A system (11) and method (50) for displaying relationships between concepts (14c, 14d) to provide classification suggestions via nearest neighbor is provided. Reference concepts (14d) previously classified and a set of uncoded concepts (14c) are provided. At
(57) Abstract (continued):
least one uncoded concept (14c) is compared with the reference concepts (14d). One or more of the reference concepts (14d) that are similar to the at least one uncoded concept (14c) are identified. Relationships between the at least one uncoded concept (14c) and the similar reference concept (14d) are depicted on a display for classifying the at least one uncoded concept (14c).
Title: DISPLAYING RELATIONSHIPS BETWEEN CONCEPTS TO PROVIDE CLASSIFICATION SUGGESTIONS VIA NEAREST NEIGHBOR

Abstract: A system (11) and method (50) for displaying relationships between concepts (14c, 14d) to provide classification suggestions via nearest neighbor is provided. Reference concepts (14d) previously classified and a set of uncoded concepts (14c) are provided. At least one uncoded concept (14c) is compared with the reference concepts (14d). One or more of the reference concepts (14d) that are similar to the at least one uncoded concept (14c) are identified. Relationships between the at least one uncoded concept (14c) and the similar reference concept (14d) are depicted on a display for classifying the at least one uncoded concept (14c).
DISPLAYING RELATIONSHIPS BETWEEN CONCEPTS TO PROVIDE
CLASSIFICATION SUGGESTIONS VIA NEAREST NEIGHBOR

TECHNICAL FIELD

This application relates in general to using documents as a reference point and, in
particular, to a system and method for displaying relationships between concepts to provide
classification suggestions via nearest neighbor.

BACKGROUND ART

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

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

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

Conventional ESI review tools have proven inadequate to providing efficient, accurate,
and consistent results. For example, DiscoverReady LLC, a Delaware limited liability company,
custom programs ESI review tools, which conduct semi-automated document review through
multiple passes over a document set in ESI form. During the first pass, documents are grouped
by category and basic codes are assigned. Subsequent passes refine and further assign codings.
Multiple pass review requires *a priori* project-specific knowledge engineering, which is only useful for the single project, thereby losing the benefit of any inferred knowledge or know-how for use in other review projects.

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

**DISCLOSURE OF THE INVENTION**

Document review efficiency can be increased by identifying relationships between reference documents and uncoded documents and providing a suggestion for classification based on the relationships. The uncoded documents for a document review project are identified and clustered. At least one of the uncoded documents is selected from the clusters and compared with the reference set based on a similarity metric. The reference documents most similar to the selected uncoded document are identified. Classification codes assigned to the similar reference documents can be used to provide suggestions for classification of the selected uncoded document. Further, a machine-generated suggestion for a classification codes can be provided with a confidence level.

An embodiment provides a system and method for displaying relationships between concepts to provide classification suggestions via nearest neighbor. Reference concepts previously classified and a set of uncoded concepts are provided. At least one uncoded concept is compared with the reference concepts. One or more of the reference concepts that are similar to the at least one uncoded concept are identified. Relationships between the at least one uncoded concept and the similar reference concept are depicted on a display for classifying the at least one uncoded concept.

Still other embodiments of the present invention will become readily apparent to those skilled in the art from the following detailed description, wherein are described embodiments by way of illustrating the best mode contemplated for carrying out the invention. As will be realized, the invention is capable of other and different embodiments and its several details are capable of modifications in various obvious respects, all without departing from the spirit and the scope of the present invention. Accordingly, the drawings and detailed description are to be regarded as illustrative in nature and not as restrictive.

**DESCRIPTION OF THE DRAWINGS**

FIGURE 1 is a block diagram showing a system for displaying relationships between concepts to provide classification suggestions via nearest neighbor, in accordance with one embodiment.
FIGURE 2 is a process flow diagram showing a method for displaying relationships between concepts to provide classification suggestions via nearest neighbor, in accordance with one embodiment.

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

FIGURE 4 is a block diagram showing, by way of example, measures for selecting a concept reference subset.

FIGURE 5 is a process flow diagram showing, by way of example, a method for comparing an uncoded concept to reference concepts for use in the method of FIGURE 2.

