



- (51) International Patent Classification:
G06F 17/21 (2006.01) *G06F 17/30* (2006.01)
- (21) International Application Number:
PCT/US2011/035666
- (22) International Filing Date:
8 May 2011 (08.05.2011)
- (25) Filing Language: English
- (26) Publication Language: English
- (71) Applicant (*for all designated States except US*): **HEW-LETT-PACKARD DEVELOPMENT COMPANY, L.P.** [US/US]; Hewlett-Packard Development Company, L.P., 11445 Compaq Center Drive West, Houston, Texas 77070 (US).
- (72) Inventor; and
(71) Applicant : **DEOLALIKAR, Vinay** [IN/US]; Hewlett-Packard Co., Mail Stop 1134, 1501 Page Mill Rd., Palo Alto, California 94304 (US).
- (72) Inventor; and
(75) Inventor/Applicant (*for US only*): **LAFFITTE, Hernan** [AR/US]; Hewlett-Packard Co., Mail Stop 1134, 1501 Page Mill Rd., Palo Alto, California 94304 (US).
- (74) Agents: **CHANG, Yeh** et al.; Hewlett-Packard Company, Intellectual Property Administration, 3404 E. Harmony Road, Mail Stop 35, Fort Collins, Colorado 80528 (US).
- (81) Designated States (*unless otherwise indicated, for every kind of national protection available*): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ,

[Continued on next page]

(54) Title: INDICATING DOCUMENTS IN A THREAD REACHING A THRESHOLD

(57) Abstract: Documents in a document thread include descriptive terms that have weights. An indication indicates when documents in the document thread reach a threshold of weight for the document thread.

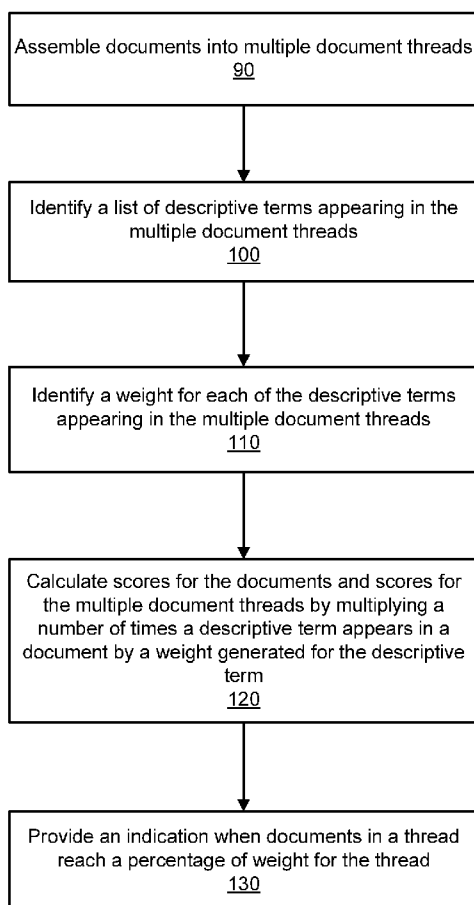


FIG. 1



CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK,

EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

- as to the identity of the inventor (Rule 4.17(i))
- as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))

Published:

- with international search report (Art. 21(3))

INDICATING DOCUMENTS IN A THREAD REACHING A THRESHOLD

BACKGROUND

- 5 A group of documents can include information on specific topics, and a reader may desire to extract this information from the documents. It can be a labor intensive task for the reader to cull through these documents and extract this information if a large number of documents exist. Furthermore, the reader may not know where the desired the information is located in the documents, or how
- 10 many of the documents to read in order to obtain the desired information.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a method for presenting documents according to a score in accordance with an example implementation.

5

Figure 2 is a method for weighting documents according to a score in accordance with an example implementation.

Figure 3 is a display showing email scores and ranks in accordance with an example implementation.

10

Figure 4A is a screenshot of email threads in clusters in accordance with an example implementation.

Figure 4B is a screenshot of a summary of email threads in a single cluster in accordance with an example implementation.

15

Figure 4C is a screenshot of an email thread in accordance with an example implementation.

20

Figure 5 is a computer with a clustering tool that calculates weights and indicates a threshold in document threads in accordance with an example implementation.

DETAILED DESCRIPTION

Example embodiments are apparatus and methods that process a thread of documents in order to remove redundant material, weight the documents
5 according to descriptive terms, and present the documents with an indication when the documents reach a threshold of weight for a thread.

Given a group of documents, example embodiments extract a list of descriptive terms from these documents and provide weights to these terms. The descriptive
10 terms and the weights come from applying a clustering algorithm to the group of documents. The documents are preprocessed to remove redundant or duplicative text, and a score is generated for each of the processed documents. This score is based on the number of descriptive terms in each of the documents and the weights for the descriptive terms. The documents are then ordered by
15 date (for example, a date when the documents were written, transmitted, or saved) and presented to a user and/or saved.

A group of documents can include thousands, hundreds of thousands, or millions of different documents, such as emails, text messages, articles, notes, etc. The
20 number and/or length of these documents may be too great for a reader to efficiently or timely review. Example embodiments remove duplicative text from these documents during preprocessing and indicate when a certain percentage of information within the documents is reached. For example, a notification is displayed when ninety percent (90%) of information in a thread of documents is
25 reached. In this example, a user would not have to read an entirety of the thread, but read a portion of the thread of documents until the notification in order to obtain ninety percent of the information in the thread. Thus, the documents are presented such that a reader can obtain knowledge of the content of the document thread by reading a portion or selection of some of the documents, as
30 opposed to reading all of the documents in the thread to obtain this knowledge.

Figure 1 is a method for presenting documents according to a score in accordance with an example implementation. As used herein, a document is something that conveys information with words. Examples of documents include, but are not limited to, emails, text messages, books, magazines, articles, notes, 5 transcriptions (such as words spoken in a video), and other information containing words (such as words written on a tangible media like paper and/or words stored in an electronic storage medium).

According to block 90, documents are assembled into multiple document 10 threads.

As used herein, a document thread is a series of documents that form a logical discussion or communication. By way of example, text messages in a text message thread form a logical discussion or communication by relating to a topic 15 in the body of the texts, by relating to a sender and/or a recipient of the texts, by relating to a subject or title of the texts, by relating to a time when the texts are sent, and/or by relating to common words or hyperlinks in the body of the texts.

Duplicative or redundant text is also removed from the multiple document threads 20 during preprocessing. This preprocessing can occur before or after the documents are assembled into the multiple document threads.

By way of example, if the document threads are text messages or email messages and include duplicative text, then this duplicative text is removed. 25 Duplicative text can occur when a user responds to an original message and includes a copy of the original message in the response. As another example, information from a first document can be copied and pasted into a second document. This information appearing in the second document is removed as duplicative text since it already appears in the first document.

According to block 100, a list of descriptive terms appearing in the multiple document threads is identified. A user can designate or input the number of descriptive terms. For example, the user can decide to consider ten descriptive terms for the documents in each cluster. These descriptive terms are used when
5 processing the document threads within that cluster. Further, the number of descriptive terms can vary according to user input, such as designating three descriptive terms, four descriptive terms, five descriptive terms, etc. Further yet, the number of descriptive terms can be based on a percentage, such as designating a word as being a descriptive term when the word has a weight of a
10 certain percentage (for example, words with a weight of one percent (1%) or more in a thread are descriptive terms).

According to block 110, a weight is identified for each of the descriptive terms appearing in the multiple document threads. For example, a user specifies a
15 weight for the descriptive terms. Alternatively, weights for descriptive terms are based on word counts, an indexing scheme that identifies a relationship between words and concepts or subjects in a document, and/or a statistical frequency with which the terms appear in the documents, such as a statistical measure using term frequency-inverse document frequency (tf-idf).

