US 20140207782A1

(19) **United States**

(12) **Patent Application Publication**

RAVID

(10) **Pub. No.: US 2014/0207782 A1**

(43) **Pub. Date: Jul. 24, 2014**

(54) **SYSTEM AND METHOD FOR COMPUTERIZED SEMANTIC PROCESSING OF ELECTRONIC DOCUMENTS INCLUDING THEMES**

(71) Applicant: **Equivio Ltd.**, Rosh Haayin (IL)

(72) Inventor: **Yiftach RAVID**, Rosh Haayin (IL)

(21) Appl. No.: **14/161,159**

(22) Filed: **Jan. 22, 2014**

Related U.S. Application Data

(60) Provisional application No. 61/755,242, filed on Jan. 22, 2013.

Publication Classification

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

(52) **U.S. Cl.**
CPC ***G06F 17/30011*** (2013.01)
USPC **707/739**

(57) **ABSTRACT**

System and method for computerized identification of themes in a large data set, the system comprising reducing the number of data set members in a large data set, using at least one computerized data set member pruning technique other than random selection; and using a computerized theme identification technique for identifying a plurality of themes in the reduced data set.

step i. Input: a set of electronic documents.

↓

Step ii Extract text from the data collection. Text extraction can be done by third party software such as: Oracle inside out, iSys, DTSearch, iFilter, etc.

↓

Step iii: Compute Near-duplicate (ND) on the dataset.

↓

Step iiia: For each document compute DuplicateSubsetID : all documents having the same DuplicateSubsetID having an identical text.

↓

Step iiib: For each document compute EquiSetID: all documents having the same EquiSetID are similar (for each document x in the set there is another document y in the set, such that the similarity between the two is greater than some threshold).

↓

Step iiic: For each document compute Pivot: 1 if the document is a representative of the set (and 0 otherwise).

↓

Step iv. Compute Email threads (ET) on the dataset.

↓

Step v. Run a topic modeling algorithm (such as LDA) on a subset of the dataset, including feature extraction. Resulting topics are defined as themes.

FIG. 1

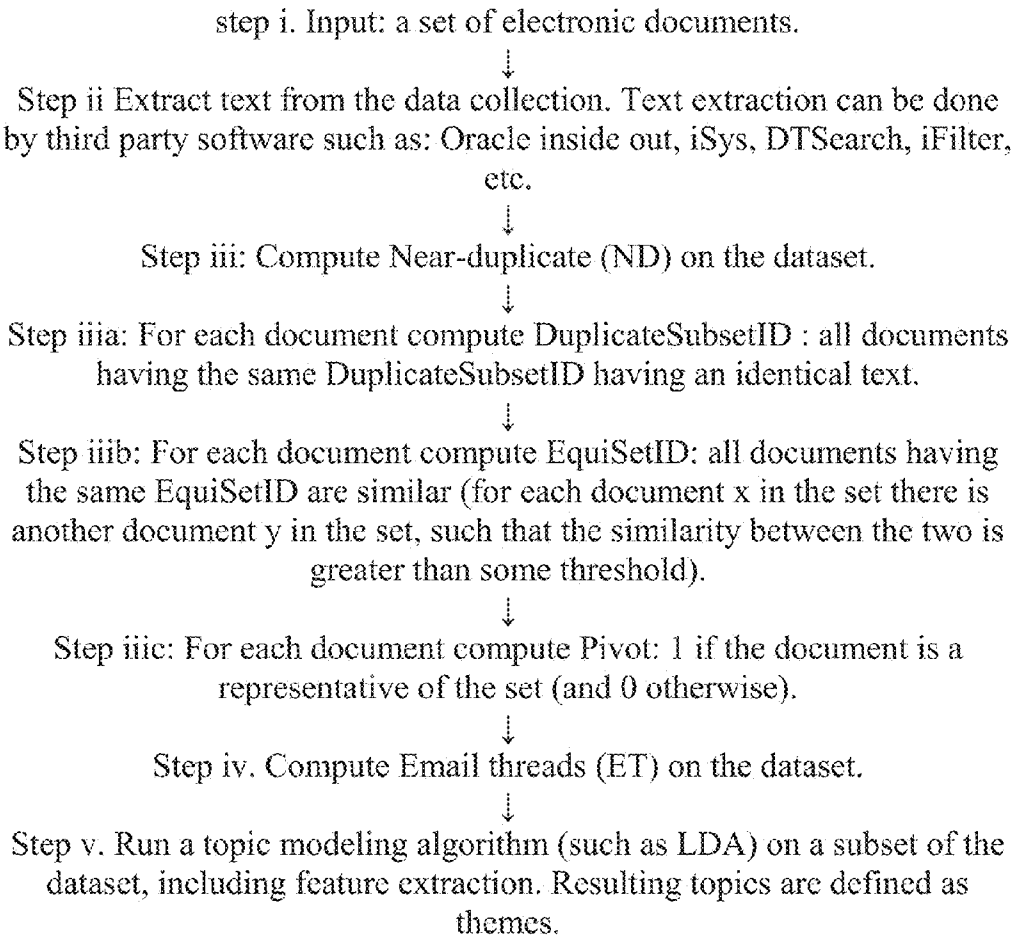
step i. Input: a set of electronic documents.
Step ii Extract text from the data collection. Text extraction can be done by third party software such as: Oracle inside out, iSys, DTSearch, iFilter, etc.
Step iii: Compute Near-duplicate (ND) on the dataset.
Step iiia: For each document compute DuplicateSubsetID : all documents having the same DuplicateSubsetID having an identical text.
Step iiib: For each document compute EquiSetID: all documents having the same EquiSetID are similar (for each document x in the set there is another document y in the set, such that the similarity between the two is greater than some threshold).
Step iiic: For each document compute Pivot: 1 if the document is a representative of the set (and 0 otherwise).
Step iv. Compute Email threads (ET) on the dataset.
Step v. Run a topic modeling algorithm (such as LDA) on a subset of the dataset, including feature extraction. Resulting topics are defined as themes.

FIG. 2

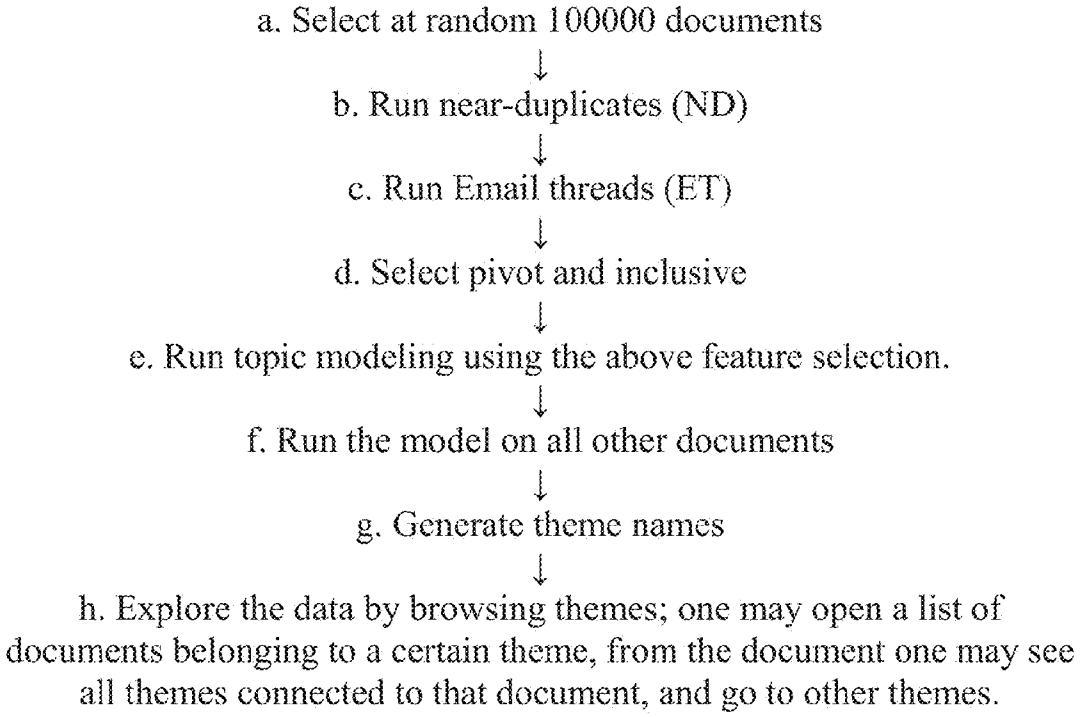
a. Select at random 100000 documents
b. Run near-duplicates (ND)
c. Run Email threads (ET)
d. Select pivot and inclusive
e. Run topic modeling using the above feature selection.
f. Run the model on all other documents
g. Generate theme names
h. Explore the data by browsing themes; one may open a list of documents belonging to a certain theme, from the document one may see all themes connected to that document, and go to other themes.

Fig. 3a

10: Provide a collection of thousands or millions of electronic documents (D) e.g. including a mixture of 1 or more of: non-emails, e-mails with attachments, e-mails without attachments

↓

20 Run Near-duplicate Identifying functionality on the collection, thereby to identify all sets of near-duplicates in the collection

↓

30 Run Email thread Identifying functionality on all emails in the collection thereby to identify all email threads in the collection

↓

40 perform step/s to pare down the collection of documents (D), thereby to yield a pared-down collection (Z):

↓

40a. Select one (say) document (e.g. pivot document) to represent each set of near-duplicate set –thereby to yield a set X1 of documents.

↓

40b. Select (only) inclusive to represent from each email thread thereby to yield a set X2 of inclusive emails

↓

40c. From the set X2 of inclusive emails select one (say) document from each near-duplicate set thereby to yield a set X3 e.g. first select all inclusives then take only one inclusive from each set of "similar" inclusive

↓

to Fig. 3b, step 50

FIG. 3b

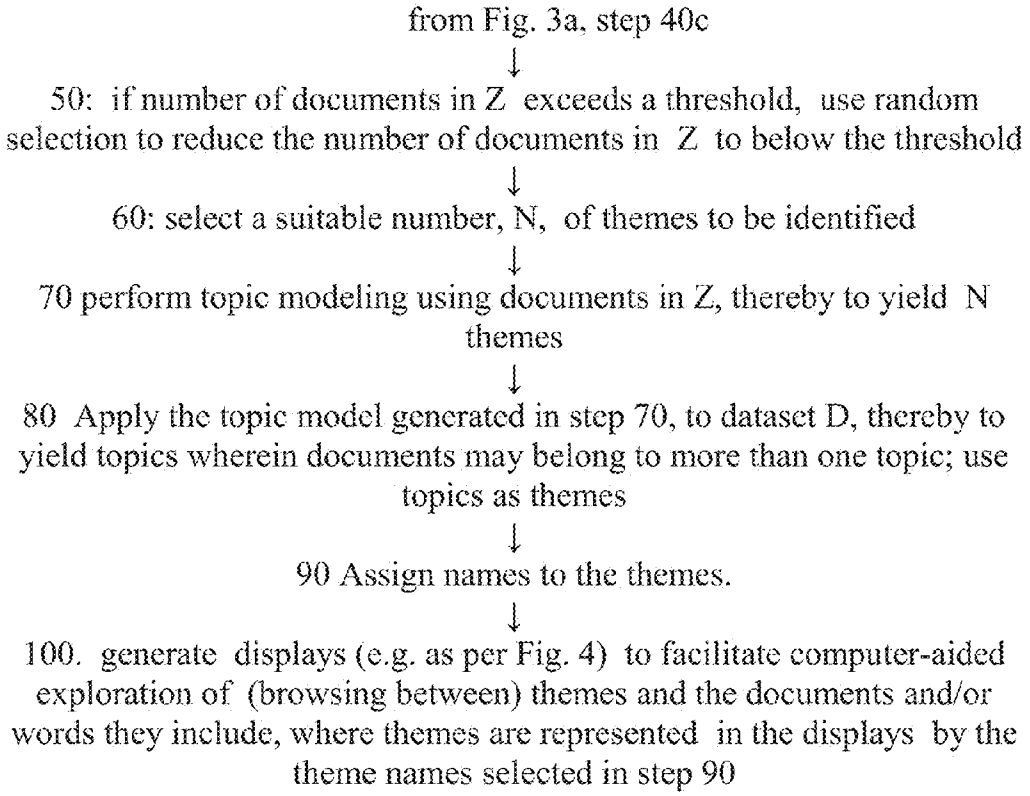
from Fig. 3a, step 40c
50: if number of documents in Z exceeds a threshold, use random selection to reduce the number of documents in Z to below the threshold
60: select a suitable number, N, of themes to be identified
70 perform topic modeling using documents in Z, thereby to yield N themes
80 Apply the topic model generated in step 70, to dataset D, thereby to yield topics wherein documents may belong to more than one topic; use topics as themes
90 Assign names to the themes.
100. generate displays (e.g. as per Fig. 4) to facilitate computer-aided exploration of (browsing between) themes and the documents and/or words they include, where themes are represented in the displays by the theme names selected in step 90

