



- (51) **International Patent Classification:**
G06Q 30/00 (2012.01)
- (21) **International Application Number:**
PCT/SG2014/000508
- (22) **International Filing Date:**
30 October 2014 (30.10.2014)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
201308079-1 30 October 2013 (30.10.2013) SG
- (71) **Applicant:** NATIONAL UNIVERSITY OF SINGAPORE [SG/SG]; 21 Lower Kent Ridge Road, Singapore 119077 (SG).
- (72) **Inventors:** CHUA, Tat-Seng; c/o National University of Singapore, School of Computing, Department of Computer Science, 21 Lower Kent Ridge Road, Singapore 119077 (SG). AMIRIEBRAHIMABADI, Hadi; c/o National University of Singapore, NUS Graduate School for Interactive Science and Engineering, 21 Lower Kent Ridge Road, Singapore 119077 (SG). CHEN, Yan; c/o National University of Singapore, School of Computing, Department of Computer Science, 21 Lower Kent Ridge Road, Singapore

119077 (SG). CUI, Anqi; c/o National University of Singapore, School of Computing, Department of Computer Science, 21 Lower Kent Ridge Road, Singapore 119077 (SG).

(74) **Agent:** VIERING, JENTSCHURA & PARTNER LLP; P.O. Box 1088, Rochor Post Office, Rochor Road, Singapore 911833 (SG).

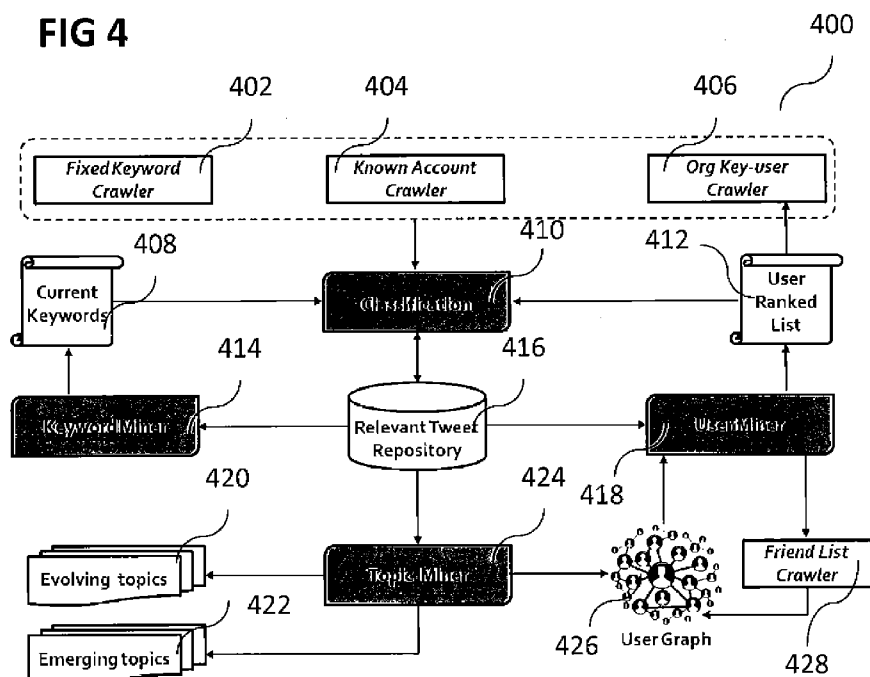
(81) **Designated States** (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, 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, SA, 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, ST, 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,

[Continued on next page]

(54) **Title:** MICROPOST BRAND MONITORING

FIG 4



(57) **Abstract:** Monitoring of social media and microblogs for information relevant to company brands and associated information by the use of web crawlers. A unified framework of fixed and dynamic keywords, known accounts, key users and friend lists are used to identify relevant microposts, tweets, on a particular organisation of interest. This organisational information is classified by relevancy, using learning algorithms, and emerging and evolving topics are identified.

WO 2015/065290 A1



LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, **Published:**

SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, — *with international search report (Art. 21(3))*
GW, KM, ML, MR, NE, SN, TD, TG).

MICROPOST BRAND MONITORING

Cross-reference to Related Applications

[0001] The present application claims the benefit of the Singapore patent application No. 201308079-1 filed on 30 October 2013, the entire contents of which are incorporated herein by reference for all purposes.

Technical Field

[0002] Embodiments relate generally to information determination devices and information determination methods.

Background

[0003] Social media portals like Twitter, Facebook, and various forum sites contain the everyday thoughts, opinions, and experiences of their online users. Parts of these user generated contents (UGCs) reflect and reveal information about organizations such as the companies, banks, government organizations, and universities etc. Thus, there may be a need for analyzing this data.

Summary

[0004] According to various embodiments, an information determination device may be provided. The information determination device may include: an account crawler

configured to determine data from at least one pre-determined user account; and an information determiner configured to determine information related to a pre-determined organization based on the data.

[0005] According to various embodiments, an information determination method may be provided. The information determination method may include: determining data from at least one pre-determined user account; and determining information related to a pre-determined organization based on the data.

Brief Description of the Drawings

[0006] In the drawings, like reference characters generally refer to the same parts throughout the different views. The drawings are not necessarily to scale, emphasis instead generally being placed upon illustrating the principles of the invention. In the following description, various embodiments are described with reference to the following drawings, in which:

FIG. 1A and FIG. 1B show information determination devices in accordance with various embodiments;

FIG. 1C shows a flow diagram illustrating an information determination method according to various embodiments;

FIG. 2 shows an illustration of websites;

FIG. 3 shows an illustration of a power law correlation;

FIG. 4 shows an illustration of an architecture according to various embodiments;

FIG. 5 shows an illustration of learning evolving and emerging topics;

FIG. 6 shows an illustration of a distribution of relevant tweets;

FIG. 7, FIG. 8, and FIG. 9 show illustrations of an effect of learning parameters;
and

FIG. 10 shows an illustration of an effect of the temporal continuity constraint.

Description

[0007] Embodiments described below in context of the devices are analogously valid for the respective methods, and vice versa. Furthermore, it will be understood that the embodiments described below may be combined, for example, a part of one embodiment may be combined with a part of another embodiment.

[0008] In this context, the information determination device as described in this description may include a memory which is for example used in the processing carried out in the information determination device. A memory used in the embodiments may be a volatile memory, for example a DRAM (Dynamic Random Access Memory) or a non-volatile memory, for example a PROM (Programmable Read Only Memory), an EPROM (Erasable PROM), EEPROM (Electrically Erasable PROM), or a flash memory, e.g., a floating gate memory, a charge trapping memory, an MRAM (Magnetoresistive Random Access Memory) or a PCRAM (Phase Change Random Access Memory).

[0009] In an embodiment, a "circuit" may be understood as any kind of a logic implementing entity, which may be special purpose circuitry or a processor executing software stored in a memory, firmware, or any combination thereof. Thus, in an embodiment, a "circuit" may be a hard-wired logic circuit or a programmable logic circuit such as a programmable processor, e.g. a microprocessor (e.g. a Complex Instruction Set Computer (CISC) processor or a Reduced Instruction Set Computer

(RISC) processor). A "circuit" may also be a processor executing software, e.g. any kind of computer program, e.g. a computer program using a virtual machine code such as e.g. Java. Any other kind of implementation of the respective functions which will be described in more detail below may also be understood as a "circuit" in accordance with an alternative embodiment.

[0010] According to various embodiments, devices and methods may be provided for online discovery of events and topics for organizations from social media. Given an organization, a unified framework may be provided to address two issues that have not been tackled to date: (a) crawling more representative distribution of relevant contents and (b) discriminating relevant from irrelevant content for the organization. The current organization or brand monitoring systems use a fixed set of known keywords to crawl micro-posts from social media. This popular strategy results in (a) many missing relevant micro-posts, and (b) many irrelevant micro-posts. The first issue is due to the dynamic nature of the social media contents, while the latter issue is due to the polysemy problem in which the acronyms of organizations are often shared by many entities. For example, NUS is shared between National University of Singapore, National Union of Students and Nu-Skin™ company.

[0011] According to various embodiments, a unified framework may be provided to address the above issues. This framework may utilize multiple aspects of organizations including fixed keywords, known accounts and automatically identified key-users to crawl more relevant data about organizations from social media. Moreover, it effectively may employ content and user information to address the polysemy problem for organizations. Given the automatically identified relevant micro-posts for an organization,

an adaptation of online sparse coding algorithms to efficiently learn the topics through time may be provided. Comprehensive experiments show promising results for three different organizations using streaming data obtained from Twitter.

[0012] According to various embodiments, devices and method for discovering topics related to a given organization by automatic identification of the relevant micro-posts (e.g. tweets) and users through time in the context of Microblogs may be provided.

[0013] According to various embodiments, devices and method for social media analytics, brand monitoring, organization monitoring, topic detection, and/ or topic evolution may be provided.

[0014] According to various embodiments, devices and methods for mining the sense of organizations in social media may be provided. Devices and methods according to various embodiments may be referred to as "OrgSense".

[0015] Live tweet streams have been previously used for topic mining and event detection in general contexts. Also, models of burst and hot topic detection have been developed, from automation to temporal patterns.

[0016] Existing approaches use keywords and hashtags (Hashtags are keywords attached to the # symbol to categorize tweets based on their context) to crawl data. While keyword based approaches work well on mining tweets about specific topics, they are restricted to a set of keywords that are maintained manually. Fixed keywords fail to discover a large fraction of relevant information simply due to missing newly-introduced terms within topics and micro-posts without known keywords. Furthermore, fixed keywords may represent several different entities and results in many irrelevant micro-posts.

[0017] Mining evolving and emerging topics in the social media content has become a hot research topic recently. An approach may be used to identify emergent keywords and to utilize them to find emerging topics. A term may be defined as emergent if it frequently occurs in the current time but not in the previous times. In another approach, it may be focused on individual users and an LDA-based (Linear discriminant analysis) approach (called Temporal-LDA) may be used that learns topic transition in a sequence of tweets posted by the same user and use it to predict the future distribution of the user's tweets. In yet a further approach, the evolution of topics may be tracked through time. It may be shown that a sparse coding algorithm with the non-negativity constraint is effective for topic modeling in the social media context. A continuity constraint may be introduced. Transient crowd, a short-lived collection of people who directly communicate with each other through social messages like reply and mention of Twitter, may be mined.

[0018] In a definition according to various embodiments, users may be part of the same community as long as they share interest on the same topic (such communities may be referred to as interest communities). Commonly used algorithms may not be effective to mine such interest communities as there may not be any direct conversation between the users in these communities.

[0019] While commonly used approaches may be effective in mining topics in general, they are not designed for addressing the lack of representative data or polysemy issues for entities like organizations. According to various embodiments, an effective framework may be provided to elicit representative amount of data about organizations and dealing with the polysemy issue for organizations in social media. Various embodiments may be focused on the above issues in Microblogs.

[0020] FIG. 1A shows an information determination device 100 according to various embodiments. The information determination device 100 may include an account crawler 102 configured to determine data from at least one pre-determined user account. The information determination device 100 may further include an information determiner 104 configured to determine information related to a pre-determined organization based on the data. The account crawler 102 and the information determiner 104 may be connected via a connection 106 (or a plurality of separate connections), for example an electrical or optical connection, for example any kind of cable or bus.

[0021] It will be understood that a crawler, or a Web crawler, as referred to herein, may be software that downloads data automatically from a network, for example from the Internet. A crawler may systematically visit web pages (for example with given configuration and strategy designed by programmers) and download the data of the web pages. The crawler may perform these repetitive tasks at a much higher rate than doing manually. According to various embodiments, the streaming API (application programming interface) of Twitter may be used to crawl tweets. A topic miner may be software that implements a method to discover human-understandable topics from the texts.

[0022] In other words, according to various embodiments, an information determination device may be provided which determines information related to a pre-determined organization based on data which are determined from one or more pre-determined user accounts.

[0023] According to various embodiments, the account crawler 102 may include or may be or may be included in a known account crawler configured to determine the data from at least one account for the organization.

[0024] According to various embodiments, the account crawler 102 may include or may be or may be included in a key-user crawler configured to determine the data from at least one account of a key user.

[0025] FIG. 1B shows an information determination device 108 according to various embodiments. The information determination device 108 may, similar to the information determination device 100 of FIG. 1A, include an account crawler 102 configured to determine data from at least one pre-determined user account. The information determination device 108 may, similar to the information determination device 100 of FIG. 1A, further include an information determiner 104 configured to determine information related to a pre-determined organization based on the data. The information determination device 108 may further include a keyword crawler 110, like will be described in more detail below. The information determination device 108 may further include a user friend list crawler 112, like will be described in more detail below. The information determination device 108 may further include a classifier 114, like will be described in more detail below. The information determination device 108 may further include a topic miner 116, like will be described in more detail below. The information determination device 108 may further include an optimization problem solver 118, like will be described in more detail below. The information determination device 108 may further include a trivial topic purging circuit 120, like will be described in more detail below. The account crawler 102, the information determiner 104 the keyword crawler

110, the user friend list crawler 112, the classifier 114, the topic miner 116, the optimization problem solver 118, and the trivial topic purging circuit 120 may be connected via a connection 122 (or a plurality of separate connections), for example an electrical or optical connection, for example any kind of cable or bus.

[0026] According to various embodiments, the keyword crawler 110 may be configured to determine further data based on at least one pre-determined keyword. The information determiner 104 may further be configured to determine the information further based on the further data.

[0027] According to various embodiments, the keyword crawler 110 may include or may be or may be included in a fixed keyword crawler configured to determine the further data based on at least one fixed keyword.

[0028] According to various embodiments, the keyword crawler 110 may include or may be or may be included in a dynamic keyword crawler configured to determine the further data based on at least one dynamic keyword. At least one dynamic keyword may be changed based on processing of the information determination device 108.

[0029] According to various embodiments, the user friend list crawler 112 may be configured to determine a user graph of users in a social relationship with at least one of the organization or each other.

[0030] According to various embodiments, the account crawler 102 may further be configured to determine the at least one pre-determined user account based on the user graph.

[0031] According to various embodiments, the classifier 114 may be configured to classify the data into data relevant to the pre-determined organization and data irrelevant to the pre-determined organization.

[0032] According to various embodiments, the information determiner 104 may further be configured to determine the information based on the data relevant to the pre-determined organization.

[0033] According to various embodiments, the topic miner 116 may be configured to detect whether the determined information related to the pre-determined organization is related to an evolving topic or to an emerging topic about the pre-determined organization.

[0034] According to various embodiments, the topic miner 116 may be configured to detect whether the determined information related to the pre-determined organization is related to an evolving topic or to an emerging topic about pre-determined organization. The dynamic keyword crawler may further be configured to determine the at least one dynamic keyword based on the information of the evolving topic and based on the information of the emerging topic.

[0035] According to various embodiments, the account crawler 102 may further be configured to determine the at least one pre-determined user account based on the data relevant to the pre-determined organization.

[0036] According to various embodiments, the classifier 114 may be configured to classify the data based on learning.

[0037] According to various embodiments, the classifier 114 may be configured to classify the data based on a support vector machine.

[0038] According to various embodiments, the topic miner 116 may be configured to detect whether the determined information related to the pre-determined organization is related to the evolving topic or to the emerging topic based on learning.