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

FIGURE 7 is an alternative visual display of the similar reference concepts and uncoded concepts.

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

BEST MODE FOR CARRYING OUT THE INVENTION

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

The tokens can include word-level, symbol-level, or character-level n-grams, raw terms, entities, or concepts. Other tokens, including other atomic parse-level elements, are possible. An n-gram is a predetermined number of items selected from a source. The items can include syllables, letters, or words, as well as other items. A raw term is a term that has not been processed or manipulated. Entities further refine nouns and noun phrases into people, places, and things, such as meetings, animals, relationships, and various other objects. Additionally, entities can represent other parts of grammar associated with semantic meanings to disambiguate different instances or occurrences of the grammar. Entities can be extracted using entity extraction techniques known in the field.

Concepts are collections of nouns and noun-phrases with common semantic meaning that can be extracted from ESI, including documents, through part-of-speech tagging. Each concept can represent one or more documents to be classified during a review. Clustering of the concepts provides an overall view of the document space, which allows users to easily identify documents sharing a common theme.
The clustering of tokens, for example, concepts, differs from document clustering, which groups related documents individually. In contrast, concept clustering groups related concepts, which are each representative of one or more related documents. Each concept can express an ideas or topic that may not be expressed by individual documents. A concept is analogous to a search query by identifying documents associated with a particular idea or topic.

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

Reference concepts are concepts that have been previously classified and can be used to influence classification of uncoded, that is unclassified, concepts. Specifically, relationships between the uncoded concepts and the reference concepts can be visually depicted to provide suggestions, for instance to a human reviewer, for classifying the visually-proximal uncoded concepts. Although tokens, such as word-level or character-level n-grams, raw terms, entities, or concepts, can be clustered and displayed, the discussion below will focus on a concept as a particular token.

Complete concept review requires a support environment within which classification can be performed. FIGURE 1 is a block diagram showing a system 10 for displaying relationships between concepts 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 14a, such as uncoded documents, in the form of structured or unstructured data, a database 30 for maintaining information about the documents, a lookup database 38 for storing many-to-many mappings 39 between documents and document features, such as concepts, and a concept document index 40, which maps documents to concepts. The storage device 13 also stores classified documents 14b, concepts 14c, and reference concepts 14d. Concepts are collections of nouns and noun-phrases with common semantic meaning. The nouns and noun-phrases can be extracted from one or more documents in the corpus for review.
Thus, a single concept can be representative of one or more documents. The reference concepts 14d 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 concepts can be hand-selected or automatically selected through guided review, which is further discussed below. Additionally, the set of reference concepts can be predetermined or can be generated dynamically, as the selected uncoded concepts are classified and subsequently added to the set of reference concepts.

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

The clustering engine 33 performs efficient concept scoring and clustering of documents, including uncoded and coded documents. Efficient scoring and clustering is described in commonly-assigned U.S. Patent No. 7,610,313, the disclosure of which is incorporated by reference. Clusters of uncoded concepts 14a 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 concepts are provided.

In one embodiment, the clusters can include uncoded and coded concepts, which are generated based on a similarity measure, as discussed in commonly-owned 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, and 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, the disclosures of which are incorporated by reference.

The similarity searcher 34 identifies the reference concepts 14d that are most similar to selected uncoded concepts 14c, clusters, or spines, as further described below with reference to FIGURE 4. For example, the uncoded concepts, reference concepts, 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 concept, cluster or spine, and the associated score for that token. Subsequently, the score vector of the uncoded concept, cluster, or spine is then compared with the score vectors of the reference concepts to identify similar reference concepts.
The classifier 35 provides a machine-generated suggestion and confidence level for classification of selected uncoded concepts 14d, clusters, or spines, as further described below with reference to FIGURE 7. The display generator 36 arranges the clusters and spines in thematic relationships in a two-dimensional visual display space, as further described below beginning with reference to FIGURE 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.

The document mapper 32 operates on uncoded 14c and coded concepts 14d, 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 and concepts 17, and a local client 18, which is coupled to a storage device 19 with documents and concepts 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 concepts 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 and concepts 26, and a remote client 27, which is coupled to a storage device 28 with documents and concepts 29. Other document sources, either local or remote, are possible.