FIG. 4

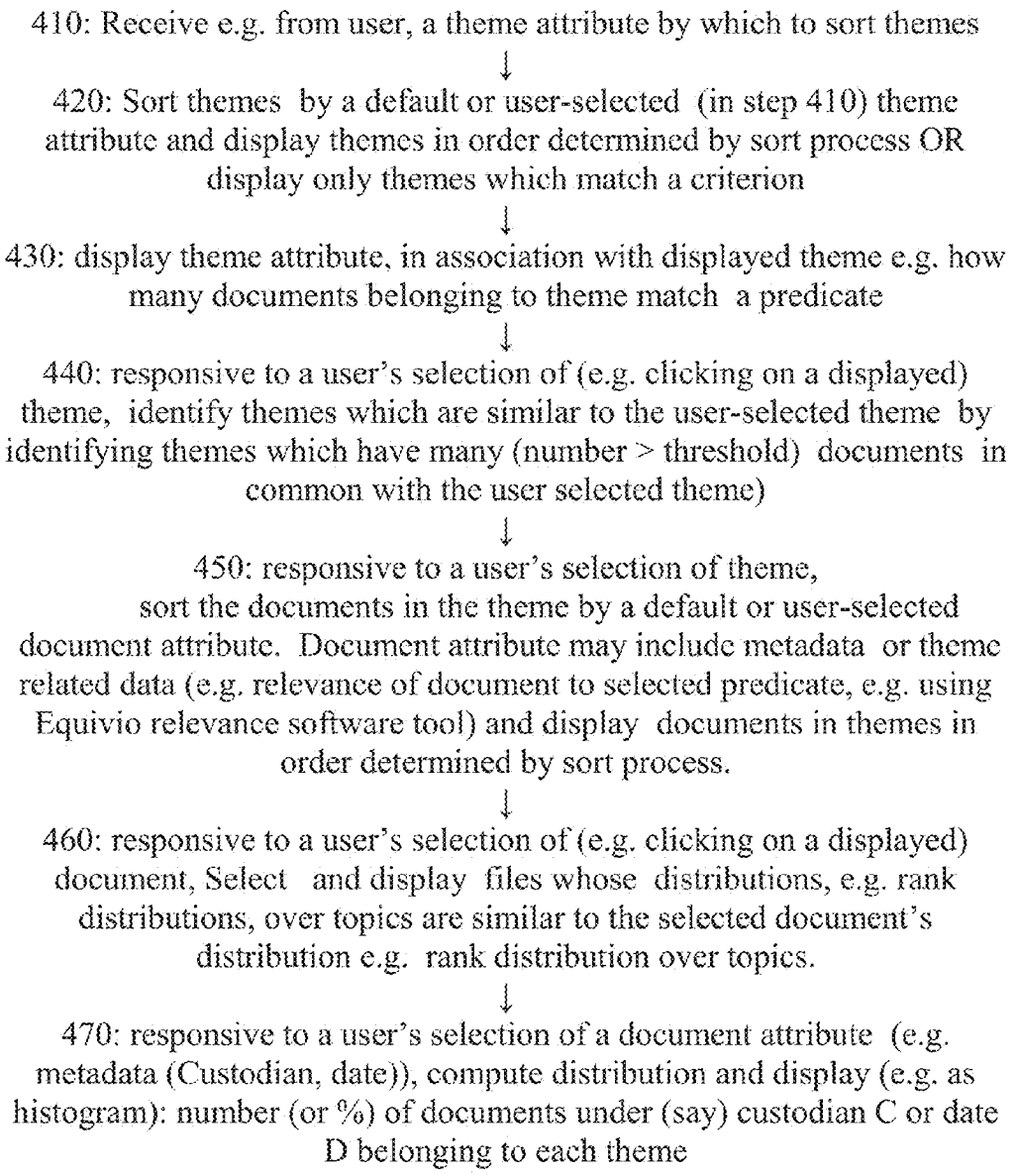
410: Receive e.g. from user, a theme attribute by which to sort themes
420: Sort themes by a default or user-selected (in step 410) theme attribute and display themes in order determined by sort process OR display only themes which match a criterion
430: display theme attribute, in association with displayed theme e.g. how many documents belonging to theme match a predicate
440: responsive to a user's selection of (e.g. clicking on a displayed) theme, identify themes which are similar to the user-selected theme by identifying themes which have many (number > threshold) documents in common with the user selected theme)
450: responsive to a user's selection of theme, sort the documents in the theme by a default or user-selected document attribute. Document attribute may include metadata or theme related data (e.g. relevance of document to selected predicate, e.g. using Equivio relevance software tool) and display documents in themes in order determined by sort process.
460: responsive to a user's selection of (e.g. clicking on a displayed) document, Select and display files whose distributions, e.g. rank distributions, over topics are similar to the selected document's distribution e.g. rank distribution over topics.
470: responsive to a user's selection of a document attribute (e.g. metadata (Custodian, date)), compute distribution and display (e.g. as histogram): number (or %) of documents under (say) custodian C or date D belonging to each theme