[0039] According to various embodiments, the topic miner 116 may be configured to detect whether the determined information related to the pre-determined organization is related to an evolving topic or to an emerging topic based on solving an optimization problem.

[0040] According to various embodiments, the optimization problem may include a temporal continuity constraint.

[0041] According to various embodiments, the optimization problem may include a sparse matching constraint.

[0042] According to various embodiments, the optimization problem solver 118 may be configured to solve the optimization problem based on a least angle regression.

[0043] According to various embodiments, the trivial topic purging circuit 120 may be configured to remove old topics from evolving topics and emerging topics.

[0044] FIG. 1C shows a flow diagram 124 illustrating an information determination method. In 126, data may be determined from at least one pre-determined user account. In 128, information related to a pre-determined organization may be determined based on the data.

[0045] According to various embodiments, the method may further include determining the data from at least one account for the organization.

[0046] According to various embodiments, the method may further include determining the data from at least one account of a key user.

[0047] According to various embodiments, the method may further include: determining further data based on at least one pre-determined keyword; and determining the information further based on the further data.

[0048] According to various embodiments, the method may further include determining the further data based on at least one fixed keyword.

[0049] According to various embodiments, the method may further include determining the further data based on at least one dynamic keyword, wherein the at least one dynamic keyword is changed based on processing of the information determination method.

[0050] According to various embodiments, the method may further include determining a user graph of users in a social relationship with at least one of the organization or each other.

[0051] According to various embodiments, the method may further include determining the at least one pre-determined user account based on the user graph.

[0052] According to various embodiments, the method may further include classifying the data into data relevant to the pre-determined organization and data irrelevant to the pre-determined organization.

[0053] According to various embodiments, the method may further include determining the information based on the data relevant to the pre-determined organization.

[0054] According to various embodiments, the method may further include detecting whether the determined information related to the pre-determined organization is related to an evolving topic or to an emerging topic.

[0055] According to various embodiments, the method may further include: detecting whether the determined information related to the pre-determined organization is related to an evolving topic or to an emerging topic; and determining the at least one dynamic keyword based on the information of the evolving topic and based on the information of the emerging topic.

[0056] According to various embodiments, the method may further include determining the at least one pre-determined user account based on the data relevant to the pre-determined organization.

[0057] According to various embodiments, the method may further include classifying the data based on learning.

[0058] According to various embodiments, the method may further include classifying the data based on a support vector machine.

[0059] According to various embodiments, the method may further include detecting whether the determined information related to the pre-determined organization is related to the evolving topic or to the emerging topic based on learning.

[0060] According to various embodiments, the method may further include detecting whether the determined information related to the pre-determined organization is related to an evolving topic or to an emerging topic based on solving an optimization problem.

[0061] According to various embodiments, the optimization problem may include a temporal continuity constraint.

[0062] According to various embodiments, the optimization problem may include a sparse matching constraint.

[0063] According to various embodiments, the method may further include solving the optimization problem based on a least angle regression.

[0064] According to various embodiments, the method may further include removing old topics from evolving topics and emerging topics.

[0065] Social media portals like Twitter, Facebook, and various forum sites contain the everyday thoughts, opinions, and experiences of their online users. Parts of these user generated contents (UGCs) reflect and reveal information about organizations such as the companies, banks, government organizations, and universities etc. According to various embodiments, devices and methods may be provided for analyzing this data.

[0066] FIG. 2 shows an illustration 200 of websites, for example Optus Online Department on Twitter, see Optus account on Twitter at <https://twitter.com/Optus>. Optus is the second largest telecommunications company in Australia. As an example, FIG. 2 shows the verified Twitter account of the Optus Telecommunication Company. The biography of this account and its activity level indicate that user-centric businesses are spending substantial resources to hear the voice of their customers. In fact, it may be invaluable for such organizations to keep track of their live feedback to discover actionable insights from social media and provide better (personalized) services to their users. According to various embodiments, sophisticated methods and devices may be provided to discover such topics about a given organization from social media contents.

[0067] There may be three key challenges in mining topics for organizations from social media, like will be described in the following:

[0068] A first key challenge may be effective data harvesting: The first challenge may be about effective crawling of a live and representative distribution of data about

organizations. Most current crawling methodologies rely on a fixed list of keywords (a few previously-known keywords) such as the name of the organization to crawl data. However such methodologies cannot cover all the relevant micro-posts and consequently topics about the organization.

[0069] To address this issue, according to various embodiments, the user community of the target organization may be automatically identified and monitored. The rationale of this approach may be based on the power law correlation between the number of users and the number of relevant tweets for organizations.

[0070] FIG. 3 shows an illustration 300 of the power law correlation between the number of users and the number of relevant tweets for three organizations, namely NUS (in plot 302), DBS (in plot 304), and StarHub (in plot 306). The statistics are obtained from 1-year tweets posted for NUS, and 6-month tweets posted for DBS and StarHub organizations.

[0071] FIG. 3 shows that a small number of users of an organization often produce the major portion of relevant content about the organization. To account for this, according to various embodiments, data may be crawled based on multiple aspects of organizations: (a) known accounts, (b) key-users, and (c) fixed keywords of the organization. The known accounts may be a few manually identified official accounts created on social media portals that broadcast news and announcements about the organization; while key-users are a dynamic list of active and influential users of the target organization that should be automatically identified (like will be described in more detail below). The above sources collectively elicit more relevant data for organizations as compared to the fixed keywords used by the current crawling methodologies.

[0072] The second key challenge may be micro-post disambiguation: The second challenge is about discriminating relevant from irrelevant micro-posts with respect to the target organization as data streams in. This is a challenging task because of the polysemy problem in which the acronyms of organizations are often shared by many entities in social media. Current systems simply return many irrelevant micro-posts as they don't disambiguate micro-post for organizations that share the same acronym. It is to be noted that users often use the acronym forms instead of the complete names of the organizations in the social media context mainly due to the length limit imposed by social media portals.

[0073] To address this challenge, according to various embodiments, the context of the target organization defined by the current relevant content (keywords and micro-posts) and the user community of the organizations may be utilized. A highly accurate classifier may be provided to predict the relevance of each incoming micro-post to the target organization based its context information.

[0074] The third key challenge may be topic discovery and monitoring: The third challenge is about online clustering of relevant streaming data into coherent set of topics. This is challenging because, with streaming data, new topics as well as the old ones can be introduced or vanished respectively at any point of time.

[0075] To address this challenge, according to various embodiments, the stream of relevant micro-posts may be clustered into emerging and evolving topics. The emerging topics may be the new topics that emerge and potentially become major in a short period of time, while the evolving ones may be those that have been detected previously and are smoothly evolving through time. For the topic modeling purpose, a novel online sparse

coding approach with temporal continuity and sparse matching constraints may be provided. The approach according to various embodiments may be linear with respect to the number of input micro-posts. Furthermore, according to various embodiments, a simple purging mechanism may be provided to detect the inactive topics to further improve the performance of topic modeling.

[0076] A dataset of tweets obtained from Twitter for three organizations was created to evaluate our approach. The three organizations are: National University of Singapore (NUS), Development Bank of Singapore (DBS), and StarHub company. The first two organizations are ambiguous (in which NUS is shared between National University of Singapore, National Union of Students, and NU Skin company, and DBS is shared between several organizations like Development Bank of Singapore and Dublin Business School etc), while the third organization (StarHub) is not ambiguous.

[0077] The evaluation results show that key-users are effective factors to elicit more relevant content about organizations from social media. We show that monitoring key-users leads to: (a) higher performance of topic modeling algorithms and (b) earlier detection of emerging topics. In addition, it can be shown that the proposed framework can discriminate relevant micro-posts with a high accuracy of 86.85% for the ambiguous organizations. Also, the performance of topic modeling further improves when topics are allowed to evolve under a temporal continuity constraint according to various embodiments.

[0078] According to various embodiments, a framework may be provided which effectively addresses the data harvesting problem for organizations. It may utilize

multiple aspects of organizations to obtain more relevant data about them from social media.

[0079] According to various embodiments, a framework may be provided which effectively resolves the polysemy issue in social media for organizations.

[0080] According to various embodiments, a framework may be provided which provides a novel adaptation of online sparse coding algorithms to mine the emerging and evolving topics for organizations.

[0081] In the following, an overview of the framework according to various embodiments, for example for mining the sense of organizations in social media, will be given.

[0082] FIG. 4 shows an illustration 400 of an architecture according to various embodiments for mining the sense of organizations from social media. A fixed keyword crawler 402, a known account crawler 404, and org key-user crawler 406 may be provided, like will be described in more detail below.

[0083] Given a target organization, the framework according to various embodiments may utilize the several crawlers to obtain potentially relevant data about the organization from social media. The resultant data is given to a classifier 410 to make a real-time judgment about their relevance to the target organization. The classifier according to various embodiments may make use of the context of the organization (both content-level information 408 and user-level information 412) provided by the keyword miner 414 and user miner 418 (using a user graph 426 and a friend list crawler 428) components respectively. The relevant data may then be stored in the relevant tweet repository 416. The topic miner component 424 may extract the current emerging topics 422 and

evolving topics 420 about the organization using the resultant relevant data. Each component will be described in more detail below, and the approach according to various embodiments for each component will be discussed.

[0084] In the following, information acquisition will be described.

[0085] In the following, the fixed keyword crawler 402 will be further described. Brand monitoring systems may make use of a few manually selected fixed keywords to crawl data for organizations. Examples of fixed known keywords for a given organization are the name of the target organization or its products, the acronym of the organization etc. The fixed keyword crawler 402 may crawl the micro-posts that contain the fixed keywords.

[0086] In the following, the known account crawler 404 will be further described. Similar to fixed keywords, a few known accounts for the target organization (such as the Optus account in FIG. 2) may be manually identified. These may be official accounts of the target organization that act as informers and usually post relevant micro-posts about their organization. These accounts may be given to the known account crawler 404 to be observed.

[0087] In the following, the org key-user crawler 406 will be further described. The org (organization) key-user crawler 406 may be provided with a dynamic list of key-users to be observed. A definition for key-users according to various embodiments will be provided below.

[0088] In the following, the user friend crawler 428 (in other words: friend list crawler 428) will be further described. The user friend list crawler 428 may be used to construct the user graph 426 of the target organization by crawling the social

relationships between users who have posted relevant data about the organization. This user graph 426 may evolve over time as new users are identified.

[0089] In the following, the keyword miner 414 will be further described. The keyword miner 414 may utilize an active learning approach to extract temporally-relevant keywords for organization from the recently seen relevant data. These keywords may be considered as dynamic keywords at each point of time and used by the classification component to determine the content-based relevance of the incoming micro-posts.

[0090] In the following, the user miner 418 will be further described. The user miner 418 may identify the user community and the key-users of the organization so that such users may be monitored in order to obtain more relevant data about the organization. The user miner 418 may utilize the user graph 426 and user activity information to rank the users and find key-users of the organization (like will be described in more detail below).

[0091] In the following, the classifier 410 will be further described. The input data obtained by different crawlers may be a mix of relevant and irrelevant data. For example, key-users may also send micro-posts about other subjects like their various life activities. The classification component 410 (in other words: the classifier 410) may utilize the context information to label the input data as relevant or irrelevant to the target organizations.

[0092] In the following, the topic miner 424 will be described. The topic miner component 424 may utilize the relevant tweets to detect and keep track of topics related to the target organization. According to various embodiments, an adaptation of online sparse coding algorithms may be provided to learn the topics in an efficient way.

[0093] In the following, mining keywords and organization users will be described and the approach according to various embodiments for mining organization context defined by its content and user community will be described.

[0094] In the following, mining dynamic keywords will be described.

[0095] Dynamic keywords may be those keywords that represent the current discussions about the target organization at each point of time. To identify such keywords, suppose we have two sets of *foreground* (S_{for}^t) and *background* (S_{bak}^t) tweets at each point of time t . Let S_{for}^t include the recently-seen relevant tweets posted in a short time window of length T , i.e. $[t-T, t]$, while S_{bak}^t includes the irrelevant tweets identified in the same time window, $[t-T, t]$. In addition, let $W' = \{w_1, w_2, \dots\}$ be the vocabulary set obtained from S_{for}^t . We define the dynamic keywords as a subset of W' words that best represent the current relevant discussions about the organization. Our aim is thus to extract such keywords from W' .

[0096] For this purpose, we identify the terms of W' that have different distributions in S_{for}^t and S_{bak}^t . A significant difference between the two distributions of a term $w_i \in W'$ in S_{for}^t and S_{bak}^t signals that w_i better represents one of these sets. Those significant terms that have rising frequency in S_{for}^t can potentially represent the dynamic keywords.

[0097] Here we utilize the chi-squared test to compare distributions as its calculation is fast and suitable for rapidly evolving social media content. To derive this value for each $w_i \in W'$, we use the following Equation:

$$\chi_i^2 = \begin{cases} \frac{(f_i - b_i)^2}{b_i} + \frac{|(100 - f_i) - (100 - b_i)|^2}{100} & \text{if } f_i > b_i \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where f_i and b_i are the normalized term frequency values of w_i in the foreground and background sets respectively and are computed as follows:

$$f_i = 100 * \frac{w_i^{for}}{\sum_{\forall i} w_i^{for}} \quad (2)$$

$$b_i = 100 * \frac{w_i^{bak}}{\sum_{\forall i} w_i^{bak}} \quad (3)$$

where w_i^{for} and w_i^{bak} are the term frequency of w_i in S_{for}^t and S_{bak}^t respectively. Equation (1) assigns higher weights to the terms that frequently occur in S_{for}^t , but rarely occur in S_{bak}^t . Thus, Equation (1) only takes into account the words w_i with $f_i > b_i$ and assign zero weight to those with $f_i \leq b_i$. We rank the terms based on their χ^2 values and consider those with χ^2 value greater than ϵ (where $\epsilon = 2.706$ which corresponds to $p=0.10$ significant level of t-test) as the dynamic keywords.

[0098] In the following, mining organization users according to various embodiments will be described.

[0099] The framework according to various embodiments may rank the more *active* and *influential* users of the organizations in the higher orders, while, in case of ambiguous organizations, discard the users of the other organizations. We define an active user of an organization as the one who sends many relevant micro-posts about the organization, and an influential user as the one who has many followers within the organization and initiates major discussions about the organization. The combination of these measures can be used to rank the users of the target organization with high accuracy. Let G^t be the user graph of the target organization at time t (compare FIG. 4) and $U^t = \{u_1, \dots, u_m\}$ be the set of nodes in G^t . We compute the score for each user $u_i \in U^t$ based on the following Equation at time t :

$$W_{u_i}^t = \text{sign}(r_{u_i}^t - I_{u_i}^t) \log \left(\tau \frac{|r_{u_i}^t - I_{u_i}^t|}{\sum_j r_{u_j}^t + I_{u_j}^t} * \varphi \frac{f_{u_i}^t + 1}{\sum_j f_{u_j}^t} * \omega \frac{q_{u_i}^t + 1}{\sum_j q_{u_j}^t} \right) \quad (4)$$

where $r_{u_i}^t$ is the total number of *relevant* tweets posted by u_i up to time t , $I_{u_i}^t$ is used in case of ambiguous organizations and indicates the total number of irrelevant tweets that contain the acronym of the target organization posted by u_i up to time t (in case of non-ambiguous organization $I_{u_i}^t$ is 0). The variable $f_{u_i}^t$ is the total number of u_i 's followers who exist in U^t , $q_{u_i}^t$ is the total number of u_i 's relevant tweets that have been re-tweeted by other users up to time t , $\text{sign}(\cdot)$ is the sign function, and τ, φ , and ω are weighting parameters such that $\tau + \varphi + \omega = 1$. For example, these parameters may be empirically set as follows for experiments: $\tau = 0.5, \varphi = 0.25, \omega = 0.25$.