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

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

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

The reference concepts are initially uncoded concepts that can represent documents
selected from the corpus or other sources of documents. The reference concepts assist in
providing suggestions for classification of the remaining uncoded concepts representative of the
document corpus based on visual relationships between the uncoded concepts and reference
concepts. The reviewer can classify one or more of the remaining uncoded concepts by
assigning a classification code based on the relationships. In a further embodiment, the reference
concepts can be used as a training set to form machine-generated suggestions for classifying the
remaining uncoded concepts, as further described below with reference to FIGURE 7.

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

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

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

Identifying relationships between the reference concepts and uncoded concepts includes clustering and similarity measures. FIGURE 2 is a process flow diagram showing a method 50 for displaying relationships between concepts to provide classification suggestions via nearest neighbor, in accordance with one embodiment. A set of concept clusters is obtained (block 51). The clusters can include uncoded concepts, and in a further embodiment, the clusters can include uncoded and coded concepts.

Clustering of the concepts provides groupings of related concepts and is based on a similarity metric using score vectors assigned to each concept. The score vectors can be generated using a matrix showing the concepts in relation to documents that contain the concepts. FIGURE 3 is a table showing, by way of example, a matrix mapping 60 of concepts 64 and documents 63. The documents 63 are listed along a horizontal dimension 61 of the matrix, while the concepts 64 are listed along a vertical dimension 62. However, the placement of the documents 63 and concepts 64 can be reversed. Each cell 65 within the matrix 60 includes a cumulative number of occurrences of each concept within a particular document 63. Score vectors can be generated for each document by identifying the concepts and associated weights within that document and ordering the concepts along a vector with the associated concept weight. In the matrix 60, the score vector 66 for a document 63 can be identified as all the concepts included in that document and the associated weights, which are based on the number of occurrences of each concept. Score vectors can also be generated for each concept by identifying the documents that contain that concept and determining a weight associated with each document. The documents and associated weights are then ordered along a vector for each concept, as the concept score vector. In the matrix 60, the score vector 67 for a concept can be identified as all the documents that contain that concept and the associated weights.

In one embodiment, the clustered uncoded concepts can represent a corpus of uncoded concepts representative of a document review project, or one or more concepts representative of at least one assignment of uncoded concepts. The concept corpus can include all uncoded concepts for a document review project, while, each assignment can include a subset of uncoded concepts that are representative of one or more 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 concept, content, document type, and date. Other criteria are possible.

Returning to the discussion of FIGURE 2, reference concepts can be identified (block 52). The reference concepts can include all reference concepts generated for a document review project, or alternatively, a subset of the reference concepts. Obtaining reference concepts is further discussed below with reference to FIGURE 4.

An uncoded concept is selected from one of the clusters in the set and compared against the reference concepts (block 53) to identify one or more reference concepts that are similar to the selected uncoded concept (block 54). The similar reference concepts are identified based on a similarity measure calculated between the selected uncoded concept and each reference concept. Comparing the selected uncoded concept with the reference concepts is further discussed below with reference to FIGURE 4. Once identified, relationships between the selected uncoded concept and the similar reference concepts can be identified (block 55) to provide classification hints, including a suggestion for the selected uncoded concept, as further discussed below with reference to FIGURE 5. Additionally, machine-generated suggestions for classification can be provided (block 56) with an associated confidence level for use in classifying the selected uncoded concept. Machine-generated suggestions are further discussed below with reference to FIGURE 7. Once the selected uncoded concept is assigned a classification code, either by the reviewer or automatically, the newly classified concept can be added to the set of reference concepts for use in classifying further uncoded concepts. Subsequently, a further uncoded concept can be selected for classification using similar reference concepts.