20 According to block 120, scores are calculated for the documents and for the multiple document threads based on the number of times a descriptive term appears in a document and the weight identified for the descriptive term. The scores are thus based on the descriptive terms found in block 100 and the
25 weights for these descriptive terms found in block 110.

For example, if a document includes three descriptive terms (term 1 with a weight of X, term 2 with a weight of Y, and term 3 with a weight of Z), then the score for this document equals (X times the number of times term 1 appears in the
30 document) + (Y times the number of times term 2 appears in the document) + (Z times the number of times term 3 appears in the document).

Each document thread can have multiple documents, with each document and each thread having a score. One example method assembles the threads and removes duplicative content that appears in more than one document (e.g., text
5 that is repeated multiple documents in the thread). The threads are clustered together, and scores are assigned to the clustered threads. Scores are also assigned to unique textual content in documents within each of the threads.

According to block 130, an indication is provided when the documents in a thread
10 reach a threshold or percentage of weight for the thread. This indication can be a visual and/or an audible indication. For example, documents are displayed in a thread until the documents in this thread reach ninety percent (90%) of the weight of the thread according to the descriptive terms and their corresponding weights. After the ninety percentile is reached, subsequent documents in the
15 thread are displayed if the user requests it. As another example, after documents in a thread reach a specified percentage of weight of the thread, subsequent documents in the thread are identified, such as being highlighted, removed from being displayed, marked with a symbol or other visual indication, and/or displayed with text indicating to the user that the documents are below a
20 threshold of weight.

By way of example, the first or earliest message in a thread is maintained in its original form (i.e., with no text removed) and displayed on a screen and/or saved. Subsequent messages in the thread are displayed beneath or after the first
25 message and are ordered according to their date. These subsequent messages have redundant textual content removed such that each subsequent message includes unique content. The subsequent messages retain unique content with respect to the other messages. Consider an example in which a user replies to an original email message, and this reply email includes the content of the
30 original email. The content of the original email appearing in the reply is considered redundant since it already appeared in the original email. Content in

the reply email (other than the content of the original email) would be considered unique content since it did not appear in the original email. Another example of redundant text is the inclusion of parts of the original message in the reply message, such as quoting text from an original email in a reply email.

5

Figure 2 is a method for weighting documents according to a score in accordance with an example implementation. For illustration, the method is discussed in connection with emails, but the method is also applicable to other types of documents. For example, this method can be applied to a corpus of email messages coming from email inboxes from a large group of users, such as employees of a company.

10

According to block 200, preprocessing occurs on a group or corpus of emails. During preprocessing, stop words, email headers, signatures, and spurious text are removed from the emails.

15

According to block 202, the group or corpus of emails is assembled into multiple email threads. For example, the emails are assembled according to a subject line of the emails or information present in the email server storing the emails, such as ordering emails according to sender, recipient, geographical location (for example, emails originating from users at a specific building), users in a workgroup, etc.

20

As used herein, an email thread is a series of emails that form a logical discussion or communication. By way of example, emails in an email thread form a logical discussion or communication by relating to a topic in the body of the emails, by relating to a sender and/or a recipient of the emails, by relating to a subject or title of the emails, by relating to a time when the emails are sent, and/or by relating to common words or hyperlinks in the body of the email messages. By way of illustration, two emails are in a thread when they include the same words in the subject line, and they include two common users as

25

30

recipients or senders of the emails. Also, email threads can be assembled by using email header information, or information present in the email server.

According to block 205, redundant or duplicative content is removed from the email threads. For example, the documents are ordered by date, and duplicative text that occurs in later documents is removed. Spurious text (such as headers, signatures, stop words, etc.) is also removed during the preprocessing.

According to block 207, duplicative inboxes are removed from the email threads so each email is included once in the email thread. A single email message can occur in multiple inboxes when the email is sent from a sender to multiple recipients. For example, if a user sends an email to five different recipients, then this email occurs in the inbox of all five recipients. This email is removed from four of the five recipients so the email occurs once in the email thread.

According to block 210, the multiple email threads are grouped into multiple clusters. As used herein, a cluster is a group of related threads.

For example, a clustering tool assembles or clusters the email threads into clusters or groups. Alternatively, the clustering tool obtains or retrieves the clusters and email threads from memory if clustering has already been performed on the threads. The number of email clusters depends on the number of emails threads and other factors that can be input from a user, such as a range of desired clusters, range of threads per cluster, desired performance/speed of the clustering tool, etc. By way of illustration, an email corpus having 150,000 different threads could be grouped into 30 – 100 clusters.

According to block 220, a list of descriptive terms is identified from the email threads for each of the clusters found in block 210. For example, the clustering tool generates labels or keywords from the text corpus of emails on the basis of how useful they were in making decisions about to which cluster a particular

thread belongs. The clustering tool generates the descriptive terms and weights from a corpus of the threads. For example, the clustering tool assigns a weight to each of the terms appearing in the documents. The descriptive terms are intuitively those words or terms of a corpus such that selecting such a term
5 maximizes the increase of similarity within the objects of each cluster. The weight associated with a descriptive term measures how much of an intra-cluster similarity can be attributed to the descriptive term.

The number of descriptive terms can vary depending, for example, on the
10 number of email threads in a cluster, number of words in the emails, and user input. By way of illustration, an email thread can include about 10 – 30 descriptive terms (though this number can increase or decrease based on conditions of the corpus and/or user input).

15 According to block 230, a weight is identified for each descriptive term found in block 220. The weight can be calculated using any one of various methods, such as those discussed in connection with block 110 in Figure 1. Further, descriptive terms with relatively low weights can be dropped (for example, drop a descriptive term when its weight is under 1% of the total weight for the descriptive terms).

20 According to block 240, a weight is calculated for each email message and each email thread based on a number of times the descriptive terms appear in each of the email messages and each of the email threads. One example embodiment (a) counts a number of times each descriptive term in the list appears in the
25 email message, (b) multiplies this number by the weight of the descriptive term, and then (c) sums up the numbers calculated in (b). This sum provides a weight for each email message. The counts obtained from (a) can be capped at a user specified number (for example, cap the number of times a single descriptive term appears in a thread or component message to the number 3, 4, 5, etc.).

30 Next, a fraction of the weight of the thread that is contributed by each individual

message is computed.

The following illustration in tables 1 – 5 provides an example of how the calculations in block 240 are executed.

5

By way of illustration, assume that a cluster of emails discussing storage technology has the following four descriptive terms: storage, SAN (storage area network), server, and disk array. A numerical weight generated for each of these terms is shown in table 1 as follows:

10

Descriptive Term	Weight
storage	30.5
SAN	21
server	14
disk array	8

Table 1

Further, assume that this cluster includes four email threads (email thread 1, email thread 2, email thread 3, and email thread 4). Table 2 shows a count of how many times the descriptive terms appear in each of the email threads.

15

Thread	storage	SAN	server	disk array
email thread 1	2	1	0	1
email thread 2	1	3	0	0
email thread 3	3	2	1	1
email thread 4	1	0	1	3

Table 2

The number of times a descriptive term appears in each email thread is multiplied by the weight for the descriptive term, as shown in table 3.

20

Thread	storage	SAN	server	disk array
email thread 1	$0 \times 30.5 = 0$	$1 \times 21 = 21$	$0 \times 14 = 0$	$1 \times 8 = 8$
email thread 2	$1 \times 30.5 = 30.5$	$3 \times 21 = 63$	$0 \times 14 = 0$	$0 \times 8 = 0$
email thread 3	$3 \times 30.5 = 91.5$	$2 \times 21 = 42$	$1 \times 14 = 14$	$1 \times 8 = 8$
email thread 4	$1 \times 30.5 = 30.5$	$0 \times 21 = 0$	$1 \times 14 = 14$	$3 \times 8 = 24$

Table 3

The sum of the weights for each email thread is calculated as shown in Table 4.

