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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
25 April 2013 (25.04.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/059290 Al

(51) International Patent Classification:
G06F 17/00 (2006.01) *G06Q 50/30* (2012.01)
G06F 77/20 (2006.01)

(21) International Application Number:
PCT/US2012/060573

(22) International Filing Date:
17 October 2012 (17.10.2012)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/548,183 17 October 201 1 (17. 10.201 1) US

(71) Applicant *(for all designated States except US):* **METAVANA, INC.** [US/US]; 575 Market Street, Suite 3025, c/o Spencer Trask & Company, San Francisco, CA 94105 (US).

(72) Inventor; and
(71) Applicant *(for US only):* **MINH, Duong-Van** [US/US]; 11256 Caminito Inocenta, San Diego, CA 92126 (US).

(74) Agents: **SU, Peter** et al; SNR Denton US LLP, P.O. Box 061080, Chicago, IL 60606-1080 (US).

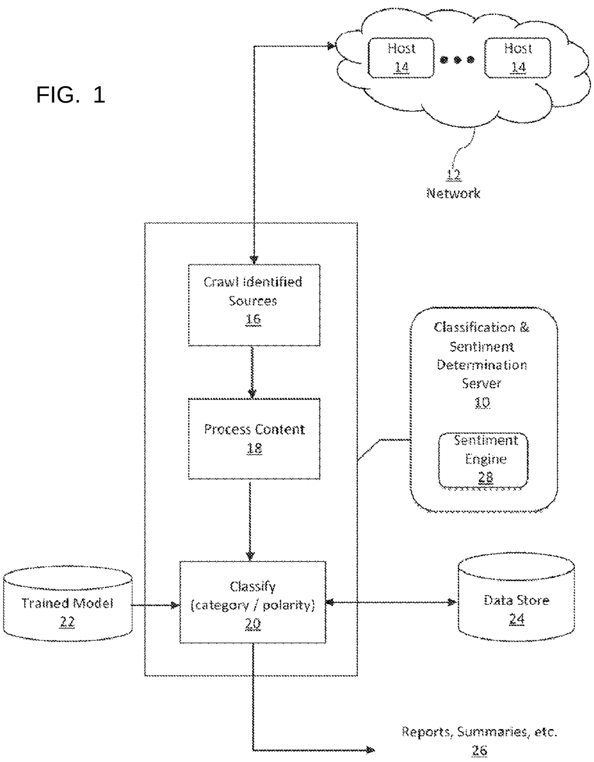
(81) Designated States *(unless otherwise indicated, for every kind of national protection available):* AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, EST, IS, JP, 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, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, 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, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, 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,

[Continued on nextpage]

(54) Title: SENTIMENT AND INFLUENCE ANALYSIS OF TWITTER TWEETS

FIG. 1



(57) Abstract: The present, invention is directed to a system, method, and article of manufacture that employs a sentiment engine for conducting sentiment and influence analysis of various types of messages from the social media hosts or websites to extract opinions on different, categories, which includes services, products or hotels, and others, collectively referred to as 44 the keyword produce. The sentiment engine includes a sentiment module configured to gather opinions or determine sentiment expressed in documents, a crawling module configured to servers of social network websites to obtain at least a subset of the documents or opinions from social media websites, a keyword module configured to extract keywords from documents, a tillering module configured to filter keywords and documents, and a classification module configured to classify documents, sentences, and/or keywords, a polarity prediction module configured to predict the polarity of a sentiment sentence, and a social media net promoter score configured to calculate a loyalty metric of users from social media websites, and a message analysis module configured to conduct analysis of a message from host social media, sites, forums, blogs and product/service providers. The message analysis module includes analyzing message from other host social media sites.

TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

SENTIMENT AND INFLUENCE ANALYSIS OF TWITTER TWEETS

Cross References to Related Patent Applications

This application claims priority to U.S. Provisional Application Sex. No. 61/548, 183 entitled "Sentiment and Influence Analysis of Twitter Tweets," filed on 1? October 201 1, the disclosure of which is incorporated herein by reference in its entirety.

Field of the Invention

[0001] The present invention relates to methodologies to extract and categorize opinion information from Twitter ™ tweets ™ and similar postings, including social media sites, and to score the influence or clout of the individuals) associated with said postings.

Background

[0002] The World Wide Web (WWW), or simply the "Web", is the well-known collection of interlinked hypertext documents hosted at a vast number of computer resources ("hosts ") communicatively coupled to one another over networks of computer networks known as the Internet. These documents, which may include text, multimedia files and images, are typically viewed as Web pages with the aid of a Web browser—a software application running on a user's computer system. Collections of related Web pages that can he addressed relative to a common uniform resource locator (URL) are known as websites, and are typically hosted on one or more Web servers accessible via the internet.

[0003] ln recent years, websites .featuring User Generated Content (UGC) that is content created and posted to websites by owners of and, sometimes, visitors to those sites, have become increasingly popular. UGC accounts for a wide variety of content including news, gossip, audio-video productions, photography and social commentary, to name but a few. Of interest to the present inventors is UOC which expresses opinions (usually, but not necessarily, of the person posting the UOC), for example of products, services, or combinations thereof (herein, the term "product" refers to mean any or all such products and/or services). Social media sites in particular have become popular places for users of those sites to post UGC thai includes opinion information.

[0004] The opinions and commentary posted to social media sites have become highly influential and many people now make purchasing decisions based on such content. Unfortunately, however, for people seeking out such content in order to inform prospective purchasing decisions and the like, the task is not always easy. Blogs, micro-blogs and social networking sites are replete with ever-changing content, and even if one can locate a review or similar post of interest, such reviews typically include .much information which is of little or no relevance to the topic and/or the purpose for which the review is being read. Further; while the UGC and opinion information can be of great value to advertisers, retailers and others, it is extremely burdensome to collect and analyze in any systematic way, and even more difficult to extract therefrom meaningful commentary or opinions which can form the basis for appropriate responses or informed decisions.

SUMMARY OF THE INVENTION

[0001] Embodiments of the present invention provide a system, method, and article of manufacture mat employs a sentiment engine for conducting sentiment and influence analysis of various types of messages (such as tweets and blogs) from the social media hosts or websites, including Twitter™, Facebook™, and Linkedin™, extract opinions on different categories, which includes services, products or hotels, and others, collectively referred to as "the keyword product". The sentiment engine includes a sentiment module configured to gather opinions or determine sentiment expressed in documents, a crawling module configured to servers of social network websites to obtain at least a subset of the documents or opinions from social media websites, a keyword module configured to extract keywords from documents, a filtering module configured to filter keywords and documents, a classification module configured to classify documents, sentences, and/or keywords, a polarity prediction module configured to predict the polarity of a sentiment sentence, a social media net promoter score (SNPS) configured to calculate a loyalty metric of users from social media websites, and a message analysis (also referred to as "tweets") module 44 configured to conduct analysis of a message (or text, graphics, or video) from host social media sites, forums, blogs and product/service providers, such as tweets™ from Twitter online message service. The message analysis module 44 includes analyzing message from other host social media sites, such as Facebook and Linkedin, Yelp™, blogs, and Sina Weibo. An influential score module is configured to compute the amount of influence that an author of a tweet has in his or her .message. The functionalities of these modules may be combined with one another or in addition to other .modules.

[0002] Broadly stated, a computer-implemented method for sentiment and influential analysis comprises receiving, by a.processor, a plurality electronic messages posted by one or more users on social media web websites; identifying, by a processor, a polarity of the sentiment-beating keywords for each electronic- message using a phase transition formula; determining, by a processor, at least one category corresponding to the at least one sentiment-bearing keyword associated with each electronic message; and determining, by a processor, an influence attribute for each electronic message based on a plurality of influence factors.

[0003] The structures and methods of the present invention are disclosed in the detailed description below. This summary does not purport to define the invention. The invention is defined by the claims. These and other embodiments, features, aspects, and advantages of the invention will become better understood with regard to the following description, appended claims and accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0004] The invention will be described with respect to specific embodiments thereof, and reference will be made to the drawings, in which;

[0005] Figure 1 is a system diagram illustrating a classification and sentiment determination server in a communication network in accordance with the present invention.

[0006] Figure 2 is a software system diagram illustrating the various modules of a sentiment engine in the classification and sentiment determination server in accordance with the present invent ion.

[0007] Figure 3 is a Sow diagram illustrating the process of extracting, categorizing, and identifying keywords in accordance with the present invention.

[0008] Figure 4 is a flow diagram illustrating a method for filtering documents from the corpus of files according to some embodiments of the present invention.

[0009] Figure 5 is a flowchart illustrating the highlights of a method for determining sentiment expressed in a sentence of a document in accordance with some embodiments of the present invention.

[0010] Figures 6A-6C are flow diagrams illustrating one embodiment of the sentiment determination and influence analysis process as applied to Twitter tweets in accordance with the present invention

[0011] Figures 7A-7B show a sample screen shot of the user interface (III) for a smart phone showing a graph with buzz feature in accordance with the present invention.

[0012] Figure 8 shows a sample buzz plot in accordance with some embodiments of the present inventsion.

[0013] Figures 9A-9B show a sample screen shot of the *VI* with the sentiment feature in accordance with some embodiments of the present invention.

[0014] Figure 10 is a block diagram of a machine in the example form of a computer system within which may be executed a set of instructions for causing the machine to perform any one or more of the methodologies discussed herein.

DETAILED DESCRIPTION

[0005] A description of structural embodiments and methods of the present invention is provided with reference to Figures I-10. It is to be understood that there is no intention to limit the invention to the specifically disclosed embodiments but that the invention may be practiced using other features, elements, methods and embodiments. Lite elements in various embodiments are commonly referred to with like reference numerals. In the following description, for purposes of explanation, numerous specific details are set forth in order to provide an understanding of various embodiments of the inventive subject matter, it will be evident, however, to those skilled in the ait that embodiments of the inventive subject matter may be practiced without these specific details. In general, well-known instruction instances, protocols, structures, and techniques have not been shown in detail.

[0015] Referring now to Figure 1, a classification and sentiment determination server 10 is communicatively coupled to a network 12 (e.g., the internet or a wireless network), which includes hosts 14 at which UGC of interest is located. The hosts may host social media sites (e.g., social networking sites), forums, blogs, product/service provider sites, etc., and the UGC of interest may include opinion-bearing content. In one particular embodiment of the invention, the UGC is Twitter tweets. The output of the classification and sentiment determination server is stored to a data store 24 (which may be included in or separate from the classification and sentiment determination server). The opinion-bearing content may be included within or with non-opinion-hearing content and non-UGC, hence the need to extract it before it can he analyzed/used.

[0016] In the illustration, the functions of classification and sentiment determination are shown as being performed by the same server; however, this need not necessarily be so. In other arrangements, the classification and sentiment determination functions may be performed by different servers and/or may be distributed across multiple servers or other computer-based platforms. The precise hardware arrangement used to perform the methods of the present invention is not necessarily critical to the invention.

[0017] in order to extract the UGC from the various content sources, customized Web crawlers are developed. in some instances it may be possible to use general purpose Web crawlers to extract UGC from the content sources, but increasingly it is the case that individual websites employ specialised formatting or other features, which makes the use of per-site custom Web crawlers appropriate. This way, the Web crawlers can be designed to extract only desired content (e.g., content which may include opinion-bearing information) and not all content at a particular site. This can reduce the burden on the analysis components discussed below. The customized crawlers are deployed to gather the content (and, optionally, associated metadata) from the identified sites 16. Content so gathered is processed (e.g., using stop-word removal) IS, the most frequent n-grams are identified, and these n-grams are then used to identify categories as part of a classification and polarity determination process 20. The categories are identified with the aid of category information obtained from a trained model 22, which identifies for each category sentiment-bearing keywords .

[0018] After the categories have been determined, the sentiment-bearing words associated with those categories are identified and their orientations (polarities) determined. ïn one embodiment of the invention, adjectives associated with each category keyword in the extracted content are identified as the opinion-bearing keywords. Keywords can be extracted automatically (e.g., for the entire training dataset or a portion thereof and using a lexicon provided to an extraction engine) and those adjectives can be manually tagged with a polarity (e.g., positive, negative, or neutral). Synonyms and antonyms of identified adjectives may be included in the sentiment-bearing words list with their polarity for the selected category.

[0019] The identified categories and associated opinion -bearing keywords from trained model .22 are used by the classification and sentiment determination server 10. As indicated above, this model is preferably constructed on a per-eategory basis so that category-appropriate polarities can be identified 20 and associated with the respective n-grant keywords.

[0020] The trained model 22 preferably associates category keywords and their respective opinion-bearing keywords, segregated or otherwise tagged by polarity. Categories may exist at a variety of granularities, for example hotels, rooms, bathrooms, etc. Within each category, adjectives or other identified opinion-bearing keywords may be segregated as positive, negative, or neutral. *In* some instances, the model will be stored locally by the classification and sentiment determination server 10, but in other eases it will be stored remotely therefrom. in instances where multiple classification and sentiment determination servers are deployed, a single instance of the trained model may be made available to each of the servers, so thai the servers all classify and determine sentiment of UGC in the same way, according to a common rule set. In other cases, different classification and sentiment determination servers may be given individual responsibilities for certain sources of UGC and each may have its own unique model, customized to that data source.

[0021] Regardless of such implementation specifics, the model 22 is used by the classification and sentiment determination server .10 to classify 20 content harvested by the Web crawlers by category and sentiment. To do so, the classification and determination server 10 processes 18 the harvested content to extract the category and sentiment keywords and then consults the trained model 22 to determine the polarity of the sentiment-bearing keywords. The output of the classification and sentiment determination server 10 is then stored to data store 24 and may later be used by the sentiment server 10 to create summaries 26 regarding the different products, and/or their features, for which UGC content was harvested.

[0022] The summaries 26 may be provided to advertisers, merchants or others and used to create/revise advertising and marketing campaigns or other for other purposes. Alternatively, or in addition, the summaries may be posted to other websites for easy review by users interested in the subject products. In still further embodiments, the summaries may be provided to search engine operators for return to users that execute searches related to the subject products. Of course, such search engines may be owned/operated by the same entity that owns/operates the sentiment server 10 and the sentiment server 10 may respond to queries executed by users of the search engine by providing pre-computed and/or computed-on-the-fly summaries concerning products which are identified in search queries.