[00100] The above equation may rank the user based on the aforementioned three criteria. The top K users may be considered as the key-users of the organization at time t . These users may be passed to the *org key-user* crawler 406 to be monitored.

[00101] In the following, relevant tweet detection will be described, and it will be elaborated on the classification approach according to various embodiments to address the polysemy problem for ambiguous organizations.

[00102] According to various embodiments, a high quality classifier may be provided to discriminate relevant content for organizations by (a) learning their content relevance and (b) their user information respectively.

[00103] In the following, a learning content-based classifier will be described.

[00104] According to various embodiments, the framework may assign a relevance score to each input data based on its content similarity with the current discussions about the organization. For this purpose, we utilize the dynamic keywords (mined as described

further below) because such keywords are good indicators of the current discussions about the organization.

[00105] Formally, let $W^t = \{w_1, \dots, w_m\}$ of arbitrary size m contains the dynamic keywords at time t . Also, as before, let S_{for}^t be the set of recently-seen relevant tweets over the time window $[t-T, t]$ and S_{for}^t be the irrelevant tweets in the same time-span $[t-T, t]$ where t is the current time. We utilize W^t as the classification features and $S_{for}^t \cup S_{bak}^t$ as training data to discriminate the input streaming data into relevant and irrelevant sets. The dynamic keywords may provide a fast way to prune the huge amount of irrelevant input data as they stream in.

[00106] We take a binary weighting schema to weight the features for each input tweet. That is, given a tweet, we create its m -dimensional feature vector using W^t as follows: the i^{th} entry of the feature vector is set to 1, if the tweet contains w_i , and 0 otherwise. Any input data with a zero feature vector may be regarded as irrelevant by default. At the end, each test tweet may be assigned a relevance score which represents the content-based relevance score of the tweet.

[00107] As the classification approach, experiments may be done with SVM (support vector machine) classifier which is an effective classifier on textual data. As the baseline, Unigram and Bigram features obtained the combination of S_{for}^t and S_{bak}^t tweets may be considered.

[00108] In the following, combining content and user information according to various embodiments will be described.

[00109] Given the context (i.e. content and user information), a final judgment may be made about the relevance of an input tweet to the target organization. The user

information for the data obtained from the fixed keyword crawler may be utilized. This may be because the data crawled from the other two crawlers (known account crawler 404 and org key-user crawlers 406) may come from the users who already have high relevance scores to the target organization and therefore it may be desired to ensure the relevance of their content.

[00110] Formally, given a test tweet s_i^t at time t obtained from the fixed keyword crawler, the final score of the tweet may be determined by the linear combination of its content and user score as follows:

$$L_i = \alpha * C_{s_i} + (1 - \alpha) * W_{u_i}^t \quad (5)$$

whereas for a tweet obtained from the other crawlers we determine its final score by solely considering its content relevance score as follows:

$$L_i = C_{s_i} \quad (6)$$

where $C_{s_i} \in [-1,1]$ may indicate the content-based relevance score of s_i^t and $W_{u_i}^t \in [-1,1]$ may indicate the relevance score of u_i as the author of s_i^t (see Equation (4)). The parameter α may control the contribution of each of the above scores in labeling the tweet. This parameter may be learnt using development data.

[00111] Any incoming tweet with $L_i > 0$ may be considered as relevant, and the rest as irrelevant. The relevant tweet may be added to the relevant tweet repository which will be then utilized in the next iterations. Table 1 illustrates an online classification algorithm according to various embodiments. The effect of the length of the time interval t on the classification performance may be analyzed.

[00112] Table 1 shows an illustration of Algorithm 1 and Classification at time t .

Input: Q^t: input test data, T: learning time interval, α: learning parameter. Output: L: classification result. 1. learn SVM classifier with labeled data seen in $[t - T, t]$ 2. for each $s_i^t \in Q^t$ do 3. use the classifier to compute C_{s_i} 4. if s_i^t contains a fixed keyword 5. compute $W_{u_i}^t$ 6. $L_i = \alpha * C_{s_i} + (1 - \alpha) * W_{u_i}^t$ 7. else 8. $L_i = C_{s_i}$ 9. end for
--

Table 1.

[00113] In the following, an optimization framework according to various embodiments designed for mining organization topics (for example mining evolving and emerging topics) will be described.

[00114] In the following, streaming input data according to various embodiments will be described.

[00115] Assume that, at each point of time t , we receive a set of relevant tweets $S^t = \{s_1, \dots, s_{n^t}\} \in \mathbb{R}^{m \times n^t}$ where n^t is the number of relevant tweets at time t and m is the size of vocabulary. We represent each $s_i \in \mathbb{R}^m$ as a term vector of length m weighted by the standard Term Frequency (TF) and Inverted Document Frequency (IDF) as follows:

$$s_{ij} = \frac{\log(TF(i, j)) * \log(IDF(j))}{C} \quad (7)$$

where C is the normalization factor, $TF(i, j)$ indicates the frequency of the term w_j in s_i , and $IDF(j)$ indicates the inverted document frequency of w_j .

[00116] In the following, live topic learning will be described.

[00117] At each point of time, the incoming tweets may be either matched with the already known topics or may potentially represent new emerging topics for the organization.

[00118] Let the non-negative matrix $D^{t-1} \in \mathbb{R}^{m \times k^{t-1}}$ represent the k^{t-1} topics found up to time $t-1$ for the target organization and $S^t \in \mathbb{R}^{m \times n}$ indicate the relevant incoming tweets at time t . Given D^{t-1} and S^t , the problem is to determine the topic matrix at time t , i.e. $D^t \in \mathbb{R}^{m \times k^t}$. This matrix may include the smooth evolution of the k^{t-1} previously known topics (evolving topics) as well as the new topics identified at time t (emerging topics).

[00119] Let $S^{ev} \in \mathbb{R}^{m \times n^{ev}}$ indicate the tweets of S^t that can be matched, to a certain level of significance, with a topic in D^{t-1} , and $S^{em} \in \mathbb{R}^{m \times n^{em}}$ be the rest of tweets (these tweets can potentially form the emerging topics) where $n^t = n^{ev} + n^{em}$. It will be explained how to decompose S^t into these two matrices further below.

[00120] FIG. 5 shows an illustration 500 of learning evolving and emerging topics at time t ; wherein the circles represent the topic learning (TL) process. In 502, the error difference between the topic of the tweet s_i and the topics that have been learned up to time $t-1$, i.e. D^{t-1} may be computed. The error difference may be called residual error. In 504, a purging method may be performed which removes a topic from topic set D^{t-1} if it is not matched with any tweet s_i for 24 hours. In 506, 508, 520 (which may be all together provided as one component), topic modeling may be performed over the evolving tweets S^{ev} to create D^{ev} the evolving topics at time t . In 514, 516 (which may be all together provided as one component), topic modeling may be performed over the emerging tweets S^{em} to create D^{em} the emerging topics at time t . In 510, the results of the above two topic

modelers, i.e. D^{ev} and D^{em} , may be concatenated to form the final set of topics at time t , i.e. D^t . In 512, the emerging tweets may be clustered into groups to form D^{em} (this component may be optional or may be removed as (514, 516) may do this, and as such, this component may be there merely for quality purposes). It will be understood that the diagram of FIG. 5 shows what is happening in one time-step (for example the transition from time $t-1$ to t), and the time is indicated by time axis 518.

[00121] As depicted in FIG. 5, given S^{ev} , S^{em} , and D^{t-1} , we need to solve the following two sub-problems to obtain D^t : (a) how to learn the evolving topics using S^{ev} and D^{t-1} (we indicate the evolving topics by $D^{ev} \in \mathbb{R}^{m \times k^{t-1}}$), and (b) how to learn the new emerging topics using S^{em} (we indicate the emerging topics by $D^{em} \in \mathbb{R}^{m \times k'}$). The topic matrix $D^t \in \mathbb{R}^{m \times k^t}$ where $k^t = k^{t-1} + k'$ may then be achieved by vertical concatenation of D^{ev} and D^{em} .

[00122] The following two constraints may be considered when learning the topic matrix D^t :

[00123] - Temporal Continuity constraint: This requirement may constrain D^{ev} to be a smooth evolution of D^{t-1} , and

[00124] - Sparse Matching constraint: This constraint may indicate that each tweet s_i can only represent a "few" topics.

[00125] The first constraint may be to prevent dramatic changes in the evolving topics in two consecutive time stamps, whereas the second constraint may be due to the limited length of the tweets. This may be because tweets are limited to 140 characters; this space may be too short to be used for writing about several topics.

[00126] Based on the above requirements, the evolving topic matrix D^{ev} may be learned by minimizing the following optimization problem

$$\begin{aligned} (D^{ev}, X^{ev}) = \arg \min_{D, X} & \| S^{ev} - DX \|_F^2 + \mu \| D - D^{t-1} \|_F^2 + \lambda \| X \|_1 \\ \text{s.t. : } & X \geq 0, D \geq 0, \| d_j \|_2^2 \leq 1 \forall j \in \{1 \dots k^{t-1}\} \end{aligned} \quad (8)$$

where $X^{ev} \in \mathbb{R}^{k^{t-1} \times n^{ev}}$ may be the weight matrix and $\lambda \in [0,1]$ and $\mu \in [0,1]$ may be the learning parameters. The first term in the above Equation may be the reconstruction error, while the second and the third terms may represent the two aforementioned constraints respectively. The above topic learning process may optimize the matrix D^{ev} with respect to D^{t-1} and S^{ev} . It is to be noted that, in this step, no new topic is introduced.

[00127] In contrast to the evolving topics, the emerging topics may be totally new and there may be no prior information about the number of emerging topics. Therefore, the X-Means approach may be utilized to find an initial set of clusters from S^{em} . X-Means may be an extension of the standard k-means that utilizes the Bayesian Information Criterion (BIC) model to estimate the best number of clusters within a given range. The resultant clusters that have sufficient number of tweets (for example, only the clusters that have more than 20 tweets may be taken into account) may be considered as the emerging topics and their centroid vector may be used to create an initial emerging topic matrix $D^{init} \in \mathbb{R}^{m \times k'}$ where k' is the number of such centroids. Then, the same approach may be followed as evolving topics to find the optimum value for D^{em} as follows:

$$\begin{aligned} (D^{em}, X^{em}) = \arg \min_{D, X} & \| S^{em} - DX \|_F^2 + \mu \| D - D^{init} \|_F^2 + \lambda \| X \|_1 \\ \text{s.t. : } & X \geq 0, D \geq 0, \| d_j \|_2^2 \leq 1 \forall j \in \{1 \dots k'\} \end{aligned} \quad (9)$$

where $X^{em} \in \mathbb{R}^{k^{t-1} \times n^{em}}$ may be the weight matrix. FIG. 5 depicts the overall procedure of learning topics at each point of time. It is to be noted that the above two processes

(learning D^{ev} and D^{em}) can be performed in parallel to speed up the overall learning process. The purging process in FIG. 5 will be explained further below.

[00128] In the following, decomposition of streaming data will be described.

[00129] Given the input matrix S^t and the topic matrix D^{t-1} , it may be desired to decompose S^t into S^{ev} and S^{em} matrices. For this, we find the best representation of each $s_i \in S^t$ in terms of D^{t-1} as follows:

$$\begin{aligned} x_i &= \arg \min_x \| s_i - D^{t-1}x \|_2^2 + \lambda \| x \|_1 \\ s.t.: x &\geq 0 \end{aligned} \quad (10)$$

[00130] The resultant vector $x_i \in \mathbb{R}^{k^{t-1}}$ may indicate the already known topics that best represent the input vector s_i . Using this vector, we compute the representation error of s_i on D^{t-1} (what we call *residual error*) as follows:

$$\mathcal{R}^*(s_i, D^{t-1}) = \| s_i - D^{t-1}x_i \|_2^2 + \lambda \| x_i \|_1 \quad (11)$$

[00131] Based on the value of the residual error, the matrix S^t may be decomposed into the two matrices as follows:

[00132] - S^{ev} : contains all $s_i \in S^t$ with a residual error equal to or smaller than a chosen threshold η , and

[00133] - S^{em} : includes other inputs, i.e. all $s_j \in S^t$ with a residual error greater than η .

[00134] In the following, purging trivial topics will be described.

[00135] As time passes, some topics may become old and no more discussions arrive about them. Such topics may be safely removed from the topic matrix D^t . There may be different approaches to accomplish this. For example one may directly *remove* the non-active topics or *replace* them with a randomly selected input data. According to various

embodiments, the first approach may be applied as it better suits the need for keeping the size of the learned topics manageable.

[00136] To do so, the most recent time that each topic is selected as the *dominant* topic for an input tweet may be stored. This time may be used as a measure to purge the topics. The dominant topic of each s_i may be the topic that has the greatest matching score with s_i as compared to all the other topics, i.e. the topic d_j such that $j = \arg \max_j x_{ij}$ where x_i is obtained from Equation (10). It is to be noted that the matching score between each d_j and each s_i may be determined by the (ij) th entry of the weight matrix X , i.e. x_{ij} , see Equations (8) and (9). According to various embodiments, all the topics that have not been selected as a dominant topic in the past 24 hours may be considered as non-active and are removed from D^t .

[00137] In the following, optimization algorithms and solutions to the resultant optimization problems will be described. A fast online approach to solve the optimization problem of Equation (8) will be described, and the same approach may be used to solve Equation (9). This optimization problem may be in general non-convex, but, it may be shown that, if one of the variables, either D or X is known, optimization with respect to the other variable will be convex. Therefore, a general solution may be to iteratively optimize the objective function by alternatingly optimizing with respect to D and X while holding the other fixed.

[00138] If D is fixed, i.e. $D = D^{t-1}$, then the problem may be equivalent to an ℓ_1 -regularized least square problem and can be efficiently solved by a least angle regression (LARS) method or an alternating direction method. However, when X is fixed, the problem may be a least square problem with quadratic constraints. To solve this problem,

an advanced version of the projected gradient approach may be provided. It may be an effective online approach that processes each input data (or a small subset of data) only once. This may be particularly important in the context of social media where the input data can potentially be large at each time.

