coded documents. The clustered uncoded documents can represent a corpus of uncoded documents for a document review project, or one or more assignments of uncoded documents. The document corpus can include all uncoded documents for a document review project, while, each assignment can include a subset of uncoded documents selected from the corpus and assigned to a reviewer. The corpus 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. Prior to, concurrent with, or subsequent to obtaining the cluster set, reference documents are identified (block 42). The reference documents can include all reference documents generated for a document review project, or alternatively, a subset of the reference documents. Obtaining reference documents is further discussed below with reference to FIG. 3.

[0034] An uncoded document is selected from one of the clusters in the set and compared against the reference documents (block 43) to identify one or more reference documents that are similar to the selected uncoded document (block 44). The similar reference documents are identified based on a similarity measure calculated between the selected uncoded document and each reference document. Comparing the selected uncoded document with the reference documents is further discussed below with reference to FIG. 4. Once identified, relationships between the selected uncoded document and the similar reference documents can be identified (block 45) to provide classification hints, including a suggestion for the selected uncoded document, as further discussed below with reference to FIG. 5. Additionally, machine-generated suggestions for classification can be provided (block 46) with an associated confidence level for use in classifying the selected uncoded document. Machine-generated suggestions are further discussed below with reference to FIG. 7. Once the selected uncoded document is assigned a classification code, either by the reviewer or automatically, the newly classified document can be added to the set of reference documents for use in classifying further uncoded documents. Subsequently, a further uncoded document can be selected for classification using similar reference documents.

[0035] In a further embodiment, similar reference documents can also be identified for a selected cluster or a selected spine along which the clusters are placed.

Selecting a Document Reference Subset

[0036] After the clusters have been generated, one or more uncoded documents can be selected from at least one of the clusters for comparing with a reference document set or subset. FIG. **3** is a block diagram showing, by way of example, measures **50** for selecting a document reference subset **51**. The subset of reference documents **51** can be previously defined **54** and maintained for related document review projects or can be specifically generated for each review project. A predefined reference subset **54** provides knowledge previously obtained during the related document review project to increase efficiency, accuracy, and consistency. Reference subsets newly generated for each review project can include arbitrary **52** or customized **53** reference subsets that are determined automatically or by a human reviewer. An arbitrary reference subset **52** includes reference documents randomly selected for inclusion in the reference subset. A customized reference subset **53** includes reference documents specifically selected for inclusion in the reference subset based on criteria, such as reviewer preference, classification category, document source, content, and review project. Other criteria are possible.

[0037] The subset of reference documents, whether predetermined or newly generated, should be selected from a set of reference documents that are representative of the document corpus for a review project in which data organization or classification is desired. Guided review assists a reviewer or other user in identifying reference documents that are representative of the corpus for use in classifying uncoded documents. During guided review, the uncoded documents that are dissimilar to all other uncoded documents are identified based on a similarity threshold. In one embodiment, the dissimilarity can be determined as the cos σ of the score vectors for the uncoded documents. Other methods for determining dissimilarity are possible. Identifying the dissimilar documents provides a group of documents that are representative of the corpus for a 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 the reference documents. Guided review can be performed by a reviewer, a machine, or a combination of the reviewer and machine.

[0038] Other methods for generating reference documents for a document review project using guided review are possible, including clustering. A set of uncoded documents to be classified is clustered, as described in commonly-assigned U.S. Pat. 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 selected uncoded documents are then assigned classification codes. In a further embodiment, sample clusters can be used to generate reference documents 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 selected for classification by assigning classification codes. The classified documents represent reference documents for the document review project. The number of reference documents can be determined automatically or by a reviewer. Other methods for selecting documents for use as reference documents are possible.

Comparing a Selected Uncoded Document to Reference Documents

[0039] An uncoded document selected from one of the clusters can be compared to the reference documents to identify similar reference documents for use in providing suggestions regarding classification of the selected uncoded document. FIG. **4** is a process flow diagram showing, by way of example, a method **60** for comparing an uncoded document to reference documents for use in the method of FIG. **2**. The uncoded document is selected from a cluster (block **61**) and applied to the reference documents (block **62**). The reference documents can include all reference documents for a document review project or a subset of the reference documents. Each of the reference documents and the selected uncoded document can be represented by a score vector having paired values of tokens occurring within that document and associated token scores. A similarity between the uncoded document and each reference document is determined (block **63**) as the cos σ of the score vectors for the uncoded document and reference document being compared and is equivalent to the inner product between the score vectors. In the described embodiment, the cos σ is calculated in accordance with the equation:

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

where cos σ_{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. Other forms of determining similarity using a distance metric are possible, as would be recognized by one skilled in the art, including using Euclidean distance.