[0023] The classification and sentiment determination server 10 includes a sentiment engine 28, which is illustrated in figure **2.** The sentiment engine 28 includes a sentiment module 30 configured to gather opinions or determine sentiment expressed in documents, a crawling module 32 configured to crawl servers (not. shown) to obtain at least a subset of the documents or opinions from social media websites 14, a keyword module 34 configured to extract keywords from documents, a filtering module 36 configured to filter keywords and documents, a classification module 38 configured to classify documents, sentences, and/or keywords, a polarity prediction module 40 configured to predict the polarity of a sentiment sentence, a social media net promoter score 42 configured to calculate a loyalty metric of users from social media websites, and a message analysis (also referred to as "tweets") module 44 configured to conduct analysis of a message (or text, graphics, or video) from host social media sites, forums, blogs and product/service providers, such as tweets from Twitter online message service. The message analysis module 44 Includes analysing message from other host social media sties, such as Facebookand Linkedin, Yelp, blogs, and Sina Werbo（新浪微博）. An influential score module 46 is configured to compute the amount of influence that ati author of a tweet*^{M} has in his or her message. The functionalities of these modules may be combined with one another or in addition to other modules. f or example, the sentiment module 30 may include the functionality of the keyword module 34 and the filtering module 36. The sentiment module 30, the crawling module 32, the keyword module 34, the filtering .module 36, and the classification module 38 are coupled to a communication bus 48.

[0024] For additional information on determining sentiment expressed in documents, see U.S. Patent Application No. 12/977,513 entitled "System and Method for Determining Sentiment Expressed in Documents", f led on December 23, 2012, and U.S. Patent Application No. 13/632,01 1 entitled "Sentiment Analysis from Social Media Content," filed *on* September 30, 2012, all owned by the assignee of this application and incorporated by reference in their entirety as if fully set forth herein .

[0025] Turning now to Figure 3, the sentiment engine 28 is configured for the keyword extraction process used to identify categories and opinion-bearing keywords 50 begins with the corpus of files or other content 52 downloaded by the crawlers and stored in the content store 54. The sentiment engine 28 retrieves corpus files at step 52, and filters and cleans the files by removing unwanted stop words 54. These files are cleaned, for example by stop-word filtering, to remove any words, phrases or other constructs that are known not to be opinion-bearing content 54. The data files extracted through the process in Figure 3 are pre-processed for keyword extraction This process is used to identify categories and opinion bearing keywords, which are then used to train the sentiment engine 28. The process of keyword extraction is described with reference to Figure 4.

[0026] Figure 4 illustrates a method 74 for filtering documents from the corpus of files according to some embodiments of the present invention. For each candidate document in the corpus 76, n-grams are extracted (78), art n-gram spectrum for the document is determined based on the extracted n-grams (80), wherein the n-gram spectrum indicates a frequency of occurrence of n-gratns as a function of a size of n-grams, and a determination *is* made as to whether the n-gram spectrum for the document conforms to a reference n-gram spectrum (82) within a predetermined threshold (84), wherein the reference n-gram spectrum is defined by a predetermined function. In some embodiments, the predetermined function is $cx^{\hat{}s} * e^{\hat{}K'';}$s wherein *x* is the ske of the n-gram, and wherein a, b, and c are predetermined values that place a peak of the predetermined function between an n-gram of size 2 and an n-gram of size 3. In some embodiments, the value of b is between 1 and 2, and the value of c is between 1 and *2*. The candidate document is retained (86) when the n-gram spectrum for the document conforms to the reference n-gram spectrum within the predetermined threshold, and discarded (88) when the n-gram spectrum for the document does not conform to the reference n-gram spectrum within the predetermined threshold.

[0027] Returning to Figure 3, keywords are then extracted from the retained documents 56. At step 56, the keyword extraction module 34 is configured to extract keywords from each document file of the plurality of documents Keywords may be regarded as those n~grams extracted during the filtering process. At step 58, for each extracted keyword, the keyword module 34 is configured to calculate a frequency, f of the keyword *in* the plurality of the documents, and a number of documents, N, that include the keyword, are calculated 58. At step 60, the keyword module 34 is configured to use a phase transition formula to calculate the relevancy of the keyword, based on its frequency in the plurality of documents and the number of documents that include the keyword. *In* one embodiment, the phase transition 6 rmula used to determine the relevancy of an individual keyword is f/N^{x}, where x>:1. At step 62, the relevancy is compared to a pre-established threshold and the keyword module adds the keyword to the list of keywords when the relevancy of the keyword exceeds that threshold. Otherwise, the subject keyword is not added to the list.

[0028] Having produced the list of relevant keywords (i.e., those with a relevancy score above the predetermined threshold), the classification module now determines unique pairs of keywords that are related to each other. For example, assume that the corpus of files or other content included *m* files, from which were extracted n keywords. Each n,^{n} keyword from an rrr* file is matched against (m-1) files, thus forming different clusters. Keywords belonging to each cluster are believed to belong to the same domain. Ousters obtained through this process are later refined and named as categories. The classification module 38 identifies sets of pairs of the keywords in which each set includes at least one keyword that is common to all of the pairs of keywords in the set 64. Nest, the classification module 38 iteratively combines the sets of the pairs of keywords In which each combined set includes at least one keyword thai is common to all of the pairs of keywords in the combined set until a predetermined termination condition is achieved 66.

[0029] Thus, the classification module determines sets of keywords that are related to each other and iteratively combines the sets to form categories. For example, the classification module may identify the following pairs of keywords front the list of keywords:

[Paris, Romance],

[Paris, City of Love],

[Paris, French],

[Dog, Beagle],

[Cat, Siamese].

The classification module may then determine that (Paris, Romance, City of Love, French} is a set of related keywords (e.g., a category) because the word "Paris" is common to the pairs (Paris, Romance), (Paris, City of Love), (Paris, French). Note that the classification module may also determine that (Paris, Romance, City of Love) is a set of related keywords. The level of specificity desired for a category determines the predetermined termination condition. The more keywords that are used to describe the category, the more specific the category is (e.g., (Paris, Romance, City of Love, French) is more specific than (Paris, Romance, City of Love)).

[0030] At step 68, The classification module 38 then obtains a plurality of (dot products) category spectruras. Ai step 70, the classification module 38 then determines at least one dot product that exceeds a predetermined threshold. A category spectrum may be represented by the pair *{WordID),Frequency}*, where the value of *WordII)* corresponds to a unique keyword and Frequency corresponds to a frequency of occurrence of the associated keyword. For example, the keyword "Paris" may have a *WordID* of 8 and a frequency of occurrence of 1002. Thus, the category spectrum includes a pair {8, 1002). These category spectrums may be visually represented. For example, on a 2-dimensional plot, one axis (e.g., the x-axis) may be *WordID* and the other (orthogonal) axis (e.g.. the y-axis) may be Frequency. At step 72, in some instances, the category spectrums may be normalized so that the area under each of the category spectrums is the same. in other words, the sentiment engine 28 is configured to identify the set of keyword pairs in which each set includes at least one keyword that is common to all pairs of keywords in the set. Doing so may reduce comparative bias between categories. Normalizing may be accomplished by normalizing the frequency of occurrence of the filtered keywords to produce the normalized category spectrum for the category.

[0031] The sentiment engine 28 is responsible for determining polarities of individual sentences in a review or other item of UGC. Therefore, in order to employ the sentiment engine, the harvested UGC content is split into sentences, which sentences may be units that are smaller or larger than the grammatical unit typically termed a sentence. That is, the sentences applied to the sentiment engine may fee grammatical sentences, portions of one or more grammatical sentences, or multiple grammatical sentences. For convenience. The term sentence refers to all such constructs which may form inputs for the sentiment engine.

[0032] As indicated above, the sentences are first processed to identify categories to which they refer or relate. Those sentences that include category keywords are passed to the sentiment engine. The sentiment engine first determines whether or not the subject sentence contains any opinion-bearing words. A positive and statistically significant correlation between adjectives and subjectivity of the opinion may be observed. Therefore, in one embodiment of the present invention, the presence of an adjective in a sentence is deemed to be a strong Indication that the sentence is subjective, i.e., sentiment-bearing. Accordingly; the present sentiment engine deems adjectives as sentiment-bearing keywords and any sentence that is classified into a category is analysed for such sentiment-bearing keywords. These sentences that are determined to contain at least one category keyword and one or more sentiment-bearing keywords are referred to as *semiment sentences.*

[0033] Fox each sentiment sentence reviewed by the sentiment engine, all adjectives in the sentence are extracted as sentiment-bearing keywords (the adjectives in the sentence being located using an adjectives lexicon provided to the sentiment engine), and the most adjacent adjective to a subject category keyword is identified as the *effective adjective* for that category. For example, in the following sentiment sentence: "The beds were nice, the sofas and chairs were comfy, and the kitchenette was stocked with the essentials.", the words *nice, comfy* and *stocked* may be identified as sentiment-beating keywords and the word *nice* is identified as the effective adjective for the category *bed* Effective adjectives are used to identify the orientation (polarity) of sentiment sentences by reference to the trained model. in this way, the category keywords and the sentiment-bearing words included in the harvested *UGC* are used to classify reviews and similar information concerning the subject product.

[0834] Various refinements for this overall method may be introduced. For example, in one embodiment of the invention for each sentence in the harvested *UGC,* category keywords may be identified (as described above) and sentiment-bearing words located. A. sentence that is found to contain at least one category keyword and one or more sentiment-bearing words may be referred to as a *sentiment candidate.* For each category. adjective keyword pair in a given sentiment candidate, the sentiment engine may compute a. distance (e.g., *in* terms of number of words) between them. If the distance is less than a predefined threshold, then the sentiment candidate is identified as a sentiment sentence for the category the subject keyword belongs to. Otherwise, the sentiment candidate is ignored.

[0835] To identify the polarity of the sentiment sentence for the identified category, we need to consider the following situations:

[0836] 1. A sentiment sentence might contain both likes and dislikes concerning some or all of the categories of the product. in such instances, the opinion words may be either positive or negative. Each opinion word is, however, likely to be closer in distance to the category keyword that it is related to than to other category keywords. Therefore, such a sentence can be listed many times for each category with respective probabilities for each sentiment: category pair. For example, in the sentence *"The staff was nice, however, the room was very small.", nice* and *§mall* are opinion words and both are mentioned. Proximities of these opinion words to the identified categories reveals the categories to which each relates; .here *nice* corresponds to a customer service category {as identified by the keyword *staff),* while*small* corresponds to a room category (as identified *by* the keyword *room).*

[0037] 2. Sentiment sentences might contain both likes and dislikes about the same category. For instance in the following sentence, *"Rooms are small mid clean",* the writer is (presumably) not happy with the size of the room, bin happy with the room being tidy and neat. Such sentences must also be captured and reported as both negative and positive.

[0038] 3. For a sentence that contains a contrastive clause (e.g., sentences that start with or include words such as "but", "however", etc.) that indicates a sentiment change for features in the clause, the effective opinion in that clause is used to identity the orientation of the categories. However, if there is no category orientation in the clause, then the polarity of the contrastive clause is identified as the opposite polarity of the remainder of the sentence.

[0839] The sentiment engine may also be configured (e.g., via the trained model) to handle manifestations of negation: if there is a negation keyword before a. sentiment-bearing keyword and its distance to the sentiment-bearing keyword is less than a predetermined threshold, then the polarity of the sentiment sentence may be determined to be the opposite of the polarity of the sentiment-bearing keyword that is associated with the category keyword. For example, *m* the sentence, *"The rooms were not large.",* the opinion-bearing keyword *large* is associated with the category keyword *room* and, ordinarily, would be deemed to express a positive sentiment. However, because the word *not* is determined to modify the sentiment-bearing keyword *large,* the sentiment engine may determine that the opposite sentiment is, in fact, being expressed.

[0040] Sentiment candidates or sentiment sentences identified as discussed above might also be determined o contain wishes, thoughts, beliefs, etc., concerning a product. As such, they may not reflect actual opinions concerning an indentified category. Accordingly, in some embodiments of the present invention the sentiment engine applies a filtering technique, wherein keywords such as *"guess", "believe",* "wish", and other terms expressing desires rather than true opinions, are treated as sentiment eliminators. Any sentiment candidates or sentiment sentences determined to contain such keywords are eliminated from the sentiment sentences list. A dictionary of such eliminators may be provided to the sentiment engine as part of the trained mode! or in addition thereto.

[0041] After identifying the orientation of a sentiment sentence, the sentiment engine identifies how strong the sentiment is. The severity of an opinion can be measured by associating each opinion-bearing keyword with a sentiment score. For example, the sentiment score for the opinion-bearing keyword "*bad*" may be -1, while the sentiment score for the opinion-bearing keyword *"horrible"* may be -3 (e.g., on a scale where the sign of the sentiment score is indicative of a positive or negative polarity and the magnitude of the sentiment score indicates the strength or severity thereof). Assigning an overall severity score may require comparison of multiple reviews and an averaging thereof.

[0042] Figure 5 is a flowchart illustrating the highlights of a method 90 for determining sentiment expressed in a sentence of a document (e.g., a harvested Web page or the like), according to embodiments of the present invention. For candidate sentences 92 (which candidates may be grammatical units larger than, equal to, or smaller than a grammatical sentence) provided to the sentiment engine 28, at step 94, a sentence that includes at least one sentiment-bearing keyword within a predetermined distance of at least one candidate keyword is identified. The sentiment-bearing keyword should be a word (e.g., an adjective) indicating an expression of sentiment. At 96, the orientation or polarity of the sentiment-bearing keyword is determined (e.g., using the trained model provided to the sentiment engine). The polarity may indicate that die sentiment-bearing keyword reflects a positive sentiment, a negative sentiment, or a neutral sentiment. At 98, the sentiment engine determines whether the assessed polarity is negated (e.g., due to the presence of any sentiment negating words in proximity to the sentiment-bearing keyword). Then, at 100, the sentiment engine classifies the sentiment of the sentence. Not shown, although an optional component of method 90 is an option to discard a candidate sentence if the sentiment engine determines that one or more sentiment eliminators are present in the candidate sentence.