[00139] If D is fixed, then Equation (8) may be converted to the following problem (for simplicity in notation and exposition, it may be assumed that $D=D^{ev}$, $S=S^{ev}$, and $X=X^{ev}$):

$$\begin{aligned} X &= \arg \min_X \| S - D^{t-1} X \|_F^2 + \lambda \| X \|_1 \\ \text{s.t. } X &\geq 0. \end{aligned} \quad (12)$$

[00140] The above Equation may find the optimal value of X and may be solved by a least angle regression (LARS) method. It is to be noted that as x_i s are independent, they may be optimized in parallel. However, if X is fixed, i.e. obtained from the above Equation, then Equation(8) may be converted to the following problem:

$$\begin{aligned} D &= \arg \min_D \| S - DX \|_F^2 + \lambda \| X \|_1 + \mu \| D - D^{t-1} \|_F^2 \\ \text{s.t. } D &\geq 0, \| d_j \|_2^2 \leq 1 \forall j \in \{1 \dots k^{t-1}\} \end{aligned} \quad (13)$$

[00141] Given S , X , and D^{t-1} , let a loss function $\mathcal{L}(D)$ be defined as follows:

$$\mathcal{L}(D) = \| S - DX \|_F^2 + \lambda \| X \|_1 + \mu \| D - D^{t-1} \|_F^2 \quad (14)$$

[00142] The projected gradient approach may solve Equation (13) by iteratively obtaining the projected gradients using the following updating rule:

$$D_{i+1} = P \left[D_i - \alpha_i \nabla_D \mathcal{L}(D)_{|D_i, X|} \right] \quad (15)$$

where D_i may indicate D at iteration i , the parameter α_i may be the step size, and $\nabla_D \mathcal{L}(D)_{|D_i, X|}$ may be the gradient of $\mathcal{L}(D)$ with respect to D , see Equation (16), evaluated

on D_i and X , and $P[\cdot]$ may be a projection function defined for the non-negativity constraint, Equation (17):

$$\nabla_D \mathcal{L}(D) = 2SX^T + DXX^T + 2\mu(D - D^{t-1}) \quad (16)$$

$$P[z] = \begin{cases} z & \text{if } z \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

[00143] The disadvantage of the above approach may be that it may be slow and may need the parameter α to be carefully chosen to obtain good results. To resolve these issues, the second order information, the Hessian matrix, may be used to make the updating rule in Equation (15) parameter free with faster convergence. The Hessian matrix may be utilized to obtain the final updating rule as follows:

$$D_{i+1} = P \left[D_i - \nabla_D \mathcal{L}(D)_{(D_i, X)} \mathcal{H}^{-1}[\mathcal{L}(D)]_{(X)} \right] \quad (18)$$

where the Hessian matrix of $\mathcal{L}(D)$ may be defined as follows:

$$\mathcal{H}[\mathcal{L}(D)] = XX^T + 2\mu I_k \quad (19)$$

and $\mathcal{H}^{-1}[\mathcal{L}(D)]_{(X)}$ may be the inverse of the Hessian matrix evaluated on X . Since the exact calculation of the inverse of the Hessian matrix may be time-consuming for large number topics, the Hessian matrix may be approximated by its diagonal line based on a diagonal approximation method. Table 2 summarizes the detail procedure of computing D^l and X^l given S^l and D^{l-1} according to various embodiments.

Input: S^t , D^{t-1} , itr: number of iterations

Output: D^t , X^t

1. Compute X^t using S^t and $D^{t-1} \rightarrow$ Equation (12)
2. $D_0^t = D^{t-1}$
3. for $i=1 : \text{itr}$ do
4. compute $\nabla_D \mathcal{L}(D_{i-1}^t) \rightarrow$ Equation (16)
5. $U = \nabla_D \mathcal{L}(D_{i-1}^t) \text{diag}^{-1}(\mathcal{H}[\mathcal{L}(D)]_{|X^t|}) + D_{i-1}^t \rightarrow$ Equation (18)
6. $D_i^t = \max(0, U)$
7. end for

Table 2: Algorithm 2, computing D^t and X^t at time t , see TL in Figure 5.

[00144] In the following, the evaluation methodology according to various embodiments will be described.

[00145] It is assessed how the approach according to various embodiments can make a real-time judgment on the relevant keywords, micro-posts, and topics about a given organizations. The framework according to various embodiments is evaluated from two perspectives: performance in (a) identifying the relevant data and (b) modeling the topics about the organization as data streams in.

[00146] In the following, an evaluation metrics for classification will be described.

[00147] The performance of the classifier according to various embodiments for the positive (relevant) and negative (irrelevant) classes may be evaluated based on Precision, Recall and F1-score metrics:

$$\begin{aligned}
 Precision^+ &= \frac{N_{correct^+}}{N_{labeled^+}}, \quad Recall^+ = \frac{N_{correct^+}}{N_{total^+}}, \quad F1^+ = \frac{2(Precision^+)(Recall^+)}{(Precision^+) + (Recall^+)} \\
 Precision^- &= \frac{N_{correct^-}}{N_{labeled^-}}, \quad Recall^- = \frac{N_{correct^-}}{N_{total^-}}, \quad F1^- = \frac{2(Precision^-)(Recall^-)}{(Precision^-) + (Recall^-)} \\
 Avg - F1 &= \frac{(F1^+) + (F1^-)}{2}
 \end{aligned} \tag{20}$$

where $N_{correct^+}$ is the number of micro-posts that were assigned correct relevant label, $N_{labeled^+}$ is the number of micro-posts that were labeled as relevant, and N_{total^+} is the

total number of relevant micro-posts (the same definition applies for the irrelevant class). $F1^+$ and $F1^-$ are the classification performances for the relevant and irrelevant classes respectively and therefore Avg-F1 indicates the average classification performance in terms of F1-score.

[00148] In the following, evaluation metrics for topic learning will be described.

[00149] Two evaluation metrics may be considered to assess the performance of the topic miner component, namely topic detection accuracy, and miss-rate at first detection. The first measure evaluates the topic detection performance in terms of precision and recall, while the second measure evaluates the amount of information (number of tweets) that has been missed before the first automatic detection of each topic. The second measure is important as we need a small miss-rate for earlier prediction of emerging topics.

[00150] Assume that the set $I = \{I_1, \dots, I_n\}$ is our topic ground-truth where each I_j represents a topic and $\phi(I_j)$ indicates the set of tweets that are related to the topic I_j . Furthermore, let o_{I_j} and l_{I_j} be the time that the first and last tweet of I_j were posted respectively (we call these times the origin and the last time for I_j respectively, thus, $[o_{I_j}, l_{I_j}]$ shows the life time of I_j). Let $d_i \in D^1$ be a topic that was detected at time t such that $o_{I_j} \leq t \leq l_{I_j}$. We define the closeness between d_i and I_j as follows:

$$\begin{aligned}
 Precision^{ij} &= \frac{|\phi(d_i) \cap \phi(I_j)|}{|d_i|} \\
 Recall^{ij} &= \frac{|\phi(d_i) \cap \phi(I_j)|}{|I_j|} \\
 F1^{ij} &= \frac{2Precision^{ij}Recall^{ij}}{Precision^{ij} + Recall^{ij}}
 \end{aligned} \tag{21}$$

where $|\cdot|$ indicates the cardinality of the corresponding set (number of tweets). The value of $F1^j$ shows the similarity between the two topics, i.e. $F1^j=1$, iff (i.e. if and only if) the two topics contain exactly the same set of tweets, and $F1^j=0$, iff they are disjoint. Topic d_i that produces the maximum value of $F1^j$ for I_j is considered as the most similar topic to I_j (i.e. the *best match*). The overall performance of topic detection for the topic set I may then be determined as follows:

$$F1 = \frac{\sum_{j=1}^n F1^j}{n} \quad (22)$$

[00151] As for the second evaluation measure, miss-rate at first detection, the fraction of $\phi(I_j)$ tweets posted before the origin time of d_i (that is the best match of I_j) may be considered as the missed tweets and their percentage determines the value of miss rate (MR) for I_j . Formally, the miss rate for I_j is determined with respect to d_i and may be defined as follows:

$$MR^j = \frac{|\{s : s \in \phi(I_j) \ \& \ timestamp(s) < b_i\}|}{|I_j|} \quad (23)$$

where b_i is the origin time of d_i . The overall miss-rate for the topic set I may then be obtained as follows:

$$MR = \frac{\sum_{j=1}^n MR^j}{n} \quad (24)$$

[00152] A good topic miner should have a high topic detection performance, $F1$, and a small miss rate, MR .

[00153] In the following, the experimental settings and results will be described. In the following, the terms micro-post, tweet, and data may be used interchangeably.

[00154] In the following, data and settings will be described. Three organizations are considered in this study. They are: National University of Singapore (NUS), Development Bank of Singapore (DBS), and StarHub company (StarHub). NUS and DBS are ambiguous organizations (DBS shares its acronym with many organizations like Dublin Business School, Doha British School, and concepts like Defensive Backs etc), while the third organization (StarHub) is not ambiguous.

[00155] The crawlers according to various embodiments utilized the streaming API of twitter to crawl data. Around 10 fixed keywords were manually identified for each of the ambiguous organizations NUS and DBS (including their acronyms) and only one fixed keyword, the term "starhub" itself, for StarHub. Furthermore, the known accounts for each organization (around 5 to 30 accounts for each organization) were manually identified. Only the top K ($k=200$) key-users of each organization (see above) may be consider and the key-user list may be updated on a daily basis according to Equation (4). Table 3 shows the number of tweets obtained from each of the crawlers and the crawling period for the three organizations.

NUS 1/1/2012-12/30/2012			DBS 6/1/2012-12/30/2012			StarHub 6/1/2012-12/30/2012		
fixed	known acc	key-user	fixed	known acc	key-user	fixed	known acc	key-user
142K	10K	2.3M	6.6K	5.5K	0.5M	9.7K	5.9K	2.3M

Table 3: Data statistics and crawling period for the three organizations NUS, DBS, and StarHub.

[00156] In the following, ground truth and settings for classification will be described.

[00157] We created a ground-truth of tweets as relevant or irrelevant for each of the three organizations. For this purpose, we considered all the tweets crawled in a time-

window of 10 continuous days for each organization and employed a semi-automatic approach to label them as relevant or irrelevant to the target organization.

[00158] To ease the annotation task, we first extracted all the hashtags from the tweet set of each organization. We manually labeled these hashtags as relevant or irrelevant to the target organizations. We then constructed a set of labeled tweets using: (a) all the tweets that contained at least one of the labeled hashtags and (b) all the tweets posted by known accounts of the organizations that contained at least one fixed keyword. We learned an SVM classifier using this training set and utilized it to label the rest of the tweets crawled in the time-window of 10 continuous days. We utilized term Unigrams and user profile information such as user's location and timezone as classification features. In case of low confidence in the classification results, we judged the tweets based on manual annotation. Overall, we obtained 2.5K, 1.5K, and 4.5K relevant tweets for NUS, DBS, and StarHub respectively.

[00159] FIG. 6 shows an illustration 600 of the distribution of the relevant tweets in the resultant ground-truth for the three organizations (NUS, DBS, StarHub, in subplots 602, 604, and 606). "Fixed-Known" indicates the number of relevant tweets obtained by the fixed keyword or known account crawlers for the organization, while "overall" indicates the total number of relevant tweets obtained by all the three crawlers. As it is clear, there are many relevant tweets crawled by the key-user crawler. Such tweets can greatly improve the performance of online topic miner algorithms by providing more content information about the topics. We should also note that there is a high overlap between the data obtained by the fixed keyword and known account crawlers. This is to

be expected as the tweets posted by the known accounts of organizations are mainly official news about the organization and usually contain the fixed keywords.

[00160] For parameter setting, we use the first three days of the ground-truth as development data to learn the time interval t and linear combination parameter α in Equation (5). We then employed the resultant values to evaluate the classification performance on the remaining seven days.

[00161] In the following, ground truth and settings for topic modeling will be described.

[00162] Similar to the above approach, we conducted a semi-automatic method to construct our topic dataset. For this purpose, we manually identified 45 topics for the three organizations (15 for each organization). For each topic, we identified the *hashtags* and all the keywords and key-phrases that uniquely identify the topic. Then, for each topic, we found the tweets that are posted within the topic life time and contain at least one topical keyword or key-phrase. We treat these tweets as the relevant tweets to that topic. Table 4 shows the examples of such topics. Our topic dataset covers different events about the organizations and range from small topics of around 50 tweets per topic to topics with more than 1000 tweets.

Organization	Topic/Event	Period
NUS	fire in engineering	08/10/12-08/30/12
	nus open house	03/09/12-03/24/12
	flagday	08/05/12-08/16/12
DBS	sudden jump in dbs profit	11/01/12-11/04/12
	dbs grant for social enterprises	10/29/12-10/31/12
	paypass facility	11/01/12-11/06/12
StarHub	poor outdoor coverage fine	12/06/12-12/10/12
	leeteuk sistar on starhub	12/16/12-12/18/12
	lunch of central comedy asia	10/31/12-11/08/12

Table 4. Some sample topics/events and their corresponding organizations.

[00163] For parameter setting, we used the first five topics of each organization as development data to tune the temporal continuity parameter μ in Equations (8) and (9) for each organization. We then utilized the resultant μ values to perform the evaluation on the rest of the topics for the target organizations. We also study the effect of this parameter on the performance of our approach. In addition, for parameter setting, we set $\lambda = 1.2/\sqrt{m}$ that is a classical normalization factor, in all the experiments where μ is the number of terms. We also empirically set the threshold for residual error to $\eta = 0.3$.

[00164] In the following, experimental results of tweet classification will be described.

[00165] To simulate live data streaming, we ran our online model over one month data (the month that includes the ground truth data) for each organization, while we restricted the evaluation to the tweets in our ground truth dataset.

[00166] As mentioned above, we used the first three days of the ground-truth to learn t and α and employed the obtained values to evaluate the classification performance on the remaining seven days. Table 5 presents the classification performance as average F1 score as described above. The fixed-kw, Unigram, and Bigram rows show the classification performance when we used fixed keywords, term Unigrams, and term Bigrams as classification features respectively (we consider them as baselines). Dynamic-kw shows the classification performance when we used dynamic keywords as the classification features, i.e. the results obtained from Equation (6), while Dynamic-kw + User represents the performance when we used dynamic keywords in conjunction with user information, as shown in Equation (5). Note that, in all the settings, if an input tweet did not contain any classification feature, we treated it as irrelevant. In addition, in all the experiments the two-tailed paired t-test with $p < 0.01$ was used for significance testing.

We use the asterisk mark (*) to indicate significant improvement over the best performing baseline.

[00167] Table 5 shows an illustration of a Classification performance in terms of Avg-F1 with different types of features and input Classification performance in terms of Avg-F1 with different types of features and input data.

	NUS	DBS	StarHub	Average
Fixed-kw	47.82, T:3	41.67, T:4	49.63, T:4	46.37
Unigram	63.24, T:6	84.76, T:3	88.92, T:7	78.97
Bigram	62.30, T:6	84.89, T:7	88.92, T:7	78.70
Dynamic-kw	65.15*, T:5	85.29*, T:4	88.94*, T:3	79.79
Dynamic-kw + User	81.08*, T:5, α :0.3	89.64*, T:3, α :0.4	89.82*, T:2	86.85

Table 5: Classification performance in terms of Avg-F1 with different types of features and input Classification performance in terms of Avg-F1 with different types of features and input data.

