METHOD AND SYSTEM FOR IDENTIFYING HIGH VALUE SOCIAL AUDIENCE

Field of the Invention

This invention relates to processes for identifying a group of high-value audience for an owner of an account in a social media. More particularly, this invention relates to processes for dynamically identifying relevant information from data retrieved over the Internet. Still more particularly, this invention relates to processes for identifying entities of interest from data retrieved over the Internet.

Prior Art

The Internet has become a very popular medium for dispersing information. Recently, with the rise in popularity of blogs and social networks, companies and organisations have been sharing content on social media such as Facebook, Twitter or the likes. However, it is increasingly difficult to find prospective customers or interested audience from the crowded social space.

A few social media analytics tools such as Klout, Twitalyzer, Twitterrank, etc. have been developed to analyse the social space. Such social media analytics tools have focused on identifying a few influencers that are related to a particular topic and depending on these influencers to achieve viral spread. However, the identified audience based on this method usually does not truly translate the audience to a customer.

As such, those skilled in the art are constantly striving to provide analytics tools that provide a more accurate result of target audience from social media in order for companies to better allocate resources when planning their marketing strategy for either customer retention or new products promotion.

Summary of the Invention

The above and other problems are solved and an advance in the art is made by a system and method of identifying a group of high-value audience for an owner of an account in a social media in accordance with this invention. A first advantage of a system and method of identifying a group of high-value audience for an owner of an account in a social

media in accordance with this invention is that the system is able to automatically generate a topic entity map based on various online resources. As this topic entity map developed is industry specific, it can also be used in other aspect such as emerging topic discovery. A second advantage of a system and method of identifying a group of high-value audience for an owner of an account in a social media in accordance with this invention is that the group of high-value audience is determined based on contents published by the audiences with the consideration of temporal effect. Hence, the result is more accurate compared to common method which is based on certain influencers. This means that the result from the system is not depending on influencers but is capable of identifying audience who are highly likely interested in the product. The method and system can also be expanded to analyse any content of any platforms and not limited to the content shared by the friends, fans or followers of an owner of an account in social media to understand the potential of a particular product or to measure the ground-feel or sentiment of a promotion or campaign. A third advantage of a system and method of identifying a group of high-value audience for an owner of an account in a social media in accordance with this invention is that the value index used can be used to segment audience into various groups such as high-value, mid-value and low-value group. Different incentive and engagement approaches can be applied to different groups so that marketing budget can be better planned through this actionable insight.

In accordance with an aspect of the invention, the system and method for determining high value audience from a social media may be performed in the following manner. The method retrieves information created by an owner of an account in the social media and generates a topic-entity map based on the information retrieved. The topic-entity map includes topics and entities, the entities being determined based on the information retrieved from the account of the owner, and each of the topics is determined based on the determined entities. The method then retrieves information from accounts of users in the social media where the users are associated with the account of the owner. Relevant information from information retrieved from accounts of users in the social media is extracted

and compared with each of the entities in the topic-entity map. The information is then assigned to one of the entities in the topic-entity map accordingly. Subsequently, a value index is determined for each of the users based on said information from the users being assigned in the topic-entity map. The users are arranged according to the value index from highest to lowest.

In accordance with an embodiment of the invention, the step of retrieving information created by the owner of an account in the media may be performed in the following manner. User information from the owner is received, wherein the user information includes username and name of social media. A cutoff date is then received from the owner. The method then extracts all information from the cutoff date for the username in the social media. The information is content being published by the username in the social media.

In accordance with an embodiment of the invention, the step to generating a topic-entity map based on said information retrieved may be performed in the following manner. The contents are pre-processed by removing symbol in the contents and normalising exaggerative words. The contents are then segmented by dividing the content into segments, wherein conjunction, determiners, stop words and punctuation in the content are being used to separate the content into segments. The segments are arranged based on frequency and the top N segments are determined as entities.

In accordance with an embodiment of the invention, the step to generating a topic-entity map based on the information retrieved may be performed in the following manner. A website is received from the owner. Data from the website is retrieved in response to receiving a website from the owner. A similarity search between each of the entities and the data retrieved is performed. A topic is determined for each of the entities based on the data retrieved.

In accordance with another embodiment of the invention, the step to generating a topic-entity map based on the information retrieved may be performed in the following manner. The method performs a search using relevant online resources using each of the entities as keywords a search result for each of the entities. The method then determines a

topic based on a category in which the entity is being classified in the search result for each of the entities. Subsequently, the topic-entity map is generated by grouping the entity that determines the topic to the entity.