[0043] By way of example for the process described with respect to Figure 5, consider an exemplary document that includes an exemplary sentence: "The room was stinky and the carpets were dirty." Assume that the words *"stinky"* and *"dirty"* are sentiment-bearing keywords expressing a negative sentiment (e.g., a negative polarity), and the words *"room"* and *"carpets"* are category (or sub-category) keywords. The sentiment engine identifies this candidate sentence as including the sentiment-bearing keywords *"stinky"* and *"dirty"* and identifies that these sentiment-bearing keywords ate in sufficient proximity to the category keywords *"room"* and *"carpets"*, respectively, hence, the candidate sentence is passed for further processing. in this example, *room* and *carpet* may be sub-categories of a broader category *"hotel room"* or may be categories of their own. In either instance, the sentiment identifi*es "dirty" and "stinky"* as sentiment-bearing keywords expressing a negative sentiment. There are no sentiment negating words, hence, the sentence is classified as one that expresses a negative sentiment concerning a hotel room (and/or a room and carpet). This sentence and its classification may be subsequently stored and statistics reflecting the classification updated.

[0644] Figures 6A-6C, which collectively represent one composite graphical representation, are flow diagrams illustrating one embodiment of the sentiment determination and influence analysis process 102 as applied to Twitter tweets (or feeds). At block 104, each of the host social media sites 14 has its own database, including GNIP database. Twitter database and Facebook database. The GNIP database, for example, is a social media application programming interfaces ***(API)*** aggregator which would supply GNIP API, which provides notification of activities (events) occurring in a variety of services including a user "tweet" (Twitter), a user "dugg" (diggX a user creating a blog post, etc. Twitter tweets, Facebook status updates and other feeds from social media APIs (e.g., GNIP APIs) may also be processed using the sentiment engine 28. Near real time results may be reported using a real time user interface. The process involved in obtaining the relevant feeds and determining the corresponding sentiments is shown in figures 6A-6C. in one embodiment, the process involves three steps:

- STEP 1: Fetching real time feeds from GNIP
- STEP 2: Processing the GNIP data
- STEP 3: Displaying data on Real Time User Interface

[0045] STEP 1: At block 106, the sentiment engine 28 is configured to fetch real time feeds from GNIP. GNIP has a streaming API from which can fetch real time tweet feeds and Facebook status feeds by querying the API with keywords. This can replace the crawlers described above, or the crawlers can be instantiated so as to provide keywords from data dictionary 108 to the APIs and retrieve the resulting streams by the sentiment engine 28 at block 11.0, The feeds are provided in JavaScript Object Notation (iSQN) format and the resulting data is queued to be read by processes described in step 2.

[0046] STEP 2, Processing the GNIP data. At block 112, the queued data is processed by the sentiment engine 28 (one can use multiple threads across multiple queues) to determine category and polarity for each feed and again queue these results. This queue is then written to a database at block 1.14. The process may be parallelized to handle high volumes of data.

[0047] The sentiment determination and influence analysis process .102 includes supplying information source 118 (circular symbol 1} between the real time Twitter and Facebook feed 106 and the real time tweets and Facebook statuses with subject, polarity, infiueoce, and other relevant metadata 114. The information source 118 can be provided using a variety of methods, including Klout score (see 2.1 below) or compute influence of a tweet (see 2.2 below).

[0048] 2.1 Klout score: A Klout score may be provided as a parameter by a GN1P API Klout scores measure the influence of the individual posting the tweet, etc., based on his/her ability to drive action, it relies on the fact that every time content is created, the poster of the content somehow influences others. The Klout score uses data from social networks in order to measure, how many people are so influenced, how those individuals are influenced, the influence of the poster's network,

[0049] 2.2 Influence computation: The present real time user interface shows the influence of the tweet and the author thereof, thus letting a user sort tweets based on influence. The influence of the tweet is computed with (he following five exemplary parameters, which alternatively could be user-defined with more or fewer parameters, initially, the message analysis module 44 is configured io stream one or more tweets at step 122. At step 124, the message analysis module 44 then is configured to extract a particular author. At step 126, the message analysis module 44 is configured to determine if a tweet includes a link, such as an URL to a website.

- At step 128 (influence factor a), the message analysis module 44 is configured to compute the number of people following the author of the subject tweet (follower).
- At step 130 (influence factor b), the message analysis module 44 is configured to compute the number of tweets on a given subject that the author of a subject tweet creates (freq).
- At step i3:2 (influence factor c), the message analysis module 44 is configured to compute the number of people re-tweeting a subject tweet (retw).
- At step 134 (influence factor d), the message analysis module 44 is configured to compute the number of people replying to the subject tweet (reply).
- Optionally, the message analysis module 44 is configured to compute the number of people the author of (he subject tweet is following (followee) as another influence factor.

a) In same blogs, at step 136 (influence factor e), the message analysis module 44 is configured to extract indicator buttons for Facebook, Twitter, Linkedin, delicious™, Dig™, and MySpace™; thus, the message analysis module 44 extracts "likes" counts in Facebook, Linkedin, delicious, Dig, and MySpace (likes).

b) Is addition the sentiment engine 28 is configured to count the number of people who use a subject blag for their tweet information (blog tweet).

[0050] in one embodiment, at step 138, an influential score module 46 is configured to compute influence formula thai lias these seven parameters, which can also be referred to as a Twitter Influence score (TIS). Aft example is:

TIS = freq x follower x retw x reply x foilowee x likes x blog tweet.

[0051] For calculating influence, the TiS score is compared with a Klout score. The TIS score takes into account the number of followers, number following, listed count and status count. A curve fitting formula may be used to derive a final influence score, for example, taking into account the content of the tweet (keywords, etc.), language of the tweet, whether or not the tweet uses profanity or other unacceptable language, etc. An example of a final influence computation is:

[0052] $$\sum_{i=1}^{n} a_i S_{feature}$$

where $S_{feature}$ is calculated on the basis of the above-mentioned parameters (some, but not all of which, may be obtained from GNIP data).

[0053] Due the fact there are many Twitter authors that tweet about many subjects, the influence scores may be skewed. To correct for such cases, the score is scaled according to the total number of tweets obtained directly from Tvvttter.com: to obtain a ratio, defined as the number of tweets in our database divided by the total number of tweets on twitter.com/author. For an Internet marketer, this ratio could be very small because the author's business requires this author to tweets about tens or hundreds subjects. This author, even having a high TIS score, should be discarded.

[0054] 2.3 Spam check: A spam check may be performed on the tweets to avoid potential problems. At step 140, the sentiment engine 28 is configured to identify link from the data source of host social media websites. At step 142, the sentiment engine 28 is configured to crawl the one or more blogs associated with the identified link. At step 144, the sentiment engine 28 is configured to parse blog pages by removing unwanted headers, footers, and advertisement. In one embodiment, the sentiment engine 28, at step 146, is configured to analyze electronic messages, such as Twitter tweets, to find certain patterns as spam. At step 146, if the sentiment engine 28 locates same patterns in some tweets then may discard such electronic messages or tweets considering them as spam. Also some tweeters are market advertisers. Setting thresholds on various filters eliminates a! least some of the tweets from them. Such irrelevant tweets are not processed by sentiment engine. If the electronic messages are not considered as sparn, the sentiment engine 28 continues, at step 150, to obtain blog content.

[0055] 2.4 Sentiment engine ref nement. in one embodiment, a sentiment engine, similar to the sentiment engine 28 but could be a more simplified version, may be specially configured for the sentiment analysts of tweets and public Faeehook status updates, both of which are shorter (i.e., limited to 140 characters) and buster than other forms of social media. For Illustration purpose, the sentiment engine 28 is used for describing the process. The sentiment engine 152 is configured to process blog content through steps 152 for sentiment analysis, 154 for identifying sentiments with subject and polarities, 156 for extracting metadata including author and date, and 158 tor computing the degree of influence.

[0056] Moreover, a LAMP architecture may be used for the application. Before applying the proposed approach, the reviews must he split into sentences, which may be units that are equal to, smaller than or larger than a grammatical sentence. Then these units are processed to identify the categories they mention as explained above. After categories are identified, the sentiment bearing keywords are extracted. These are then are expanded to a full opinion-hearing keywords list as described above. The polarity of each sentiment sentence Is identified as discussed above.

[0057] STEP 3. Displaying data on the real time user interface

[0058] At block 116, the real time user interface (UI) queries the database to provide user-selected filters, generate different types of near real time buzz and polarity plots, and enable Boolean search of the database. The UI has a reply feature attached to each tweet wherein the author of the tweet can be replied to directly from the UI by authenticating with Twitter.

[0059] Figures 7A-7B show the screen shot of the UI for a smart phone (e.g., an tPhone™) showing a graph with *bmz* feature. The graph can be switched between buzz and sentiment view. The 111 shows the search feature where tweets can be searched on the basis of keywords. The table displays the tweets at that instant of time. The table contains a.tweet, its polarity and category. In addition to the above, the screen displays Klout score and influence (computed in accordance with the above-described process). Figure 8 shows a buzz plot. Figures 9A-9B show a screen shot of the U I with the sentiment feature.

[0060] Figure 10 is a block diagram of a machine in the example form of a computer system 160 within which may be executed a set of instructions for causing the machine to perform any one or more of the methodologies discussed herein. In alternative embodiments, the machine operates as a standalone device or may be connected (e.g., networked) to other machines. In a networked deployment, the machine may operate in the capacity of a server or a client machine in a server-client network environment or as a peer machine in a peer-to-peer (pr distributed) network environment.

[0061] The machine is capable of executing a set of instructions (sequential or otherwise) that specify actions to be taken by that machine. Further, while only a single machine is illustrated, the term "machine" shall also be taken to include any collection of machines thai individually or jointly execute a set (or multiple sets) of instructions to perform any one or more of the methodologies discussed herein.

[0062] The example of the computer system 160 includes a. processor 162 (e.g., a central processing unit (CPU), a graphics processing unit (GPU) or both), and memory 164, which communicate with each other via bus 168. Memory 164 includes volatile memory devices (e.g., DRAM, SRAM, DDR RAM, or other volatile solid state memory devices), non-volatile memory devices (e.g., magnetic disk memory devices, optical disk memory devices, .flash memory devices, tape drives, or other non-volatile solid state memory devices), or a combination thereof Memory 164 may optionally include one or snore storage devices -remotely located from the computer system 160. The computer system 160 may further include video display unit .166 (e.g., a plasma display, a liquid crystal display (LCD) or a cathode ray tube (CRT)). The computer system 160 also includes input devices 170 (e.g., keyboard, mouse, trackball, touchscreen display, etc.), output devices 172 (e.g., speakers), and a network interface device **174**. The aforementioned components of the computer system 160 may be located within a single housing or ease (e.g., as depicted by the dashed lines in Figure 6), Alternatively, a subset of the components may be located outside of the housing. For example, the video display unit 166, the input devices 1.70, and the output device 172 may exist outside of the housing, but be coupled to the bus .168 via external ports or connectors accessible on the outside of the housing.

[0063] Memory 164 includes a machine-readable medium 176 on which is stored one or more sets of data structures and instructions 178 (eg., software) embodying or utilized by any one or more of the methodologies or functions described herein. The one or more sets of data structures may store data. Note that a machine-readable medium refers to a storage medium thai is readable by a machine (e.g., a computer-readable storage medium) . The data structures and instructions 178 may also reside, completely or at least partially, within memory 164 and/or within the processor 162 during execution thereof by computer system 160, with memory 164 and processor 162 also constituting machine-readable, tangible media.

[0064] The data structures and instructions 178 may further be transmitted or received over a network ISO via network interface device 174 utilizing any one of a number of well-known transfer protocols Hypertext Transfer Protocol (HTTP)). Network 180 can generally include any type of wired or wireless communication channel capable of coupling together computing nodes (e.g., the computer system 160). This includes, but is not limited to, a local area network, a wide area network, or a combination of networks . In some embodiments, network 180 includes the Internet

[0065] Certain embodiments are described herein as including logic or a number of components, modules, or mechanisms. Modules may constitute either software modules (e.g , code and/or instructions embodied on a machine-readable medium or in a transmission signal) or hardware modules. A hardware module is a tangible unit capable of performing certain operations and may be configured or arranged in a certain manner. In example embodiments, one or more computer systems (e.g., the computer system 160) or one or more hardware modules of a computer system (e.g., a processor 162 or a group of processors) may be configured by software an application or application portion) as a hardware module that operates to perform certain operations as described herein.

[0066] In various embodiments, a hardware module may be implemented mechanically or electronically. For example, a hardware module may comprise dedicated circuitry or logic that is permanently configured (e.g., as a special-purpose processor, such as a field programmable gate array (f PGA) or an application-specific integrated circuit (ASIC)) to perform cer ain operations. A hardware module may also comprise programmable logic or circuitry (e.g., as encompassed within a general-purpose processor 162 or other programmable processor) that is temporarily configured by software to perform certain operations. It will be appreciated that the decision to implement a hardware module mechanically, in dedicated and permanently, configured circuitry, or in temporarily configured circuitry (e.g., configured by software) may be driven by cost and time considerations.

[0067] Accordingly, the term "hardware module" should be understood to encompass a tangible entity, be that an entity that is physically constructed, permanently configured (e.g., hardwired) or temporarily configured (e.g., programmed) to operate in a certain manner and/or to perform certain operations described herein. Considering embodiments in which hardware modules are temporarily configured (e.g., programmed), each of the hardware modules need not be configured or instantiated at any one instance in time. For example, where the hardware modules comprise a genera!-purpose processor 162 configured using software, the general-purpose processor 162 may be configured as respective different hardware modules at different times. Software may accordingly configure a processor 162, for example, to constitute a particular hardware module at one instance of time and to constitute a different hardware module at a different instance of time.

[0068] Modules can provide information to, and receive information from, other modules. For example, the described modules may be regarded as being communicatively coupled. Where multiples of such hardware modules exist contemporaneously, communications may he achieved through signal transmission (e.g., over appropriate circuits and buses) that connect the modules. In embodiments in which multiple modules are configured or instantiated at different times, communications between such modules may be achieved, for example, through the storage and retrieval of information in memory structures to which the multiple modules have access. For example, one module may perform an operation and store the output of that operation in a memory device to which it is communicatively coupled. A further module may then, at a later time, access the memory device to retrieve and process the stored output. Modules may also initiate communications with input or output devices, and can operate on a resource (e.g., a collection of information).