[00168] The results show that the dynamic keywords alone significantly improve the classification performance over the best performing baseline. This indicates that our keyword mining algorithm is able to effectively discriminate current relevant keywords from irrelevant ones for each organization. We should also note that since the dynamic keyword model has fewer number of features (as compared to the total number of terms or Unigrams), the classification can be performed very fast which is desirable in online settings. In addition, adding user information significantly improves the classification performance. This is because we utilize the entire user activity with respect to each organization, see Equation (4), to judge its input data.

[00169] It is to be noted that the value of α is smaller than 0.5 for both NUS and DBS. This was expected because the parameter α only affects tweets with fixed keywords (see Algorithm 1) and for such tweets the weight of the user score, i.e. $1-\alpha$, is expected to be high. Also, the classification performance is invariant to the parameter α in case of

StarHub: as we mentioned above, the parameter α only affects tweets with fixed keywords. Such tweets are considered as relevant for non-ambiguous organizations by default (see subplot 604 in FIG. 6).

[00170] FIG. 7 shows an illustration 700 of the effect of learning parameters T and α of the classification performance for NUS.

[00171] FIG. 8 shows an illustration 800 of the effect of learning parameters T and α of the classification performance for DBS.

[00172] FIG. 9 shows an illustration 900 of the effect of learning parameters T and α of the classification performance for StarHub.

[00173] FIG. 7, FIG. 8, and FIG. 9 (in subplots 702, 704, 802, 804, 902, and 904) show the effect of the learning parameters t and α on our model, Dynamic-kw + User, evaluated over the entire ground truth dataset. In each case, we fixed one of the parameters and investigated the effect of the other one. For the fixed parameter, we used the value obtained from the development set. We restricted the interval time to 7 hours, i.e. $1 \leq t \leq 7$, and the parameter α to $0.1 \leq \alpha \leq 1$ with learning steps set to 1 hour and 0.1 for t and α respectively.

[00174] As FIG. 7, FIG. 8, and FIG. 9 show, greater time intervals (t) increase the classification performance for NUS but causes great reduction in the classification performance for DBS and StarHub. The life time of the topics happening about the organization may affect the classification performance. If the topics are long, increasing the time interval t may not reduce the performance as the old topics are still active, whereas for short topics, increasing t reduces the performance as the old discussions are not active anymore and thus the dynamic keywords extracted from such topics are not

useful features to classify the current input data. As FIG. 8, and FIG. 9 show, for NUS and DBS as ambiguous organizations, smaller values of α (i.e. giving less weight to the content relevance score and higher weight to the user score for the tweets with fixed keywords) leads to better performance. This result indicates the important role of user scores to classify tweets with fixed keywords. As mentioned above, the classification performance for non-ambiguous organizations like StarHub is invariant to the parameter α , but will be affected by the learning time interval.

[00175] In the following, experimental results of topic modeling will be described.

[00176] For these experiments, the online topic modeling method according to various embodiments may be applied over the entire dataset for each organization and only restrict the evaluation to the topic dataset.

[00177] We first tune the learning parameter μ (see Equation (8)) for each organization using the development dataset as described above. We then employ the resultant μ to evaluate the topic modeling performances of different approaches on the test topics. In these experiments, we consider the basic Non-Negative Matrix Factorization (NMF) algorithm without the temporal continuity and sparse matching constraints as the baseline (i.e. for the baseline, we set $\lambda = 0$ and $\mu = 0$ in our optimization framework to obtain the baseline performance).

[00178] Table 6 shows the evaluation results for topic detection in terms of F1 performance (Equation (22)). The Overall column shows the performance when we perform the evaluation over all the relevant input data for the topic modeling purpose, while the Known column shows the corresponding performance when we only use the relevant tweets obtained from fixed keyword or known account crawlers.

Organization	Baseline (NMF)		Optimization Framework	
	known	overall	known	overall
NUS	39.30	40.03*	37.56, $\mu : 0.3$	39.82, $\mu:0.3$
DBS	64.67	61.42	65.63, $\mu : 0.3$	80.64*, $\mu:0.4$
StarHub	42.88	46.84	43.07, $\mu : 0.3$	51.78*, $\mu:0.3$
Average	48.95	49.43	48.75	57.41

Table 6: Topic detection F1 performance with known and overall input, the higher values show better performances.

[00179] In almost all the case (except DBS baseline), the *overall* data results in a higher performance as compared to known data. This improvement was indeed expected as there are many relevant tweets that are not covered by the fixed keywords and known accounts of organizations. The average performance of baseline and our optimization framework increases from 48.95% to 49.43% and 48.75% to 57.41% respectively by utilizing these relevant tweets.

[00180] The optimization framework according to various embodiments outperforms the baseline for DBS and StarHub while its performance for NUS is comparable with the baseline. The average improvement over the baseline is 7.98%, i.e. from 49.43% to 57.41%, when we utilize the overall input data for topic modeling.

[00181] It is to be noted that the lower F1 performance for NUS and StarHub (as compared to DBS) could be due to the longer life time of NUS' and StarHub's topics as compared to DBS's topics in our dataset. The long topics may reduce the topic modeling performance as such topics may be divided into sub-topics by different algorithms (mainly because of shifts in topics through time). This is while we only have one best match for each topic.

[00182] Comparing the average performances of the baseline and the optimization framework according to various embodiments, it may be concluded that topic detection and tracking is more effective if we use the sparsity and temporal continuity constraints for topic mining. In fact, the temporal continuity constraint helps the system to utilize the past information about topics to make a better judgment about the current topics. FIG. 10 shows the effect of this constraint on the overall topic modeling performance.

[00183] FIG. 10 shows an illustration 1000 (including subplots 1002, 1004, and 1006) of an effect of the temporal continuity constraint on the performance of topic modeling for three organizations. The experiments were performed with $\mu = 0; 0.3; 0.6; 0.9$.

[00184] Table 7 shows the evaluation results for the miss-rate at first detection metric (Equation (24)). The lower values of miss-rate indicate that the topic modeling algorithm is able to identify the emerging topics earlier. As the Figure shows, the average miss-rate is lower when we use the overall data instead of only tweets obtained by the fixed keyword or known account crawlers. This suggests that we can detect emerging topic earlier, if we make use of more (relevant) tweets. It may be concluded that the key-user crawler according to various embodiments is an effective resource for early prediction of emerging topics about organizations. The results show that our approach outperforms the baseline by 4.83% reduction in the average miss-rate (from 27.11% to 22.28%).

Organization	Baseline (NMF)		Optimization Framework	
	known	overall	known	overall
NUS	36.06	37.83	28.17	26.78
DBS	15.88	15.87	15.88	16.96
StarHub	40.27	27.62	40.27	23.11
Average	30.74	27.11	28.11	22.28

Table 7: Miss rate results for fixed keywords and overall input, the lower values show better performances.

[00185] According to various embodiments, given an organization, a framework may be provided to automatically identify relevant micro-posts, topics and users of the organization from social media.

[00186] Previous brand monitoring systems are not designed to address the lack of representative data or polysemy issues for entities like organizations. Various embodiments provide an effective framework to elicit representative amount of data about organizations and dealing with the polysemy issue for organizations in social media.

[00187] In fact, previous systems use a fixed set of known keywords to crawl micro-posts about a given organization from social media. This popular strategy results in (a) many missing relevant micro-posts, and (b) many irrelevant micro-posts. The first issue is due to the dynamic nature of the social media contents, while the latter issue is due to the polysemy problem in which the acronyms of organizations are often shared by many entities (for example, NUS is shared between National University of Singapore, National Union of Students and Nu-Skin™ company).

[00188] The framework according to various embodiments may provide methods to address the above issues by

[00189] - utilizing multiple aspects of organizations to crawl more relevant data. This includes crawling data through fixed keywords, known accounts and automatically identified key-users of the organizations, and

[00190] employing both content and user information to address the polysemy problem for organizations.

[00191] The above approaches lead to effective relevant data harvesting for organizations from social media. This in turn results in (a) higher performance of topic

modeling algorithms, and (b) earlier detection of emerging topics for organizations. Also, given the automatically identified relevant micro-posts of an organization, the framework according to various embodiments provide a method to efficiently learn the topics of the organization through time.

[00192] User-centric organizations and businesses are spending substantial resources to identify their customers (user community) and hear their voice (current discussions and topics). The system according to various embodiments may provide live feedback to organizations by automatic discovery of the relevant content about them from social media, identifying their user community in social media, and listening to their key-users. This information may be invaluable for user-centric organizations as they utilize such information to obtain actionable insights from social media.

[00193] The system according to various embodiments may be fed by the social media portals like Twitter and Facebook.

[00194] The following description pertains to further embodiments.

[00195] According to various embodiments, an information determination device may be provided. The information determination device may include: an account crawler configured to determine data from at least one pre-determined user account; and an information determiner configured to determine information related to a pre-determined organization based on the data.

[00196] In other words, according to various embodiments, an information determination device may be provided which determined information related to a pre-determined organization based on data which are determined from one or more pre-determined user accounts.

[00197] According to various embodiments, the account crawler may include or may be a known account crawler configured to determine the data from at least one account for the organization.

[00198] According to various embodiments, the account crawler may include or may be a key-user crawler configured to determine the data from at least one account of a key user.

[00199] According to various embodiments, the information determination device may further include a keyword crawler configured to determine further data based on at least one pre-determined keyword, and the information determiner may further be configured to determine the information further based on the further data.

[00200] According to various embodiments, the keyword crawler may include or may be a fixed keyword crawler configured to determine the further data based on at least one fixed keyword.

[00201] According to various embodiments, the keyword crawler may include or may be a dynamic keyword crawler configured to determine the further data based on at least one dynamic keyword, and at least one dynamic keyword may be changed based on processing of the information determination device.

[00202] According to various embodiments, the information determination device may further include a user friend list crawler configured to determine a user graph of users in a social relationship with at least one of the organization or each other.

[00203] According to various embodiments, the account crawler may further be configured to determine the at least one pre-determined user account based on the user graph.

[00204] According to various embodiments, the information determination device may further include: a classifier configured to classify the data into data relevant to the pre-determined organization and data irrelevant to the pre-determined organization.

[00205] According to various embodiments, the information determiner may further be configured to determine the information based on the data relevant to the pre-determined organization.

[00206] According to various embodiments, the information determination device may further include a topic miner configured to detect whether the determined information related to the pre-determined organization is related to an evolving topic or to an emerging topic about the pre-determined organization.

[00207] According to various embodiments, the information determination device may further include: a topic miner configured to detect whether the determined information related to the pre-determined organization is related to an evolving topic or to an emerging topic about pre-determined organization, and the dynamic keyword crawler may further be configured to determine the at least one dynamic keyword based on the information of the evolving topic and based on the information of the emerging topic.

[00208] According to various embodiments, the account crawler may further be configured to determine the at least one pre-determined user account based on the data relevant to the pre-determined organization.

[00209] According to various embodiments, the classifier may be configured to classify the data based on learning.

[00210] According to various embodiments, the classifier may be configured to classify the data based on a support vector machine.

[00211] According to various embodiments, the topic miner may be configured to detect whether the determined information related to the pre-determined organization is related to the evolving topic or to the emerging topic based on learning.

[00212] According to various embodiments, the topic miner may be configured to detect whether the determined information related to the pre-determined organization is related to an evolving topic or to an emerging topic based on solving an optimization problem.

[00213] According to various embodiments, the optimization problem may include a temporal continuity constraint.

[00214] According to various embodiments, the optimization problem may include a sparse matching constraint.

[00215] According to various embodiments, the information determination device may further include an optimization problem solver configured to solve the optimization problem based on a least angle regression.

[00216] According to various embodiments, the information determination device may further include a trivial topic purging circuit configured to remove old topics from evolving topics and emerging topics.

[00217] According to various embodiments, an information determination method may be provided. The information determination method may include: determining data from at least one pre-determined user account; and determining information related to a pre-determined organization based on the data.

[00218] According to various embodiments, the information determination method may further include determining the data from at least one account for the organization.

[00219] According to various embodiments, the information determination method may further include determining the data from at least one account of a key user.

[00220] According to various embodiments, the information determination method may further include determining further data based on at least one pre-determined keyword; and determining the information further based on the further data.

[00221] According to various embodiments, the information determination method may further include determining the further data based on at least one fixed keyword.

[00222] According to various embodiments, the information determination method may further include determining the further data based on at least one dynamic keyword, wherein the at least one dynamic keyword is changed based on processing of the information determination method.

[00223] According to various embodiments, the information determination method may further include determining a user graph of users in a social relationship with at least one of the organization or each other.

[00224] According to various embodiments, the information determination method may further include determining the at least one pre-determined user account based on the user graph.

[00225] According to various embodiments, the information determination method may further include classifying the data into data relevant to the pre-determined organization and data irrelevant to the pre-determined organization.

[00226] According to various embodiments, the information determination method may further include determining the information based on the data relevant to the pre-determined organization.

[00227] According to various embodiments, the information determination method may further include detecting whether the determined information related to the pre-determined organization is related to an evolving topic or to an emerging topic.

[00228] According to various embodiments, the information determination method may further include detecting whether the determined information related to the pre-determined organization is related to an evolving topic or to an emerging topic; and determining the at least one dynamic keyword based on the information of the evolving topic and based on the information of the emerging topic.

[00229] According to various embodiments, the information determination method may further include determining the at least one pre-determined user account based on the data relevant to the pre-determined organization.

[00230] According to various embodiments, the information determination method may further include classifying the data based on learning.

[00231] According to various embodiments, the information determination method may further include classifying the data based on a support vector machine.

[00232] According to various embodiments, the information determination method may further include detecting whether the determined information related to the pre-determined organization is related to the evolving topic or to the emerging topic based on learning.

[00233] According to various embodiments, the information determination method may further include detecting whether the determined information related to the pre-determined organization is related to an evolving topic or to an emerging topic based on solving an optimization problem.

[00234] According to various embodiments, the optimization problem may include a temporal continuity constraint.

[00235] According to various embodiments, the optimization problem may include a sparse matching constraint.

[00236] According to various embodiments, the information determination method may further include solving the optimization problem based on a least angle regression.

[00237] According to various embodiments, the information determination method may further include removing old topics from evolving topics and emerging topics.

[00238] While the invention has been particularly shown and described with reference to specific embodiments, it should be understood by those skilled in the art that various changes in form and detail may be made therein without departing from the spirit and scope of the invention as defined by the appended claims. The scope of the invention is thus indicated by the appended claims and all changes which come within the meaning and range of equivalency of the claims are therefore intended to be embraced.

Claims

What is claimed is:

1. An information determination device comprising:
an account crawler configured to determine data from at least one pre-determined user account; and
an information determiner configured to determine information related to a pre-determined organization based on the data.
2. The information determination device of claim 1,
wherein the account crawler comprises a known account crawler configured to determine the data from at least one account for the organization.
3. The information determination device of claim 1 or 2,
wherein the account crawler comprises a key-user crawler configured to determine the data from at least one account of a key user.
4. The information determination device of any one of statements 1 to 3, further comprising:
a keyword crawler configured to determine further data based on at least one pre-determined keyword;

wherein the information determiner is further configured to determine the information further based on the further data.