In accordance with another embodiment of the invention, the step to retrieving information from accounts of users in the social media may be performed in the following manner. A list of users associated to the owner is retrieved. Information from each of the users after the cutoff date is retrieved, wherein the information is content published by each of the users.

In accordance with another embodiment of the invention, the step to extracting relevant information from information retrieved from accounts of users in the social media may be performed in the following manner. Each of the contents being published by the users are pre-processed by removing symbol and platform-specific term in the contents and normalising exaggerative words. Subsequently, each of the contents is segmented by dividing the content into segments, wherein conjunction, determiners, stop words and punctuation in the content are being used to separate the content into segments.

In accordance with another embodiment of the invention, the step to assigning the information to at least one of said entity in the topic-entity map may be performed in the following manner. A similarity score for each of the segments against each of the entities in the topic-entity map is determined. A content is assigned to the entity in the topic-entity map if the similarity score of one of the segments of the content is above a certain threshold value

In accordance with another embodiment of the invention, the value index for each of the users may be determined in the following manner. The users for each of the contents being grouped to one of the topic in said topic entity map are determined. All contents retrieved for each of the determined users are retrieved. A recency index for each of the determined users is determined with the following expression,

$$\text{Recency} = \text{invert}((LR - R + 1)/\text{max})$$

where LR is the latest post of the determined user, R is the latest post date of the topic by the determined user and max is the time period between the earliest relevant content and the most recent relevant content from the whole corpus where the corpus is all the content from the friends, fans and followers of the owner of the account in social media after the cutoff date. A frequency index for each of the determined users is then determined with the following expression,

Frequency = Nr / total number of contents after cutoff date

where Nr is the total number of contents related to the topic and total number of contents after cutoff date is the total number of contents published by all of the determined users. The recency and frequency of each of the users are then normalised. The value index for each of the users is then determined with the following expression, Value index = (Recency + Frequency)/2.

Brief Description of the Drawings

The above and other features and advantages of a method and/or system in accordance with this invention are described in the following Detailed Description and are shown in the following drawings:

Figure 1 illustrates a network to provide or perform a system and/or method in accordance with embodiments of this invention;

Figure 2 illustrates a processing system of a server performing processes to provide a system in accordance with this invention;

Figure 3 illustrates a processing system of a mobile device performing processes to provide a system in accordance with this invention;

Figure 4 illustrates a process for determining high value audience for an owner in accordance with an embodiment of this invention;

Figure 5 illustrates a process for retrieving information created by the owner in accordance with an embodiment of this invention;

Figure 6 illustrates a process for determining the entity from the information retrieved in accordance with an embodiment of this invention;

Figure 7 illustrates a process for processing the content in accordance with an embodiment of this invention;

Figure 8 illustrates a process for determining the topic and generating a topic entity map in accordance with an embodiment of this invention;

Figure 9 illustrates a process for retrieving contents published by users associated to the owner in accordance with an embodiment of this invention;

Figure 10 illustrates a process for extracting relevant information of contents retrieved from users associated to the owner in accordance with an embodiment of this invention;

Figure 11 illustrates a process for assigning content to the topic entity map in accordance with an embodiment of this invention; and

Figure 12 illustrates a process for determining a value index for each of the users according to a topic in the topic entity map in accordance with an embodiment of this invention.

Detailed Description

This invention relates to processes for identifying a group of high-value audience for an owner of an account in a social media. More particularly, this invention relates to processes for dynamically identifying relevant information from data retrieved over the Internet. Still more particularly, this invention relates to processes for identifying entities of interest from data retrieved over the Internet.

It is envisioned that a system and/or method in accordance with embodiments of this invention may be used to identify a group of high-value audience so that resources can be better planned through this actionable insights. The system adopts a content-based approach to manage dynamisms of the social and not rely on any fixed ontology so that the latest and updated topics and entities can be extracted and used. This is especially important for various different industries as every industry has their own product or service timeline and this temporal effect needs to be captured to attract the prospective audience who are interested in the products or services.

Popularity of blogs and social networks has triggered companies and organisations to share content via social media such as Facebook, Twitter or the likes. Users of the social media who wish to follow the content shared by the companies typically link their account with the account of the companies. For example, a product company may have a Facebook account and Facebook users who wish to follow products or promotion offered by the product company will link their account with the account of the product company by clicking on “Likes” so that future posts from the product company will be shared to the users. For twitter accounts, users can link their account to the account of the product company by “following” the product company so that future tweets will be shared to the users. By analyzing the contents (i.e. posts or message in Facebook, or tweets or retweets in Twitter) from the product company and the contents of users, the system is able to determine high-value audience who are highly likely interested in the product or service. The system and method of extracting and analyzing of contents published in social media to determine high value audiences will now be described below.