In one embodiment, the classified concepts can be used to classify those documents represented by that concept. For example, in a product liability lawsuit, the plaintiff claims that a wood composite manufactured by the defendant induces and harbors mold growth. During discovery, all documents within the corpus for the lawsuit and relating to mold should be identified for review. The concept for mold is clustered and includes a "responsive" classification code, which indicates that the noun phrase mold is related to the legal matter. Upon selection of the mold concept, all documents that include the noun phrase mold can be identified using the mapping matrix, which is described above with reference to FIGURE 3. The responsive classification code assigned to the concept can be used as a suggestion for the document classification. However, if the document is represented by multiple concepts with different classification codes, each different code can be considered during classification of the document.
In a further embodiment, the concept clusters can be used with document clusters, which are described in commonly-owned in U.S. Patent 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; and 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, the disclosures of which are incorporated by reference. For example, selecting a concept in the concept cluster display can identify one or more documents with a common idea or topic. Further selection of one of the documents represented by the selected cluster in the document concept display can identify documents that are similarly related to the content of the selected document. The identified documents can be the same or different as the other documents represented by the concept.

In an even further embodiment, the documents identified from one of the concepts can be classified automatically as described in commonly-assigned U.S. Patent Application Serial No. 12/833,769, entitled “System and Method for Providing a Classification Suggestion for Electronically Stored Information,” filed July 9, 2010, pending, the disclosure of which is incorporated by reference.

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

After the clusters have been generated, one or more uncoded concepts can be selected from at least one of the clusters for comparing with a reference concept set or subset. FIGURE 4 is a block diagram showing, by way of example, measures 70 for selecting a concept reference subset 71. The subset of reference concepts 71 can be previously defined 74 and maintained for related document review projects or can be specifically generated for each review project. A predefined reference subset 74 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 72 or customized 73 reference subsets that are determined automatically or by a human reviewer. An arbitrary reference subset 72 includes reference concepts randomly selected for inclusion in the reference subset. A customized reference subset 73 includes reference concepts 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.

The subset of reference concepts, whether predetermined or newly generated, should be selected from a set of reference concepts that are representative of documents in 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 concepts that are representative of the corpus for use in classifying uncoded concepts. During guided review, the uncoded concepts that are dissimilar to all other uncoded concepts are identified based on a similarity threshold. In one embodiment, the dissimilarity can be determined as the cosine of the score vectors for the uncoded concepts. Other methods for determining dissimilarity are possible. Identifying the dissimilar concepts provides a group of concepts that are representative of the document in a corpus for a review project. Each identified dissimilar concept is then classified by assigning a particular classification code based on the content of the associated documents to collectively generate the reference concepts. Guided review can be performed by a reviewer, a machine, or a combination of the reviewer and machine.

Other methods for generating reference concepts 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. Patent No. 7,610,313, the disclosure of which is incorporated by reference. A plurality of the clustered uncoded concepts are selected based on selection criteria, such as cluster centers or sample clusters. The cluster centers can be used to identify uncoded concepts in a cluster that are most similar or dissimilar to the cluster center. The selected uncoded concepts are then assigned classification codes. In a further embodiment, sample clusters can be used to generate reference concepts by selecting one or more sample clusters based on cluster relation criteria, such as size, content, similarity, or dissimilarity. The uncoded concepts in the selected sample clusters are then selected for classification by assigning classification codes. The classified concepts represent reference concepts for the document review project. The number of reference concepts can be determined automatically or by a reviewer. Other methods for selecting concepts for use as reference concepts are possible.

An uncoded concept selected from one of the clusters can be compared to the reference concepts to identify similar reference concepts for use in providing suggestions regarding classification of the selected uncoded concept. FIGURE 5 is a process flow diagram showing, by way of example, a method 80 for comparing an uncoded concept to reference concepts for use in the method of FIGURE 2. The uncoded concept is selected from a cluster (block 81) and applied to the reference concepts (block 82). The reference concepts can include all reference
concepts for a document review project or a subset of the reference concepts. Each of the
reference concepts and the selected uncoded concept can be represented by a score vector having
paired values of documents associated with that concept and associated scores. A similarity
between the uncoded concept and each reference concept is determined (block 83) as the $\cos \sigma$
of the score vectors for the uncoded concept and reference concept being compared and is
equivalent to the inner product between the score vectors. In the described embodiment, the $\cos \sigma$
is calculated in accordance with the equation:

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

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

One or more of the reference concepts that are most similar to the selected uncoded
concept, based on the similarity metric, are identified. The most similar reference concepts can
be identified by satisfying a predetermined threshold of similarity. Other methods for
determining the similar reference concepts are possible, such as setting a predetermined absolute
number of the most similar reference concepts. The classification codes of the identified similar
reference concepts can be used as suggestions for classifying the selected uncoded concept, as
further described below with reference to FIGURE 8. Once identified, the similar reference
concepts can be used to provide suggestions regarding classification of the selected uncoded
concept, as further described below with reference to FIGURES 6 and 7.