5. The information determination device of any one of claims 1 to 4, wherein the keyword crawler comprises a fixed keyword crawler configured to determine the further data based on at least one fixed keyword.
6. The information determination device of any one of claims 1 to 5, wherein the keyword crawler comprises a dynamic keyword crawler configured to determine the further data based on at least one dynamic keyword, wherein at least one dynamic keyword is changed based on processing of the information determination device.
7. The information determination device of any one of claims 1 to 6, further comprising:
a user friend list crawler configured to determine a user graph of users in a social relationship with at least one of the organization or each other.
8. The information determination device of claim 7, wherein the account crawler is further configured to determine the at least one pre-determined user account based on the user graph.

9. The information determination device of any one of claims 1 to 8, further comprising:

a classifier configured to classify the data into data relevant to the pre-determined organization and data irrelevant to the pre-determined organization.
10. The information determination device of claim 9,

wherein the information determiner is further configured to determine the information based on the data relevant to the pre-determined organization.
11. The information determination device of any one of claims 1 to 10, further comprising:

a topic miner configured to detect whether the determined information related to the pre-determined organization is related to an evolving topic or to an emerging topic about the pre-determined organization.
12. The information determination device of claim 6, further comprising:

a topic miner configured to detect whether the determined information related to the pre-determined organization is related to an evolving topic or to an emerging topic about pre-determined organization;

wherein the dynamic keyword crawler is further configured to determine the at least one dynamic keyword based on the information of the evolving topic and based on the information of the emerging topic.

13. The information determination device of claim 9,
wherein the account crawler is further configured to determine the at least one pre-determined user account based on the data relevant to the pre-determined organization.
14. The information determination device of any one of claims 9, 10 or 13,
wherein the classifier is configured to classify the data based on learning.
15. The information determination device of any one of claims 9, 10, 13 or 14,
wherein the classifier is configured to classify the data based on a support vector machine.
16. The information determination device of claim 11,
wherein the topic miner is configured to detect whether the determined information related to the pre-determined organization is related to the evolving topic or to the emerging topic based on learning.
17. The information determination device of claim 11 or 16,
wherein the topic miner is configured to detect whether the determined information related to the pre-determined organization is related to an evolving topic or to an emerging topic based on solving an optimization problem.
18. The information determination device of claim 17,

wherein the optimization problem comprises a temporal continuity constraint.

19. The information determination device of claim 17 or 18,
wherein the optimization problem comprises a sparse matching constraint.
20. The information determination device of any one of claims 17 to 19, further comprising:
an optimization problem solver configured to solve the optimization problem based on a least angle regression.
21. The information determination device of any one of claims 11, 17, 18, 19, or 20 further comprising:
a trivial topic purging circuit configured to remove old topics from evolving topics and emerging topics.
22. An information determination method comprising:
determining data from at least one pre-determined user account; and
determining information related to a pre-determined organization based on the data.
23. The information determination method of claim 22, further comprising:
determining the data from at least one account for the organization.

24. The information determination method of claim 22 or 23, further comprising:
determining the data from at least one account of a key user.
25. The information determination method of any one of claims 22 to 24, further comprising:
determining further data based on at least one pre-determined keyword; and
determining the information further based on the further data.
26. The information determination method of any one of claims 22 to 25, further comprising:
determining the further data based on at least one fixed keyword.
27. The information determination method of any one of claims 22 to 26, further comprising:
determining the further data based on at least one dynamic keyword, wherein the at least one dynamic keyword is changed based on processing of the information determination method.
28. The information determination method of any one of claims 22 to 27, further comprising:
determining a user graph of users in a social relationship with at least one of the organization or each other.

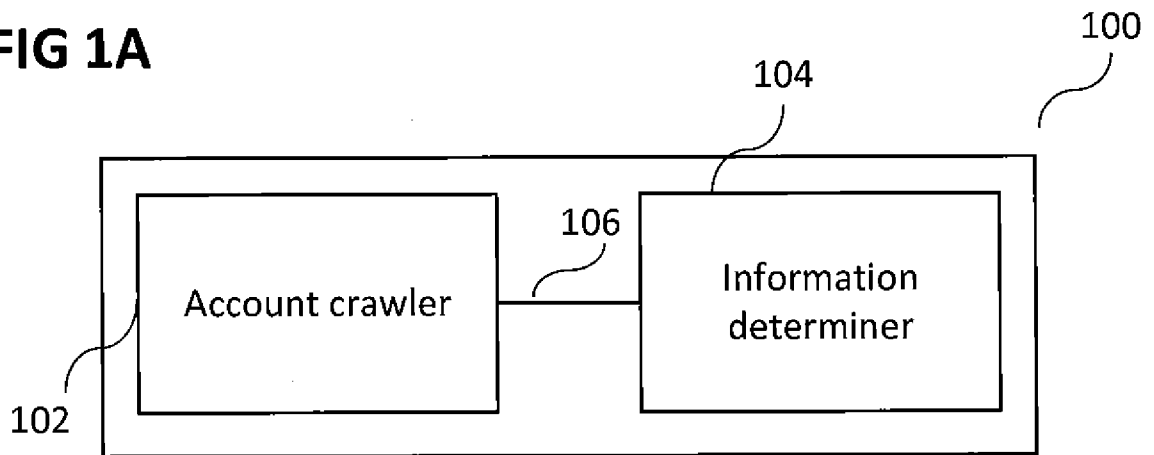
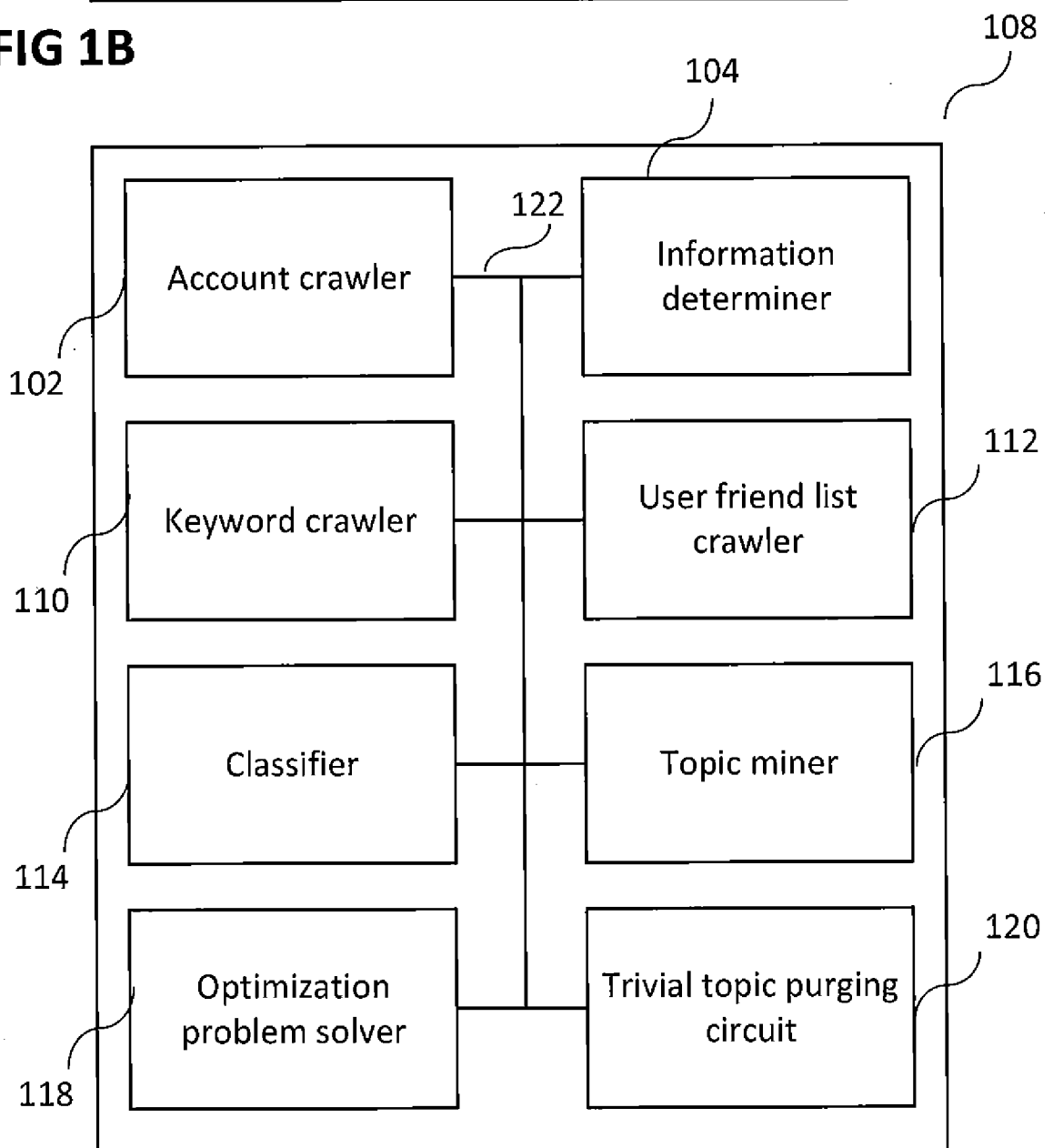
29. The information determination method of claim 28, further comprising:
determining the at least one pre-determined user account based on the user graph.
30. The information determination method of any one of claims 22 to 29, further comprising:
classifying the data into data relevant to the pre-determined organization and data irrelevant to the pre-determined organization.
31. The information determination method of claim 30, further comprising:
determining the information based on the data relevant to the pre-determined organization.
32. The information determination method of any one of claims 22 to 31, further comprising:
detecting whether the determined information related to the pre-determined organization is related to an evolving topic or to an emerging topic.
33. The information determination method of claim 27, further comprising:
detecting whether the determined information related to the pre-determined organization is related to an evolving topic or to an emerging topic; and
determining the at least one dynamic keyword based on the information of the evolving topic and based on the information of the emerging topic.

34. The information determination method of claim 30, further comprising:
determining the at least one pre-determined user account based on the data relevant to the pre-determined organization.
35. The information determination method of any one of claims 30, 31 or 34, further comprising:
classifying the data based on learning.
36. The information determination method of any one of claims 30, 31, 34 or 35, further comprising:
classifying the data based on a support vector machine.
37. The information determination method of claim 32, further comprising:
detecting whether the determined information related to the pre-determined organization is related to the evolving topic or to the emerging topic based on learning.
38. The information determination method of claim 32 or 37, further comprising:
detecting whether the determined information related to the pre-determined organization is related to an evolving topic or to an emerging topic based on solving an optimization problem.
39. The information determination method of claim 38,

wherein the optimization problem comprises a temporal continuity constraint.

40. The information determination method of claim 38 or 39,
wherein the optimization problem comprises a sparse matching constraint.
41. The information determination method of any one of claims 38 to 40, further comprising:
solving the optimization problem based on a least angle regression.
42. The information determination method of any one of claims 32, 38, 39, 40, or 41
further comprising:
removing old topics from evolving topics and emerging topics.

1/9

FIG 1A**FIG 1B**

2/9

FIG 1C

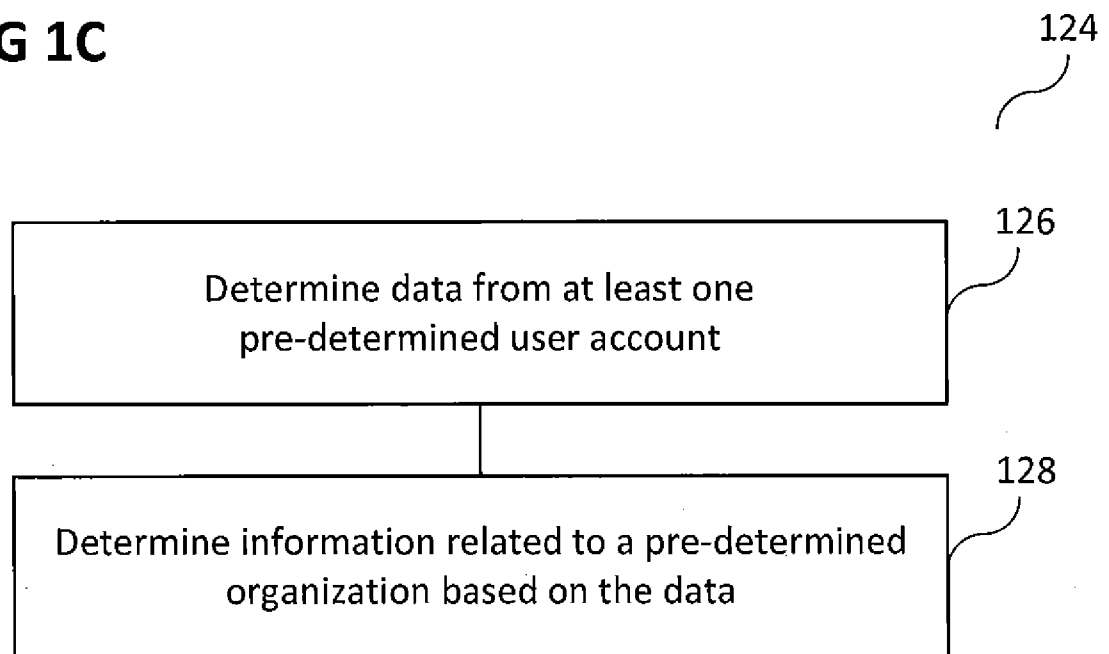


FIG 2

200

 **Optus** ✓
@Optus

We're here to hear from you. If you've got questions or just something you'd like to tell us, we're online Monday-Friday from 9am to 8pm & Saturday 9am-5pm AEST
Australia · <http://www.optus.com.au>