Figure 1 illustrates a network 100 that includes a server 110 and/or mobile device 115 for performing the processes of an embodiment of this invention. Network 100 is a network such as the Internet that allows processing systems to communicate with one another.

Server 110 is a typical processing system such as a desktop computer, laptop computer, or other computer terminal. Server 110 is communicatively connected to network 100 via either a wired or wireless connection to retrieve data from social media 140 and/or online resources 150. Server 110 is also communicatively connected to mobile device 115 via network 100 via either a wired or wireless connection. One skilled in the art will recognise that more than one server 110 may be used without departing from the invention. Further details of the server 110 will be described below with reference to figure 2.

Mobile device 115 as shown in figure 3 is a representation of a tablet. One skilled in the art will recognise that although mobile device 115 shown in figure 3 is a representation of a tablet, the mobile device 115 may also be a personal digital assistants (PDAs), mobile phones, phablets, or other similar mobile devices without departing from the invention.

Mobile device 115 is communicatively connected to network 100 via either a wired or wireless connection to retrieve data from social media 140 and/or online resources 150. Alternatively, mobile device 115 is communicatively connected to server 110 via network 100 to execute instructions to perform the processes in accordance with this invention remotely. Further details of mobile device 115 will be described with reference to figure 3 below.

Online resources 150 are resources that are publicly available from the Internet via network 100. Social media 140 are virtual communities that allow users to interact with one another by sharing and exchanging of information via network 100. Some common virtual communities include Facebook, Twitter, Youtube, Wikipedia, etc.

Processes stored as instructions in a media that are executed by a processing system in server 110 provide the method and/or system in accordance with this invention. The instructions may be stored as firmware, hardware, or software. Figure 2 illustrates processing system 200 such as the processing system in server 110 that execute the instructions to perform the processes for providing a method and/or system in accordance with this invention. One skilled in the art will recognize that the exact configuration of each processing system may be different and the exact configuration of the processing system in each device may vary. Thus, processing system 200 shown in Figure 2 is given by way of example only.

Processing system 200 includes Central Processing Unit (CPU) 205. CPU 205 is a processor, microprocessor, or any combination of processors and microprocessors that execute instructions to perform the processes in accordance with the present invention. CPU 205 connects to memory bus 210 and Input/ Output (I/O) bus 215. Memory bus 210 connects CPU 205 to memories 220 and 225 to transmit data and instructions between the memories and CPU 205. I/O bus 215 connects CPU 205 to peripheral devices to transmit data between CPU 205 and the peripheral devices. One skilled in the art will recognize that I/O bus 215 and memory bus 210 may be combined into one bus or subdivided into many other busses and the exact configuration is left to those skilled in the art.

A non-volatile memory 220, such as a Read Only Memory (ROM), is connected to memory bus 210. Non-volatile memory 220 stores instructions and data needed to operate various sub-systems of processing system 200 and to boot the system at start-up. One skilled in the art will recognize that any number of types of memory may be used to perform this function.

A volatile memory 225, such as Random Access Memory (RAM), is also connected to memory bus 210. Volatile memory 225 stores the instructions and data needed by CPU 205 to perform software instructions for processes such as the processes required for providing a system in accordance with this invention. One skilled in the art will recognize that any number of types of memory may be used as volatile memory and the exact type used is left as a design choice to those skilled in the art.

I/O device 230, keyboard 235, display 240, memory 245, network device 250 and any number of other peripheral devices connect to I/O bus 215 to exchange data with CPU 205 for use in applications being executed by CPU 205. I/O device 230 is any device that transmits and/or receives data from CPU 205. Keyboard 235 is a specific type of I/O that receives user input and transmits the input to CPU 205. Display 240 receives display data from CPU 205 and display images on a screen for a user to see. Memory 245 is a device that transmits and receives data to and from CPU 205 for storing data to a media. Network device 250 connects CPU 205 to a network for transmission of data to and from other processing systems.

Figure 3 illustrates an example of a processing system 300 in the mobile device 115. Processing system 300 represents the processing systems in the mobile device 115 that execute instructions to perform the processes described below in accordance with embodiments of this invention. One skilled in the art will recognize that the instructions may be stored and/or performed as hardware, firmware, or software without departing from this invention. Further one skilled in the art will recognise that the instructions may be installed as a software application that can be retrieved from a third party provider, such as, App Store managed by Apple Inc. or Google Play managed by Google Inc. without departing from this invention. One skilled in the art will recognize that the exact configuration of each

processing system may be different and the exact configuration executing processes in accordance with this invention may vary and processing system 300 shown in figure 3 is provided by way of example only.