[0069] The various operations of example methods described herein may be performed, at least partially, by one or more processors 162 that are temporarily configured (e.g., by software, code, and/or instructions stored in a machine-readable medium) or permanently configured to perform the relevant operations. Whether temporarily or permanently configured, such processors 162 may constitute processor-implemented (or computer-implemented) modules that operate to perform one or more operations or functions . The modules referred to herein may, ̀m some example embodiments, comprise processor-imp! emented (or computer-implemented) modules,

[0070] Moreover, the methods described herein may be at least partially processor-implemented (or computer-implemented) and/or processor-executable (or computer-executable). For example, at least some of the operations of a method may he performed by one or more processors 162 or processor-implemented (or computer-implemented) modules. Similarly., at least some of the operations of a method may be governed by instructions that are stored in a computer readable storage medium and executed by one or more processors 162 or processor-Implemented (or computer-implemented) modules. The performance of certain of the operations may be distributed among the one or more processors 162, not only residing within a single machine, but deployed across a number of machines. in some example embodiments, the processors 1002 may be located in a single location (e.g., within a home environment, an office envi o nment or as a server farm), while in other embodiments the processors 162 may be distributed across a number of locations.

[0071] While the embodiments) is (are) described with reference to various implementations and exploitations, it will be understood that these embodiments are illustrative and that the scope of the embodiments) is not limited to them. In general, the embodiments described herein may be implemented with facilities consistent with any hardware system or hardware systems defined herein. Many variations, modifications, additions, and improvements are possible.

[0072] Plural instances may be provided for components, operations or structures described herein as a single instance. Finally, boundaries between various components, operations, and data stores are somewhat arbitrary, and particular operations are illustrated in the context of specific illustrative configurations. Other allocations of functionality are envisioned and may fail within the scope of the embodiment s). In general, structures and functionality presented as separate components in the exemplary configurations may be implemented as a combined structure or component. Similarly, structures and functionality presented as a single component may be implemented as separate components. These and other variations, modifications, additions, and .improvements fall within the scope of the embodiment(s).

[0073] As used herein any reference to "one embodiment" or "an embodiment" means that a particular element, feature, structure, or characteristic described in connection with the embodiment is included in at least one embodiment. The appearances of the phrase "in one embodiment" in various places in the specification are not necessarily all referring to the same embodiment.

[0074] Some embodiments may be described using the expression "coupled" and "connected" along with their derivatives. It should be understood that these terms are not intended as synonyms for each other. For example, some embodiments may be described using the term "connected" to indicate that two or more elements are in direct physical or electrical contact with each other. in another example, some embodiments may be described using the term "coupled" to indicate that two or more elements are in direct physical or electrical contact. The term "coupled." however, may also mean that two or more elements are not in direct contact with each other, but yet still co-operate or interact with each other. The embodiments are not limited in this context.

[0075] As used herein, the terms "comprises," "comprising," "includes," "including," "has," "having" or any other variation thereof, are intended to cover a non-exclusive inclusion. For example, a process, method, article, or apparatus that comprises a list of elements is not necessarily limited to only those dements but may include other elements not expressly listed or inherent to such process, method, article, or apparatus Further, unless expressly stated to the contrary, "or" refers to an inclusive or and not to an exclusive or. For example, a condition A or B is satisfied by any one of the following: A is true (or present) and B is false (or not present), A is false (or not present) and B is true (or present), and both A and 8 are true (or present).

[0076] The terms "&" or "an," as used herein, are defined as one or more than one. The term "plurality," as used herein, is defined as two or more than two. The term "another," as used herein, is defined as at least a second or more.

[0077] The foregoing description, for purpose of explanation, has been described with reference to specific embodiments. However, the illustrative discussions above are not intended to be exhaustive or to limit the embodiments to the precise forms disclosed. Many modifications and variations are possible in view of the above teachings. The embodiments were chosen and described in order to best explain the principles and its practical applications, to thereby enable others skilled in the art to best utilize the embodiments and various embodiments with various modifications as are suited to the particular use contemplated.

CLAIMS

What is claimed and desired to be secured by Letters .Patent of the United States is:

1. A computer-implemeuted method for sentiment and influential analysis, comprising:

receiving, by a processor, a plurality electronic messages posted by one or more users on social media web websites,

identifying, by a processor, a polarity of the sentiment-bearing keywords for each electronic message using a phase transition formula,

determining, by a processor, at least one category corresponding to the at least one sentiment-bearing keyword associated with each electronic message, and

determining, by a processor, an influence attribute for each electronic message based on a.plurality of influence factors.

2. The method of claim 1, prior to the receiving step, further comprising crawling, by a processor, a plurality of social media websites to obtain electronic messages.

3. The method of claim .1 , prior to the receiving step, further comprising crawling, by a processor, a plurality of websites to obtain metadata from social media websites.

4. The method of claim i, after the determining at least one category step, determining at least one sentiment corresponding to the at. least one category based on the at least one sentiment-bearing keyword.

5. The method of claim 1, wherein the plurality of influence factors in the influence attribute comprise determining the number of people following an author associated with each message.

6. The method of claim I, wherein the plurality of influence factors in the influence attribute comprise the number of electronic messages that an author has created

7. The method of claim 1, wherein the plurality of influence factors in the influence attribute comprise the number of resending or forwarding a particular electronic message.

8. The method of claim 1, wherein the plurality of influence factors in the influence attribute comprise the number of people replying to a particular electronic message.

9. the method of claim 1, wherein the plurality of influence factors in the influence attribute comprise extracting information from the social media websites the number of people that expressed liking a particular electronic message.

10. the method of claim 1, wherein the plurality of influence factors in the influence attribute comprise determining the number of people an author of a particular electronic message is following.

11. The method of claim 1, wherein the electronic messages comprises message feeds from the social media websites.

12. The method of claim 1, wherein the extracting step comprises filtering the sentiment-bearing keywords with sentiment eliminators.

13. The method claim 1, wherein the extracting step comprises filtering the sentiment-bearing keywords by associating with a sentiment score.

14. The method of claim 1, wherein extracting step comprises extracting opinion bearing keywords from social media content;

for each keyword,

calculating a frequency, f, of the keyword in the plurality of documents and a number of documents, A', that include the keyword;

using the phase transition formula to calculate the relevancy of the keyword based on the frequency of the keyword in the plurality of documents and the number of documents that include the keyword; and

adding the keyword to the list of keywords when the relevancy of the keyword exceeds a predetermined threshold.

15. The computer-implemented method of claim 4, wherein the phase transition formula is $\frac{1}{N^x}$, wherein $x \geq$ I.

16. The method of claim 1, wherein prior to determining step, the method farther comprises generating the list of categories by:

determining pairs of keywords in the list of keywords that are related to each other, wherein the pairs of keywords are unique pairs of keywords;

Identifying sets of the pairs of the key words in which each set includes at least one keyword that is common to all of the pairs of keywords in the set; and

until a predetermined termination condition is achieved, iterative!}' combining the set of the pairs of keywords in which each combined set includes at least one key word that is common to all of the pairs of keywords in the combined set.

17. The method of claim 1, wherein determining the at least one category corresponding to the at least one keyword of the sentence includes using a neural network to determine the at least one category corresponding to the at least one keyword of the sentence.

18. The method of claim I, wherein determining the at least one category corresponding to the at least one keyword of the sentence includes:

obtaining a plurality of category spectrum s, a respective category spectrum including a frequency of occurrence of keywords in the list of keywords that corresponds to a respective category;

determining a category spectrum for the sentence based on the at least one keyword;

calculating dot products of the category spectrum for the sentence and each category spectrum in the plurality of category spectrum s: and

determining the at least one category as a category corresponding to at least one dot product that exceeds a predetermined threshold.

19. The method of claim 18, wherein prior to obtaining the plurality of category spectrums, the method further comprises for each category, determining a category spectrum for the category by:

obtaining a corpus of documents corresponding to the category ;

extracting keywords from each document in the corpus of documents;

filtering the keywords using the phase transition formula to produce filtered keywords;

determining the frequency of occurrence of the filtered keywords in the corpus of documents; and

normalizing the frequency of occurrence of the filtered keywords to produce the category spectrum for the category.

20. A system to determine sentiment expressed in a document, comprising:

at least one processor;

memory; and

at least one program stored in the memory, the at least one program comprising instructions to:

receiving, by a processor, a plurality electronic messages posted by one or more users on social media web websites;

identifying, by a processor, a polarity of the sentiment-bearing keywords for each electronic message using a phase transition formula;

determining, by a processor, at least one category corresponding to the at least one sentiment-bearing keyword associated with each electronic message; and

determining, by a processor, an influence attribute for each electronic message based on a plurality of influence factors.

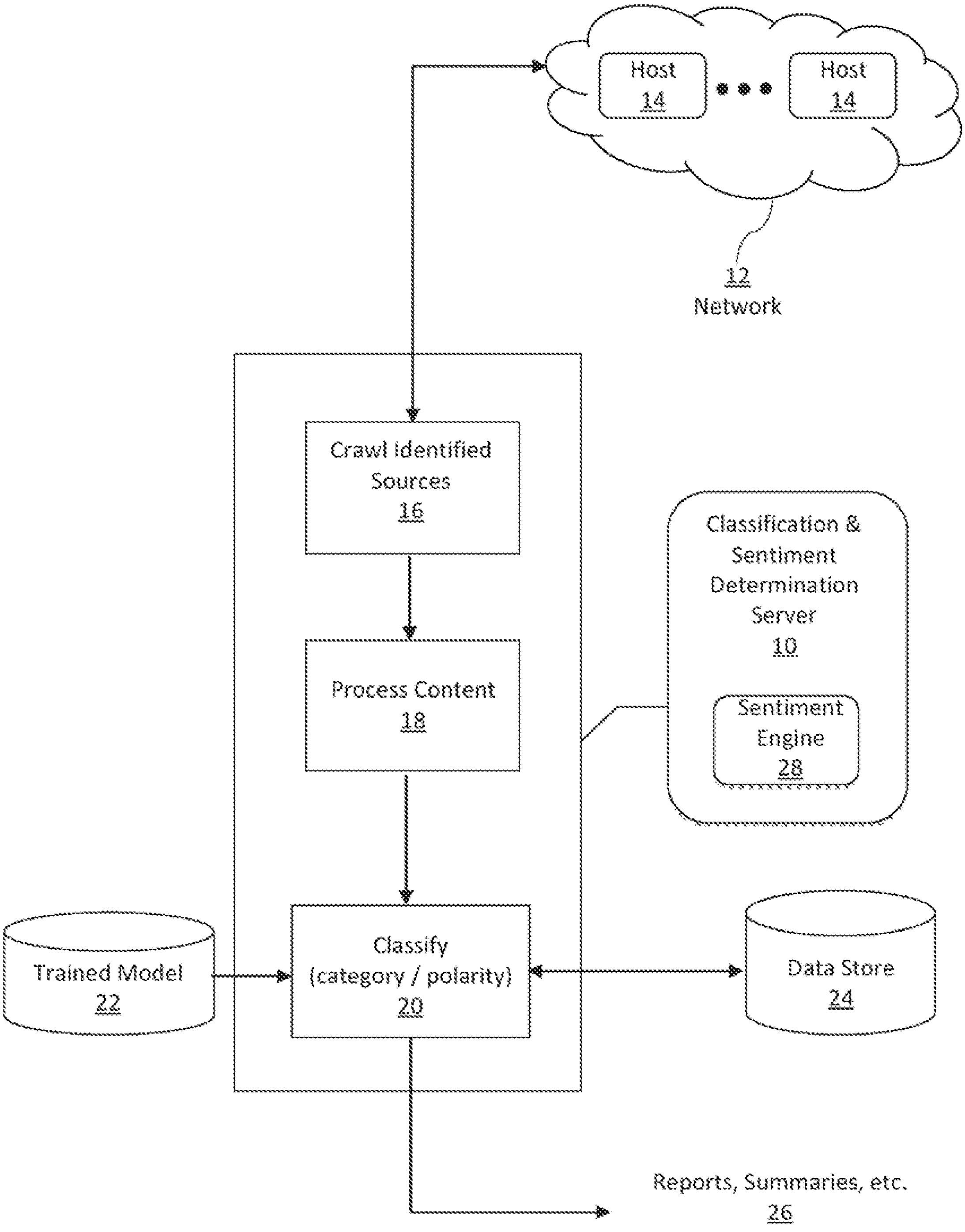

Host
14
Host
14
12
Network
Crawl Identified
Sources
16
Process Content
18
Classification &
Sentiment
Determination
Server
10
Sentiment
Engine
28
Trained Model
22
Classify
(category / polarity)
20
Data Store
24
Reports, Summaries, etc.
26