Fig. 5

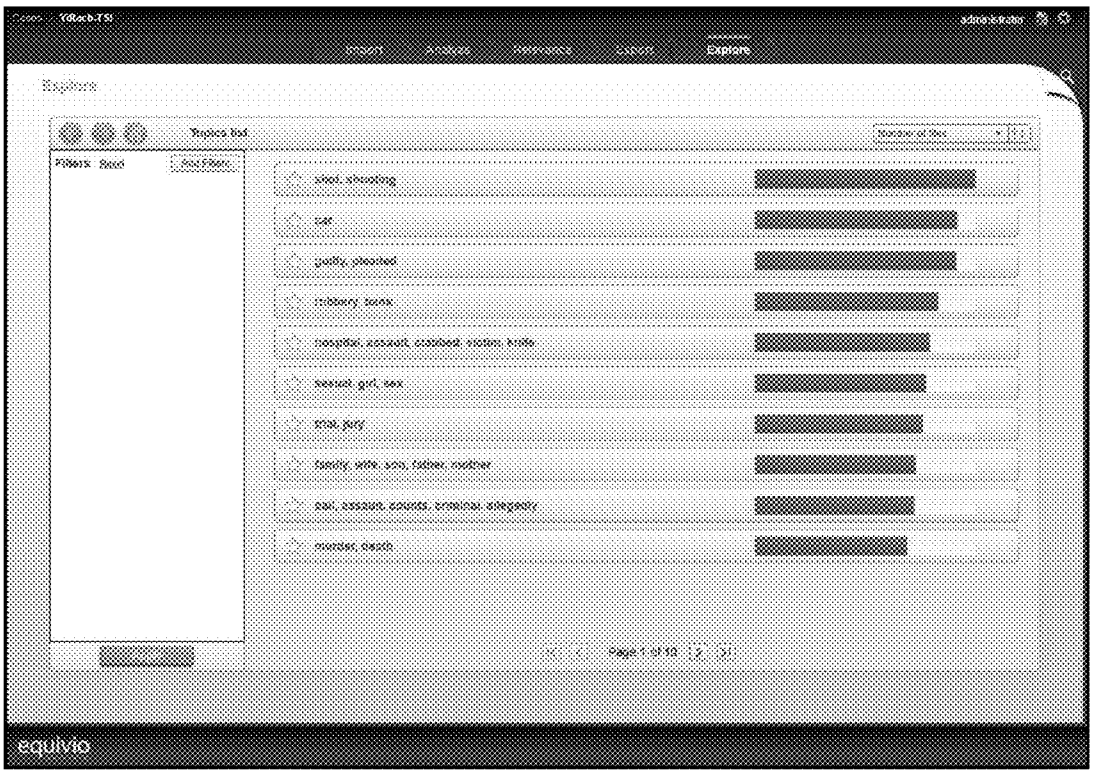

Fig. 6

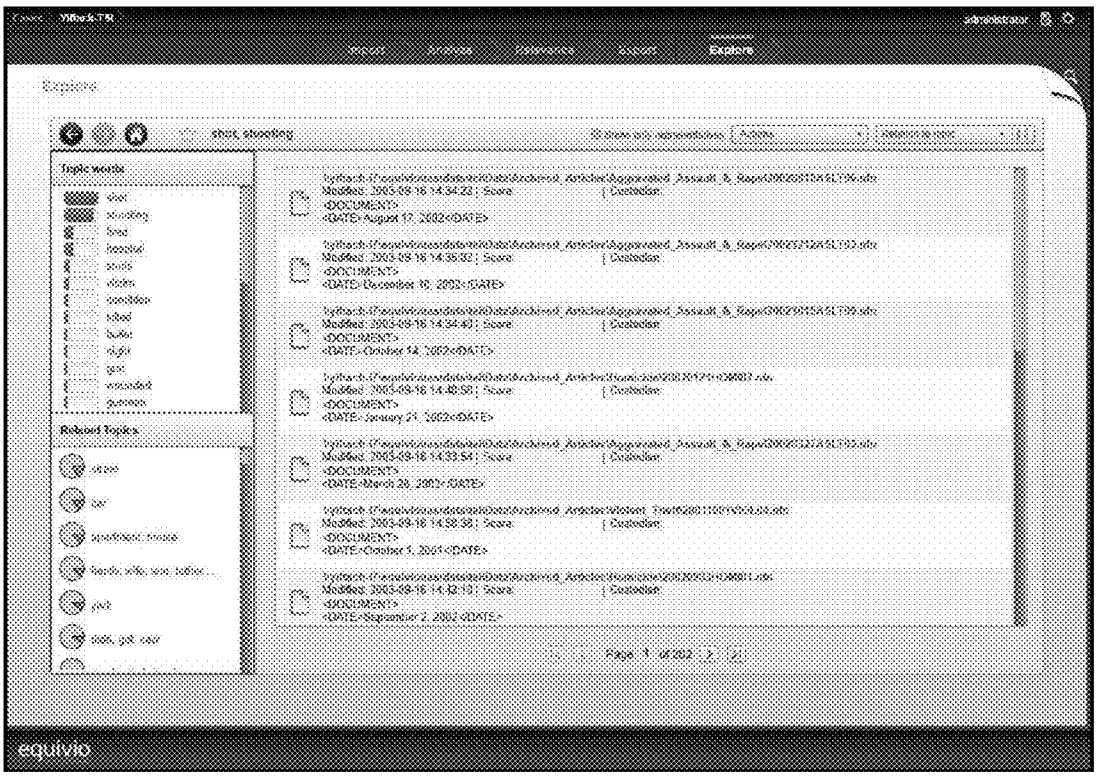

Fig. 7

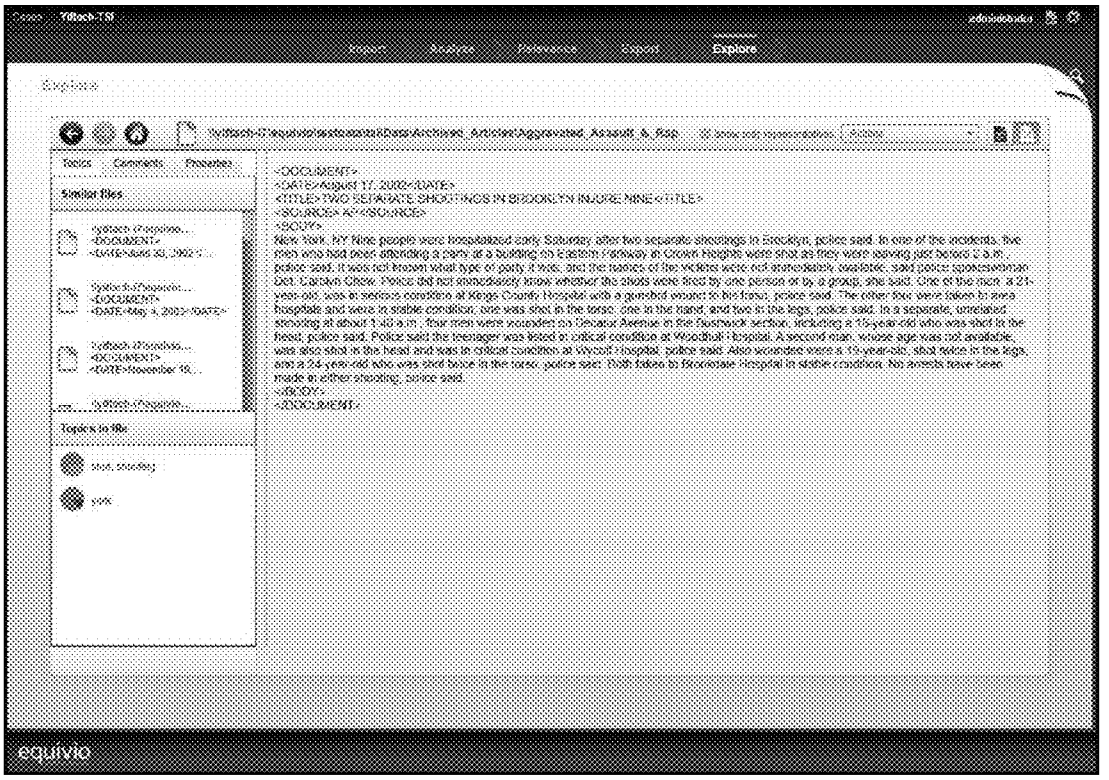
Explore
Topics
Comments
Properties
Similar files
Topics in file
<DOCUMENT>
<DATE>August 17, 2002</DATE>
<TITLE>TWO SEPARATE SHOOTINGS IN BROOKLYN INJURE NINE</TITLE>
<SOURCE> AP</SOURCE>
<BODY>
New York, NY Nine people were hospitalized early Saturday after two separate shootings in Brooklyn, police said. In one of the incidents, five men who had been attending a party at a building on Eastern Parkway in Crown Heights were shot as they were leaving just before 2 a.m., police said. It was not known what type of party it was, and the names of the victims were not immediately available, said police spokeswoman Det. Carolyn Drew. Police did not immediately know whether the shots were fired by one person or by a group, she said. One of the men, a 21-year-old, was in serious condition at Kings County Hospital with a gunshot wound to his torso, police said. The other four were taken to area hospitals and were in stable condition; one was shot in the torso, one in the hand, and two in the legs, police said. In a separate, unrelated shooting at about 1:40 a.m., four men were wounded on Decatur Avenue in the Bushwick section, including a 19-year-old who was shot in the head, police said. Police said the teenager was listed in critical condition at Woodhull Hospital. A second man, whose age was not available, was also shot in the head and was in critical condition at Wyckoff Hospital, police said. Also wounded were a 19-year-old, shot twice in the legs, and a 24-year-old who was shot twice in the torso, police said. Both taken to Brookdale Hospital in stable condition. No arrests have been made in either shooting, police said.
</BODY>
</DOCUMENT>
equivio

Fig. 8

1010. use computerized Themes functionality to identify an initial "seed" set of (say 5 – 30) relevant documents in a large sparse collection of electronic documents

↓

1020. generate a training set of documents including the initial "seed" set of relevant documents and at least an equal number of documents randomly selected from the large sparse collection of electronic documents

↓

1030. operate computerized relevant document identification system, e.g. Equivio Zoom's Relevance functionality on the training set, thereby to successfully identify the rare relevant documents in the large sparse collection

SYSTEM AND METHOD FOR COMPUTERIZED SEMANTIC PROCESSING OF ELECTRONIC DOCUMENTS INCLUDING THEMES

[0001] Priority is claimed from U.S. Provisional Patent Application No. 61/755,242, entitled "Computerized systems and methods for use of themes in e-discovery" and filed Jan. 22, 2013, the entire contents of which being hereby incorporated herein by reference.

FIELD OF THIS DISCLOSURE

[0002] The present invention relates generally to computerized processing of electronic documents and more particularly to computerized semantic processing of electronic documents.

BACKGROUND FOR THIS DISCLOSURE

[0003] Wikipedia on "Data_deduplication" states that "In computing, data deduplication is a specialized data compression technique for eliminating duplicate copies of repeating data. Related and somewhat synonymous terms are intelligent (data) compression and single-instance (data) storage . . . . For example a typical email system might contain 100 instances of the same 1 MB (megabyte) file attachment. Each time the email platform is backed up, all 100 instances of the attachment are saved, requiring 100 MB storage space. With data deduplication, only one instance of the attachment is actually stored; the subsequent instances are referenced back to the saved copy for deduplication ratio of roughly 100 to 1."

[0004] Also according to Wikipedia, the term deduplication may refer to Data deduplication, as above, or to Record linkage, in databases, i.e. finding entries that refer to the same entity in two or more files. 'DeDuping' may involve removing duplicates in Customer and Address records in a Database or Spreadsheet.

[0005] The importance of topic modeling for browsing is known, e.g. at the following http-www-linked publication: cs.princeton.edu/~blei/topicmodeling.html.

[0006] It is also known that "Clustering can be used to assist browsing. Browsing tools complement search tools" e.g. as described at the following http-linked publication: pages.cs.wisc.edu/~pradheep/Clust-LDA.pdf.

[0007] Other state of the art related technologies are described inter alia in:

[0008] 1. Papadimitriou, Christos; Raghavan, Prabhakar; Tamaki, Hisao; Vempala, Santosh (1998). "Latent Semantic Indexing: A probabilistic analysis" (Postscript). Proceedings of ACM PODS. http://www.cs.berkeley.edu/~christos/ir.ps.

[0009] 2. Hofmann, Thomas (1999). "Probabilistic Latent Semantic Indexing" (PDF). Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval. http://www.cs.brown.edu/~th/papers/Hofmann-SIGIR99.pdf.

[0010] 3. Blei, David M.; Ng, Andrew Y.; Jordan, Michael I; Lafferty, John (January 2003). "Latent Dirichlet allocation". Journal of Machine Learning Research 3: 993-1022. doi:10.1162/jmlr 2003.3.4-5.993. http://jmlr.csail.mit.edu/papers/v3/blei03a.html.

[0011] 4. Blei, David M. (April 2012). "Introduction to Probabilistic Topic Models" (PDF). Comm. ACM 55 (4): 77-84. doi:10.1145/2133806.2133826. http://www.cs.princeton.edu/~blei/papers/Blei2011.pdf

[0012] 5. Sanjeev Arora; Rong Ge; Ankur Moitra (April 2012). "Learning Topic Models-Going beyond SVD". arXiv:1204.1956.

[0013] 6. Girolami, Mark; Kaban, A. (2003). "On an Equivalence between PLSI and LDA". Proceedings of SIGIR 2003. New York: Association for Computing Machinery. ISBN 1-58113-646-3.

[0014] 7. Griffiths, Thomas L.; Steyvers, Mark (Apr. 6, 2004). "Finding scientific topics". Proceedings of the National Academy of Sciences 101 (Suppl. 1): 5228-5235. doi:10.1073/pnas.0307752101. PMC 387300. PMID 14872004.

[0015] 8. Minka, Thomas; Lafferty, John (2002). "Expectation-propagation for the generative aspect model". Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence. San Francisco, Calif.: Morgan Kaufmann ISBN 1-55860-897-4.

[0016] 9. Blei, David M.; Lafferty, John D. (2006). "Correlated topic models". Advances in Neural Information Processing Systems 18.

[0017] 10. Blei, David M.; Jordan, Michael I.; Griffiths, Thomas L.; Tenenbaum; Joshua B (2004). "Hierarchical Topic Models and the Nested Chinese Restaurant Process". Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference. MIT Press. ISBN 0-262-20152-6.

[0018] 11. Quercia, Daniele; Harry Askham, Jon Crowcroft (2012). "TweetLDA: Supervised Topic Classification and Link Prediction in Twitter". ACM WebSci.

[0019] 12. Li, Fei-Fei; Perona, Pietro. "A Bayesian Hierarchical Model for Learning Natural Scene Categories". Proceedings of the 2005 IEEE Computer Society Conference on Computer VISION and Pattern Recognition (CVPR'05) 2: 524-531.

[0020] 13. Wang, Xiaogang; Grimson, Eric (2007). "Spatial Latent Dirichlet Allocation". Proceedings of Neural Information Processing Systems Conference (NIPS).

Topic modeling (Wikipedia): In machine learning and natural language processing, a topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. An early topic model was described by Papadimitriou, Raghavan, Tamaki and Vempala in 1998. [1] Another one, called Probabilistic latent semantic indexing (PLSI), was created by Thomas Hofmann in 1999.[2] Latent Dirichlet allocation (LDA), perhaps the most common topic model currently in use, is a generalization of PLSI developed by David Blei, Andrew Ng, and Michael Jordan in 2002, allowing documents to have a mixture of topics.[3] Other topic models are generally extensions on LDA, such as Pachinko allocation, which improves on LDA by modeling correlations between topics in addition to the word correlations which constitute topics. Although topic models were first described and implemented in the context of natural language processing, they have applications in other fields such as bioinformatics.

Topics in LDA (Wikipedia): In LDA, each document may be viewed as a mixture of various topics. This is similar to probabilistic latent semantic analysis (pLSA), except that in LDA the topic distribution is assumed to have a Dirichlet prior. In practice, this results in more reasonable mixtures of topics in a document. It has been noted, however, that the

pLSA model is equivalent to the LDA model under a uniform Dirichlet prior distribution.[12]

[0021] The disclosures of all publications and patent documents mentioned in the specification, and of the publications and patent documents cited therein directly or indirectly, are hereby incorporated by reference. Materiality of such publications and patent documents to patentability is not conceded.