[0040] One or more of the reference documents that are most similar to the selected uncoded document, based on the similarity metric, are identified. The most similar reference documents can be identified by satisfying a predetermined threshold of similarity. Other methods for determining the similar reference documents are possible, such as setting a predetermined absolute number of the most similar reference documents. The classification codes of the identified similar reference documents can be used as suggestions for classifying the selected uncoded document, as further described below with reference to

[0041] FIG. **5**. Once identified, the similar reference documents can be used to provide suggestions regarding classification of the selected uncoded document, as further described below with reference to FIGS. **5** and **7**.

Displaying the Reference Documents

[0042] The similar reference documents can be displayed with the clusters of uncoded documents. In the display, the similar reference documents can be provided as a list, while the clusters can be can be organized along spines of thematically related clusters, as described in commonly-assigned

U.S. Pat. No. 7,271,804, the disclosure of which is incorporated by reference. The 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. Pat. No. 7,610,313, the disclosure of which is incorporated by reference. Other displays of the clusters and similar reference documents are possible.

[0043] 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. FIG. **5** is a screenshot **70** showing, by way of example, a visual display **71** of similar reference documents **74** and uncoded documents **74**. Clusters **72** of the uncoded documents **73** can be located along a spine, which is a vector, based on a similarity of the uncoded documents **73** in the clusters **72**. The uncoded documents **73** are each represented by a smaller circle within the clusters **72**.

[0044] Similar reference documents **74** identified for a selected uncoded document **73** can be displayed in a list **75** by document title or other identifier. Also, classification codes **76** associated with the similar reference documents **74** can be displayed as circles having a diamond shape within the boundary of the circle. The classification codes **76** can include "privileged," "responsive," and "non-responsive" codes, as well as other codes. The different classification codes **76** can each be represented by a color, such as blue for "privileged" reference documents and yellow for "non-responsive" reference documents. Other display representations of the uncoded documents, similar reference documents, and classification codes are possible, including by symbols and shapes.

[0045] The classification codes **76** of the similar reference documents **74** can provide suggestions for classifying the selected uncoded document based on factors, such as a number of different classification codes for the similar reference documents and a number of similar reference documents associated with each classification code. For example, the list of reference documents includes four similar reference documents identified for a particular uncoded document. Three of the reference documents are classified as "privileged," while one is classified as "non-responsive." In making a decision to assign a classification code to a selected uncoded document, the reviewer can consider classification factors based on the similar reference documents, such as such as a presence or absence of similar reference documents with different classification codes and a quantity of the similar reference documents for each classification code. Other classification factors are possible. In the current example, the display **81** provides suggestions, including the number of "privileged" similar reference documents, the number of "non-responsive" similar reference documents, and the absence of other classification codes of similar reference documents. Based on the number of "privileged" similar reference documents compared to the number of "non-responsive" similar reference documents, the reviewer may be more inclined to classify the selected uncoded documents as "privileged." Alternatively, the reviewer may wish to further review the selected uncoded document based on the multiple classification codes of the similar reference documents. Other classification codes and combinations of classification codes are possible. The reviewer can utilize the suggestions provided by the similar reference documents to assign a classification to the selected uncoded document. In a further embodiment, the now classified and previously uncoded document can be added to the set of reference documents for use in classifying other uncoded documents.

[0046] In a further embodiment, similar reference documents can be identified for a cluster or spine to provide suggestions for classifying the cluster and spine. For a cluster, the similar reference documents are identified based on a comparison of a score vector for the cluster, which is representative of the cluster center and the reference document score vectors. Meanwhile, identifying similar reference documents for a spine is based on a comparison between the score vector for the spine, which is based on the cluster center of all the clusters along that spine, and the reference document score vectors. Once identified, the similar reference documents are used for classifying the cluster or spine.

[0047] In an even further embodiment, the uncoded documents, including the selected uncoded document, and the similar reference documents can be displayed as a document list. FIG. **6** is a screenshot **80** showing, by way of example, an alternative visual display of the similar reference documents **85** and uncoded documents **82**. The uncoded documents **82** can be provided as a list in an uncoded document box **81**, such as an email inbox. The uncoded documents **82** can be identified and organized using uncoded document factors, such as file name, subject, date, recipient, sender, creator, and classification category **83**, if previously assigned.