FIG. 1

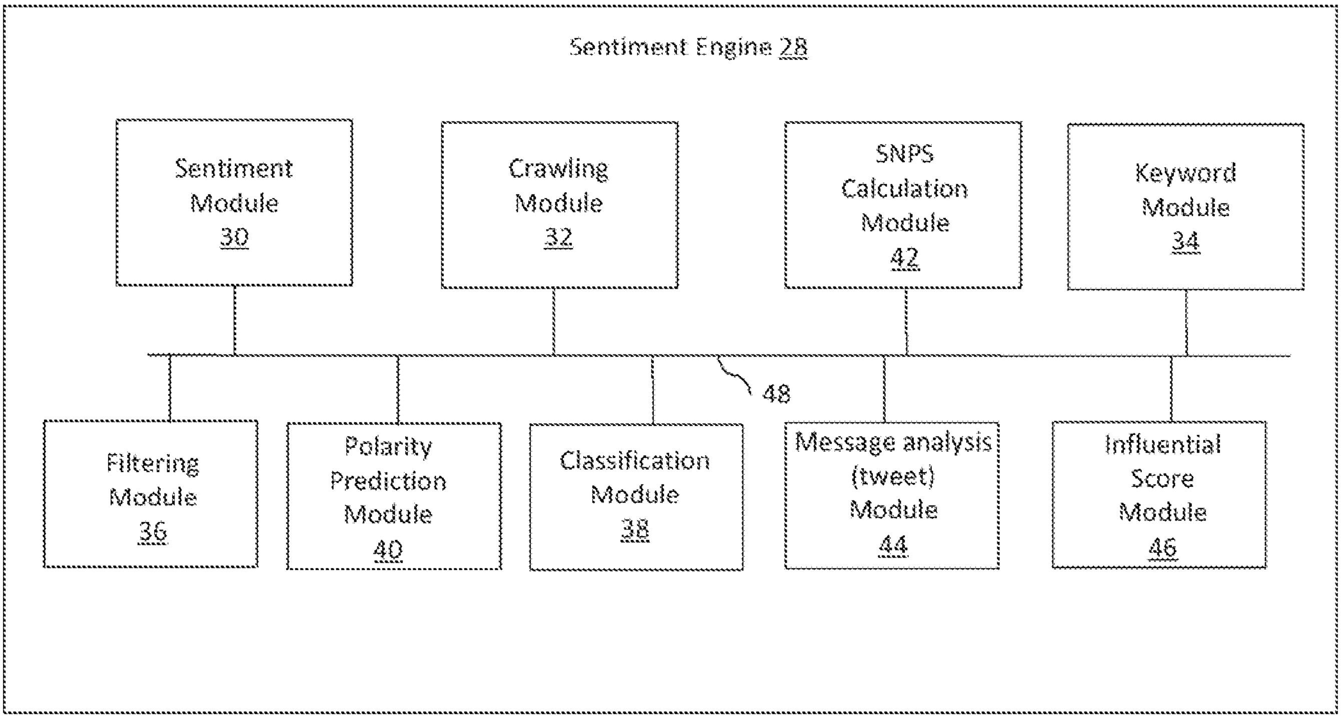

Sentiment Engine 28
Sentiment Module 30
Crawling Module 32
SNPS Calculation Module 42
Keyword Module 34
48
Filtering Module 36
Polarity Prediction Module 40
Classification Module 38
Message analysis (tweet) Module 44
Influential Score Module 46

FIG. 2

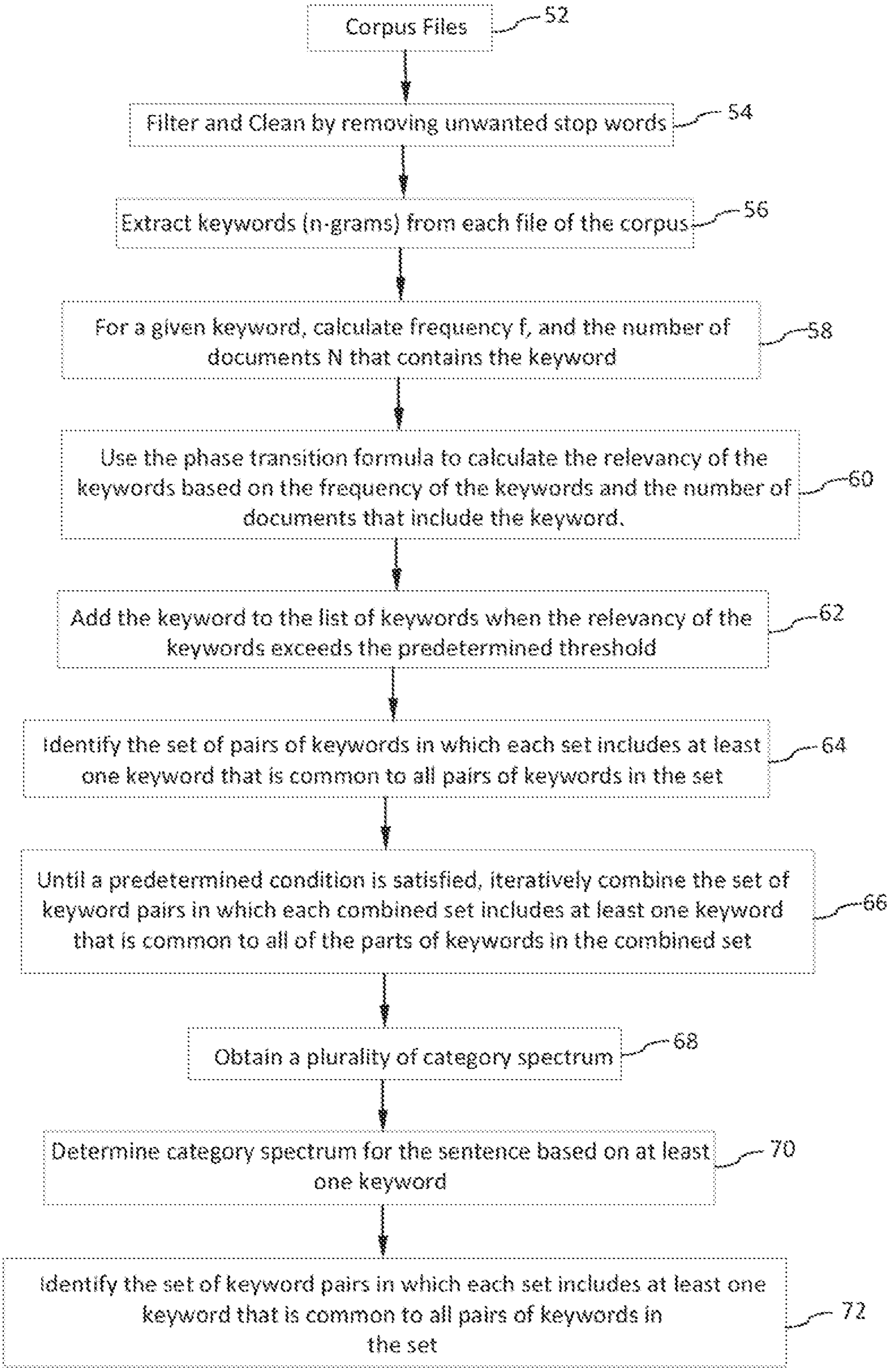

Corpus Files
52
Filter and Clean by removing unwanted stop words
54
Extract keywords (n-grams) from each file of the corpus
56
For a given keyword, calculate frequency f, and the number of documents N that contains the keyword
58
Use the phase transition formula to calculate the relevancy of the keywords based on the frequency of the keywords and the number of documents that include the keyword.
60
Add the keyword to the list of keywords when the relevancy of the keywords exceeds the predetermined threshold
62
Identify the set of pairs of keywords in which each set includes at least one keyword that is common to all pairs of keywords in the set
64
Until a predetermined condition is satisfied, iteratively combine the set of keyword pairs in which each combined set includes at least one keyword that is common to all of the parts of keywords in the combined set
66
Obtain a plurality of category spectrum
68
Determine category spectrum for the sentence based on at least one keyword
70
Identify the set of keyword pairs in which each set includes at least one keyword that is common to all pairs of keywords in the set
72
50

FIG. 3

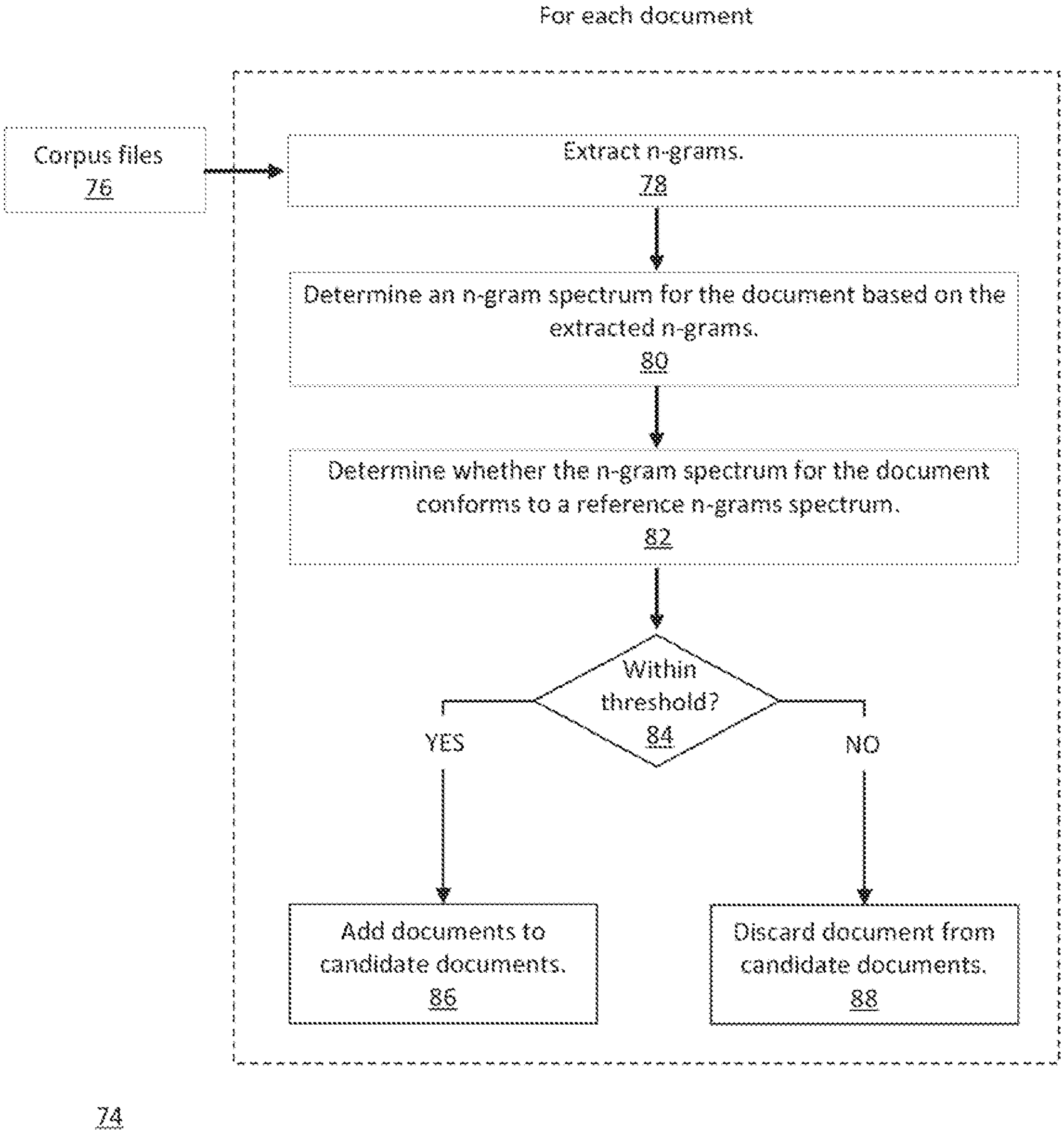

For each document
Corpus files
76
Extract n-grams.
78
Determine an n-gram spectrum for the document based on the extracted n-grams.
80
Determine whether the n-gram spectrum for the document conforms to a reference n-grams spectrum.
82
Within threshold?
84
YES
NO
Add documents to candidate documents.
86
Discard document from candidate documents.
88
74

FIG. 4

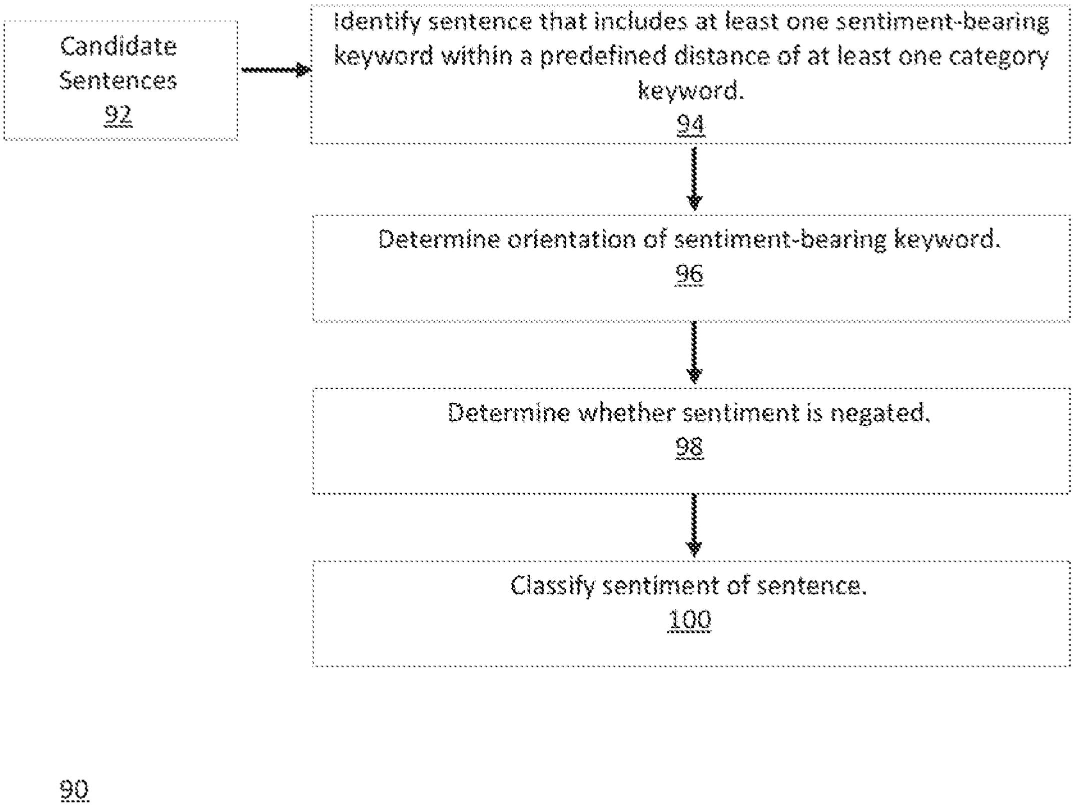
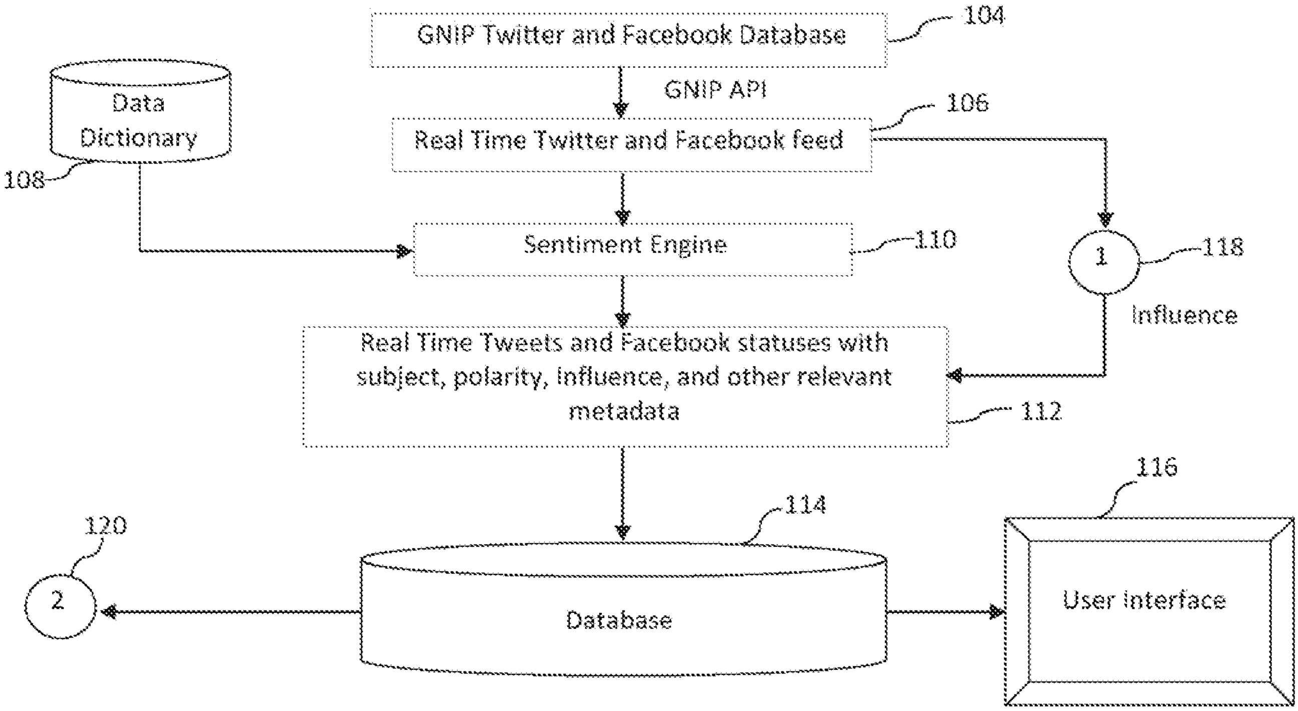