SUMMARY OF CERTAIN EMBODIMENTS

[0022] The following terms may be construed either in accordance with any definition thereof appearing in the prior art literature or in accordance with the specification, or as follows:

[0023] Document score: the significance of an individual theme in a particular document. For example, if a topic modeling process defines a topic as a distribution over a fixed vocabulary and assumes that each document includes various topics each with different proportions determined by a per-document distribution over topics then a document's "score" for a particular theme may be the document's level of probability given that document's distribution over a universe of topics; this "score" is typically generated in the course of performing conventional topic-modeling processes. In other words, each topic x has some probability θ_yx of being in document y [Blei, D., Ng, A., Jordan, M. (2003) Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993-1022].

[0024] Duplicate types:

[0025] Entirely Exact duplicate: Two documents which have the same bits in the same order.

[0026] Text exact duplicate: Two documents whose extracted texts are exact duplicates.

[0027] Near-duplicate: Two documents are near-duplicate if the resemblance between them is above a threshold. For example, conventional w-shingling techniques may be employed.

[0028] File: electronic document

[0029] Inclusive: an e-mail whose subject and/or body is not contained in any other e-mail in a given set of emails. This definition implies that if an email is not an inclusive, its subject and/or body is contained in one of the "inclusive" emails defined for that set of emails. Often, an inclusive culminates and includes an entire email thread.

[0030] Overlap: Themes are considered related, or similar, if they overlap. "overlap" may be computed by noting that each document can be assigned to more than one theme/s. Suppose X documents are in theme T_1, and Y documents in T_2; The overlap between T_1 and T_2 is (size of intersection of X and Y)/(size X). The definition of overlap need not be symmetrical; the number of documents that belong to both themes may be divided by the size of the topic of interest. For example if one theme (say, "parrots") is included in another theme (say, "birds") all documents in the small theme are typically also included in the second theme. Then the overlap from the viewpoint of the small theme is 100% and from the viewpoint of the larger theme is less than 100%.

[0031] Pivot: a document which is a representative in some sense, of a set of near-duplicates. Any suitable application-specific policy may be employed to define which document is representative e.g. the g., the document in the set which has the highest/lowest/median number of words.

[0032] Equivio Zoom's Relevance functionality is a commercially available software tool that uses "supervised" machine learning, hence there is a typically human expert that trains the system. In the themes functionality described herein there is typically no supervision or, in some versions, topics may be "semi-supervised". Themes functionality as described herein is useful inter alia in training a system by allowing a human trainer to search for relevancy by browsing through a large collection of electronic documents e.g. as described herein. Since in many cases, finding relevant documents to a specific issue is no simple task, themes functionality described herein may facilitate the process of finding such documents.

[0033] pruning: reducing size of (number and/or size of members in) a data set by removing some members of the set e.g. to achieve a predetermined number/total size of members in the set, including prioritizing removal of set members known to be superfluous e.g. duplicates, vis a vis removal of set members not known to be superfluous which is lower priority.

[0034] Similarity/relatedness:

[0035] similar/related documents/files: Various operational definitions (metrics) are possible for similar documents e.g.

[0036] (a.) documents A, B respectively belonging to theme set A and theme set B where many themes are common to theme sets A and B, or, more generally, that the distributions of the two documents over all themes are close; or

[0037] (b.) The text of the documents are similar (near-duplicate).

[0038] similar/related themes: various operational definitions (metrics) are possible for "similarity" of themes, such as but not limited to themes which have many/few documents in common, or themes whose names have many/few words in common, or

[0039] themes which "overlap" to a considerable degree (over a threshold e.g.).

[0040] Theme: A set of documents which relate to a single subject; each document may simultaneously relate to several subjects hence be included in several themes. For example, some computerized topic modeling methods yield models in which certain documents have a mixture of topics.

[0041] topic: Typically, "topic" as used herein refers to output by conventional topic modeling whereas "theme" typically refers to that output as further processed in accordance with embodiments of the present invention.

[0042] Unique documents: Documents in a collection that do not have any other documents which are near-duplicates.

[0043] Word score: the significance of an individual word to an individual theme. For example, if a topic modeling process defines a topic as a distribution over a fixed vocabulary and assumes that each document includes various topics each with different proportions determined by a per-document distribution over topics, then each word's "score" may be its level of probability given the distribution over the fixed vocabulary defined by the topic; this "score" is typically generated in the course of performing conventional topic-modeling processes. In other words, each word x has some probability β_yx of being in theme y [Blei, D., Ng, A., Jordan, M. (2003) Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993-1022].

[0044] Certain embodiments of the present invention seek to provide computerized systems and methods for use of themes in e-discovery and other semantic tasks.

[0045] Certain embodiments of the present invention seek to provide methods for computerized identification of themes in a large data set.

[0046] Certain embodiments of the present invention seek to provide methods for use of multi-topic modeling in e-discovery and other semantic tasks.

[0047] Embodiments include:

Embodiment 1

[0048] A method for computerized identification of themes in a large data set, the system comprising:

[0049] reducing the number of data set members in a large data set, using at least one computerized data set member pruning technique other than random selection; and

[0050] using a computerized theme identification technique for identifying a plurality of themes in the reduced data set.

Embodiment 2

[0051] A method according to Embodiment 1 wherein the computerized data set member pruning technique comprises thinning out at least one document which passes a document similarity criterion relative to at least one other document not being thinned out, thereby to combat skewing as a result of over-influence of similar, hence over-represented, documents upon the theme identification technique.

Embodiment 3

[0052] A method according to Embodiment 2 wherein the thinning out at least one document which passes a document similarity criterion comprises replacing a plurality of emails forming an email thread, with at least one inclusive email, thereby to thin out emails which are included in the inclusive email hence are deemed to pass the document similarity criterion with regard to the inclusive.

Embodiment 4

[0053] A method according to Embodiment 2 wherein the thinning out at least one document which passes a document similarity criterion comprises identifying and discarding near-duplicates thereby to thin out at least one document which is deemed to pass the document similarity criterion with regard to a set of near-duplicates of the document, at least one of which is not being thinned out.

Embodiment 5

[0054] A method according to Embodiment 1 wherein the computerized theme identification technique comprises topic modeling.

Embodiment 6

[0055] A method according to Embodiment 5 wherein the topic modeling allows documents to have a plurality of topics.

[0056] According to Wikipedia, "An early topic model was described by Papadimitriou, Raghavan, Tamaki and Vempala in 1998. [1] Another one, called Probabilistic latent semantic indexing (PLSI), was created by Thomas Hofmann in 1999. [2] Latent Dirichlet allocation (LDA), perhaps the most common topic model currently in use, is a generalization of PLSI developed by David Blei, Andrew Ng, and Michael I. Jordan in 2002, allowing documents to have a mixture of topics.[3] Other topic models are generally extensions on LDA, such as Pachinko allocation, which improves on LDA by modeling correlations between topics in addition to the word correlations which constitute topics."

Embodiment 7

[0057] A browsing system operative in conjunction with a stored representation of a multiplicity of electronic documents and their distribution over a plurality of themes, the system comprising:

[0058] theme-to-document flitting apparatus for retrieving and presenting to a user, documents whose document score for at least one user-selected theme; is high; and

[0059] document-level browsing apparatus for retrieving and presenting to a user, documents whose distributions over the plurality of themes are similar to the distribution of a user-selected document over the plurality of themes.

Embodiment 8

[0060] A system according to Embodiment 7 and also comprising:

[0061] theme-to-word flitting apparatus for retrieving and presenting to a user, words whose word score for at least one user-selected theme; is high;

[0062] word-level browsing apparatus for retrieving and presenting to a user, words whose distributions over the plurality of themes are similar to the distribution of a user-selected word over the plurality of themes,

[0063] thereby to provide 3-tier browsing apparatus facilitating browsing at word, document and topic levels responsive to user-initiated flitting between the levels.

Embodiment 9

[0064] A method according to Embodiment 1 and also comprising:

[0065] facilitating theme-to-word flitting by retrieving and presenting to a user, words whose word score for at least one user-selected theme; is high.

Embodiment 10

[0066] A method according to Embodiment 1 and also comprising:

[0067] facilitating theme-to-document flitting for retrieving and presenting to a user, documents whose document score for at least one user-selected theme is high.

Embodiment 11

[0068] A method according to Embodiment 1 and also comprising:

[0069] facilitating document-level browsing for retrieving and presenting to a user, documents whose distributions over the plurality of themes are similar to the distribution of a user-selected document over the plurality of themes.

Embodiment 12

[0070] A method according to Embodiment 1 and also comprising:

[0071] facilitating word-level browsing for retrieving and presenting to a user, words whose distributions over the plurality of themes are similar to the distribution of a user-selected word over the plurality of themes.

Embodiment 13

[0072] A method according to Embodiment 1 wherein the number of data set members in the large data set is further reduced subsequent to the using step and prior to a manual review process.

Embodiment 14

[0073] A method according to Embodiment 1 wherein the reducing is effected using:

[0074] random selection; and

[0075] at least one computerized data set member pruning technique other than random selection.

Embodiment 15

[0076] A method according to Embodiment 14 wherein the random selection is performed after the computerized data set member pruning technique.

Embodiment 16

[0077] A method according to Embodiment 14 wherein the random selection is performed before the computerized data set member pruning technique.

Embodiment 17

[0078] A method according to Embodiment 5 wherein the topic modeling which allows documents to have a plurality of topics comprises one of the following computerized techniques: Latent Dirichlet allocation (LDA), PLSI, and Pachinko allocation.

Embodiment 18

[0079] A method according to Embodiment 3 wherein the thinning out at least one document which passes a document similarity criterion comprises replacing a plurality of emails forming an email thread, with a single inclusive email.

Embodiment 19

[0080] A method according to Embodiment 4 wherein the identifying and discarding near-duplicates is effected using Equivio Zoom near-duplicate functionality

[0081] The present invention also typically includes at least the following embodiments:

Embodiment a1

[0082] An e-discovery method comprising:

Step 1. Input: a set of electronic documents

Step 2: Extract text from the data collection.

Step 3: Compute Near-duplicate (ND) on the dataset.

Step 4. Compute Email threads (ET) on the dataset.

Step 5. Run a topic modeling on a subset of the dataset, including data manipulation

Embodiment a2

[0083] A method according to Embodiment a1 wherein the output of step 3 includes all documents having the same DuplicateSubsetID having an identical text.

Embodiment a3

[0084] A method according to Embodiment a wherein the output of step 3 includes all documents x in the set for which there is another document y in the set, such that the similarity between the two is greater than some threshold.

Embodiment a4

[0085] A method according to Embodiment a1 wherein the output of step 3 includes at least one pivot document selected by a policy such as maximum words in the document.

Embodiment a5

[0086] A method according to Embodiment a1 wherein the subset includes inclusives of Email threads (ET).

Embodiment a6

[0087] A method according to Embodiment a1 wherein the subset includes Pivots from documents and attachments, but not emails.

Embodiment a7

[0088] A method according to Embodiment a1 wherein the data manipulation includes, if the document is an e-mail, removing all e-mail headers in the document, but keeping the subject line and the body of the e-mail.

Embodiment a8

[0089] A method according to Embodiment a7 and also comprising multiplying the subject line to set some weight to the subject words.

Embodiment a9

[0090] A method according to Embodiment a1 wherein the data manipulation includes Tokenization of the text using separators.

Embodiment a10

[0091] A method according to Embodiment a1 wherein the data manipulation includes ignoring the following features:

[0092] Words with length less than (parameter)

[0093] Words with length greater than (parameter)

[0094] Words that do not start with an alpha character.

[0095] (Optionally)—words that contain digits

[0096] (Optionally)—words that contain non-AlphaNumeric characters, optionally excluding some subset characters such as '_'.

[0097] Words that are stop words.

[0098] Words that appear more than (parameter) times number of words in the document.

[0099] Words that appear less than (parameter) times number of documents.