5

Thread	Sum of weights
email thread 1	$0 + 21 + 0 + 8 = 29$
email thread 2	$30.5 + 63 + 0 + 0 = 93.5$
email thread 3	$91.5 + 42 + 14 + 8 = 155.5$
email thread 4	$30.5 + 0 + 14 + 24 = 68.5$

Table 4

Table 4 shows that email thread 3 has the highest score of 155.5; email thread 2 has the second highest score of 93.5; email thread 4 has the third highest score of 68.5; and email thread 1 has the lowest score of 29.

10

A fraction or percentage of weight for each email in each email thread is computed. For this illustration, assume that email thread 1 has 3 emails; email thread 2 has 5 emails; email thread 3 has 6 emails; and email thread 4 has 2 emails. Table 5 below shows the fraction of weight that each email contributed to the overall weight for its respective email thread. In Table 5, the term "NA" designates not applicable (i.e., the email thread did not include this number of email messages), and a zero percentage (i.e., 0%) indicates that the email message did not include one of the descriptive terms.

15

20

Thread	Email 1	Email 2	Email 3	Email 4	Email 5	Email 6
email thread 1	21/29 72.4%	0/29 0 %	8/29 27.6%	NA	NA	NA
email thread 2	0/93.5 0%	30.5/93.5 32.6%	63/93.5 67.4%	0/93.5 0%	0/93.5 0%	NA
email thread 3	91.5/155.5 58.8%	42/155.5 27%	14/155.5 9%	0/155.5 0%	8/155.5 5.2%	0/155.5 0%
email thread 4	38.5/68.5 56.2%	30/68.5 43.8%	NA	NA	NA	NA

Table 5

Table 5 shows that the first email (Email 1) in email thread 1 has a highest
 5 relevancy (72.4%) to the descriptive terms. The third email (Email 3) in this
 thread has the second highest relevancy (27.6%), and the second email (Email
 2) does not include one of the descriptive terms. This table also shows the
 relevancy of emails for email threads 2 – 4.

10 According to block 250, the email threads in each cluster are ordered according
 to their respective scores.

Once the email threads are assigned a score, the threads are ordered by score
 within each cluster. The email thread with the highest score is displayed first; the
 15 email thread with the second highest score is displayed second; etc. Further, the
 emails in each email thread are displayed and sorted by date. The first email is
 shown in an original or unaltered state, and subsequent emails are shown with
 duplicative or redundant information removed. For example, if a subsequent
 email includes the textual content of the first email, then this textual content is
 20 removed since it is already presented on the display in the first email.

According to the scores calculated in Table 4, email thread 3 has the highest

score of 155.5; email thread 2 has the second highest score of 93.5; email thread 4 has the third highest score of 68.5; and email thread 1 has the lowest score of 29.

- 5 The documents are processed such that each document is scored according to the number of descriptive terms and weights for these terms. Additionally processing can also occur. For example, the following is executed for each thread: normalize a score of the thread to 100, start from the top of the thread, and compute a cumulative weight at each component document. A user is
10 notified once a point score of ninety (90) is obtained.

According to block 260, the emails in a thread are displayed until the weight of emails being displayed reaches a specified threshold of a weight for the thread. Emails in a thread are displayed until the emails reach a predetermined
15 percentage of the total weight of the thread. For example, the emails in a thread are displayed until the emails being displayed represent a specified percentage of a total weight for the thread. This specified percentage can be user input (such as eighty percent, eight-five percent, ninety percent, etc.). Subsequent emails can be removed from the thread and not displayed. Alternatively, the subsequent
20 emails can be displayed and visually marked to indicate that they are not within the threshold of weight for the thread.

Subsequent emails in a thread are shown until the sum of the weights of these emails reaches a predetermined value of the total weight of the thread (for
25 example, display emails in a thread until the weights reach 90% of the total weight of the thread). The first lines of each email are displayed along with a list of the inboxes where the email messages were found. Alternatively, a summary of the email can be shown (for example, show the sentences from the email that contain the highest number of descriptive terms).

30

By way of example, according to Tables 1 – 5, the email threads and

corresponding emails are displayed as follows: (1) Email Thread 3: Email 1, Email 2, and Email 3 (Emails 4 – 6 are removed from being displayed); (2) Email Thread 2: Email 1, Email 2, and Email 3 (Emails 4 and 5 are removed from being displayed, and Email 1 is displayed even though it has a low score since it is the first email in the thread); (3) Email Thread 4: Email 1 and Email 2; (4) Email thread 1: Email 1 and Email 3 (Email 2 is removed from being displayed).

Figure 3 is a display 300 showing email scores and ranks in accordance with an example implementation. For illustration, some data shown in Figure 3 is taken from Tables 1 – 5. A clustering tool scores and ranks email threads and generates output for the display 300.

A cluster includes four email threads (for example, Email Thread 1 to Email Thread 4 shown in Table 5). The email threads are ranked and scored according to the number of descriptive terms appearing in the emails of each cluster. The total weight of descriptive terms from Table 4 is $29 + 93.5 + 155.5 + 68.5 = 346.5$. The respective scores for each email thread are calculated by dividing the weight for each thread over the total weight of the threads. Thus, Email Thread 3 has first rank since it has a score of $155.5/346.5$ (44.9%). Email Thread 2 has a second rank since it has a score of $93.5/346.5$ (26.9%). Email Thread 4 has a third rank since it has a score of $68.5/346.5$ (19.8%). Email thread 1 has the fourth rank since it has a score of $29/346.5$ (8.4%).

Since Email Thread 3 has the highest rank, the emails in this thread are presented first, as shown at 320.

Display 300 provides a list of descriptive terms for Email Thread 3, shown at 330. These terms include storage (having 3 occurrences in Email Thread 3 with a total weight of 91.5), SAN (having 2 occurrences in Email Thread 3 with a total weight of 42), server (having 1 occurrence in Email Thread 3 with a total weight of 14); and disk array (having 1 occurrence in Email Thread 3 with a total weight of 8).

The email messages in Email Thread 3 are ordered by date and presented on the display 300 with the earliest email presented first. Email 1 has the highest score of 58.8%. The contents or a portion thereof of the actual email are

5 reproduced at 340 along with a list of inboxes or links 342 to where the email originated (such as link to the inboxes of users that received or sent the email).

Also, the descriptive terms 345 found in this email are displayed simultaneously with and adjacent to the email. Email 2 has the second highest score of 27%.

The contents of the actual email are reproduced at 350 along with a list of

10 inboxes or links 352 to where the email originated (such as links to the inboxes of users that received or sent the email). The descriptive terms for Email 2 are

shown at 355. Email 3 has the third highest score. The contents of the actual

email are reproduced at 360 along with a list of inboxes or links 362 to where the email originated (such as a link to the inbox of a user that received or sent the

15 email). The descriptive terms of Email 3 are shown at 365.

Figure 3 shows contents of emails being reproduced at 340, 350, and 360. The entire contents of an email can be reproduced or a selection of the email can be reproduced. For example, the first five non-quoted lines of each email are

20 reproduced. Alternatively, a summary of the email is reproduced.

Emails and email threads can each have multiple descriptive terms that are displayed adjacent to and simultaneously with the contents of an email message.

For example, emails in a thread can have multiple descriptive terms (such as the

25 descriptive terms “storage” and “SAN” appearing in both Email 1 and Email 2 in Figure 3).

Display 300 also includes a link 370 to each email in Email Thread 3. This link navigates the display to show the actual email.