Candidate Sentences
92
Identify sentence that includes at least one sentiment-bearing keyword within a predefined distance of at least one category keyword.
94
Determine orientation of sentiment-bearing keyword.
96
Determine whether sentiment is negated.
98
Classify sentiment of sentence.
100
90

FIG. 5


GNIP Twitter and Facebook Database
104
GNIP API
Data Dictionary
108
Real Time Twitter and Facebook feed
106
Sentiment Engine
110
1
118
Influence
Real Time Tweets and Facebook statuses with subject, polarity, influence, and other relevant metadata
112
116
120
114
2
Database
User interface

102

FIG. 6A

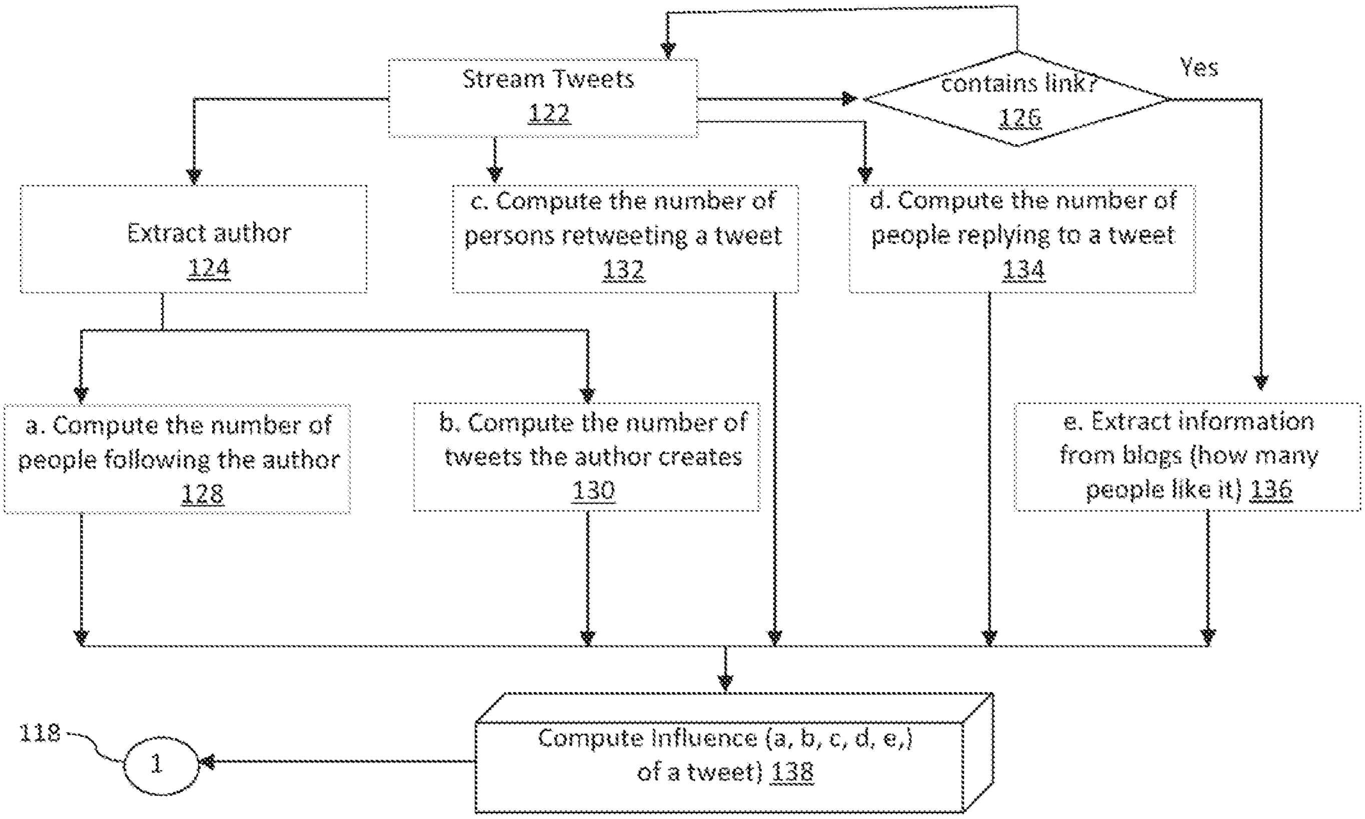

Stream Tweets
122
contains link?
126
Yes
Extract author
124
c. Compute the number of persons retweeting a tweet
132
d. Compute the number of people replying to a tweet
134
a. Compute the number of people following the author
128
b. Compute the number of tweets the author creates
130
e. Extract information from blogs (how many people like it) 136
Compute influence (a, b, c, d, e,) of a tweet) 138
118
1

FIG. 6B

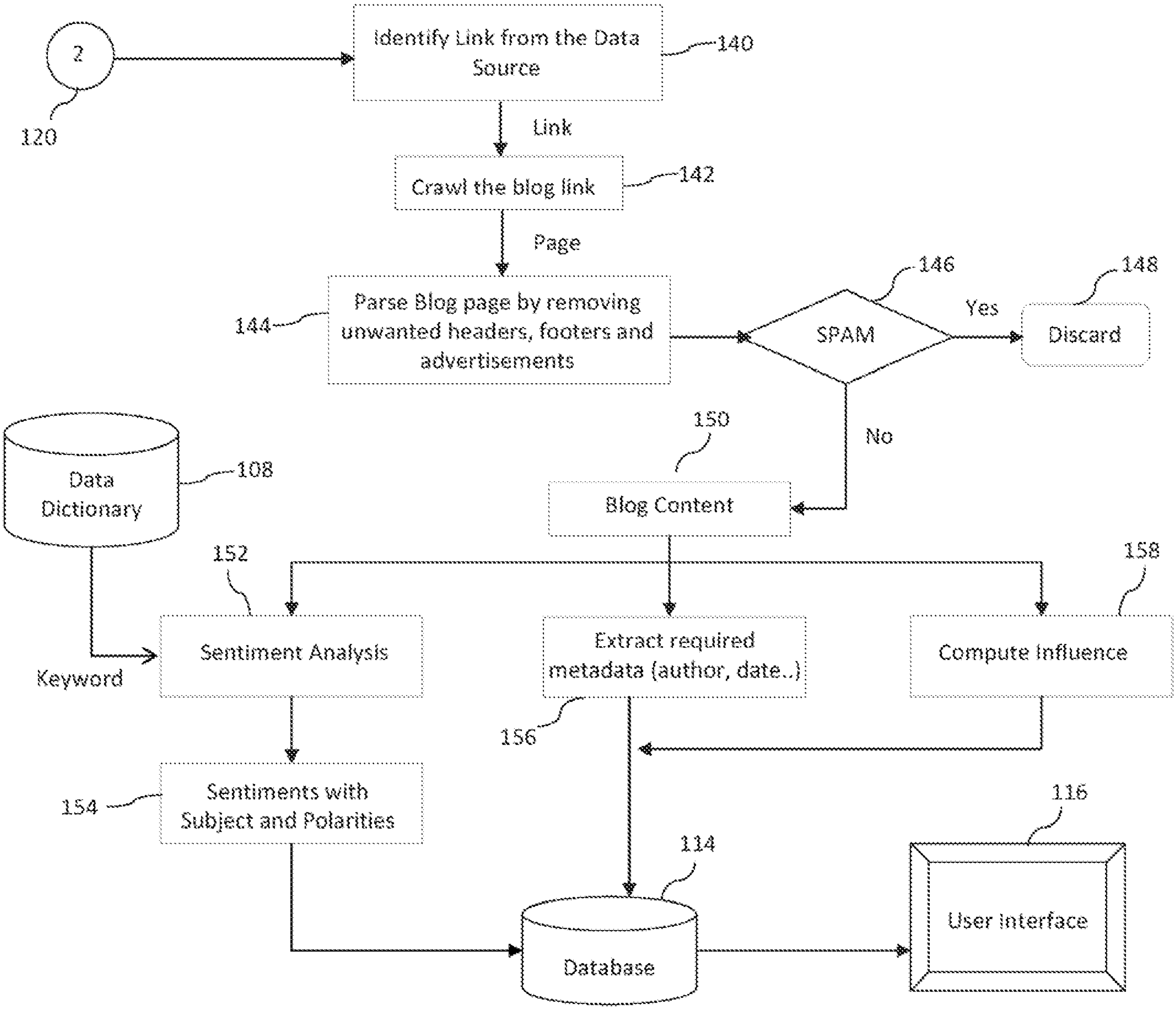

2
120
Identify Link from the Data Source
140
Link
Crawl the blog link
142
Page
Parse Blog page by removing unwanted headers, footers and advertisements
144
SPAM
146
Yes
Discard
148
No
Data Dictionary
108
150
Blog Content
152
158
Sentiment Analysis
Keyword
Extract required metadata (author, date..)
156
Compute Influence
Sentiments with Subject and Polarities
154
116
114
Database
User Interface