[0100] Words that appear more than (parameter) times number of documents.

Embodiment a11

[0101] A method according to Embodiment a1 wherein the output of step 5 includes an assignment of documents to the themes, and an assignment of words (features) to themes and each feature x has some probability P_xy of being in theme y and wherein the P matrix is used to construct names for at least one theme.

Embodiment a12

[0102] A method for computerized Early Case Assessment comprising:

a. Select at random a set of documents
b. Run near-duplicates (ND)
c. Run Email threads (ET)
d. Select pivot and inclusive
e. Run topic modeling;
g. Generate theme names; and
h. Explore the data by browsing themes.

Embodiment a13

[0103] A method according to Embodiment a1 wherein, for Post Case Assessment rather than using an entire dataset, only the documents that are relevant to the case are used.

Embodiment a14

[0104] A method according to Embodiment a1 and also comprising displaying for each theme the list of documents that are related to that theme.

Embodiment a15

[0105] A method according to Embodiment a14 wherein the user has an option to select a meta-data and the system will display for each theme the percentage of that meta-data in that theme.

[0106] Also provided, excluding signals, is a computer program comprising computer program code means for performing any of the methods shown and described herein when the program is run on a computer; and a computer program product, comprising a typically non-transitory computer-usable or -readable medium e.g. non-transitory computer-usable or -readable storage medium, typically tangible, having a computer readable program code embodied therein, the computer readable program code adapted to be executed to implement any or all of the methods shown and described herein. It is appreciated that any or all of the computational steps shown and described herein may be computer-implemented. The operations in accordance with the teachings herein may be performed by a computer specially constructed for the desired purposes or by a general purpose computer specially configured for the desired purpose by a computer program stored in a typically non-transitory computer readable storage medium. The term "non-transitory" is used herein to exclude transitory, propagating signals or waves, but to otherwise include any volatile or non-volatile computer memory technology suitable to the application.

[0107] Any suitable processor, display and input means may be used to process, display e.g. on a computer screen or other computer output device, store, and accept information such as information used by or generated by any of the methods and apparatus shown and described herein; the above processor, display and input means including computer programs, in accordance with some or all of the embodiments of the present invention. Any or all functionalities of the invention shown and described herein, such as but not limited to steps of flowcharts, may be performed by a conventional personal computer processor, workstation or other programmable device or computer or electronic computing device or processor, either general-purpose or specifically constructed, used for processing; a computer display screen and/or printer and/or speaker for displaying; machine-readable memory such as optical disks, CDROMs, DVDs, BluRays, magneto-optical discs or other discs; RAMs, ROMs, EPROMs, EEPROMs, magnetic or optical or other cards, for storing, and keyboard or mouse for accepting. The term "process" as used above is intended to include any type of computation or manipulation or transformation of data represented as physical, e.g. electronic, phenomena which may occur or reside e.g. within registers and/or memories of a computer or processor. The term processor includes a single processing unit or a plurality of distributed or remote such units.

[0108] The above devices may communicate via any conventional wired or wireless digital communication means, e.g. via a wired or cellular telephone network or a computer network such as the Internet.

[0109] The apparatus of the present invention may include, according to certain embodiments of the invention, machine readable memory containing or otherwise storing a program of instructions which, when executed by the machine, implements some or all of the apparatus, methods, features and functionalities of the invention shown and described herein. Alternatively or in addition, the apparatus of the present invention may include, according to certain embodiments of the invention, a program as above which may be written in any conventional programming language, and optionally a machine for executing the program such as but not limited to a general purpose computer which may optionally be configured or activated in accordance with the teachings of the present invention. Any of the teachings incorporated herein may wherever suitable operate on signals representative of physical objects or substances.

[0110] The embodiments referred to above, and other embodiments, are described in detail in the next section.

[0111] Any trademark occurring in the text or drawings is the property of its owner and occurs herein merely to explain or illustrate one example of how an embodiment of the invention may be implemented.

[0112] Unless specifically stated otherwise, as apparent from the following discussions, it is appreciated that throughout the specification discussions, utilizing terms such as, "processing", "computing", "estimating", "selecting", "ranking", "grading", "calculating", "determining", "generating", "reassessing", "classifying", "generating", "producing", "stereo-matching", "registering", "detecting", "associating", "superimposing", "obtaining" or the like, refer to the action and/or processes of a computer or computing system, or processor or similar electronic computing device, that manipulate and/or transform data represented as physical, such as electronic, quantities within the computing system's registers and/or memories, into other data similarly represented as physical quantities within the computing system's memories, registers or other such information storage, transmission or display devices. The term "computer" should be broadly construed to cover any kind of electronic device with data processing capabilities, including, by way of non-limiting example, personal computers, servers, computing system, communication devices, processors (e.g. digital signal processor (DSP), microcontrollers, field programmable gate array (FPGA), application specific integrated circuit (ASIC), etc.) and other electronic computing devices.

[0113] The present invention may be described, merely for clarity, in terms of terminology specific to particular programming languages, operating systems, browsers, system versions, individual products, and the like. It will be appreciated that this terminology is intended to convey general prin-

ciples of operation clearly and briefly, by way of example, and is not intended to limit the scope of the invention to any particular programming language, operating system, browser, system version, or individual product.

[0114] Elements separately listed herein need not be distinct components and alternatively may be the same structure.

[0115] Any suitable input device, such as but not limited to a sensor, may be used to generate or otherwise provide information received by the apparatus and methods shown and described herein. Any suitable output device or display may be used to display or output information generated by the apparatus and methods shown and described herein. Any suitable processor may be employed to compute or generate information as described herein e.g. by providing one or more modules in the processor to perform functionalities described herein. Any suitable computerized data storage e.g. computer memory may be used to store information received by or generated by the systems shown and described herein. Functionalities shown and described herein may be divided between a server computer and a plurality of client computers. These or any other computerized components shown and described herein may communicate between themselves via a suitable computer network.

BRIEF DESCRIPTION OF THE DRAWINGS

[0116] Certain embodiments of the present invention are illustrated in the following drawings:

[0117] FIG. **1** is a simplified flowchart illustration of a method for use of themes in e-discovery, according to certain embodiments.

[0118] FIG. **2** is a simplified flowchart illustration of a method for early case assessment, according to certain embodiments.

[0119] FIGS. **3***a*-**3***b*, taken together, is a simplified flowchart illustration of a method for associating topics with documents, according to certain embodiments.

[0120] FIG. **4** is a simplified flowchart illustration of a "navigating" or browsing method for generating suitable displays to facilitate computer-aided theme exploration, suitable e.g. for implementing step **100** in FIGS. **3***a*-**3***b*, taken together, according to certain embodiments.

[0121] FIG. **5** is a simplified screenshot illustration of an example display screen generated by a system constructed and operative in accordance with certain embodiments. The screen display facilitates theme-level browsing, according to certain embodiments.

[0122] FIG. **6** is a simplified screenshot illustration of an example display screen generated by a system constructed and operative in accordance with certain embodiments. As shown, flitting from document-level to theme-level or word-level is facilitated.

[0123] FIG. **7** is a simplified screenshot illustration of an example display screen generated by a system constructed and operative in accordance with certain embodiments. As shown, document-level browsing is facilitated.

[0124] FIG. **8** is a simplified flowchart illustration, according to certain embodiments, of a method for utilizing computerized themes functionality under these circumstances.

[0125] The methods of the flowchart figures each include some or all of the illustrated steps, suitably ordered e.g. as shown.

[0126] Computational components described and illustrated herein can be implemented in various forms, for example, as hardware circuits such as but not limited to custom VLSI circuits or gate arrays or programmable hardware devices such as but not limited to FPGAs, or as software program code stored on at least one tangible or intangible computer readable medium and executable by at least one processor, or any suitable combination thereof. A specific functional component may be formed by one particular sequence of software code, or by a plurality of such, which collectively act or behave or act as described herein with reference to the functional component in question. For example, the component may be distributed over several code sequences such as but not limited to objects, procedures, functions, routines and programs and may originate from several computer files which typically operate synergistically.

[0127] Data can be stored on one or more tangible or intangible computer readable media stored at one or more different locations, different network nodes or different storage devices at a single node or location.

[0128] It is appreciated that any computer data storage technology, including any type of storage or memory and any type of computer components and recording media that retain digital data used for computing for an interval of time, and any type of information retention technology, may be used to store the various data provided and employed herein. Suitable computer data storage or information retention apparatus may include apparatus which is primary, secondary, tertiary or off-line; which is of any type or level or amount or category of volatility, differentiation, mutability, accessibility, addressability, capacity, performance and energy use; and which is based on any suitable technologies such as semiconductor, magnetic, optical, paper and others.

DETAILED DESCRIPTION OF CERTAIN EMBODIMENTS

[0129] FIGS. **3***a*-**3***b*, taken together, is a simplified flowchart illustration of a method for associating topics with documents, according to certain embodiments. The method of FIGS. **3***a*-**3***b* typically include some or all of the following steps, suitably ordered e.g. as shown:

[0130] **10**: Provide a collection of thousands or millions of electronic documents (D) e.g. including a mixture of 1 or more of:

[0131] non-emails

[0132] e-mails with attachments

[0133] e-mails without attachments

[0134] **20** Run Near-duplicate Identifying functionality on the collection, thereby to identify all sets of near-duplicates in the collection

[0135] **30** Run Email thread Identifying functionality on all emails in the collection thereby to identify all email threads in the collection

[0136] **40** perform one, some or all of the following steps to pare down the collection of documents (D), thereby to yield a pared-down collection (Z):

[0137] **40***a*. Select one (say) document (e.g. pivot document) to represent each set of near-duplicate set—thereby to yield a set X1 of documents.

[0138] **40***b*. Select (only) inclusive to represent from each email thread thereby to yield a set X2 of inclusive emails

[0139] **40***c*. From the set X2 of inclusive emails select one (say) document from each near-duplicate set thereby to yield a set X3 e.g. first select all inclusives then take only one

inclusive from each set of "similar" inclusives (e.g. sets defined as near-duplicates by Equivio Zoom near-duplicate functionality)

[0140] **50**: if number of documents in Z exceeds a threshold, use random selection to reduce the number of documents in Z to below the threshold

[0141] **60**: select a suitable number, N, of themes to be identified

[0142] **70**: perform topic modeling using documents in Z, thereby to yield N themes

[0143] **80**: Apply the topic model generated in step **70**, to dataset D, thereby to yield topics wherein documents may belong to more than one topic; use topics as themes

[0144] **90**: Assign names to the themes. Each word in the set of all words in all documents has some probability to be in a theme; this probability may comprise the "word score". Typically, the M (predetermined integer e.g. 5) top scoring words are selected to represent the theme i.e. to constitute the theme's name. According to certain embodiments, a name may comprise one or more of the words most frequently found in the documents pertaining to the topic and less frequently or infrequently found in documents not pertaining to the topic.

[0145] **100**: Generate displays (e.g. as per FIG. **4**) to facilitate computer-aided exploration of (browsing between) themes and the documents and/or words they include, where themes are represented in the displays by the theme names selected in step **90**.

[0146] Step **40**A may be performed only on non-emails or may be performed on all documents e-mails and non-emails (e.g. e-mails are considered documents).

[0147] Step **40***b* is typically performed on e-mail bodies i.e. without their attachments.

[0148] Step **40***c* is typically performed on e-mails without attachments. After identifying inclusives, near duplicate is applied to these and typically, just one or just a few e-mail/s from each group of text-similar e-mails is/are selected. For example: if an email thread has several inclusives, only one of them might be selected.