The similar reference concepts can be displayed with the clusters of uncoded concepts.
In the display, the similar reference concepts 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. Patent 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. Patent No. 7,610,313, the disclosure of
which is incorporated by reference. Other displays of the clusters and similar reference
documents are possible.

Organizing the clusters into spines and groups of cluster spines provides an individual
reviewer with a display that presents the concepts according to a theme while maximizing the
number of relationships depicted between the concepts. FIGURE 6 is a screenshot 90 showing,
by way of example, a visual display 91 of similar reference concepts 94 and uncoded concepts 94. Clusters 92 of the uncoded concepts 93 can be located along a spine, which is a vector, based on a similarity of the uncoded concepts 93 in the clusters 92. The uncoded concepts 93 are each represented by a smaller circle within the clusters 92.

Similar reference concepts 94 identified for a selected uncoded concept 93 can be displayed in a list 95 by document title or other identifier. Also, classification codes 96 associated with the similar reference concepts 94 can be displayed as circles having a diamond shape within the boundary of the circle. The classification codes 96 can include "privileged," "responsive," and "non-responsive" codes, as well as other codes. The different classification codes 96 can each be represented by a color, such as blue for "privileged" reference documents and yellow for "non-responsive" reference concepts. Other display representations of the uncoded concepts, similar reference concepts, and classification codes are possible, including by symbols and shapes.

The classification codes 96 of the similar reference concepts 94 can provide suggestions for classifying the selected uncoded concept based on factors, such as a number of different classification codes for the similar reference concepts and a number of similar reference concepts associated with each classification code. For example, the list of reference concepts includes four similar reference concepts identified for a particular uncoded concept. Three of the reference concepts are classified as "privileged," while one is classified as "non-responsive." In making a decision to assign a classification code to a selected uncoded concept, the reviewer can consider classification factors based on the similar reference concepts, such as a presence or absence of similar reference concepts with different classification codes and a quantity of the similar reference concepts for each classification code. Other classification factors are possible. In the current example, the display 91 provides suggestions, including the number of "privileged" similar reference concepts, the number of "non-responsive" similar reference concepts, and the absence of other classification codes of similar reference concepts. Based on the number of "privileged" similar reference concepts compared to the number of "non-responsive" similar reference concepts, the reviewer may be more inclined to classify the selected uncoded concepts as "privileged." Alternatively, the reviewer may wish to further review the selected uncoded concept based on the multiple classification codes of the similar reference concepts. Other classification codes and combinations of classification codes are possible. The reviewer can utilize the suggestions provided by the similar reference concepts to assign a classification to the selected uncoded concept. In a further embodiment, the now
classified and previously uncoded concept can be added to the set of reference concepts for use in classifying other uncoded concepts.

In a further embodiment, similar reference concepts can be identified for a cluster or spine to provide suggestions for classifying the cluster and spine. For a cluster, the similar reference concepts are identified based on a comparison of a score vector for the cluster, which is representative of the cluster center and the reference concept score vectors. Meanwhile, identifying similar reference concepts 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 concept score vectors. Once identified, the similar reference concepts are used for classifying the cluster or spine.

In an even further embodiment, the uncoded concepts, including the selected uncoded concept, and the similar reference concepts can be displayed as a concept list. FIGURE 7 is a screenshot 100 showing, by way of example, an alternative visual display of the similar reference concepts 105 and uncoded concepts 102. The uncoded concepts 102 can be provided as a list in an uncoded concept box 101, such as an email inbox. The uncoded concepts 102 can be identified and organized based on metadata about the uncoded concept or information provided in the associated documents.