FIG. 6C

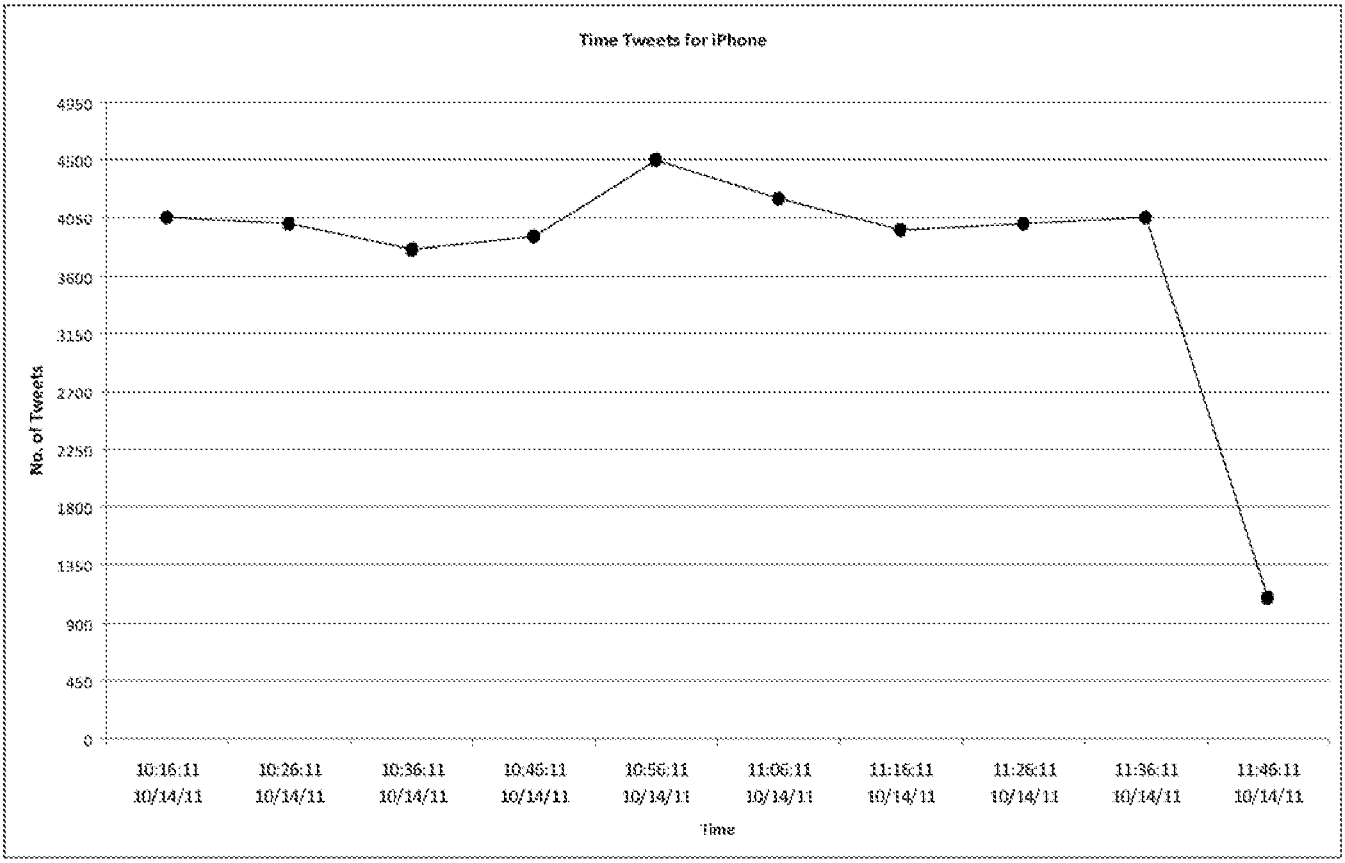
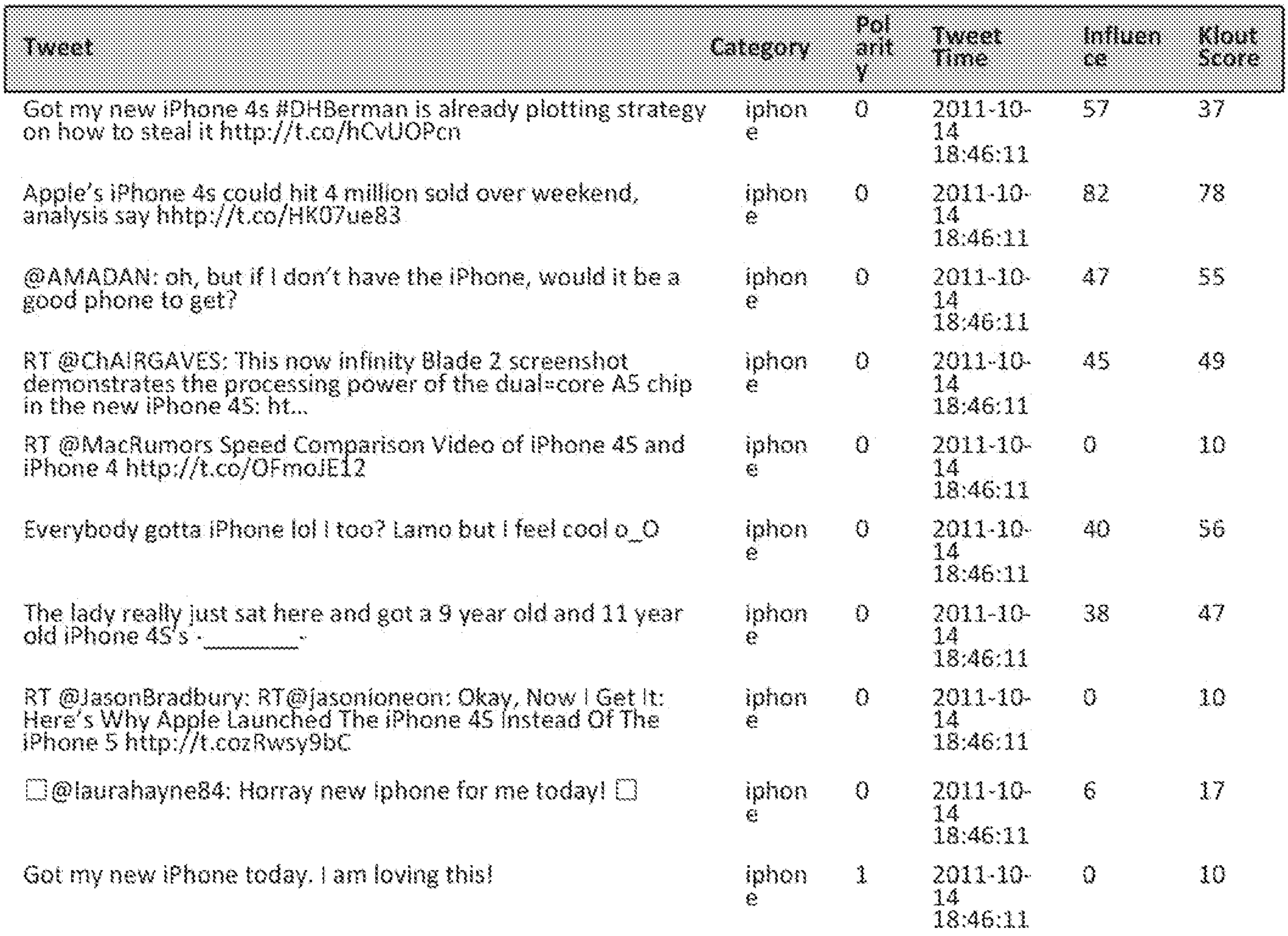

Time Tweets for iPhone
No. of Tweets
4950
4500
4050
3600
3150
2700
2250
1800
1350
900
450
0
10:16:11 10/14/11
10:26:11 10/14/11
10:36:11 10/14/11
10:46:11 10/14/11
10:56:11 10/14/11
11:06:11 10/14/11
11:16:11 10/14/11
11:26:11 10/14/11
11:36:11 10/14/11
11:46:11 10/14/11
Time

FIG. 7A



Tweet	Category	Polarity	Tweet Time	Influence	Klout Score
Got my new iPhone 4s #DHBerman is already plotting strategy on how to steal it http://t.co/hCvUOPcn	iphone	0	2011-10-14 18:46:11	57	37
Apple's iPhone 4s could hit 4 million sold over weekend, analysis say hhtp://t.co/HK07ue83	iphone	0	2011-10-14 18:46:11	82	78
@AMADAN: oh, but if I don't have the iPhone, would it be a good phone to get?	iphone	0	2011-10-14 18:46:11	47	55
RT @ChAIRGAVES: This now infinity Blade 2 screenshot demonstrates the processing power of the dual=core A5 chip in the new iPhone 4S: ht...	iphone	0	2011-10-14 18:46:11	45	49
RT @MacRumors Speed Comparison Video of iPhone 4S and iPhone 4 http://t.co/OFmoJE12	iphone	0	2011-10-14 18:46:11	0	10
Everybody gotta iPhone lol I too? Lamo but I feel cool o_O	iphone	0	2011-10-14 18:46:11	40	56
The lady really just sat here and got a 9 year old and 11 year old iPhone 4S's -________-	iphone	0	2011-10-14 18:46:11	38	47
RT @JasonBradbury: RT@jasonioneon: Okay, Now I Get It: Here's Why Apple Launched The iPhone 4S Instead Of The iPhone 5 http://t.cozRwsy9bC	iphone	0	2011-10-14 18:46:11	0	10
□@laurahayne84: Horray new iphone for me today! □	iphone	0	2011-10-14 18:46:11	6	17
Got my new iPhone today. I am loving this!	iphone	1	2011-10-14 18:46:11	0	10

FIG. 7B

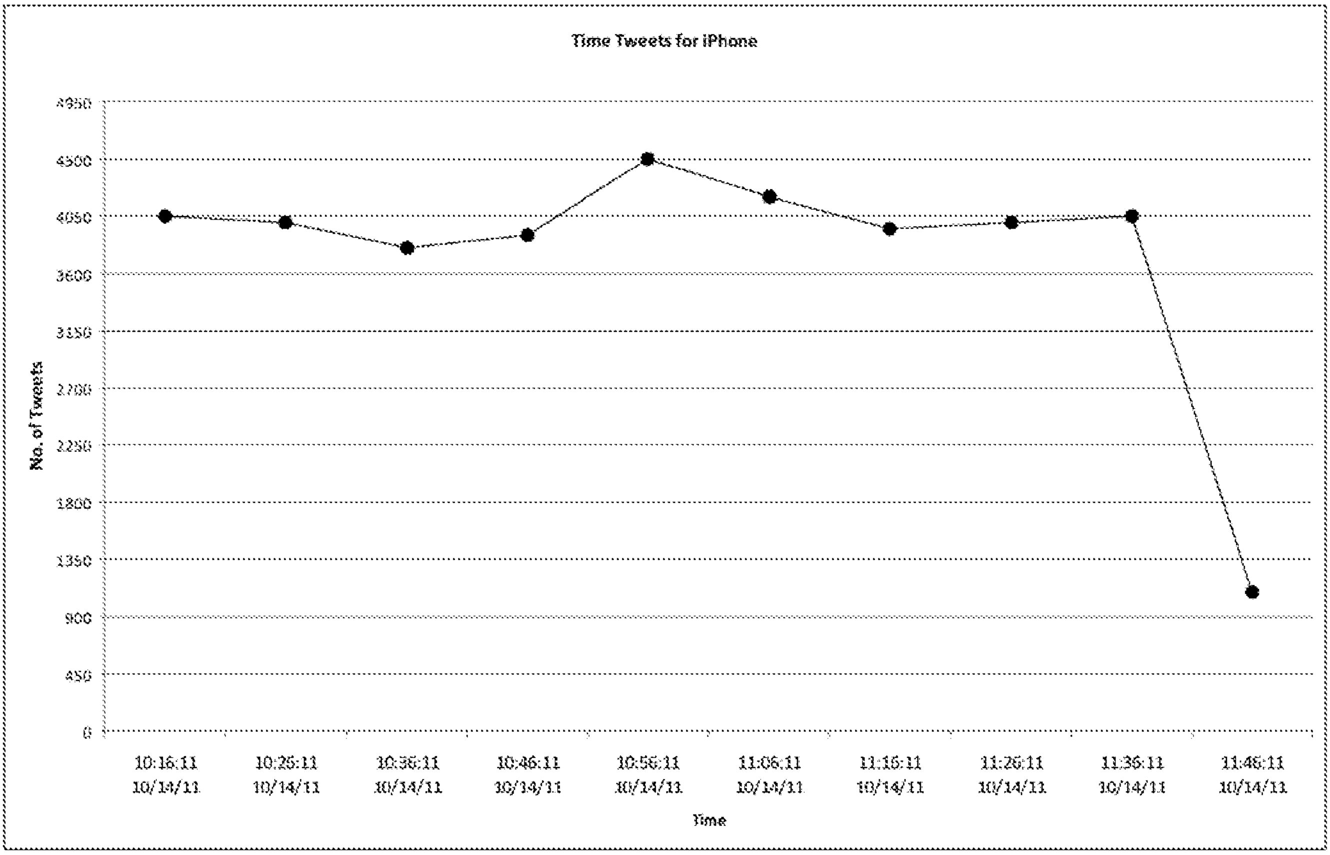

Time Tweets for iPhone
No. of Tweets
4950
4500
4050
3600
3150
2700
2250
1800
1350
900
450
0
10:16:11 10/14/11
10:26:11 10/14/11
10:36:11 10/14/11
10:46:11 10/14/11
10:56:11 10/14/11
11:06:11 10/14/11
11:16:11 10/14/11
11:26:11 10/14/11
11:36:11 10/14/11
11:46:11 10/14/11
Time