Following

87,844 TWEETS
3,439 FOLLOWING
25,692 FOLLOWERS

FIG 3

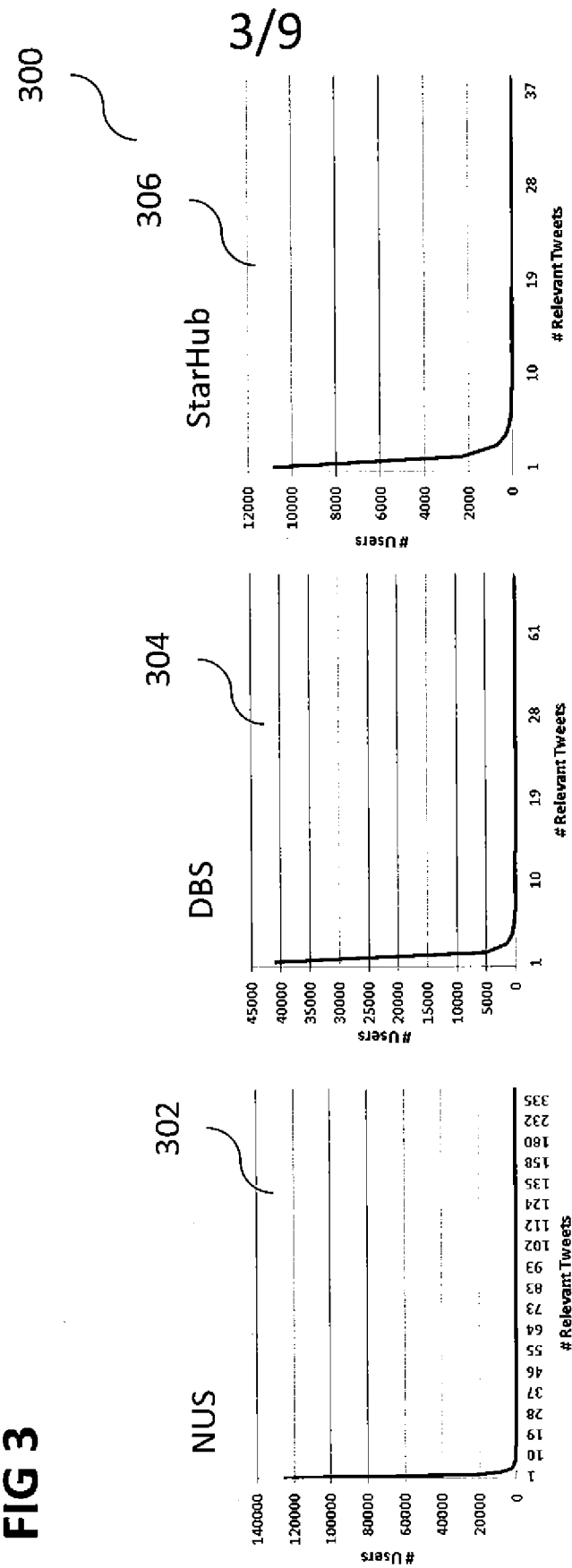


FIG 4

4/9

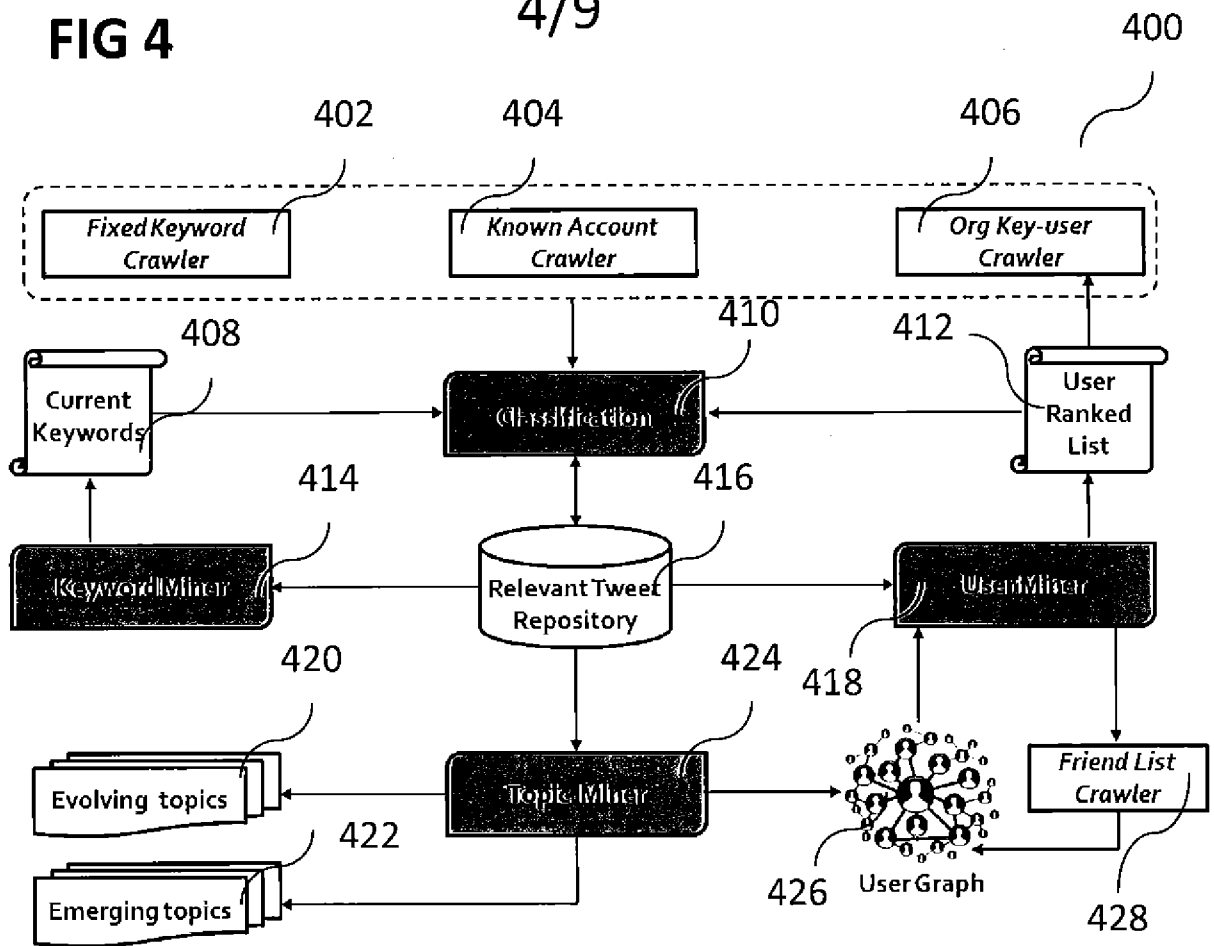


FIG 5

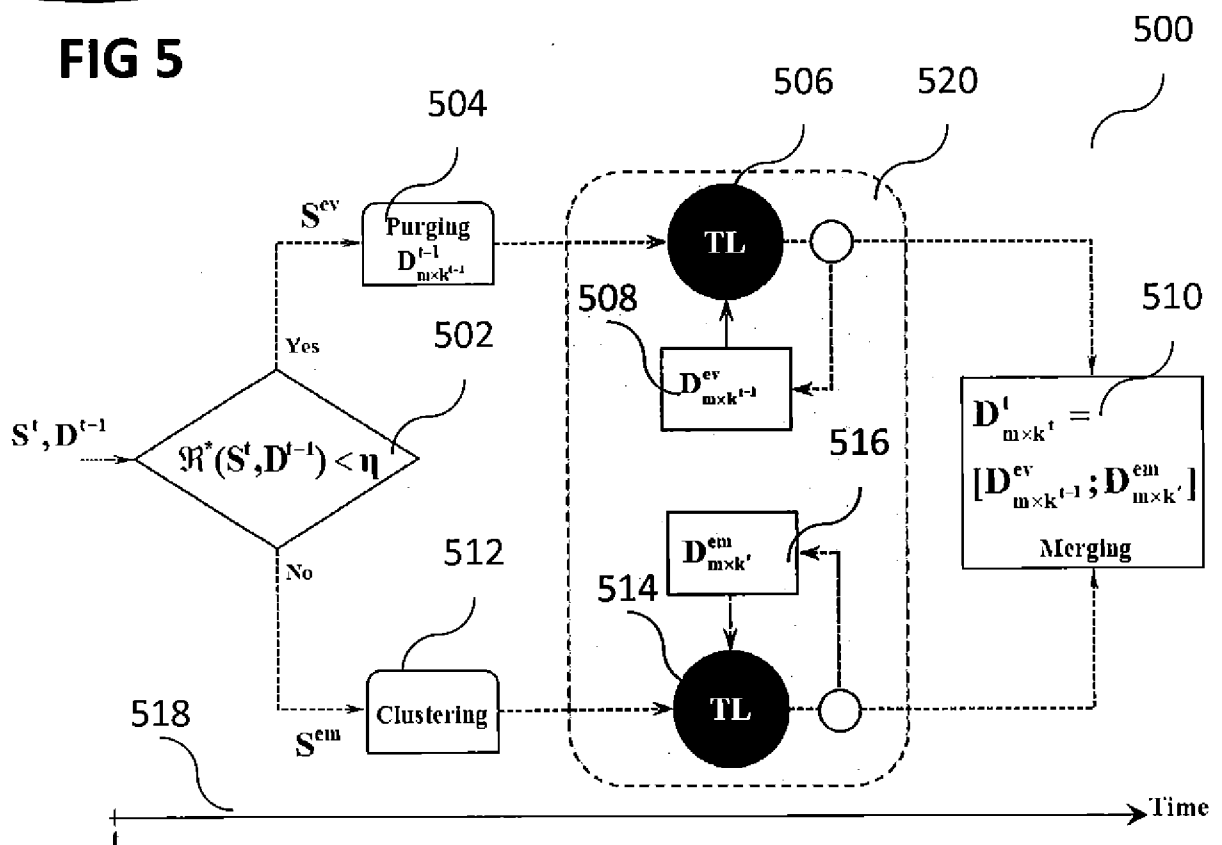


FIG 6

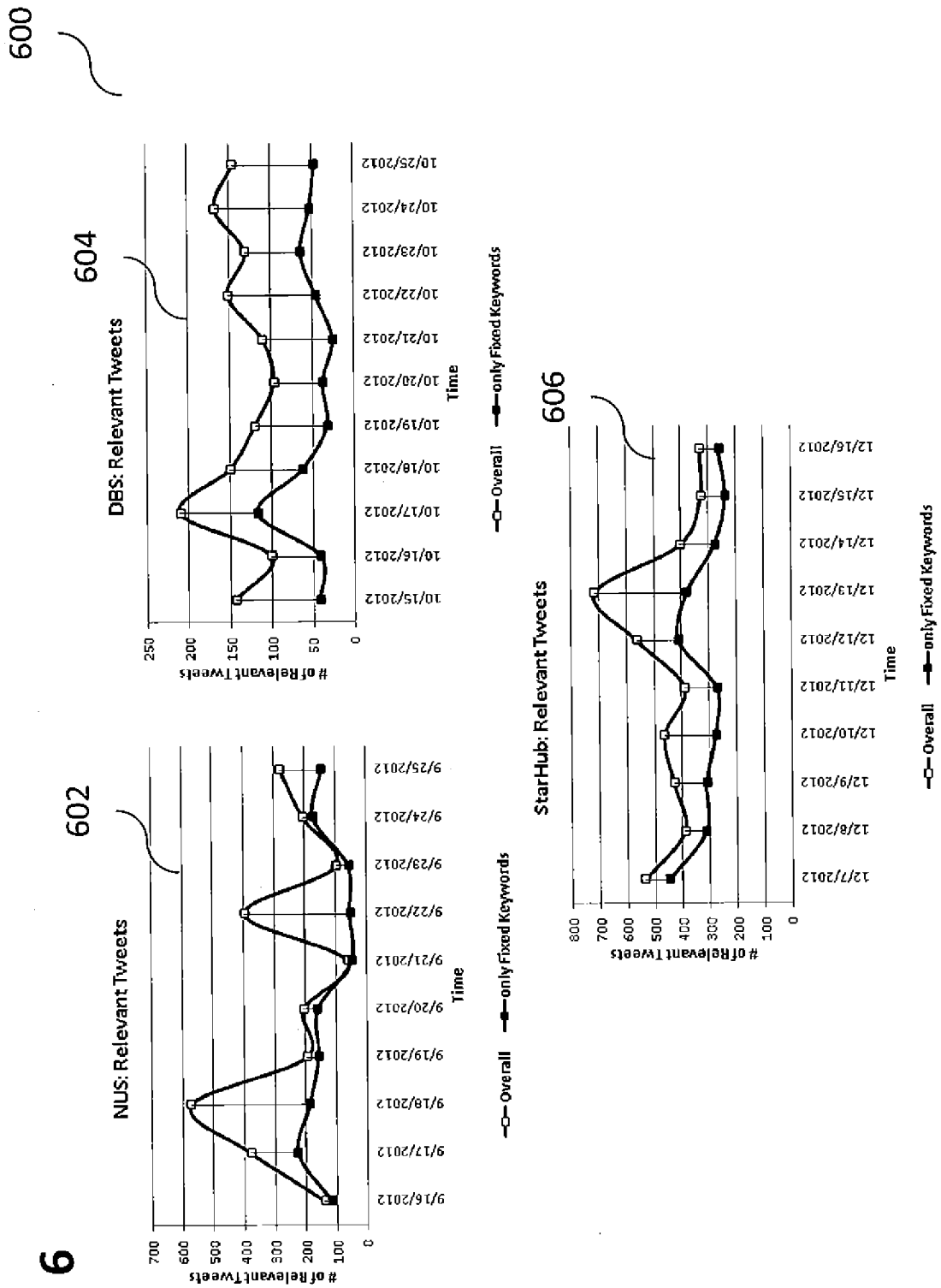


FIG 7

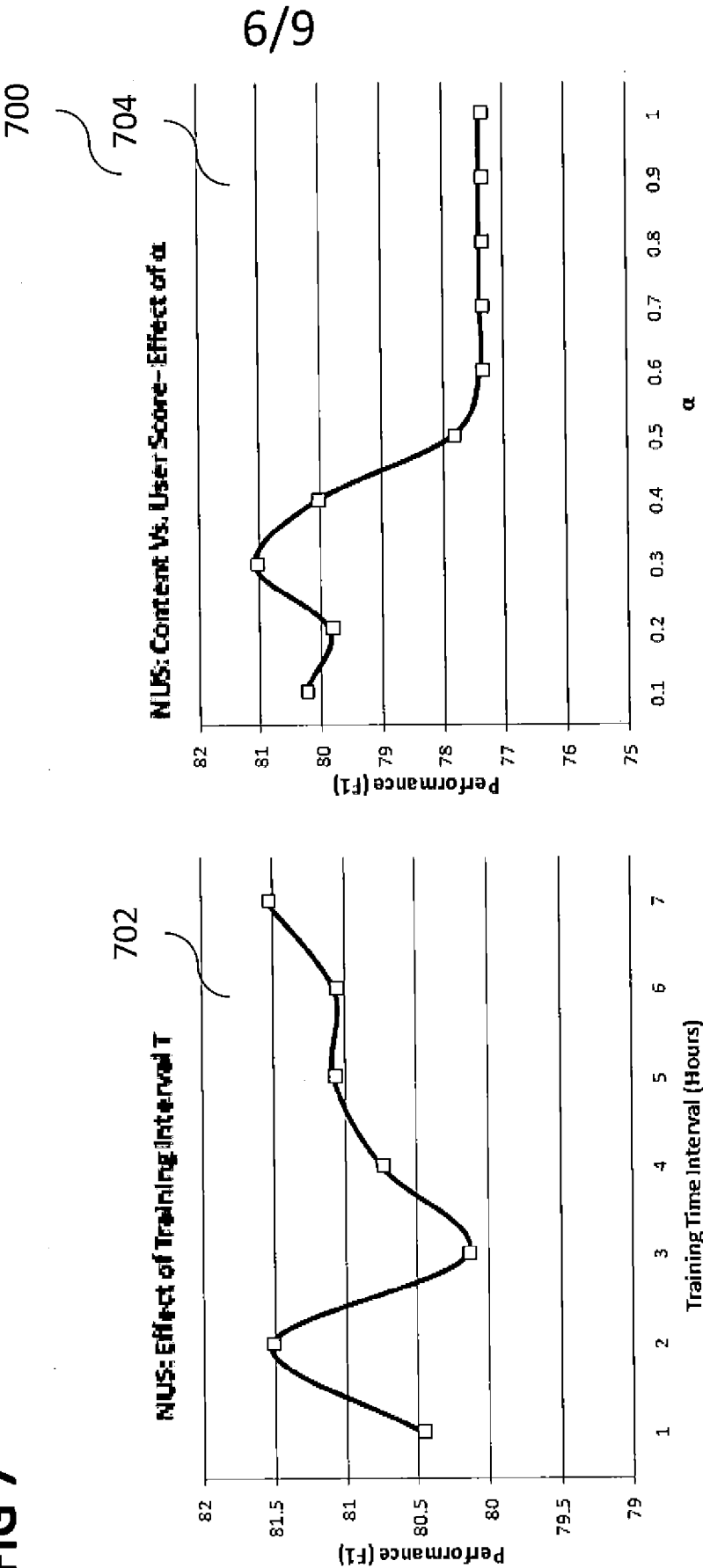


FIG 8

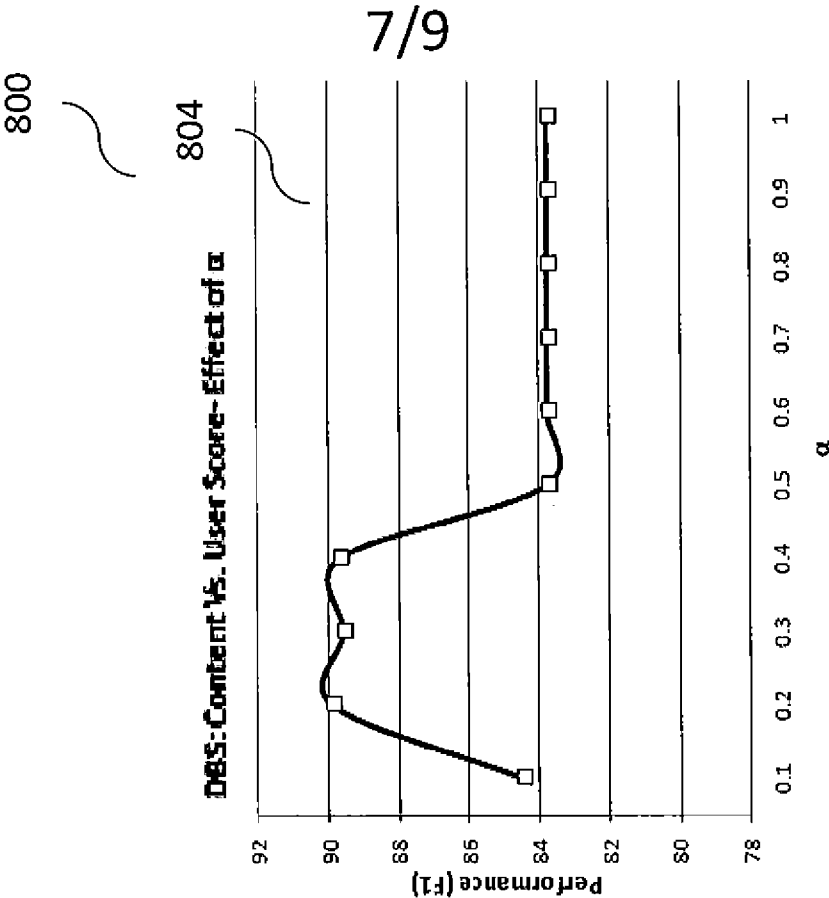
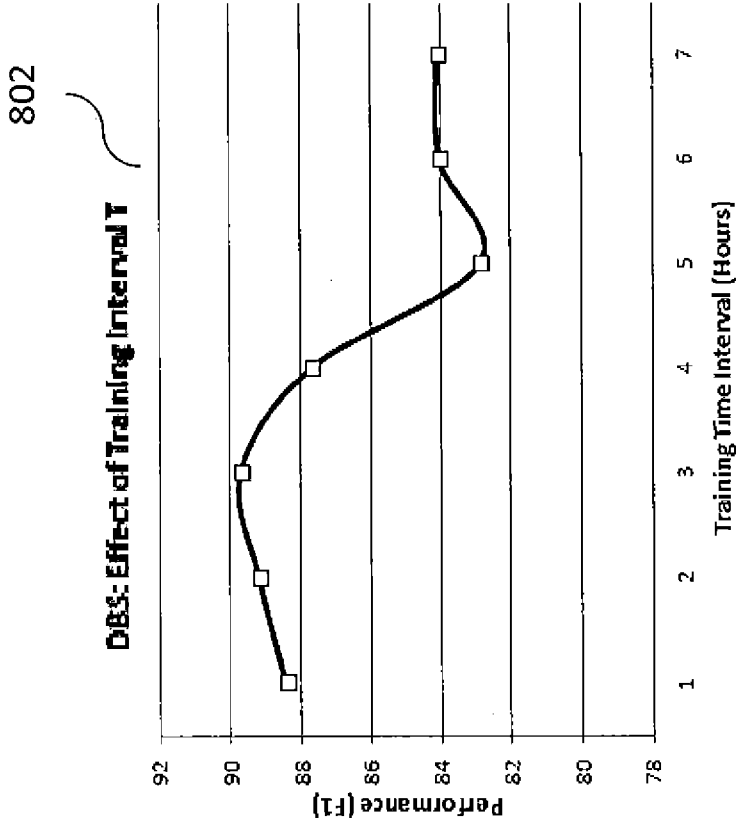


FIG 9

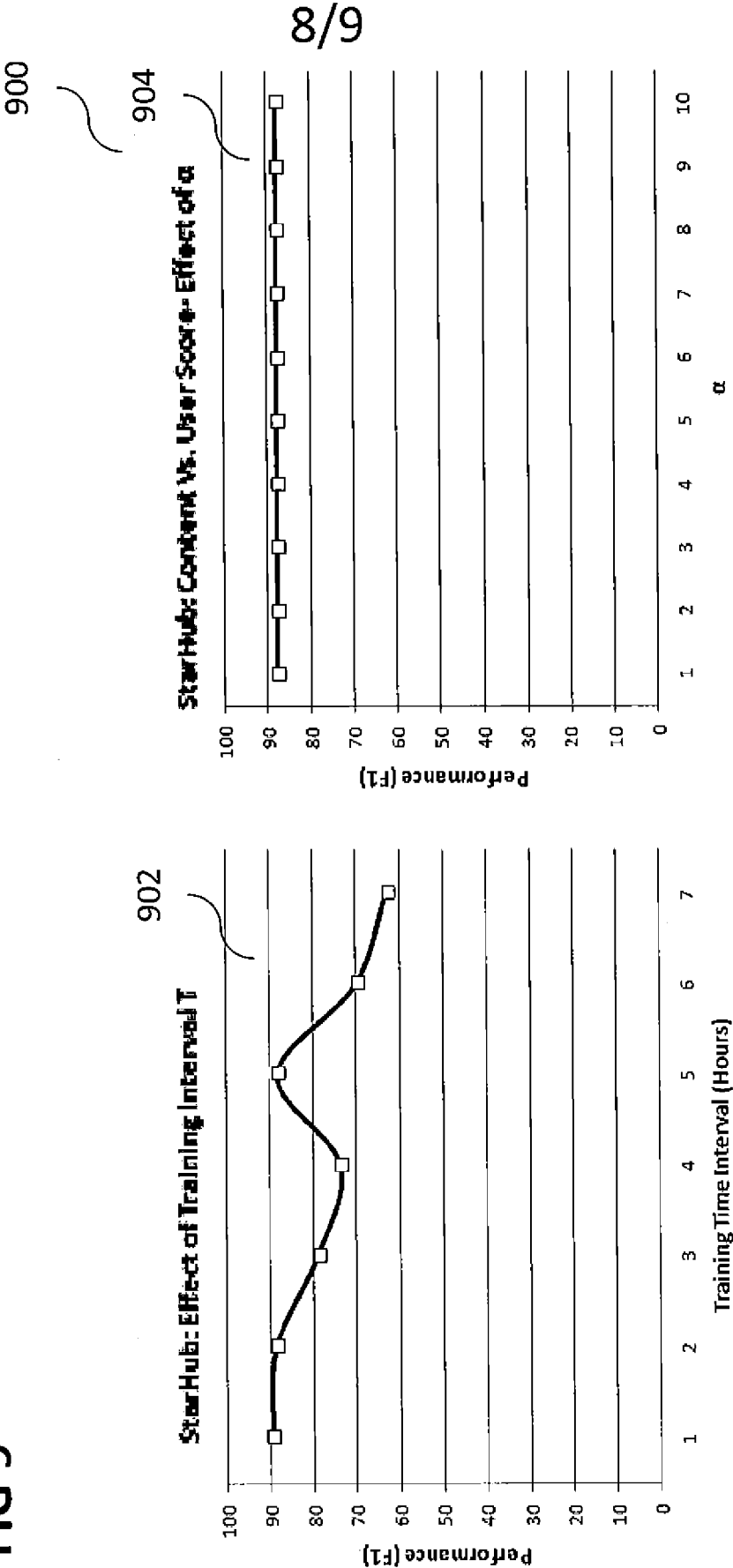


FIG 10

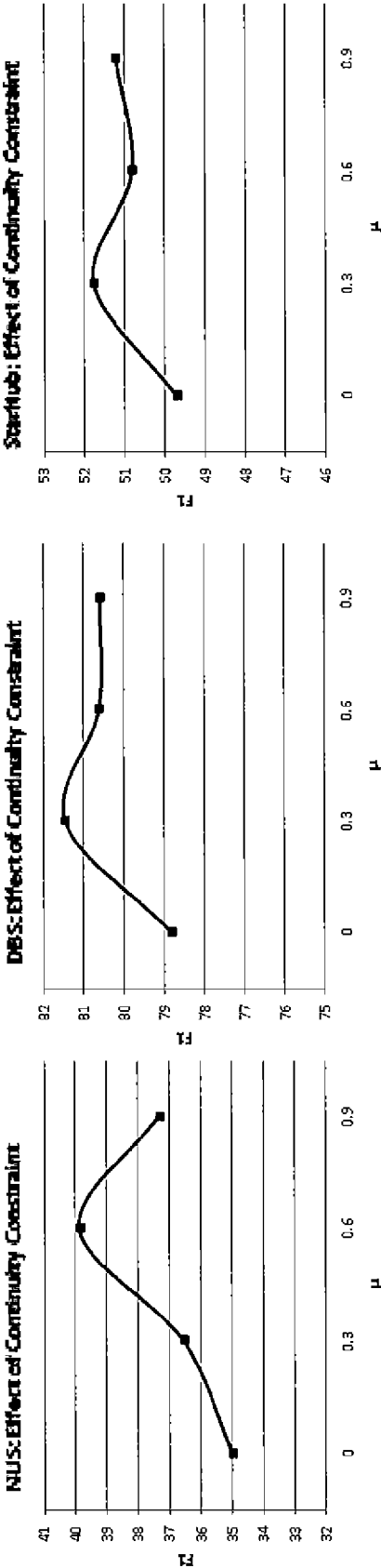
1000

1006

1004

1002

9/9



INTERNATIONAL SEARCH REPORT

 International application No.
PCT/SG2014/000508

A. CLASSIFICATION OF SUBJECT MATTER

G06Q 30/00 (2012.01)

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)

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

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

Database : AusPat : & Keywords (amiribrahimabadi, anqi, chen, chua, cui, hadi, tat-seng, yan) & like terms. Database : PAMS : & Keywords (amiribrahimabadi, anqi, chen, chua, cui, hadi, tat-seng, yan) & like terms. Database : Espacenet : & Keywords (amiribrahimabadi, anqi, chen, chua, cui, hadi, singapore, tat-seng, yan) & like terms. Database : WPI : & Keywords (account, blog, brand, class, constraint, crawler, data, emerge, evolve, facebook, filter, friend, graph, info, keyword, learn, message, microblog, micropost, miner, monitor, organ, parse, post, predetermine, purge, regress, search, social media, social network, tweet, twitter, user, vector, weblog, weibo) & like terms. Database : TXTUS5, TXTUS4, TXTUS3, TXTUS2, TXTUS1, TXTUS0, TXTEP1, TXTGB1, TXTWO1, TXTAU1, TXTCA1, TXTSG1 : & Keywords (account, angle, classify, configure, constraint, continuity, crawler, data, determine, dynamic, emerge, evolve, fixed, friend, graph, inform, key user, keyword, known, learn, list, match, miner, optimum, organ, predetermine, problem, purge, regress, relevant, remove, support, topic, user, vector) & like terms.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	Documents are listed in the continuation of Box C	



Further documents are listed in the continuation of Box C



See patent family annex

* "A"	Special categories of cited documents: document defining the general state of the art which is not considered to be of particular relevance	"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
"E"	earlier application or patent but published on or after the international filing date	"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
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"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