Processing system 300 includes a processor 310, a radio transceiver 320, an image capturing device 330, a display 340, a keypad 350, a memory 360, an audio module 370, a Near Field Communication (NFC) module 380, and an I/O device 390.

The radio transceiver 320, image capturing device 330, display 340, keypad 350, memory 360, audio module 370, NFC module 380, I/O device 390 and any number of other peripheral devices connect to processor 310 to exchange data with processor 310 for use in applications being executed by processor 310.

The radio transceiver 320 is connected to an antenna which is configured to transmit outgoing voice and data signals and receive incoming voice and data signals over a radio communication channel. The radio communication channel can be a digital radio communication channel such as a CDMA, GSM or LTE channels that employs both voice and data messages in a conventional techniques.

The image capturing device 330 is any device capable of capturing still and/or moving images such as complementary metal-oxide semiconductor (CMOS) or charge-coupled sensor (CCD) type cameras. The display 340 receives display data from processor 310 and display images on a screen for a user to see. The display 340 may be a liquid crystal display (LCD) or organic light-emitting diode (OLED) display. The keypad 350 receives user input and transmits the input to processor 310. In some embodiments, the display 340 may be a touch sensitive surface that functions as a keypad to receive user input.

The memory 360 is a device that transmits and receives data to and from processor 310 for storing data to a memory. The audio module 370 may include a microphone, an earpiece and a headset. A microphone is a device that transmits audio data to processor 310. An earpiece is a device that receives audio data from the processor. The headset is a

device that transmits and receives audio data to and from the processor 310. The NFC module 380 is a module that allows mobile device 115 to establish radio communication with another similar device by touching them together or by bringing the devices within a close proximity.

Other peripheral devices that may be connected to processor 310 include a Bluetooth transceiver, a Wi-Fi transceiver and a Global Positioning System (GPS).

The processor 310 is a processor, microprocessor, or any combination of processors and microprocessors that execute instructions to perform the processes in accordance with the present invention. The processor has the capability to execute various application programs that are stored in the memory 360. These application programs can receive inputs from the user via the display 340 having a touch sensitive surface or directly from a keypad 350. Some application programs stored in the memory 360 that can be performed by the processor 310 are application programs developed for iPhone, Android, Windows Mobile, Blackberry or other mobile platforms.

In accordance with this invention, a system and method are provided for identifying high value audiences that are linked or associated to an owner of an account of a social media. The system retrieves information created by the owner having an account in the social media. The information may be messages, tweets, or posts being created by the owner. This information is processed to generate a topic-entity map dynamically while considering the temporal effect of different industry. The system then retrieves information created by users who are associated or linked to the owner in the social media. This information is compared using the topic-entity map generated and assigned accordingly. The system then determines the value index for all the users based on recency and frequency of the assigned information. The users are then arranged according to the value index in ascending order where the user with the highest value index is arranged at the top while the user with the lowest value index is arranged at the bottom. Further details of the invention will now be described below.

Figure 4 illustrates a flow diagram of process 400 performed by the processor in server 110 or mobile device 115 in accordance with an embodiment of this invention. Process 400 begins with step 405 by retrieving information created by the owner of an account in the social media. Using Twitter as an example, tweets created by the owner of a Twitter account are retrieved based on temporal effect related to the topic of interest. Further details of retrieving information created by the owner of a social media account will be described below with reference to figure 5.

In step 410, the information retrieved is processed in order to generate a topic-entity map. Briefly, the topic-entity map includes entities which determined based on the information retrieved and topics which are determined based on the determined entities. Further details of generating of the topic-entity will be described below with reference to figures 6-8.

In step 415, process 400 retrieves information from users in the same social media associated with the owner. Using Twitter as an example, the tweets created by other users who are following the owner are retrieved. This information is processed and relevant information is extracted in step 420. Further details of retrieving information from users in the same social media associated with the owner will be described below with reference to figure 9. Further details of extracting relevant information will be described below with reference to figure 10.

The extracted information is compared with the topic-entity map in step 425. The extracted information is then assigned according to the topic-entity map in step 430. Further details of assigning of the extracted information will be described below with reference to figure 11. Process 400 ends after step 430.