[0149] Typically, random step **50** is performed after near-duplicate and inclusive steps **20**, **30** and **40**, to enable a user to ascertain that further random pruning is necessary since it is possible that steps **20**, **30**, **40** reduce the size of the data set sufficiently without requiring any random pruning. However, alternatively or in addition, random pruning may occur before steps **20**, **30**, and **40**.

[0150] Typically, random step **50** is performed only when it is desired to reduce processing time whereas for a small set of documents, e.g. less than 400 thousand documents, step **50** may be omitted. Optionally, the system computes cost (monetary or in terms of time) of topic modeling both with random selection and without. The system may for example compute the time or cost to compute a topic model on a random sample which is, say, 50%/10%/1% the size of the original data set.

[0151] FIG. **4** is a "Navigating" or browsing method for generating suitable displays to facilitate computer-aided theme exploration, suitable e.g. for implementing step **100** in FIGS. **3***a*-**3***b*, taken together. Alternatively, the method of FIG. **4** may be employed to facilitate computer-aided exploration of any set of themes, which need not have been generated using any or all of steps **10**-**90** in FIGS. **3***a*-**3***b*, taken together. The method of FIG. **4** may include some or all of the following steps, suitably ordered e.g. as shown:

[0152] **410**: Receive e.g. from user, a theme attribute by which to sort themes, e.g.

[0153] Number of documents in theme

[0154] Document Score-related attribute e.g. theme's average or median or mode document score

[0155] How many times has theme been accessed in the past, using stored history of user/group of users

[0156] Theme name (can be sorted in alphabetical order)

[0157] % (richness) or absolute number of documents belonging to theme which match a predicate (e.g. are relevant to a predicate, e.g. using Equivio relevance software tool). A predicate is a logical combination of conditions that the documents must satisfy. Examples of conditions: specific document-types, specific languages, above/below a relevance score generated e.g. by Equivio Zoom's relevance functionality A predicate may be user-selected e.g. via a suitable GUI.

[0158] **420**: Sort themes by a default or user-selected (in step **410**) theme attribute and display themes in order determined by sort process OR display only themes which match a criterion (example criteria: more than 85% of documents in theme are relevant to user-selected predicate, theme name includes "Kennedy", theme includes more than 1000 documents).

[0159] **430**: display theme attribute, in association with displayed theme e.g. how many documents belonging to theme match a predicate (e.g. are relevant to a predicate, e.g. using Equivio relevance software tool).

[0160] **440**: responsive to a user's selection of (e.g. clicking on a displayed) theme, identify themes which are similar to the user-selected theme by identifying themes which have many (number>threshold) documents in common with the user selected theme).

[0161] **450**: responsive to a user's selection of (e.g. clicking on a displayed) theme,

[0162] sort the documents in the theme by a default or user-selected document attribute. Document attribute may include metadata (Custodian, date) or theme related data (e.g. relevance of document to selected predicate, e.g. using Equivio relevance software tool) and display documents in themes in order determined by sort process.

[0163] **460**: responsive to a user's selection of (e.g. clicking on a displayed) document,

[0164] Select and display files whose distributions, e.g. rank distributions, over topics are similar to the selected document's distribution e.g. rank distribution over topics. For example, take the vector of scores of the selected document over all themes e.g., for 5 themes, (0.4, 0.01, 0.7, 0, 0); then display all documents whose distance from the above is less than a constant. Any suitable distance metric or function may be employed such as but not limited to Euclidean distance, L-infinity distance (max entry distance), L−1 distance, and Manhattan distance.

[0165] **470**: responsive to a user's selection of a document attribute (e.g. metadata (Custodian, date)), compute distribution and display (e.g. as histogram): number (or %) of documents under (say) custodian C or date D belonging to each theme.

[0166] The Themes functionality herein is particularly useful for identifying relevant documents in a large collection of electronic documents which is sparse in that only a small number of documents are relevant to a particular issue. This is

especially the case if it is not possible to identify keywords which can be used to tag relevant documents on the basis of a simple keyword search.

[0167] Computerized systems for identifying relevant documents in a large collection of electronic documents exist, such as Equivio Zoom's Relevance functionality.

[0168] However, for a sparse document set, it is sometimes necessary to seed the initial training with pre-identified relevant documents, rather than randomly selecting a training set which might include a tiny or zero amount of relevant documents. For example, the current Equivio Zoom user guide describes (in section 6.3, from page 58 onward) a process of Adding Seed Files to an Issue.

[0169] FIG. **5** is a simplified screenshot illustration of an example display screen generated by a system constructed and operative in accordance with certain embodiments. As shown, each theme is presented together with a bar (right side of screen) indicating relevance to a user-selected issue. As shown, order of presentation of the screens is in accordance with length of the bar. The screen display facilitates theme-level browsing, according to certain embodiments. For example, the bars may indicate the number of files per theme, that are Relevant (e.g. as determined manually, or automatically e.g. by Equivio Zoom's Relevance functionality) to a user-selected issue which may if desired be shown on, and selected via, a suitable GUI (not shown). The screen display facilitates theme-level browsing, according to certain embodiments.

[0170] FIG. **6** is a simplified screenshot illustration of an example display screen generated by a system constructed and operative in accordance with certain embodiments. As shown, documents are presented along with (on the left) words in the documents and their word scores, relative to an individual theme, as well as themes related to the individual theme. The words and themes may be presented in descending order of their word scores and relatedness to the individual theme, respectively. If a related theme or word is selected (e.g. clicked upon), a different "semantic view" is generated; for example, of all documents in the selected related theme, using a screen format which may be similar to that of FIG. **6**. As shown, flitting from document-level to theme-level or word-level is facilitated.

[0171] FIG. **7** is a simplified screenshot illustration of an example display screen generated by a system constructed and operative in accordance with certain embodiments. As shown, document-level browsing is facilitated.

[0172] FIG. **8** is a simplified flowchart illustration, according to certain embodiments, of a method for utilizing computerized Themes functionality under these circumstances. The method of FIG. **8** typically includes some or all of the following steps, suitably ordered e.g. as shown:

[0173] **1010**. use computerized Themes functionality to identify an initial "seed" set of (say 5-30) relevant documents in a large sparse collection of electronic documents.

[0174] **1020**. generate a training set of documents including the initial "seed" set of relevant documents and at least an equal number of documents randomly selected from the large sparse collection of electronic documents.

[0175] **1030**. operate computerized relevant document identification system, e.g. Equivio Zoom's Relevance functionality on the training set, thereby to successfully identify the rare relevant documents in the large sparse collection.

[0176] It is appreciated that step **1010** may be performed in any suitable manner. For example, if at least one relevant document is known, step **1010** may comprise:

[0177] a. running the "themes" functionality to obtain a "thematic distribution" for the relevant document e.g. an indication of the significance of each of the various themes to the document. Some topic modeling software provides "document scores" for each document relative to each theme, indicating significance of each of the various themes to each document. Alternatively, if the top key words on the key word list of a theme occur relatively frequently in some documents and relatively infrequently in others, the theme can be regarded as highly significant to the former documents and less significant to the latter documents.

[0178] b. selecting documents within the large collection of electronic documents whose "thematic distribution" is similar, using a suitable metric, to the "thematic distribution" of the document known to be relevant. A suitable metric for similarity between Document 1's "thematic distribution" and Document 2's "thematic distribution" may for example be a Euclidean distance (sum of squares-based e.g.) between the document scores of Document 1, summed over all themes, and the document scores of Document 2, summed over all themes. Other distance metrics may also be employed e.g. L-infinity distance (max entry distance), L−1 distance, and Manhattan distance.

[0179] It is appreciated that computerized processing tends to generate clusters (and topics) that are artifactual. For example—presence of the word "weekend" might trigger definition of a cluster of documents which, upon inspection, would be found to include a mass of emails about an unrelated variety of subjects united only by the fact that the emails were written on a Friday hence include an exhortation to "have a nice weekend". In multi-topic processing (e.g. topic modeling in which one document can be assigned to several topics), this is of less relevance: of the many topics found, some are safely ignored as artifactual and the system as a whole remains workable. In clustering (in which each document can belong to only one topic) however, important documents can be assigned to an artifactual cluster and thereby effectively disappear since disregarding the artifactual cluster tends to lead to disregarding documents assigned thereto.

[0180] It is appreciated that the systems and methods shown and described herein enable a S-tier browsing system to be generated, in which a user can browse at the word, document/file or topic level, and can move from one level to another. For example, a user may look at a presentation of topics, arranged say by relevance to an issue, and the system may present to her or him, words or documents whose score for the theme/s the user has selected, are high. The system may for example compute word scores or document scores for all words or documents, sort the words or documents, and present to the user only those whose word or document scores is high. The user may then select one of those words or documents, thereby browsing to a different level. When s/he does select, say, a document scoring high for the topic s/he previously was viewing, the system then shows the document, and also identifies and displays indications of themes to which the document is strongly related, and words whose document scores are high for the themes to which the document is strongly related. The system may do this by computing the degree of relatedness of the document to all themes (each of the themes), sorting the themes on this basis, and presenting to the user only those themes for which the docu-

ment's degree of relatedness is high. Again the user can change level, from the document level up to the topic level or down to the word level, or the user may continue to browse at the document level, e.g. to documents whose distribution over the identified themes is similar (using a suitable distance metric) to the distribution over the identified themes of the document of previous interest. To support this, the system may compute all documents' distributions over all themes identified, and may also compute the distances between these distributions, either in advance for all document pairs, or in real time for a user-designated document. The system may then present, responsive to a user request for documents similar to document D, the top few documents from a list of documents sorted in accordance with the documents' respective distances from Document D. Alternatively, a user may perform "word-level" browsing by moving from one word to another word which has a similar distribution over N identified topics. To support this, the system may compute all words' distributions over all themes identified, and may also compute the distances between these distributions, either in advance for all word pairs, or in real time for a user-designated word. The system may then present, responsive to a user request for words similar to an individual word W of interest, the top few words from a list of words sorted in accordance with the words' respective distances from Word W.

[0181] Another embodiment of the invention, e.g. as described above with reference to FIG. **4**, is a browsing system operative in conjunction with a stored representation of a multiplicity of electronic documents and their distribution over a plurality of themes, the system comprising some or all of the following:

[0182] theme-to-word flitting apparatus for retrieving and presenting to a user, words whose word score for at least one user-selected theme; is high;

[0183] theme-to-document flitting apparatus for retrieving and presenting to a user, documents whose document score for at least one user-selected theme; is high;

[0184] document-level browsing apparatus for retrieving and presenting to a user, documents whose distributions over the plurality of themes are similar to the distribution of a user-selected document over the plurality of themes

[0185] word-level browsing apparatus for retrieving and presenting to a user, words whose distributions over the plurality of themes are similar to the distribution of a user-selected word over the plurality of themes,

[0186] thereby to provide 2- or 3-tier browsing apparatus facilitating browsing at word, document and topic levels responsive to user-initiated flitting between the levels.

[0187] It is appreciated that any suitable parameters and work-processes may be employed. For example, a set of electronic documents comprising thousands, tens or hundreds of thousands, or millions of electronic documents may be processed as described herein.

[0188] Typically, the number of themes to identify is selected by a user and any suitable number of themes may be requested by the user such as 10, 20, 50, 100, 200 or 500 themes. For example, the number of themes selected may be, perhaps, 200 themes for a collection of a few hundred thousand electronic documents, and proportionally more or less themes if the number of documents in the collection is proportionally larger or smaller.

Any suitable "view" of themes may be provided, such as themes sorted by number of files or meta-data attributes of the files, themes sorted by various attributes of the words in the theme name, themes sorted by relevance to an issue and so forth.