30

Display 300 also includes an indication 380 when emails displayed in a thread reach a threshold of unique information of the thread. For example, a visual indication, such as text or indicia displayed on the display, is provided when ninety percent (90%) or more by weight of information in the email thread is displayed. As shown on display 300, the content of Emails 1 – 3 include 94.8 %
5 of unique information for Email Thread 3 (Email 1 with a score of 58.8% plus Email 2 with a score of 27% plus Email 3 with a score of 9%).

Figure 4A is a screenshot 400 of email threads in clusters in accordance with an example implementation. Several email threads in each cluster are shown side-by-side. Further information is displayed for each cluster. For example, Clusters #0 - #4 include a number of threads in each cluster, descriptive terms and scores for these terms, subjects of threads by weight, dates of emails, etc.

Figure 4B is a screenshot 430 of a summary of email threads in a single cluster in accordance with an example implementation. Specifically, Figure 4B shows the summary of email threads for Cluster 0 from Figure 4A. As shown in Figure 4B, Cluster 0 has labels or descriptive terms and corresponding scores of “carol (57.7)” and “clair (35.8).” The threads are displayed with subject, date, number of
15 20 messages, and weight. For example, thread “Update” has a date of 30 June 2000, has 34 email messages, and has a weight of 3148.9.

Figure 4C is a screenshot 460 of an email thread in accordance with an example implementation. Specifically, Figure 4C shows the email thread “MEGA
25 Assignment” from Figure 4B. As shown in Figure 4C, this email thread includes a list of the descriptive terms 462, a number of messages in the email thread 464, the actual email messages in the email thread 466 (which includes sender of the email, date of the email, unique lines in the email, and unique words in the email), and further information at 468 (which includes links to inboxes where the
30 documents originated and relevant words in the email message),

Figure 5 is a computer 500 with a clustering tool that scores and orders documents in accordance with an example implementation. The computer 500 includes memory 530, a clustering tool that calculates weights for documents and document threads and indicates a threshold in the document threads 540, a display 550, a processing unit 560, and buses or communication paths 570. The clustering tool 540 generates the output shown in display 300 of Figure 3, generates screenshots of Figures 4A-4C, and assists in executing blocks shown in Figures 1 and 2.

The processor unit includes a processor (such as a central processing unit, CPU, microprocessor, application-specific integrated circuit (ASIC), etc.) for controlling the overall operation of memory 530 (such as random access memory (RAM) for temporary data storage, read only memory (ROM) for permanent data storage, and firmware). The processing unit 560 communicates with memory 530 and clustering tool 540 to perform operations identified in Figures 1 – 3 and 4A – 4C. The memory 530, for example, stores applications, data, and programs (including software to implement or assist in implementing example embodiments) and other data.

Example embodiments can be used in a wide range of applications, such as personal email management, corporate level eDiscovery, and applications that rank and/or score documents.

Blocks or steps discussed herein can be automated and executed by a computer or electronic device. The term “automated” means controlled operation of an apparatus, system, and/or process using computers and/or mechanical/electrical devices without the necessity of human intervention, observation, effort, and/or decision.

The methods in accordance with example embodiments are provided as examples, and examples from one method should not be construed to limit

examples from another method. Further, methods or steps discussed within different figures can be added to or exchanged with methods of steps in other figures. Further yet, specific numerical data values (such as specific quantities, numbers, categories, etc.) or other specific information should be interpreted as illustrative for discussing example embodiments. Such specific information is not provided to limit example embodiments.

In some example embodiments, the methods illustrated herein and data and instructions associated therewith are stored in respective storage devices, which are implemented as computer-readable and/or machine-readable storage media, physical or tangible media, and/or non-transitory storage media. These storage media include different forms of memory including semiconductor memory devices such as DRAM, or SRAM, Erasable and Programmable Read-Only Memories (EPROMs), Electrically Erasable and Programmable Read-Only Memories (EEPROMs) and flash memories; magnetic disks such as fixed, floppy and removable disks; other magnetic media including tape; optical media such as Compact Disks (CDs) or Digital Versatile Disks (DVDs). Note that the instructions of the software discussed above can be provided on computer-readable or machine-readable storage medium, or alternatively, can be provided on multiple computer-readable or machine-readable storage media distributed in a large system having possibly plural nodes. Such computer-readable or machine-readable medium or media is (are) considered to be part of an article (or article of manufacture). An article or article of manufacture can refer to any manufactured single component or multiple components.

CLAIMS

What is claimed is:

- 1) A method executed by a computer, comprising:
 - assembling, by the computer, documents into multiple document threads;
 - identifying, by the computer, a list of descriptive terms appearing in the multiple document threads and weights for the descriptive terms;
 - calculating, by the computer, scores for the documents and scores for the multiple document threads by multiplying a number of times a descriptive term appears in a document by a weight generated for the descriptive term;
 - and
 - indicating, by the computer, when the documents in the multiple document threads reach a percentage of weight for the multiple document threads.
- 2) The method of claim 1 further comprising:
 - calculating a weight for a document thread;
 - displaying an earliest document in a document thread;
 - displaying subsequent documents in the document thread until a weight for a document reaches ninety percent of the weight for the document thread.
- 3) The method of claim 1 further comprising:
 - removing duplicative text in the documents;
 - displaying both the documents and inboxes with links to where the documents originated.
- 4) The method of claim 1 further comprising, displaying with one of the multiple document threads, a list of the descriptive terms appearing in the one of the multiple document threads, a number of times each of the list of descriptive terms appears in the one of the multiple document threads, and a subject for the one of the multiple document threads.

- 5) The method of claim 1 further comprising, computing a fraction of a weight of a multiple document thread that is contributed by each document in the multiple document thread.
- 6) A non-transitory computer readable storage medium comprising instructions that when executed causes a computer to:
 - assemble emails threads of emails into clusters;
 - identify, for each of the clusters, a list of descriptive terms from the email threads and weights for each of the descriptive terms;
 - calculate a weight for each of the emails and each of the email threads based on a number of times the descriptive terms appear in each of the emails and the email threads; and
 - display the emails in each of the email threads with an indication when the emails being displayed reach a threshold of weight of the email threads.
- 7) The non-transitory computer readable storage medium of claim 6 that when executed further causes the computer to:
 - display a list of the clusters with descriptive terms and subjects of some threads in each cluster that include top the top threads by weight;
 - order the emails threads in each of the clusters according to the scores for the email threads.
- 8) The non-transitory computer readable storage medium of claim 6 that when executed further causes the computer to:
 - display the email threads according to ranks based on weights of the email threads.
- 9) The non-transitory computer readable storage medium of claim 6 that when executed further causes the computer to:

remove, from being displayed, emails in a thread that have a weight below the threshold, wherein the emails removed from the thread do not include a sufficient number of the descriptive terms.

10) The non-transitory computer readable storage medium of claim 6 that when executed further causes the computer to:

cap, to a number three, a number of times a single descriptive term appears in an email.

11) A computer, comprising:

a clustering tool; and

a processor to execute the clustering tool to:

obtain a group of emails threads;

calculate weights for emails in the email threads and weights for the email threads based on a number of times descriptive terms appear in the emails and in the email threads; and

indicate when the emails in the email threads reach a threshold of weight for the email threads.

12) The computer of claim 11, wherein the processor further executes the clustering tool to:

display scores for the email threads and rankings of the email threads with respect to each other;

display scores for the emails and rankings of the emails in an email thread with respect to each other.

13) The computer of claim 11, wherein the processor further executes the clustering tool to:

display an indication when ninety percent (90%) or more of information in the email thread is displayed.

- 14) The computer of claim 11, wherein the processor further executes the clustering tool to:
- remove an email from an email thread when a score for the email is below a value;
 - display the email thread with the email removed;
 - display, with the email thread, a link to the email that is removed.
- 15) The computer of claim 11, wherein the processor further executes the clustering tool to:
- display an email in an email thread;
 - display a score for the email with respect to other emails in the email thread; and
 - display, adjacent to the email, the descriptive terms that appear in the email.