FIG. 8

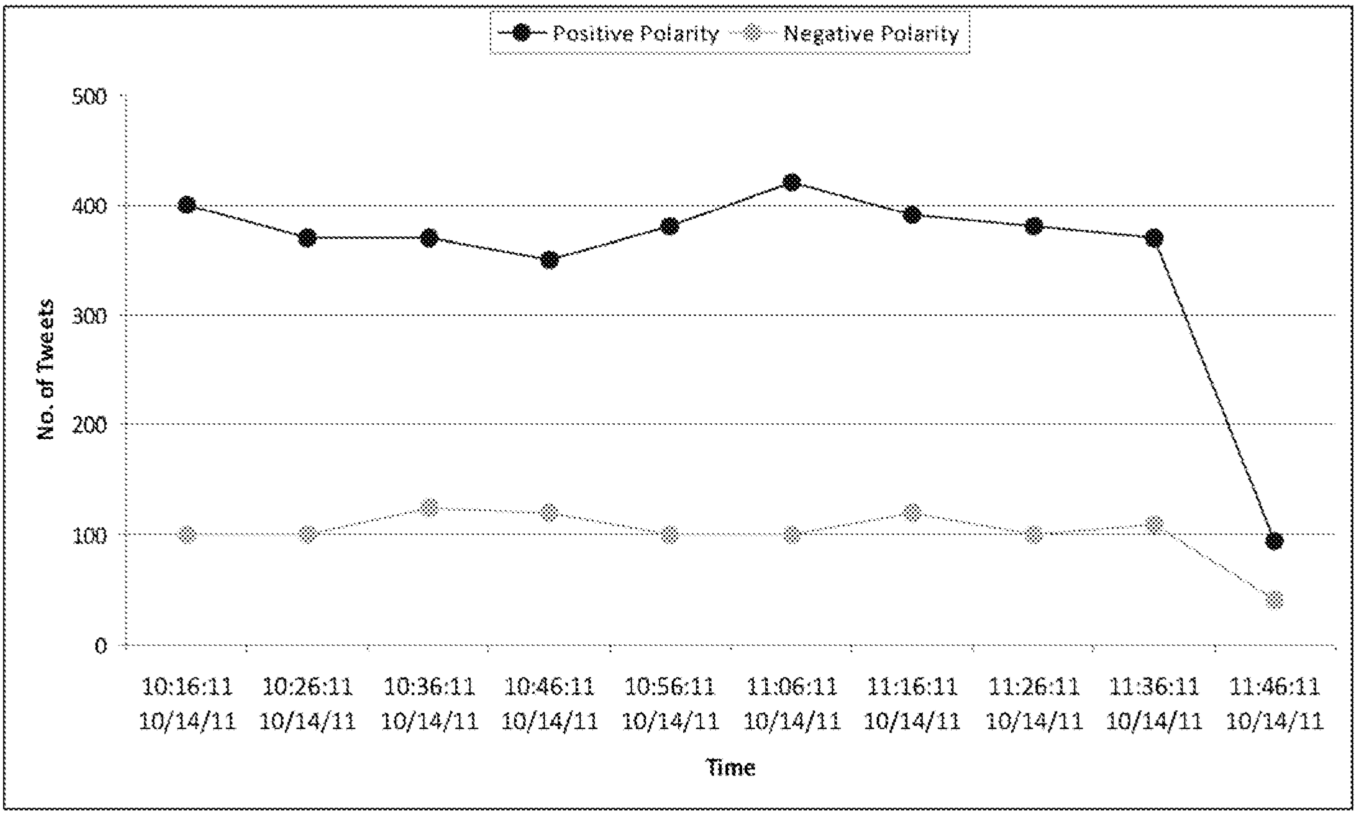
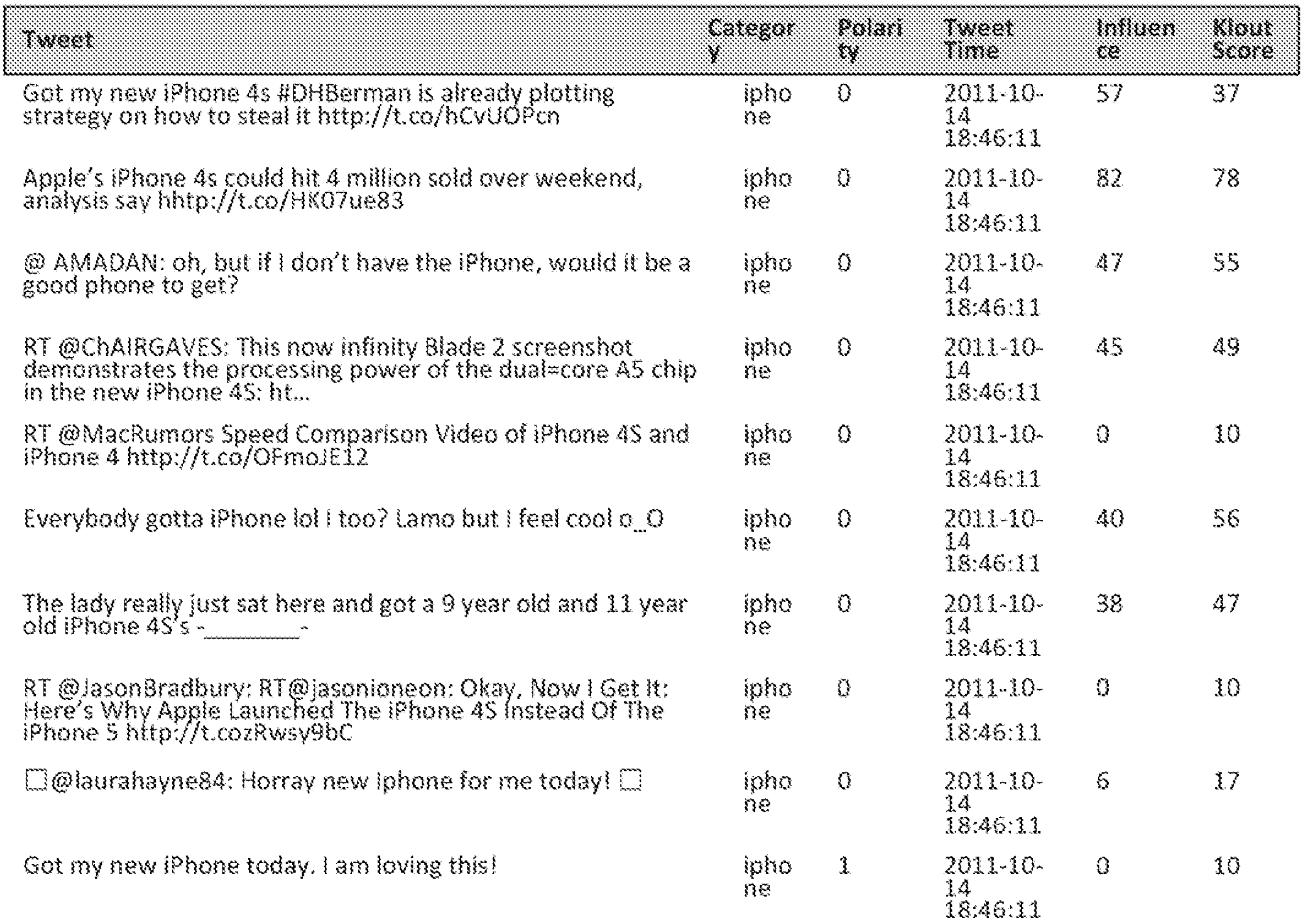

Positive Polarity
Negative Polarity
500
400
300
200
100
0
No. of Tweets
10:16:11 10/14/11
10:26:11 10/14/11
10:36:11 10/14/11
10:46:11 10/14/11
10:56:11 10/14/11
11:06:11 10/14/11
11:16:11 10/14/11
11:26:11 10/14/11
11:36:11 10/14/11
11:46:11 10/14/11
Time

FIG. 9A



Tweet	Category	Polarity	Tweet Time	Influence	Klout Score
Got my new iPhone 4s #DHBerman is already plotting strategy on how to steal it http://t.co/hCvUOPcn	iphone	0	2011-10-14 18:46:11	57	37
Apple's iPhone 4s could hit 4 million sold over weekend, analysis say hhtp://t.co/HK07ue83	iphone	0	2011-10-14 18:46:11	82	78
@ AMADAN: oh, but if I don't have the iPhone, would it be a good phone to get?	iphone	0	2011-10-14 18:46:11	47	55
RT @ChAIRGAVES: This now infinity Blade 2 screenshot demonstrates the processing power of the dual=core A5 chip in the new iPhone 4S: ht...	iphone	0	2011-10-14 18:46:11	45	49
RT @MacRumors Speed Comparison Video of iPhone 4S and iPhone 4 http://t.co/OFmoJE12	iphone	0	2011-10-14 18:46:11	0	10
Everybody gotta iPhone lol I too? Lamo but I feel cool o_O	iphone	0	2011-10-14 18:46:11	40	56
The lady really just sat here and got a 9 year old and 11 year old iPhone 4S's -_________-	iphone	0	2011-10-14 18:46:11	38	47
RT @JasonBradbury: RT@jasonioneon: Okay, Now I Get It: Here's Why Apple Launched The iPhone 4S Instead Of The iPhone 5 http://t.cozRwsy9bC	iphone	0	2011-10-14 18:46:11	0	10
□@laurahayne84: Horray new iphone for me today! □	iphone	0	2011-10-14 18:46:11	6	17
Got my new iPhone today. I am loving this!	iphone	1	2011-10-14 18:46:11	0	10

FIG. 9B

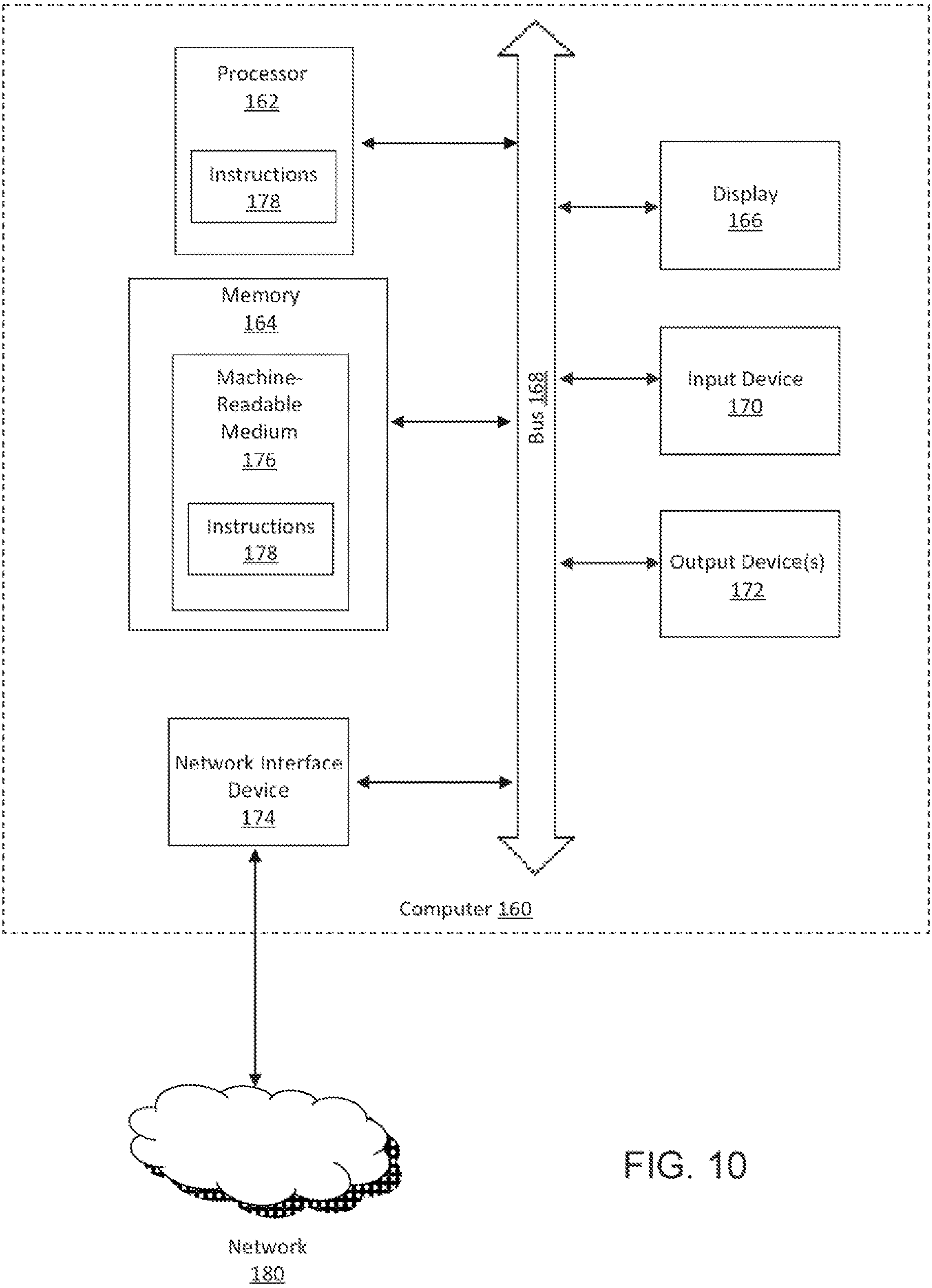

Processor
162
Instructions
178
Memory
164
Machine-
Readable
Medium
176
Instructions
178
Network Interface
Device
174
Bus 168
Display
166
Input Device
170
Output Device(s)
172
Computer 160
Network
180

FIG. 10

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2012/060573

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/00(2006.01)i, G06F 17/20(2006.01)1, G06Q 50/30(2012.01)1

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 17/00; G06F 3/048; G06F 15/16; G06F 17/30; G06Q 30/00; G06F 17/27; G06Q 10/00; G06N 5/02

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: receive, website, message, identify, sentiment, polarity, formular, influence, factor, and similar terms.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2011-0161071 Al (DUONG-VAN, MINH) 30 June 2011 See paragraphs [0021H0032] and [0035] -[0042] ; claims 1-3 , 6, and 8-10 ; and 14-16 ; and figures 2-4 and 6-8 .	1-11 , 16-20
Y	US 2010-0070485 Al (PARSONS, TODD A. et al.) 18 March 2010 See paragraphs [0043]-[0050] and [0088] -[0094] ; claims 18 and 20; and figures 2A, 3A, and 7D.	1-11 , 16-20
A	US 2008-0133488 Al (BANDARU, NAGARAJU et al.) 05 June 2008 See paragraphs [0034]-[0040] ; claims 1 and 11; and figure 1.	1-11 , 16-20
A	US 2009-0319342 Al (SHILMAN, MICHAEL et al.) 24 December 2009 See paragraphs [0051]-[0052] ; claims 1 and 7; and figure 7.	1-11 , 16-20
A	US 2011-0125793 Al (ERHART, GEORGE et al.) 26 May 2011 See paragraphs [0067]-[0076] ; claims 1 and 4; and figure 5A.	1-11 , 16-20
A	KR 10-2009-0068803 A (HAN, SOUNG J00) 29 June 2009 See paragraphs [0029]-[0049] ; claim 3; and figure 2.	1-11 , 16-20

□ 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 29 March 2013 (29.03.2013)	Date of mailing of the international search report **29 March 2013 (29.03.2013)**
Name and mailing address of the ISA/KR Korean Intellectual Property Office 189 Cheongsa-ro, Seo-gu, Daejeon Metropolitan City, 302-70 1, Republic of Korea Facsimile No. 82-42-472-7140	Authorized officer NHO, Ji Myong Telephone No. 82-42-481-8528

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. □ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☒ Claims Nos.: **12-15**
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

Claims 12-15 are too unclear to make meaningful search.

3. □ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. Ill Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. □ As all required addtional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. □ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. □ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. □ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

□ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

□ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

□ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2012/060573

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2011-0161071 A1	30. 06,,2011	Wo 2011-079311 A1	30. 06,,2011
us 2010-0070485 A1	18. 03,,2010	AU 2007-219997 A1	07. 09,,2007
		EP 1989639 A2	12. 11,,2008
		JP 2009-528639 A	06. 08,,2009
		US 2007-0214097 A1	13. 09,,2007
		US 2009-0119173 A1	07. 05,,2009
		Wo 2007-101263 A2	07. 09,,2007
us 2008-0133488 A1	05. 06,,2008	EP 2095219 A2	02. 09,,2009
		EP 2095219 A4	03. 02,,2010
		US 7930302 B2	19. 04,,2011
		WO 2008-066675 A2	05. 06,,2008
		wo 2008-066675 A3	31. 07,,2008
us 2009-0319342 A1	24. 12,,2009	AU 2009-260033 A1	23. 12,,2009
		EP 2304660 A2	06. 04,,2011
		JP 2011-530729 A	22. 12,,2011
		WO 2009-155375 A2	23. 12,,2009
us 2011-0125793 A1	26. 05,,2011	EP 2328328 A2	01. 06,,2011
		EP 2328328 A3	17. 08,,2011
		GB 2477839 A	17. 08,,2011
		GB 2479825 A	26. 10,,2011
		US 2011-0123015 A1	26. 05,,2011
		US 2011-0125550 A1	26. 05,,2011
		US 2011-0125580 A1	26. 05,,2011
		US 2011-0125697 A1	26. 05,,2011
		US 2011-0125826 A1	26. 05,,2011
		US 8331550 B2	11. 12,,2012
KR 10-2009-0068803 A	29. 06,,2009	US 2010-0262597 A1	14. 10,,2010
		WO 2009-082100 A2	02. 07,,2009