At least one of the uncoded concepts can be selected and displayed in a concept viewing box 104. The selected uncoded concept can be identified in the list 101 using a selection indicator (not shown), including a symbol, font, or highlighting. Other selection indicators and uncoded concept factors are possible. Once identified, the selected uncoded concept can be compared to a set of reference concepts to identify the reference concepts 85 most similar. The identified similar reference concepts 105 can be displayed below the concept viewing box 104 with an associated classification code 103. The classification code of the similar reference concept 105 can be used as a suggestion for classifying the selected uncoded concept. After assigning a classification code, a representation 103 of the classification can be provided in the display with the selected uncoded concept. In a further embodiment, the now classified and previously uncoded concept can be added to the set of reference concepts.

Similar reference concepts can be used as suggestions to indicate a need for manual review of the uncoded concepts, when review may be unnecessary, and hints for classifying the uncoded concepts, clusters, or spines. Additional information can be generated to assist a reviewer in making classification decisions for the uncoded concepts, such as a machine-generated confidence level associated with a suggested classification code, as described in common-assigned U.S. Patent Application Serial No. 12/844,785, entitled "System and Method
for Providing a Classification Suggestion for Concepts," filed on July 27, 2010, pending, the
disclosure of which is incorporated by reference.

The machine-generated suggestion for classification and associated confidence level can be
determined by a classifier. FIGURE 8 is a process flow diagram 110 showing, by way of
example, a method for classifying uncoded concepts by a classifier for use in the method of
FIGURE 2. An uncoded concept is selected from a cluster (block 111) and compared to a
neighborhood of x-similar reference concepts (block 112) to identify those similar reference
concepts that are most relevant to the selected uncoded concept. The selected uncoded concept
can be the same as the uncoded concept selected for identifying similar reference concepts or a
different uncoded concept. 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 concepts for the cluster or spine.

The neighborhood of x-similar reference concepts is determined separately for each
selected uncoded concept and can include one or more similar reference concepts. During
neighborhood generation, a value for x-similar reference concepts is first determined
automatically or by an individual reviewer. The neighborhood of similar reference concepts can
include the reference concepts, which were identified as similar reference concepts according to
the method of FIGURE 5, or reference concepts 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 concepts nearest to the selected uncoded concept are identified.
Finally, the identified x-number of similar reference concepts are provided as the neighborhood
for the selected uncoded concept. In a further embodiment, the x-number of similar reference
concepts are defined for each classification code, rather than across all classification codes.
Once generated, the x-number of similar reference concepts in the neighborhood and the selected
uncoded concept are analyzed by the classifier to provide a machine-generated classification
suggestion for assigning a classification code (block 113). A confidence level for the machine-
generated classification suggestion is also provided (block 114).

The machine-generated analysis of the selected uncoded concept and x-number of similar
reference concepts can be based on one or more routines performed by the classifier, such as a
nearest neighbor (NN) classifier. The routines for determining a suggested classification code
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 concept includes identifying a neighbor that is the closest distance to the
selected uncoded concept and assigning the classification code of the closest neighbor as the suggested classification code for the selected uncoded concept. The closest neighbor is determined by comparing the score vectors for the selected uncoded concept with each of the x-number of similar reference concepts in the neighborhood as the cos σ to determine a distance metric. The distance metrics for the x-number of similar reference concepts are compared to identify the similar reference concept closest to the selected uncoded concept as the closest neighbor.

The minimum average distance classification measure includes calculating an average distance of the similar reference concepts for each classification code. The classification code of the similar reference concepts having the closest average distance to the selected uncoded concept is assigned as the suggested classification code. The maximum count classification measure, also known as the voting classification measure, includes counting a number of similar reference concepts for each classification code and assigning a count or “vote” to the similar reference concepts based on the assigned classification code. The classification code with the highest number of similar reference concepts or “votes” is assigned to the selected uncoded concept as the suggested classification code. The distance weighted maximum count classification measure includes identifying a count of all similar reference concepts for each classification code and determining a distance between the selected uncoded concept and each of the similar reference concepts. Each count assigned to the similar reference concepts is weighted based on the distance of the similar reference concept from the selected uncoded concept. The classification code with the highest count, after consideration of the weight, is assigned to the selected uncoded concept as the suggested classification code.