1/7

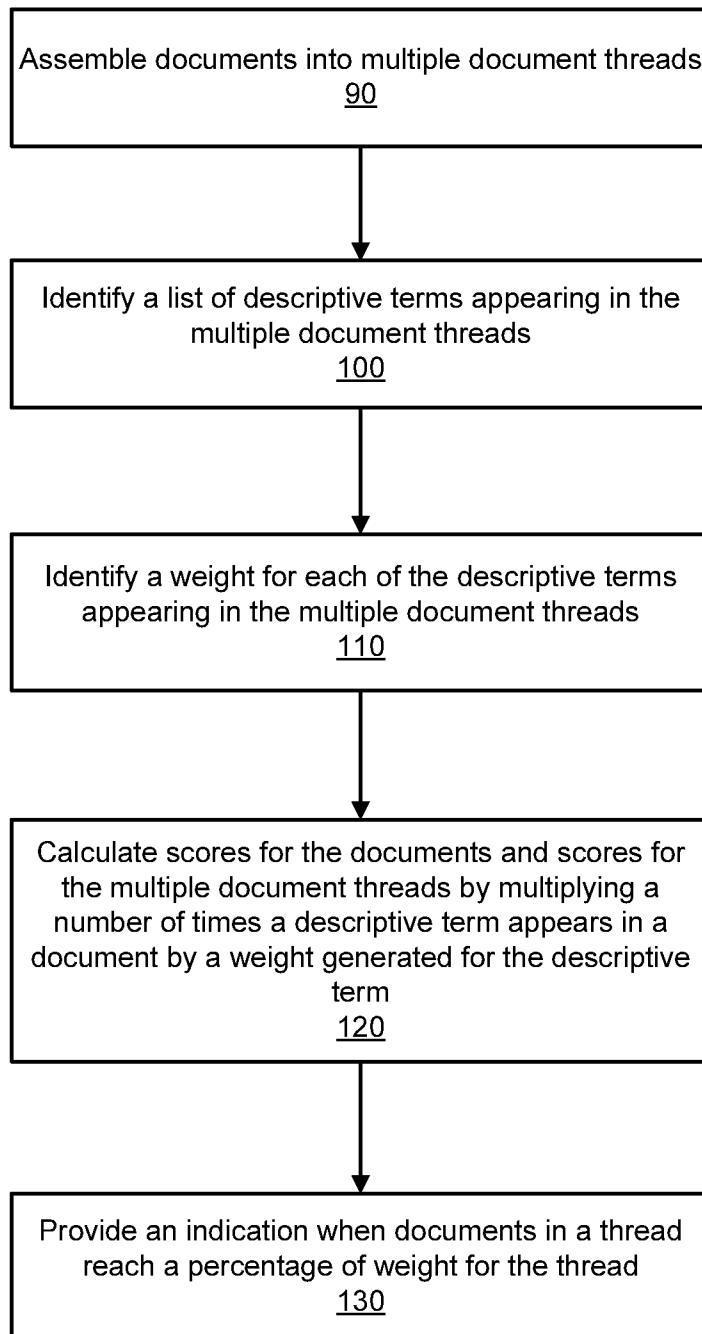
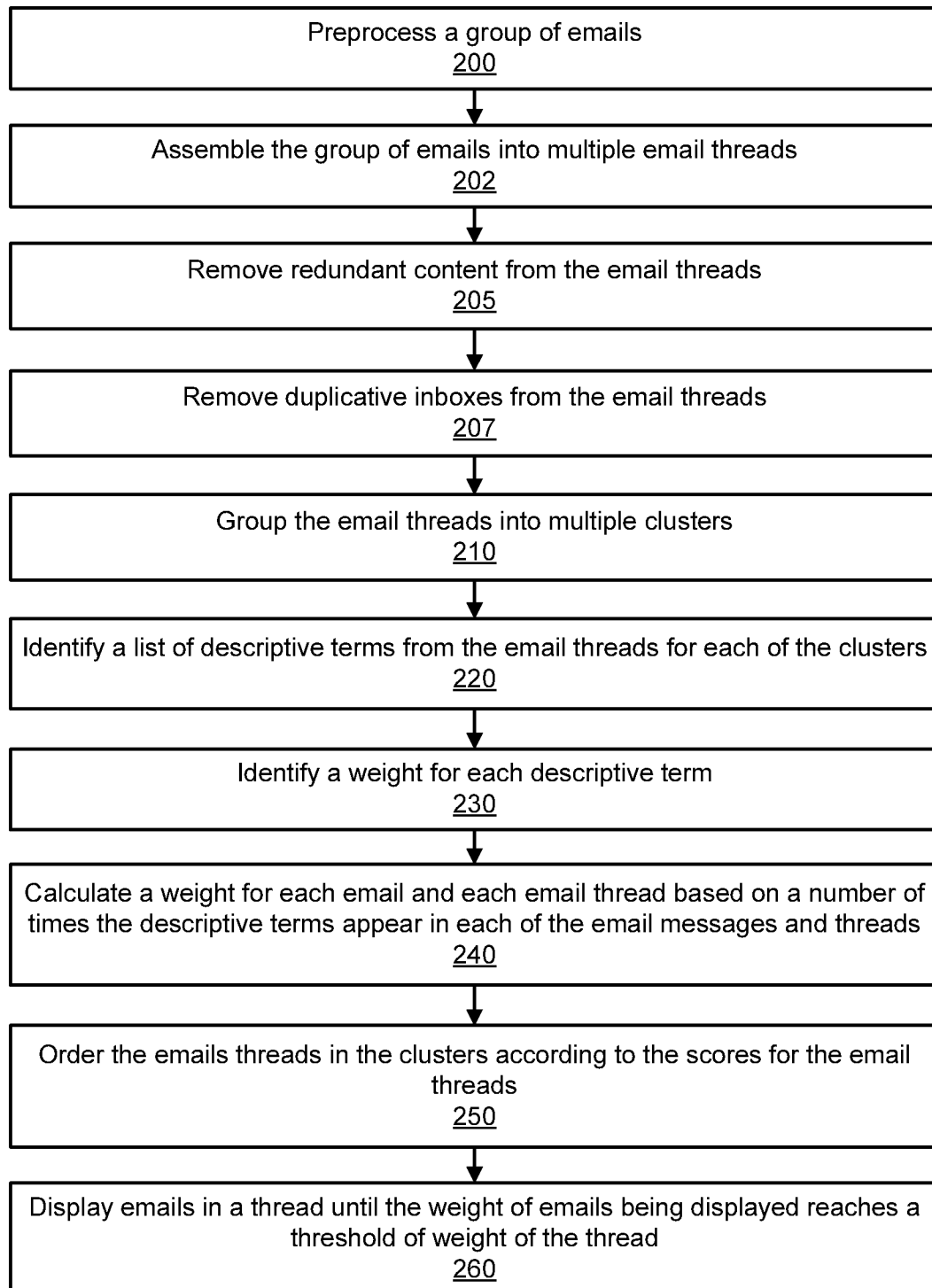