[0048] At least one of the uncoded documents can be selected and displayed in a document viewing box **84**. The selected uncoded document can be identified in the list **81** using a selection indicator (not shown), including a symbol, font, or highlighting. Other selection indicators and uncoded document factors are possible. Once identified, the selected uncoded document can be compared to a set of reference documents to identify the reference documents **85** most similar. The identified similar reference documents **85** can be displayed below the document viewing box **84** with an associated classification code **83**. The classification code of the similar reference document **85** can be used as a suggestion for classifying the selected uncoded document. After assigning a classification code, a representation **83** of the classification can be provided in the display with the selected uncoded document. In a further embodiment, the now classified and previously uncoded document can be added to the set of reference documents.

Machine Classification of Uncoded Documents

[0049] Similar reference documents can be used as suggestions to indicate a need for manual review of the uncoded documents, when review may be unnecessary, and hints for classifying the uncoded documents, clusters, or spines. Additional information can be generated to assist a 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 commonassigned U.S. Pat. No. 8,635,223, issued Jan. 21, 2014, the disclosure of which is incorporated by reference.

[0050] The machine-generated suggestion for classification and associated confidence level can be determined by a classifier. FIG. **7** is a process flow diagram **90** showing, by way of example, a method for classifying uncoded documents by a classifier for use in the method of FIG. **2**. An uncoded document is selected from a cluster (block **91**) and compared to a neighborhood of x-similar reference documents (block **92**) to identify those similar reference documents that are

most relevant to the selected uncoded document. The selected uncoded document can be the same as the uncoded document selected for identifying similar reference documents or a different uncoded document. In a further embodiment, a machine-generated suggestion can be provided for a cluster or spine by selecting and comparing the cluster or spine to a neighborhood of x-reference documents for the cluster or spine.

[0051] The neighborhood of x-similar reference documents is determined separately for each selected uncoded document and can include one or more similar reference documents. During neighborhood generation, a value for x similar reference documents is first determined automatically or by an individual reviewer. The neighborhood of similar reference documents can include the reference documents, which were identified as similar reference documents according to the method of FIG. **4**, or reference documents located in one or more clusters, such as the same cluster as the selected uncoded document or in one or more files, such as an email file. Next, the x-number of similar reference documents nearest to the selected uncoded document are identified. Finally, the identified x-number of similar reference documents are provided as the neighborhood for the selected uncoded document. In a further embodiment, the x-number of similar reference documents are defined for each classification code, rather than across all classification codes. Once generated, the x-number of similar reference documents in the neighborhood and the selected uncoded document are analyzed by the classifier to provide a machine-generated classification suggestion for assigning a classification code (block **93**). A confidence level for the machine-generated classification suggestion is also provided (block **94**).

[0052] The machine-generated analysis of the selected uncoded document and x-number of similar 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 for a selected uncoded document 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 similar reference documents in the neighborhood as the cos σ to determine a distance metric. The distance metrics for the x-number of similar reference documents are compared to identify the similar reference document closest to the selected uncoded document as the closest neighbor.

[0053] The minimum average distance classification measure includes calculating an average distance of the similar reference documents for each classification code. The classification code of the similar 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 similar reference documents for each classification code and assigning a count or "vote" to the similar reference documents based on the assigned classification code. The classification code with the highest number of similar reference documents or "votes" is assigned to the selected uncoded document as the suggested classification code. The distance weighted maximum count classification measure includes identifying a count of all similar reference documents for each classification code and determining a distance between the selected uncoded document and each of the similar reference documents. Each count assigned to the similar reference documents is weighted based on the distance of the similar 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 code.

[0054] The machine-generated suggested 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 code. 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.

[0055] Machine classification can also occur on a cluster or spine level once one or more documents in the cluster have been classified. 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 FIG. **4**. A neighborhood for the selected cluster can be determined based on a distance metric. The x-number of similar reference documents that are closest to the cluster center can be selected for inclusion in the neighborhood, as described above. Each document in the selected cluster is associated with a score vector from which the cluster center score vector is generated. The distance is then determined by comparing the score vector of the cluster center with the score vector for each of the similar reference documents to determine an x-number of similar reference documents that are closest to the cluster center. However, other methods for generating a neighborhood are possible. Once determined, one of the classification routines is applied to the neighborhood to determine a suggested classification code and confidence level for the selected cluster. The neighborhood of x-number of reference documents is determined for a spine by comparing a spine score vector with the vector for each similar reference document to identify the neighborhood of similar documents that are the most similar.