The machine-generated suggested classification code is provided for the selected uncoded concept 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 concept. Alternatively, the x-NN classifier can automatically assign the suggested classification code. In one embodiment, the x-NN classifier only assigns an uncoded concept 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.

Machine classification can also occur on a cluster or spine level once one or more concepts 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 FIGURE 5. A neighborhood for the selected cluster can be determined based on a
distance metric. The x-number of similar reference concepts that are closest to the cluster center can be selected for inclusion in the neighborhood, as described above. Each concept 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 concepts to determine an x-number of similar reference concepts 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 concepts is determined for a spine by comparing a spine score vector with the vector for each similar reference concept to identify the neighborhood of similar concepts that are the most similar.

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

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

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

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.
1. A method (50) for displaying relationships between concepts (14c, 14d) to provide classification suggestions via nearest neighbor, comprising:
   providing reference concepts (14d) each associated with a classification code (96) and a set of uncoded concepts (14c), wherein each of the reference concepts (14d) and the uncoded concepts (14c) comprises one or more nouns extracted from a plurality of documents (14a, 17, 20, 26, 29);
   comparing at least one uncoded concept (14c) with the reference concepts (14d) and identifying one or more of the reference concepts (14d) that are similar to the at least one uncoded concept (14c); and
   depicting relationships between the at least one uncoded concept (14c) and the similar reference concepts (14d) for classifying the at least one uncoded concept (14c), wherein the steps are performed on a suitably programmed computer.

2. A method (50) according to Claim 1, further comprising:
   classifying the at least one uncoded concept by assigning a classification code (96) based on the relationships between the at least one uncoded concept and the similar reference concepts (14d).

3. A method (50) according to Claim 2, further comprising:
   adding the classified at least one uncoded concept to the reference concepts (14d).

4. A method (50) according to Claim 2, further comprising:
   providing a confidence level for the classification code (96) of the at least one uncoded concept.

5. A method (50) according to Claim 2, further comprising:
   identifying those documents (14a, 17, 20, 26, 29) associated with the classified at least one uncoded concept; and
   assigning the classification code (96) for the classified at least one uncoded concept to one or more of the associated documents (14a, 17, 20, 26, 29).
6. A method (50) according to Claim 5, wherein the documents (14a, 17, 20, 26, 29) are identified using a matrix comprising a mapping of concepts and related documents (14a, 17, 20, 26, 29).

7. A method (50) according to Claim 1, further comprising:
   generating the reference concepts (14d) from a set of concepts, comprising at least one of:
   identifying the concepts that are dissimilar from each other concept in the set of concepts and assigning the classification code (96) to each of the dissimilar concepts, as the reference concepts (14d); and
   grouping the set of concepts into clusters (92), selecting one or more of the concepts in at least one cluster, and assigning the classification code (96) to each of the selected concepts, as the reference concepts (14d).

8. A method (50) according to Claim 1, further comprising:
   determining the similar reference concepts (14d), comprising:
   forming a score vector for each uncoded concept and each reference concept; and
   calculating a similarity metric by comparing the score vectors for the at least one uncoded concept and each of the reference concepts (14d); and
   selecting the reference concepts (14d) with the highest similarity metrics as the similar reference concepts (14d).

9. A method (50) according to Claim 1, further comprising:
   determining the similar reference concepts (14d), comprising:
   determining a measure of similarity between the at least one uncoded concept and each of the reference concepts (14d) based on the comparison;
   applying a threshold to the measures of similarity; and
   selecting those reference concepts (14d) that satisfy the threshold as the similar reference concepts (14d).

10. A method (50) according to Claim 1, further comprising:
clustering the uncoded concepts (14c) and displaying the clusters (92);

displaying the similar reference concepts (14d) in a list adjacent to the
clusters (92).

11. A system (10) for displaying relationships between concepts to
provide classification suggestions via nearest neighbor, comprising:
a database to maintain reference concepts (14d) each associated with a
classification code (96) and a set of uncoded concepts (14c), wherein each of
the reference concepts (14d) and the uncoded concepts (14c) comprises one or
more nouns extracted from a plurality of documents (14a, 17, 20, 26, 29);
a similarity module to compare at least one uncoded concept with the
reference concepts (14d) and to identify one or more of the reference concepts
(14d) that are similar to the at least one uncoded concept; and
a display to depict relationships between the at least one uncoded
concept and the similar reference concepts (14d) for classifying the at least
one uncoded concept.