[0189] A particular advantage of certain embodiments is that documents which are known to be mutually similar or near duplicates are "thinned" so that they do not over-influence or skew the topic modeling process.

[0190] It is appreciated that thinning need not result in retaining only a single pivot or only a single inclusive email, instead one may, if appropriate, reduce the influence of repeated or highly related materials without eliminating the repetition entirely.

[0191] Regarding topic-modeling steps herein e.g. step v of FIG. **2**, step **3** of FIG. **2**, step **70** of FIG. **3**:

[0192] A topic model is a computational functionality analyzing a set of documents and yielding "topics" that occur in the set of documents typically including (a) what the topics are and (b) what each document's balance of topics is. According to Wikipedia, "Intuitively, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently: "dog" and "bone" will appear more often in documents about dogs, "cat" and "meow" will appear in documents about cats, and "the" and "is" will appear equally in both. A document typically concerns multiple topics in different proportions". Topic models may analyze large volumes of unlabeled text and each "topic" may consist of a cluster of words that occur together frequently.

[0193] Another definition, from the following http location:

[0194] faculty.washington.edu/jwilker/559/SteyversGriffiths.pdf, is that topic modeling functionality proceeds from an assumption "that documents are mixtures of topics, where a topic is a probability distribution over words. A topic model is a generative model for documents: it specifies a simple probabilistic procedure by which documents can be generated. To make a new document, one chooses a distribution over topics. Then, for each word in that document, one chooses a topic at random according to this distribution, and draws a word from that topic. Standard statistical techniques can be used to invert this process, inferring the set of topics that were responsible for generating a collection of documents."

Topic modeling as used herein includes any or all of the above, as well as any computerized functionality which inputs text/s and uses a processor to generate and output a list of semantic topics which the text/s are assumed to pertain to, wherein each "topic" comprises a list of keywords assumed to represent a semantic concept.

[0195] Topic modeling includes but is not limited to any and all of: the Topic modeling functionality described by Papadimitriou, Raghavan, Tamaki and Vempala in 1998; Probabilistic latent semantic indexing (PLSI), created by Thomas Hofmann in 1999; Latent Dirichlet allocation (LDA), developed by David Blei, Andrew Ng, and Michael I. Jordan in 2002 and allowing documents to have a mixture of topics; extensions on LDA, such as but not limited to Pachinko allocation;

[0196] Griffiths & Steyvers Topic modeling e.g. as published in 2002, 2003, 2004; Hofmann Topic modeling e.g. as published in 1999, 2001; topic modeling using the synchronic approach; topic modeling using the diachronic approach; Topic modeling functionality which attempts to fit appropriate model parameters to the data corpus using heuristic/s for

maximum likelihood fit, topic modeling functionality with provable guarantees; topic modeling functionality which uses singular value decomposition (SVD), topic modeling functionality which uses the method of moments, topic modeling functionality which uses an algorithm based upon non-negative matrix factorization (NMF); and topic modeling functionality which allows correlations among topics. Topic modeling implementations may for example employ Mallet (software project), Stanford Topic Modeling Toolkit, or GenSim—Topic Modeling for Humans.

[0197] Earlier presented embodiments are now described for use either independently or in suitable combination with the embodiments described above:

[0198] When enhancing expert-based computerized analysis of a set of digital documents, a system for computerized derivation of leads from a huge body of data may be provided, the system comprising:

[0199] an electronic repository including a multiplicity of accesses to a respective multiplicity of electronic documents and metadata including metadata parameters having metadata values characterizing each of the multiplicity of electronic documents;

[0200] a relevance rater using a processor to run a first computer algorithm on the multiplicity of electronic documents which yields a relevance score which rates relevance of each of the multiplicity of electronic documents to an issue; and

[0201] a metadata-based relevant-irrelevant document discriminator using a processor to rapidly run a second computer algorithm on at least some of the metadata which yields leads, each lead comprising at least one metadata value for at least one metadata parameter, which value correlates with relevance of the electronic documents to the issue.

[0202] The application is operative to find outliers of a given metadata and relevancy score (i.e. relevant, not relevant). When theme-exploring is used, the system can identify themes with high relevancy score based on the given application. The above system, without theme-exploring, may compute the outlier for a given metadata, and each document appears one in each metadata. In the theme-exploring settings for a given set of themes the same document might fall into several of the metadata.

[0203] Method for Use of Themes in e-Discovery (FIG. **1**): step i. Input: a set of electronic documents. The documents could be in:

Text format, Native files (PDF, Word, PPT, etc.), ZIP files, PST, Lotus notes, MSG, etc.

Step ii Extract text from the data collection. Text extraction can be done by third party software such as: Oracle inside out, iSys, DTSearch, iFilter, etc.

Step iii: Compute Near-duplicate (ND) on the dataset.

The following teachings may be used: U.S. Pat. No. 8,015,124, entitled "A Method for Determining Near Duplicate Data Objects"; and/or WO 2007/086059, entitled "Determining Near Duplicate "Noisy" Data Objects", and/or suitable functionalities in commercially available e-discovery systems such as those of Equivio.

[0204] For each document compute the following:

Step iiia: DuplicateSubsetID: all documents having the same DuplicateSubsetID having an identical text.

Step iiib: EquiSetID: all documents having the same EquiSetID are similar (for each document x in the set there is another document y in the set, such that the similarity between the two is greater than some threshold).

Step iiic: Pivot: 1 if the document is a representative of the set (and 0 otherwise). Typically, for each EquiSet only one document is selected as Pivot. The pivot document can be selected by a policy for example (maximum words number of words, minimum number of words, median number of words, minimum docid, etc.) When using theme networking (TN) it is recommended to use maximum words in documents as pivot policy as it is desirable for largest documents to be in the model.

Step iv. Compute Email threads (ET) on the dataset. The following teachings may be used: WO 2009/004324, entitled "A Method for Organizing Large Numbers of Documents" and/or suitable functionalities in commercially available e-discovery systems such as those of Equivio.

The output of this phase is a collection of trees, and all leafs of the trees are marked as inclusive. Note, that family information is accepted (to group e-mails with their attachments).

Step v. Run a topic modeling algorithm (such as LDA) on a subset of the dataset, including feature extraction. Resulting topics are defined as themes. The subset includes the following documents:

[0205] Inclusive from Email threads (ET)

[0206] Pivots from all documents that are not e-mails. i.e. pivots from documents and attachments.

[0207] The data collection include less files (usually the size is 50% of the total size); and the data do not include similar documents, therefore if a document appears many times in the original data collection it will have the same weight as if it appears once.

[0208] In building the model documents were used with more than 25 (parameter) words and less than 20,000 words. The idea behind this limitation was to improve performance, and not be influenced by high words frequency when the document has few features.

[0209] If the dataset is extremely large, at most 100,000 (parameter) documents may be selected at random to build the model, and after building the model, it may be applied on all other documents.

[0210] The first step in the topic modeling algorithm is to extract features from each document.

[0211] A method suitable for the Feature extraction of step v may include obtaining features as follows:

[0212] A topic modeling algorithm uses features to create the model for the topic-modeling step v above. The features are words; to generate a list of words from each word one may do the following:

If the document is an e-mail, remove all e-mail headers in the document, but keep the subject line and the body. One may multiply the subject line to set some weight to the subject words. Tokenize the text using separators such as, spaces, semicolon, colon, tabs, new line etc. Ignore the following features:

Words with length less than 3 (parameter)

Words with length greater than 20 (parameter)

Words that do not start with an alpha character.

Words that are stop words.

Words that appear more than 0.2 times number of words in the document. (parameter)

Words that appear in less than 0.01 times number of documents. (Parameter)

Words that appear in more than 0.2 times number of documents. (Parameter)

Stemming, part-of-speech—as features.

Step viii. Theme names. The output of step v includes an assignment of documents to the themes, and an assignment of words (features) to themes. Each feature x has some probability P_xy of being in theme y. Using the P matrix, construct names to the themes.

[0213] In e-discovery one may use the following scenarios: Early Case Assessment, Post Case Assessment and provision of helpful User Interfaces.

Early Case Assessment (FIG. **2**, Including Some or all of the Following Steps a-h):

a. Select at random 100000 documents

b. Run near-duplicates (ND)

c. Run Email threads (ET)

d. Select pivot and inclusive

e. Run topic modeling using the above feature selection. The input of the topic modeling is a set of documents. The first phase of the topic modeling is to construct a set of features for each document. The feature getting method described above may be used to construct the set of features.

f. Run the model on all other documents (optional).

g. Generate theme names e.g. using step viii above.

h. Explore the data by browsing themes; one may open a list of documents belonging to a certain theme, from the document one may see all themes connected to that document, and go to other themes.

The list of documents might be filtered by a condition set by the user. For example filter all documents by dates, relevancy, file size, etc.

The above procedure assists users in early case assessment when the data is known and one would like to know what is in the data, and assess the contents of the data collection.

In early case assessment one may randomly sample the dataset to get results faster.

[0214] Post Case Assessment This process uses some or all of steps I-v above, but in this setting an entire dataset is not used, but rather, only the documents that are relevant to the case. If near-duplicates (ND) and Email threads (ET) have already run, there is no need to re-run them.

[0215] 1^{st} pass review is a quick review of the documents that can be handled manually or by an automatic predictive coding software; the user wishes to review the results and get an idea on the themes of the documents that passed that review. This phase is essential because the number of such documents might be extremely large. Also, there are cases in which, in some sub-issues, there are only a few documents.

[0216] The above building block can generate a procedure for such cases. Here, g only documents that passed the 1^{St} review phase are taken, and themes are calculated for them.

[0217] User Interface using the output of steps I-v and displaying results thereof. Upon running the topic modeling each resulting topic is defined as a theme, and for each theme the list of documents is displayed that are related to that theme. The user has an option to select a meta-data (for example is the document relevant to an issue, custodian, date-range, file type, etc.) and the system will display for each theme the percentage of meta-data in that theme. Such presentation would assist the user while evaluating the theme.

[0218] An LDA model might have themes that can be classified as CAT_related and DOG_related. A theme has probabilities of generating various words, such as milk, meow, and kitten, which can be classified and interpreted by the viewer as "CAT_related". The word cat itself will have high probability given this theme. The DOG_related theme likewise has probabilities of generating each word: puppy, bark, and bone might have high probability. Words without special relevance, such as the (see function word), will have roughly even probability between classes (or can be placed into a separate category). A theme is not strongly defined, neither semantically nor epistemologically. It is identified on the basis of supervised labeling and (manual) pruning on the basis of their likelihood of co-occurrence. A lexical word may occur in several themes with a different probability, however, with a different typical set of neighboring words in each theme.

Each document is assumed to be characterized by a particular set of themes. This is akin to the standard bag of words model assumption, and makes the individual words exchangeable.

[0219] Processing a large data set requires time and space, in the context of the current invention N documents are selected to create the model, and then the model is applied on the remaining documents.

When selecting the documents to build the model, a few options may be possible:

[0220] O1. Take all documents.

[0221] O2. Take one documents for each set of exact duplicate documents

[0222] O3. Take one documents from each EquiSet (e.g. as per U.S. Pat. No. 8,015,124, entitled "A Method for Determining Near Duplicate Data Objects"; and/or WO 2007/086059, entitled "Determining Near Duplicate "Noisy" Data Objects").

[0223] O4. Take the inclusive from the data collection. Another option is to randomly sample X documents from the collection, as described above.

Steps 02, 03, 04 aim to create themes that are known to the user, and also not to weight documents that already appear in a known set.

The input for the algorithm is a text documents that can be parsed to a bag-of-words. When processing an e-mail, one may notice that the e-mail contains a header (From, to, CC, Subject); and a body. The body of an e-mail can be a formed by a series of e-mails.