FIG. 1

2/7**FIG. 2**

3/7

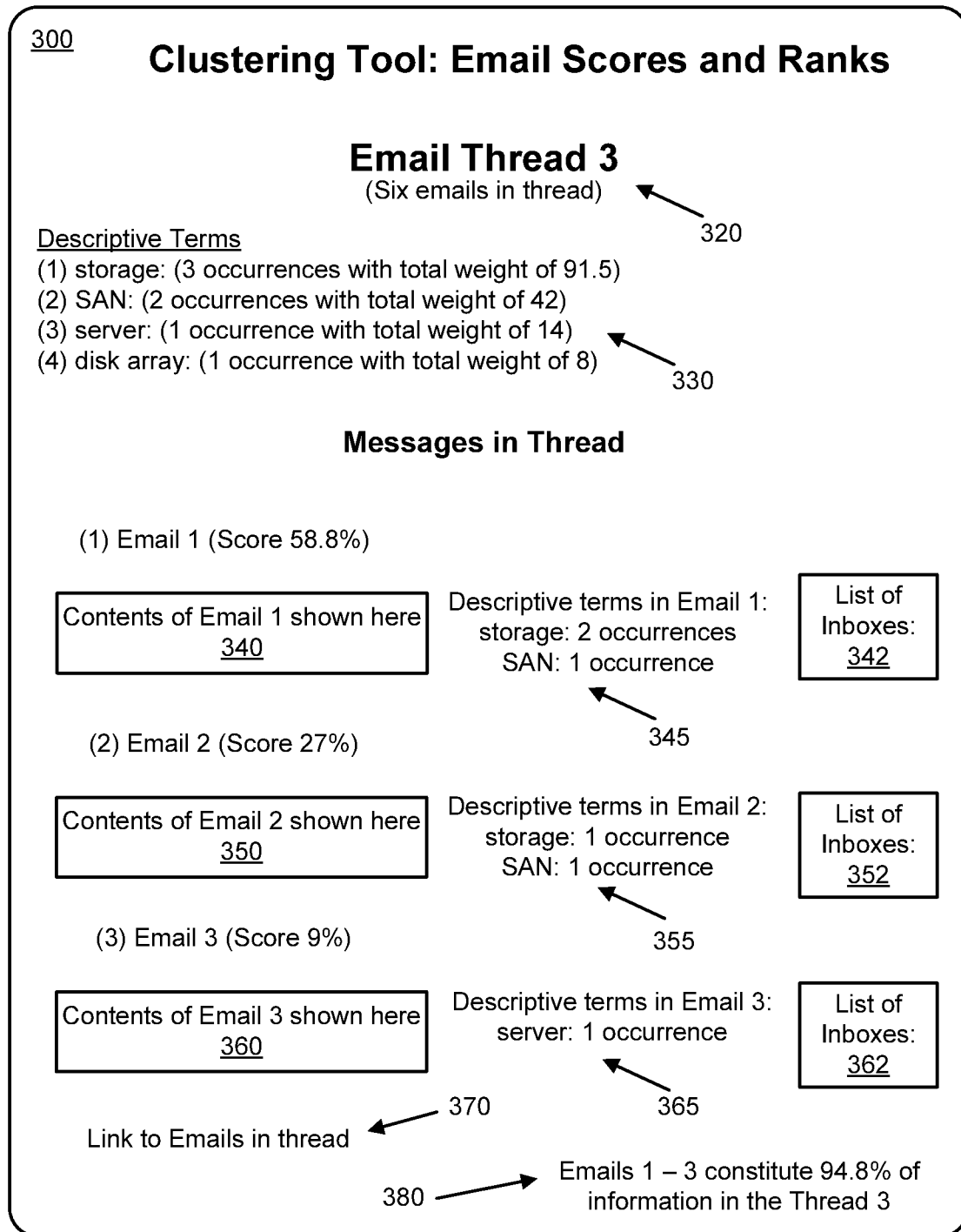
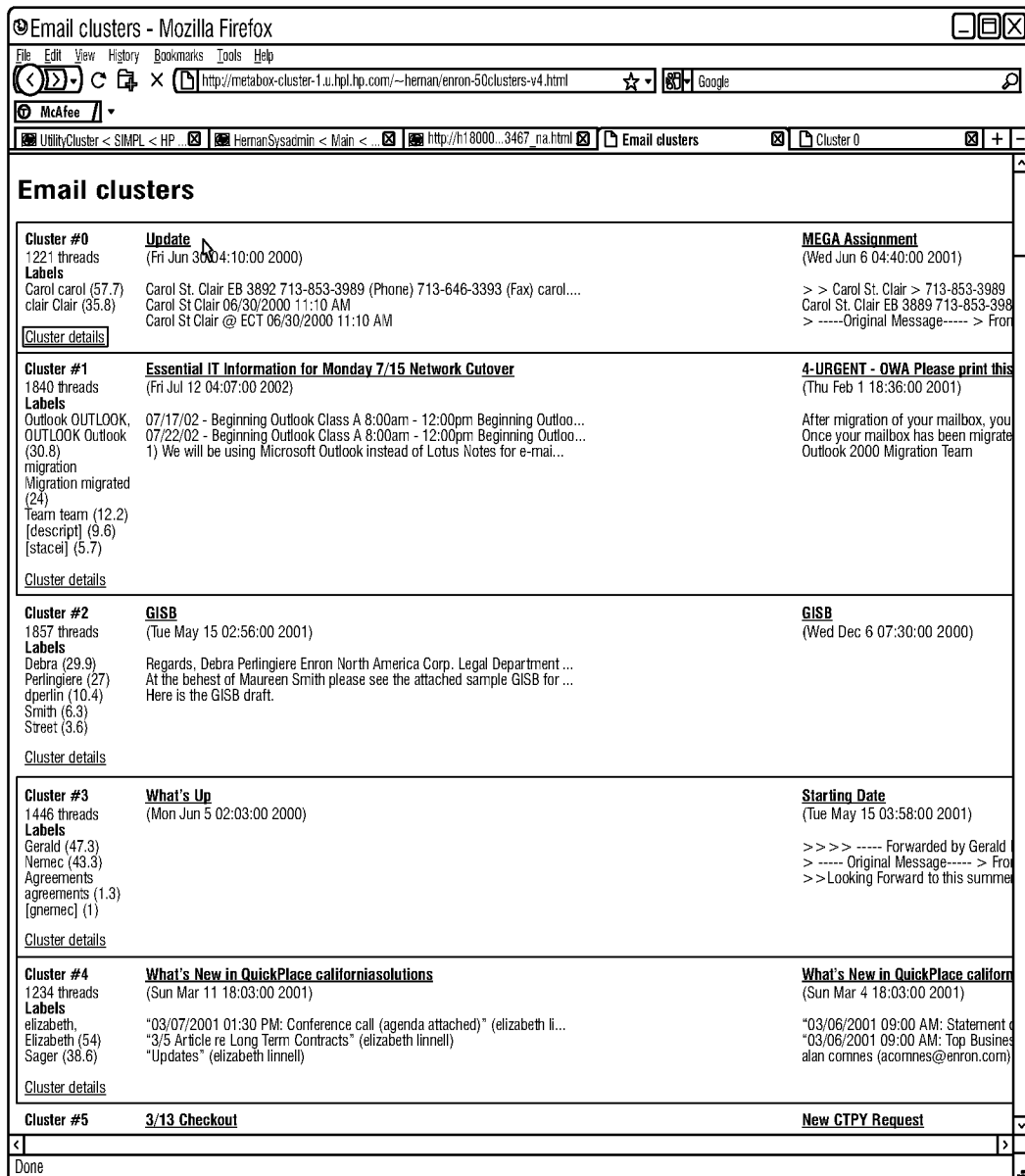


FIG. 3

4/7

400



Email clusters - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://metabox-cluster-1.u.hpl.hp.com/~herman/enron-50clusters-v4.html

McAfee

UtilityCluster < SIMPL < HP... HermanSysadmin < Main < http://h18000...3467...na.html Email clusters Cluster 0