12. A system (10) according to Claim 11, further comprising:
a classification module to classify the at least one uncoded concept by
assigning a classification code (96) based on the relationships between the at
least one uncoded concept and the similar reference concepts (14d).

13. A system (10) according to Claim 12, further comprising:
a reference module to add the classified at least one uncoded concept
to the reference concepts (14d).

14. A system (10) according to Claim 12, wherein the classification
module provides a confidence level for the classification code (96) of the at
least one uncoded concept.

15. A system (10) according to Claim 12, further comprising:
a document classification module to identify those documents (14a, 17,
20, 26, 29) associated with the classified at least one uncoded concept and to
assign the classification code (96) for the classified at least one uncoded
concept to one or more of the associated documents (14a, 17, 20, 26, 29).
16. A system (10) according to Claim 15, wherein the documents (14a, 17, 20, 26, 29) are identified using a matrix comprising a mapping of concepts and related documents (14a, 17, 20, 26, 29).

17. A system (10) according to Claim 11, further comprising:
   a reference set module to generate the reference concepts (14d) from a set of concepts, comprising at least one of:
   a comparison module to identify the concepts that are dissimilar from each other concept in the concept set and to assign the classification code (96) to each of the dissimilar concepts, as the reference concepts (14d); and
   a reference clustering module to group the set of concepts into one or more clusters (92), to select one or more of the concepts in at least one cluster, and to assign the classification code (96) to each of the selected concepts, as the reference concepts (14d).

18. A system (10) according to Claim 11, further comprising:
   a concept similarity module to determine the similar reference concepts (14d), comprising:
   a vector module to form a score vector for each uncoded concept and each reference concept; and
   a similarity measurement module to calculate a similarity metric by comparing the score vectors for the at least one uncoded concept and each of the reference concepts (14d) and to select the reference concepts (14d) with the highest similarity metrics as the similar reference concepts (14d).

19. A system (10) according to Claim 11, further comprising:
   a concept similarity module to determine the similar reference concepts (14d), comprising:
   a similarity measurement module to determine a measure of similarity between the at least one uncoded concept and each of the reference concepts (14d) based on the comparison; and
7 a threshold module to apply a threshold to the measures of
8 similarity and to select those reference concepts (14d) that satisfy the
9 threshold as the similar reference concepts (14d).

20. A system (10) according to Claim 11, further comprising:
   a clustering module to cluster the uncoded concepts (14c); and
   the display to present the clusters (92) and the similar reference
   concepts (14d) in a list adjacent to the clusters (92).
Fig. 2.

50

Start

51

Obtain concept cluster set

52

Identify reference concept

53

Compare an uncoded concept with the reference set

54

Identify similar reference concepts

55

Identify relationships between the similar reference concepts and the uncoded concept

56

Classify the uncoded concept

End
Fig. 3.

\[ \text{Concepts (62)} \]

\[ \begin{array}{c|c|c|c|c}
C_1 & C_2 & C_3 & \cdots & C_m \\
2 & 12 & 4 & \cdots & 21 & 0 & 2 \\
\end{array} \]

\[ \text{Documents (61)} \]

\[ D_1, D_2, D_3, \ldots, D_n \]
Fig. 4.

70

74 Predetermined

74 Arbitrary

71 Reference Concept Set

73 Customized

Fig. 5.

80

Start

Select uncoded concept

Apply uncoded concept to reference set

Determine similarity between the uncoded concept and each reference concept

End
Fig. 7.

---

From: Michael Guenther
Sent: 2/2/2002 3:39 PM
To: LoriS@reademnweep.com
Cc:
Bcc:

Subject: acme

I saw acme construction here 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

ace
Fig. 8.

110

Classification

Select an uncoded concept

Compare the selected uncoded concept to a neighborhood of reference concepts

Provide a suggested classification for the selected uncoded concept based on the comparison

Provide a confidence level for the suggested classification

End