For example:

From: A
To: B
Subject: CCCCC
Body1 Body1
From: B
To: A
Subject: CCCCC
Body2 Body2

While processing e-mails for topic modeling one can consider removing all e-mail headers within the body, and by setting a weight to the subject by using a multiple subject line. In the above example the processed text would be:

CCCCC
CCCCC
CCCCC
Body1 Body1
Body2 Body2

Step viii (Theme names) is now described in detail:

Let P(w_i,t_j) the probability that the feature w_i belongs to theme t_j. In known implementations the theme name is a list

of words with the highest probability. The solution is good when the dataset is sparse, i.e. the vocabulary of the themes is different from each other. In e-discovery the issues are highly connected and therefore, there are cases when the "top" words appeared in two or more themes. In settings of the problem "stable marriage" was used as in an algorithm, to pair words to themes. The algorithm may include:

Order the theme by some criteria (Size, Quality, #of relevant documents, etc.); i.e. theme_3 is better than theme_4.

(1) Create an empty set S
(2) Sort themes by some criteria
(3) For j=0 ; j < maximum words in theme name; j++
(4) For I = 0 ; I < #number of themes; i++) do
(5) For theme_i, assign the word with the highest score that is not in S, and add that word to S

[0224] After X words are assigned for each theme, the number of words can be reduced by, for example, taking only those words in each theme that are bigger than the maximum word rank in that theme, divided by some constant.

[0225] Typically, electronic documents do not bear, or do not need to bear, any pre-annotation or labeling or meta-data, or if they do, such is not employed by the topic modeling which instead is derived by analyzing the actual texts.

[0226] A particular advantage of certain embodiments of the invention is that collections of electronic documents are hereby analyzed semantically by a processor on a scale that would be impossible manually. Output of topic modeling may include the n most frequent words from the m most frequent topics found in an individual document.

[0227] It is appreciated that when presenting documents, it need not be the case that all documents whose document score for at least one user-selected theme; is high in a defined sense e.g. over a certain threshold are displayed. Similarly, it need not be the case that all documents whose distributions over the plurality of themes are similar in a defined sense to the distribution of a user-selected document over the plurality of themes, are described. Instead, only a subset of the documents may be displayed, e.g. only such documents as answer at least one individual criterion. So, for example, a search engine could be used on the data collection, and then results of the search query might be presented using the embodiments shown and described herein. Alternatively or in addition, a predicate may be used as a criterion e.g. presenting only documents in English or only documents relevant to a given issue.

[0228] The methods shown and described herein are particularly useful in processing or analyzing or sorting or searching bodies of knowledge including hundreds, thousands, tens of thousands, or hundreds of thousands of electronic documents or other computerized information repositories, some or many of which are themselves at least tens or hundreds or even thousands of pages long. This is because practically speaking, such large bodies of knowledge can only be processed, analyzed, sorted, or searched using computerized technology.

[0229] It is appreciated that terminology such as "mandatory", "required", "need" and "must" refer to implementation choices made within the context of a particular implementation or application described herewithin for clarity and are not intended to be limiting since in an alternative implantation, the same elements might be defined as not mandatory and not required or might even be eliminated altogether.

[0230] It is appreciated that software components of the present invention including programs and data may, if desired, be implemented in ROM (read only memory) form including CD-ROMs, EPROMs and EEPROMs, or may be stored in any other suitable typically non-transitory computer-readable medium such as but not limited to disks of various kinds, cards of various kinds and RAMs. Components described herein as software may, alternatively, be implemented wholly or partly in hardware and/or firmware, if desired, using conventional techniques, and vice-versa. Each module or component may be centralized in a single location or distributed over several locations.

[0231] Included in the scope of the present invention, inter alia, are electromagnetic signals carrying computer-readable instructions for performing any or all of the steps or operations of any of the methods shown and described herein, in any suitable order including simultaneous performance of suitable groups of steps as appropriate; machine-readable instructions for performing any or all of the steps of any of the methods shown and described herein, in any suitable order; program storage devices readable by machine, tangibly embodying a program of instructions executable by the machine to perform any or all of the steps of any of the methods shown and described herein, in any suitable order; a computer program product comprising a computer useable medium having computer readable program code, such as executable code, having embodied therein, and/or including computer readable program code for performing, any or all of the steps of any of the methods shown and described herein, in any suitable order; any technical effects brought about by any or all of the steps of any of the methods shown and described herein, when performed in any suitable order; any suitable apparatus or device or combination of such, programmed to perform, alone or in combination, any or all of the steps of any of the methods shown and described herein, in any suitable order; electronic devices each including a processor and a cooperating input device and/or output device and operative to perform in software any steps shown and described herein; information storage devices or physical records, such as disks or hard drives, causing a computer or other device to be configured so as to carry out any or all of the steps of any of the methods shown and described herein, in any suitable order; a program pre-stored e.g. in memory or on an information network such as the Internet, before or after being downloaded, which embodies any or all of the steps of any of the methods shown and described herein, in any suitable order, and the method of uploading or downloading such, and a system including server/s and/or client/s for using such; a processor configured to perform any combination of the described steps or to execute any combination of the described modules; and hardware which performs any or all of the steps of any of the methods shown and described herein, in any suitable order, either alone or in conjunction with software. Any computer-readable or machine-readable media described herein is intended to include non-transitory computer- or machine-readable media.

[0232] Any computations or other forms of analysis described herein may be performed by a suitable computerized method. Any step described herein may be computer-implemented. The invention shown and described herein may include (a) using a computerized method to identify a solution to any of the problems or for any of the objectives described herein, the solution optionally includes at least one of a decision, an action, a product, a service or any other

information described herein that impacts, in a positive manner, a problem or objectives described herein; and (b) outputting the solution.

[0233] The system may, if desired, be implemented as a web-based system employing software, computers, routers and telecommunication equipment as appropriate.

[0234] Any suitable deployment may be employed to provide functionalities e.g. software functionalities shown and described herein. For example, a server may store certain applications, for download to clients, which are executed at the client side, the server side serving only as a storehouse. Some or all functionalities e.g. software functionalities shown and described herein may be deployed in a cloud environment. Clients e.g. mobile communication devices such as smartphones may be operatively associated with, but external to the cloud.

[0235] The scope of the present invention is not limited to structures and functions specifically described herein and is also intended to include devices which have the capacity to yield a structure, or perform a function, described herein, such that even though users of the device may not use the capacity, they are, if they so desire, able to modify the device to obtain the structure or function.

[0236] Features of the present invention which are described in the context of separate embodiments may also be provided in combination in a single embodiment.

[0237] For example, a system embodiment is intended to include a corresponding process embodiment. Also, each system embodiment is intended to include a server-centered "view" or client centered "view", or "view" from any other node of the system, of the entire functionality of the system, computer-readable medium, apparatus, including only those functionalities performed at that server or client or node.

[0238] Conversely, features of the invention, including method steps, which are described for brevity in the context of a single embodiment or in a certain order may be provided separately or in any suitable subcombination or in a different order. "e.g." is used herein in the sense of a specific example which is not intended to be limiting. Devices, apparatus or systems shown coupled in any of the drawings may in fact be integrated into a single platform in certain embodiments or may be coupled via any appropriate wired or wireless coupling such as but not limited to optical fiber, Ethernet, Wireless LAN, HomePNA, power line communication, cell phone, PDA, Blackberry GPRS, Satellite including GPS, or other mobile delivery. It is appreciated that in the description and drawings shown and described herein, functionalities described or illustrated as systems and sub-units thereof can also be provided as methods and steps therewithin, and functionalities described or illustrated as methods and steps therewithin can also be provided as systems and sub-units thereof. The scale used to illustrate various elements in the drawings is merely exemplary and/or appropriate for clarity of presentation and is not intended to be limiting.

1. A method for computerized identification of themes in a large data set, the system comprising:

reducing the number of data set members in a large data set, using at least one computerized data set member pruning technique other than random selection; and

using a computerized theme identification technique for identifying a plurality of themes in the reduced data set.

2. A method according to claim **1** wherein said computerized data set member pruning technique comprises thinning out at least one document which passes a document similarity criterion relative to at least one other document not being thinned out, thereby to combat skewing as a result of over-influence of similar, hence over-represented, documents upon said theme identification technique.

3. A method according to claim **2** wherein said thinning out at least one document which passes a document similarity criterion comprises replacing a plurality of emails forming an email thread, with at least one inclusive email, thereby to thin out emails which are included in said inclusive email hence are deemed to pass the document similarity criterion with regard to said inclusive.

4. A method according to claim **2** wherein said thinning out at least one document which passes a document similarity criterion comprises identifying and discarding near-duplicates thereby to thin out at least one document which is deemed to pass the document similarity criterion with regard to a set of near-duplicates of said document, at least one of which is not being thinned out.

5. A method according to claim **1** wherein said computerized theme identification technique comprises topic modeling.

6. A method according to claim **5** wherein said topic modeling allows documents to have a plurality of topics.

7. A browsing system operative in conjunction with a stored representation of a multiplicity of electronic documents and their distribution over a plurality of themes, the system comprising:

theme-to-document flitting apparatus for retrieving and presenting to a user, documents whose document score for at least one user-selected theme; is high; and

document-level browsing apparatus for retrieving and presenting to a user, documents whose distributions over the plurality of themes are similar to the distribution of a user-selected document over the plurality of themes.

8. A system according to claim **7** and also comprising:

theme-to-word flitting apparatus for retrieving and presenting to a user, words whose word score for at least one user-selected theme; is high;

word-level browsing apparatus for retrieving and presenting to a user, words whose distributions over the plurality of themes are similar to the distribution of a user-selected word over the plurality of themes,

thereby to provide 3-tier browsing apparatus facilitating browsing at word, document and topic levels responsive to user-initiated flitting between the levels.

9. A method according to claim **1** and also comprising:

facilitating theme-to-word flitting by retrieving and presenting to a user, words whose word score for at least one user-selected theme; is high.

10. A method according to claim **1** and also comprising:

facilitating theme-to-document flitting for retrieving and presenting to a user, documents whose document score for at least one user-selected theme is high.

11. A method according to claim **1** and also comprising:

facilitating document-level browsing for retrieving and presenting to a user, documents whose distributions over the plurality of themes are similar to the distribution of a user-selected document over the plurality of themes.

12. A method according to claim **1** and also comprising:

facilitating word-level browsing for retrieving and presenting to a user, words whose distributions over the plurality of themes are similar to the distribution of a user-selected word over the plurality of themes.

13. A method according to claim **1** wherein the number of data set members in the large data set is further reduced subsequent to said using step and prior to a manual review process.

14. A method according to claim **1** wherein said reducing is effected using:

random selection; and

at least one computerized data set member pruning technique other than random selection.

15. A method according to claim **14** wherein said random selection is performed after said computerized data set member pruning technique.

16. A method according to claim **14** wherein said random selection is performed before said computerized data set member pruning technique.

17. A method according to claim **5** wherein said topic modeling which allows documents to have a plurality of topics comprises one of the following computerized techniques: Latent Dirichlet allocation (LDA), PLSI, and Pachinko allocation.

18. A method according to claim **3** wherein said thinning out at least one document which passes a document similarity criterion comprises replacing a plurality of emails forming an email thread, with a single inclusive email.

19. A method according to claim **4** wherein said identifying and discarding near-duplicates is effected using Equivio Zoom near-duplicate functionality.

20. A method according to claim **9** and wherein said facilitating comprises retrieving and presenting to a user, only those words whose word score for at least one user-selected theme;

is high and which answer to at least one additional criterion.

* * * * *