[0056] Providing classification suggestions and suggested classification codes has been described in relation to uncoded documents and reference documents. However, in a further embodiment, classification suggestions and suggested classification codes can be provided for the uncoded documents based on a particular token identified within the uncoded documents. The token can include concepts, n-grams, raw terms, and entities. In one example, the uncoded tokens, which are extracted from uncoded documents, can be clustered. A token can be selected from one of the clusters and compared with reference tokens. Relationships between the uncoded token and similar reference tokens can be displayed

to provide classification suggestions for the uncoded token. The uncoded documents can then be classified based on the classified tokens.

[0057] 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.

What is claimed is:

1. A computer-implemented system for providing classification suggestions, comprising:

a database to maintain a set of uncoded documents;

a selection module to select one of the uncoded documents and to compare the selected uncoded document with a set of reference documents each associated with a classification;

a similarity module to identify those reference documents that are similar to the uncoded document;

an identification module to identify relationships between the uncoded document and each reference document comprising counting a number of similar reference documents associated with each different classification; and

a suggestion module to suggest for the selected uncoded document the classification having a highest count of similar reference documents.

2. A system according to claim **1**, further comprising:

a vector module to generate a score vector for each uncoded document and each reference document, wherein the score vectors each comprise one or more terms occurring in that document and a score for each term; and

a comparison module to determine a similarity value for each reference document and the uncoded document by comparing the score vectors of that reference document to the score vector of the uncoded document

3. A system according to claim **2**, further comprising:

a threshold module to apply a predetermined threshold to the similarity values and to identify those reference documents with similarity values that satisfy the predetermined threshold as the reference documents similar to the uncoded document.

4. A system according to claim **1**, further comprising:

a display to display the similar reference documents with the uncoded documents.

5. A system according to claim **1**, further comprising:

a display to display the classifications with the similar reference documents.

6. A system according to claim **5**, further comprising:

a classification display module to differentiate different types of the classifications via at least one of color, symbol, and shape.

7. A system according to claim **1**, further comprising:

a placement module to add the selected uncoded document with the suggested classification to the set of reference documents.

8. A system according to claim **1**, further comprising:

a display to display the uncoded documents as a list, wherein at least the selected uncoded document is displayed with the suggested classification.

9. A system according to claim **8**, further comprising:

a reference selection module to select a further uncoded document from the set displayed in the list;

a similarity display module to display the reference documents similar to the further selected uncoded document; and

a classification receipt module to receive for the further uncoded document one of the classifications associated with one or more of the reference documents similar to the further selected uncoded document.

10. A system according to claim **1**, further comprising:

a distance determination module to determine a distance between the selected uncoded document and each of the similar reference documents; and

a weighting module to weigh the count of similar reference documents associated with each classification based on the distances of the associated similar reference documents.

11. A computer-implemented method for providing classification suggestions, comprising:

maintaining a set of uncoded documents;

selecting one of the uncoded documents and comparing the selected uncoded document with a set of reference documents each associated with a classification;

identifying those reference documents that are similar to the uncoded document;

identifying relationships between the uncoded document and each reference document comprising counting a number of similar reference documents associated with each different classification; and

suggesting for the selected uncoded document the classification having a highest count of similar reference documents.

12. A method according to claim **11**, further comprising:

generating a score vector for each uncoded document and each reference document, wherein the score vectors each comprise one or more terms occurring in that document and a score for each term; and

determining a similarity value for each reference document and the uncoded document by comparing the score vectors of that reference document to the score vector of the uncoded document

13. A method according to claim **12**, further comprising:

applying a predetermined threshold to the similarity values; and

identifying those reference documents with similarity values that satisfy the predetermined threshold as the reference documents similar to the uncoded document.

14. A method according to claim **11**, further comprising:

displaying the similar reference documents with the uncoded documents.

15. A method according to claim **11**, further comprising:

displaying the classifications with the similar reference documents.

16. A method according to claim **15**, further comprising:

differentiating different types of the classifications via at least one of color, symbol, and shape.

17. A method according to claim **11**, further comprising:

adding the selected uncoded document with the suggested classification to the set of reference documents.

18. A method according to claim **11**, further comprising:

displaying the uncoded documents as a list, wherein at least the selected uncoded document is displayed with the suggested classification.

19. A method according to claim **18**, further comprising:

selecting a further uncoded document from the set displayed in the list;

displaying the reference documents similar to the further selected uncoded document; and

receiving for the further selected uncoded document one of the classifications associated with one or more of the reference documents similar to the further selected uncoded document.

20. A method according to claim **11**, further comprising:

determining a distance between the selected uncoded document and each of the similar reference documents; and

weighing the count of similar reference documents associated with each classification based on the distances of the associated similar reference documents.

* * * * *