Email clusters

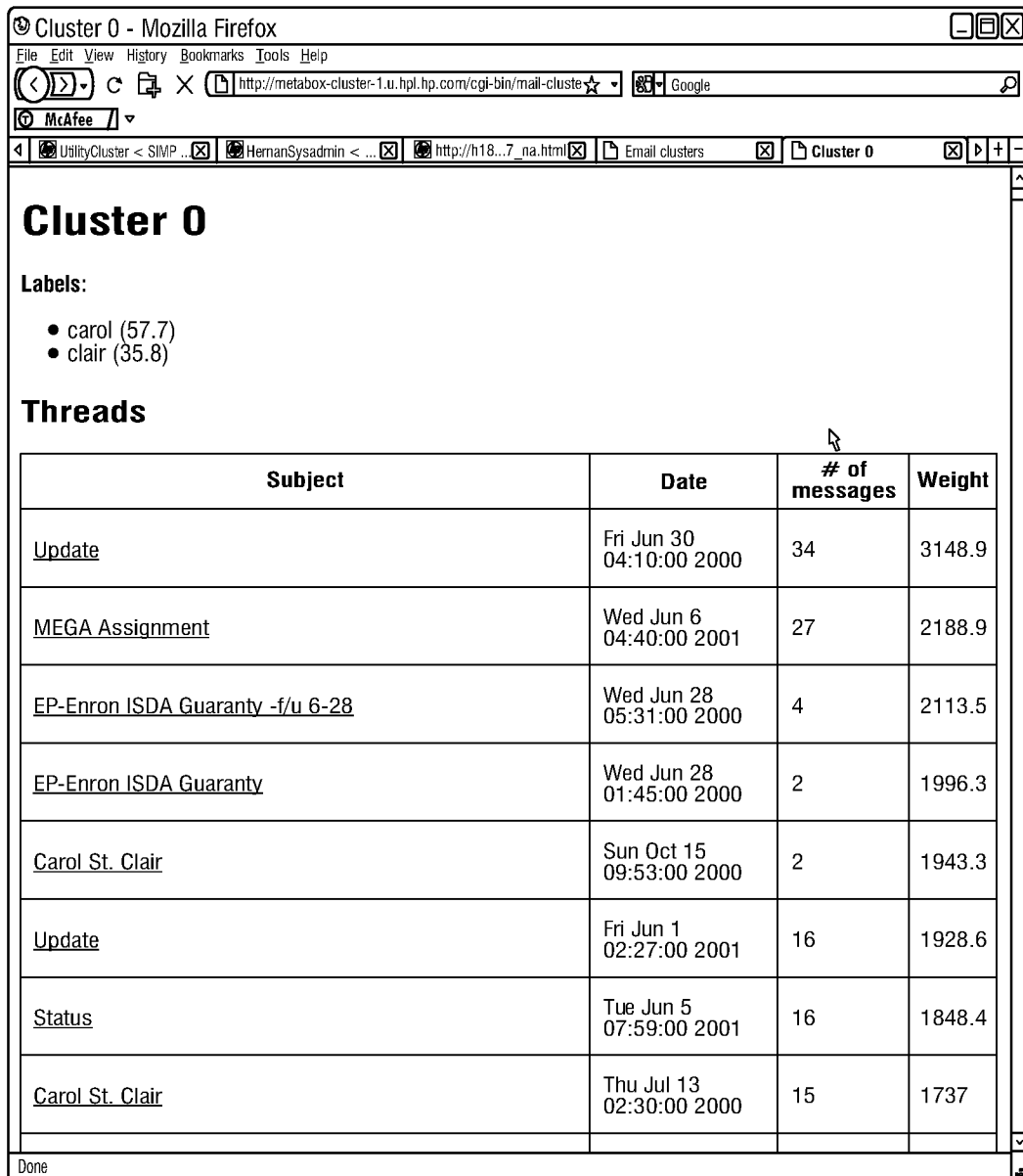
Cluster #0 1221 threads Labels Carol carol (57.7) clair Clair (35.8) Cluster details	Update (Fri Jun 30 04:10:00 2000) Carol St. Clair EB 3892 713-853-3989 (Phone) 713-646-3393 (Fax) carol... Carol St Clair 06/30/2000 11:10 AM Carol St Clair @ ECT 06/30/2000 11:10 AM	MEGA Assignment (Wed Jun 6 04:40:00 2001) > > Carol St. Clair > 713-853-3989 Carol St. Clair EB 3889 713-853-3989 > -----Original Message----- > For
Cluster #1 1840 threads Labels Outlook OUTLOOK, OUTLOOK Outlook (30.8) migration Migration migrated (24) Team team (12.2) [descript] (9.6) [stacei] (5.7) Cluster details	Essential IT information for Monday 7/15 Network Cutover (Fri Jul 12 04:07:00 2002) 07/17/02 - Beginning Outlook Class A 8:00am - 12:00pm Beginning Outloo... 07/22/02 - Beginning Outlook Class A 8:00am - 12:00pm Beginning Outloo... 1) We will be using Microsoft Outlook instead of Lotus Notes for e-mai...	4-URGENT - OWA Please print this (Thu Feb 1 18:36:00 2001) After migration of your mailbox, you Once your mailbox has been migrate Outlook 2000 Migration Team
Cluster #2 1857 threads Labels Debra (29.9) Perlingiere (27) dperlin (10.4) Smith (6.3) Street (3.6) Cluster details	GISB (Tue May 15 02:56:00 2001) Regards, Debra Perlingiere Enron North America Corp. Legal Department ... At the behest of Maureen Smith please see the attached sample GISB for ... Here is the GISB draft.	GISB (Wed Dec 6 07:30:00 2000)
Cluster #3 1446 threads Labels Gerald (47.3) Nerrec (43.3) Agreements (1.3) [gnerrec] (1) Cluster details	What's Up (Mon Jun 5 02:03:00 2000)	Starting Date (Tue May 15 03:58:00 2001) > > > ----- Forwarded by Gerald > ----- Original Message----- > Fro > > Looking Forward to this summer
Cluster #4 1234 threads Labels elizabeth, Elizabeth (54) Sager (38.6) Cluster details	What's New in QuickPlace californiasolutions (Sun Mar 11 18:03:00 2001) "03/07/2001 01:30 PM: Conference call (agenda attached)" (elizabeth li... "3/5 Article re Long Term Contracts" (elizabeth linnell) "Updates" (elizabeth linnell)	What's New in QuickPlace californ (Sun Mar 4 18:03:00 2001) "03/06/2001 09:00 AM: Statement d "03/06/2001 09:00 AM: Top Busines alan cornes (acomnes@enron.com)
Cluster #5 3/13 Checkout	New CTPY Request	

< Done >

FIG. 4A

5/7

430



The screenshot shows a Mozilla Firefox browser window titled 'Cluster 0 - Mozilla Firefox'. The address bar displays 'http://metabox-cluster-1.u.hpl.hp.com/cgi-bin/mail-cluster-1.cgi'. The browser has several tabs open, including 'UtilityCluster < SIMP...', 'HernanSysadmin < ...', 'http://h18...7_na.html', 'Email clusters', and 'Cluster 0'. The main content area is titled 'Cluster 0' and includes a 'Labels:' section with two items: 'carol (57.7)' and 'clair (35.8)'. Below this is a 'Threads' section containing a table with four columns: 'Subject', 'Date', '# of messages', and 'Weight'.