"O"	document referring to an oral disclosure, use, exhibition or other means	"&"	document member of the same patent family
"P"	document published prior to the international filing date but later than the priority date claimed		

 Date of the actual completion of the international search
 4 February 2015

 Date of mailing of the international search report
 04 February 2015

Name and mailing address of the ISA/AU

 AUSTRALIAN PATENT OFFICE
 PO BOX 200, WODEN ACT 2606, AUSTRALIA
 Email address: pct@ipaustalia.gov.au

Authorised officer

 Ross Stopford
 AUSTRALIAN PATENT OFFICE
 (ISO 9001 Quality Certified Service)
 Telephone No. 0262832177

INTERNATIONAL SEARCH REPORT C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		International application No. PCT/SG2014/000508
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2007/101263 A2 (BUZZLOGIC, INC.) 07 September 2007 [00103], [00105], [00106], [0018], [00213], [00222], [0044], [0045], [0046], [0047], [0050], [0056], [0058], [0067], [0075], [0076], claim 30	1 - 42
X	WO 2008/045792 A2 (TECHNORATI, INC.) 17 April 2008 [00108], [00112], [00124], [0018], [0028], [0036], [0038], [0040], [0042], [0053], [0075], [0076], [0084], [0086], [0088]	1 - 42
A	US 2013/0103490 A1 (ABE et al) 25 April 2013 Abstract	1 - 42
A	US 2010/0063948 A1 (VIRKAR et al) 11 March 2010 Abstract	1 - 42

INTERNATIONAL SEARCH REPORT		International application No.	
Information on patent family members		PCT/SG2014/000508	
This Annex lists known patent family members relating to the patent documents cited in the above-mentioned international search report. The Australian Patent Office is in no way liable for these particulars which are merely given for the purpose of information.			
Patent Document/s Cited in Search Report		Patent Family Member/s	
Publication Number	Publication Date	Publication Number	Publication Date
WO 2007/101263 A2	07 September 2007	AU 2007219997 A1	07 Sep 2007
		EP 1989639 A2	12 Nov 2008
		JP 2009528639 A	06 Aug 2009
		US 2010070485 A1	18 Mar 2010
		US 8682723 B2	25 Mar 2014
		US 2007214097 A1	13 Sep 2007
		US 2009119173 A1	07 May 2009
		US 2014297403 A1	02 Oct 2014
WO 2008/045792 A2	17 April 2008	US 2008147487 A1	19 Jun 2008
US 2013/0103490 A1	25 April 2013	US 2007174105 A1	26 Jul 2007
		US 8359226 B2	22 Jan 2013
		US 2008177621 A1	24 Jul 2008
		US 8452639 B2	28 May 2013
US 2010/0063948 A1	11 March 2010	US 8386401 B2	26 Feb 2013
		US 2013238533 A1	12 Sep 2013
		WO 2010030794 A1	18 Mar 2010
End of Annex			
Due to data integration issues this family listing may not include 10 digit Australian applications filed since May 2001.			