Subject	Date	# of messages	Weight
Update	Fri Jun 30 04:10:00 2000	34	3148.9
MEGA Assignment	Wed Jun 6 04:40:00 2001	27	2188.9
EP-Enron ISDA Guaranty -f/u 6-28	Wed Jun 28 05:31:00 2000	4	2113.5
EP-Enron ISDA Guaranty	Wed Jun 28 01:45:00 2000	2	1996.3
Carol St. Clair	Sun Oct 15 09:53:00 2000	2	1943.3
Update	Fri Jun 1 02:27:00 2001	16	1928.6
Status	Tue Jun 5 07:59:00 2001	16	1848.4
Carol St. Clair	Thu Jul 13 02:30:00 2000	15	1737

Done

FIG. 4B

6/7

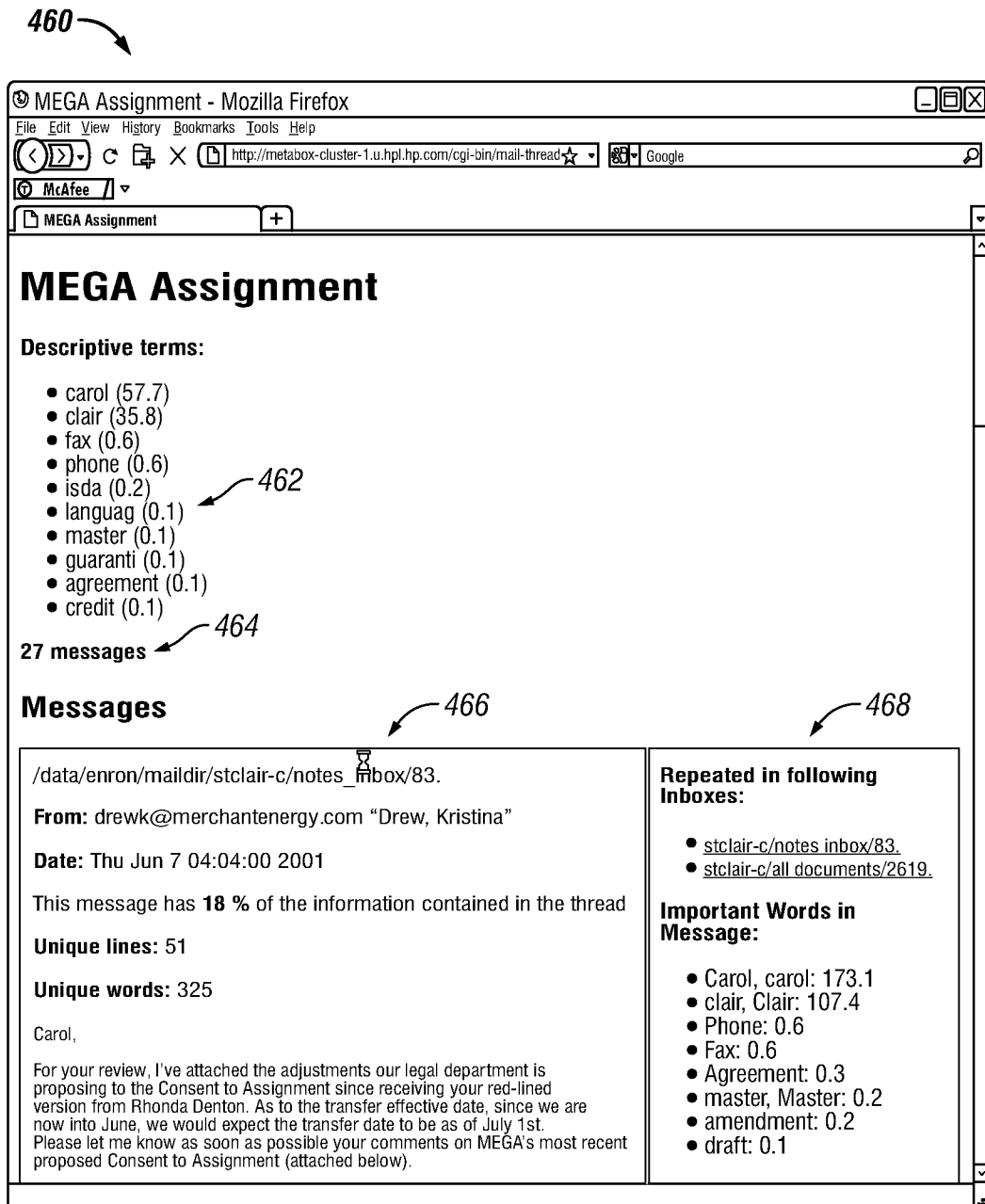
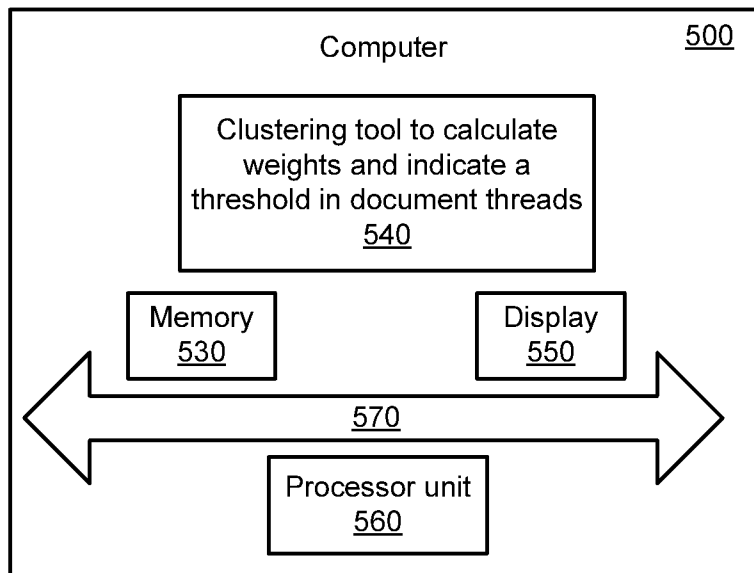


FIG. 4C

7/7

**FIG. 5**

A. CLASSIFICATION OF SUBJECT MATTER**G06F 17/21(2006.01)i, G06F 17/30(2006.01)i**

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F 17/21

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models

Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS(KIPO internal) & Keywords:document,weight,frequency,score

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 6507839 B1 (PONTE, JAY MICHAEL) 14 January 2003 See abstract and claims.	1-15
A	US 7747555 B2 (REGIER JEFFREY) 29 June 2010 See abstract and claims.	1-15
A	KR 10-1997-0076328 A1 (MATSUSHITA ELECTRIC INDUSTRIAL CO., LTD.) 12 December 1997 See abstract and claims.	1-15



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

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

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

30 JANUARY 2012 (30.01.2012)

Date of mailing of the international search report

08 FEBRUARY 2012 (08.02.2012)

Name and mailing address of the ISA/KR

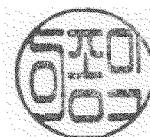
Korean Intellectual Property Office
Government Complex-Daejeon, 189 Cheongsu-ro,
Seo-gu, Daejeon 302-701, Republic of Korea

Facsimile No. 82-42-472-7140

Authorized officer

LEE, Jong Ick

Telephone No. 82-42-481-8373



INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2011/035666

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 6507839 B1	14.01.2003	AU 2000-43280 A1	16.10.2000
		AU 4328000 A	16.10.2000
		US 06496818 B1	17.12.2002
		US 06507839 B1	14.01.2003
		US 06826559 B1	30.11.2004
		US 06850935 B1	01.02.2005
		US 07024416 B1	04.04.2006
		US 07024627 B2	04.04.2006
		US 07047242 B1	16.05.2006
		US 07386795 B2	10.06.2008
		US 07725424 B1	25.05.2010
		US 2002-0060695 A1	23.05.2002
		US 2006-0179460 A1	10.08.2006
		US 2007-0027902 A1	01.02.2007
		WO 2000-058863 A1	05.10.2000
US 7747555 B2	29.06.2010	US 2007-0282780 A1	06.12.2007
KR 10-1997-0076328 A1	12.12.1997	CN 1133127 C0	31.12.2003
		CN 1172994 A0	11.02.1998
		EP 0810535 A2	03.12.1997
		JP 10-049549 A	20.02.1998
		JP 2005-063468 A	10.03.2005
		KR 10-0295354 B1	17.09.2001
		US 06154737 A	28.11.2000