Figure 5 illustrates a flow diagram of process 500 performed by the processor in server 110 or mobile device 115 in accordance with an embodiment of this invention. Process 500 is a process to further elaborate step 405 in process 400. Process 500 beings with step 505 by requesting for user information. The user information required from the owner is a name of a social media that the owner wishes to retrieve information from and the

relevant username. For example, if the owner has a Twitter account and wishes to determine the high value audience from its followers, the owner has to select Twitter and input its Twitter's username. If the owner has a Facebook account or Facebook fanpage and wishes to determine the high value audience from its friends or users or fans who "Likes" their Facebook account, the owner has to select Facebook and input its Facebook's username or page name. One skill in the art will recognise that various social media platforms may be requested and the system may be modified to retrieve information from one or more social media platforms to identify target audiences from various social media concurrently without departing from the invention.

In step 510, process 500 receives user information from the owner via keypad 235 of processing system 200 or keypad 350 or touch sensitive surface incorporated with display 340 of processing system 300.

In step 515, process 500 requests for cutoff date. The cutoff date is the date where the system will retrieve information having a date later than the cutoff date. The cutoff date is dependent on the industry or the specification of the user. For example, it would be logical to consider only the last six months of data for a mobile industry, as the changes in mobile industry is rather fast, but a longer period for tourism industry is considered, as the changes in tourism industry takes longer. This is also known as considering the temporal effect based on industry. The information may be content such as messages, tweets or posts published by the owner.

In step 520, process 500 receives cutoff date from the owner via keypad 235 of processing system 200 or keypad 350 or touch sensitive surface incorporated with display 340 of processing system 300.

In step 525, process 500 accesses the social media with the username and extracts all contents published by the owner after the cutoff date. Process 500 ends after step 525.

Figure 6 illustrates a flow diagram of process 600 performed by the processor in server 110 or mobile device 115 in accordance with an embodiment of this invention.

Process 600 is a process to further elaborate step 410 in process 400. Process 600 beings with step 605 by pre-processing each of the extracted contents in step 525 in process 500. The pre-processing step is required to create a string that can be later used for searching and analysing. Further details of pre-processing of contents will be described further with reference to figure 7 below.

In step 610, process 600 generates segments for each of the contents. The segments are generated by using the common words (such as conjunctions or determiners), stop words and punctuation as separator or delimiter. For example, if a content is “Samsung is holding a galaxy contest!”, three segments will be generated for the content as follows: (samsung) | (holding) | (galaxy contest).

In step 615, the segments are ranked based on frequency. The segment with the highest number of occurrence will be ranked the highest while the segment with the lowest number of occurrence will be ranked the lowest.

In step 620, the top N segments will be determined as the potential entities. Top N segments are adopted as potential entities and N can be customised according to user’s preference. N may be set to a value of between 10 and 20. However, one skilled in the art will recognise that a stringent N will be very specific but will also limit the number of potential high value audience discovered while a more relaxed N may have implication on the system performance and hence the exact value of N is left as a design choice to those skilled in the art.

In step 625, process 600 determines the topic for each of the segments. Further details of determining the topics will be described further with reference to figure 8 below. Process 600 ends after step 625.

Figure 7 illustrates a flow diagram of process 700 performed by the processor in server 110 or mobile device 115 in accordance with an embodiment of this invention. Process 700 is a process to further elaborate step 606 in process 600. Process 700 beings

with step 705 by removing URL or other specific non-informative platform dependent symbol such as #, etc or platform-specific term such as @username, etc.

In step 710, process 700 normalise exaggerative words such as “goooooood” and “noooooo” via regular expressions. Process 700 ends after step 710.

Figure 8 illustrates a flow diagram of process 800 performed by the processor in server 110 or mobile device 115 in accordance with an embodiment of this invention. Process 800 is a process to further elaborate step 625 in process 600. Process 800 beings with step 805 by requesting owner for a website hosted by the owner. In step 810, if the owner has a website and enters one or more websites, process 800 proceeds to step 815. If the owner does not have a website, process 800 proceeds to step 830.

In step 815, process 800 retrieves the data from the website owner has provided. In step 820, process 800 performs a string similarity comparison between each of the determined potential entities and the data from the website. Words or phrases in the data that is similar to the determined entities with a score above a threshold of at least 0.85 will be extracted. One skilled in the art will note that a higher threshold would return a more accurate match while a lower threshold would return a less accurate match and hence the exact value of the threshold is left as a design choice to those skilled in the art.

In step 825, process 800 identifies the topic associated to the extracted words or phrases. This step is repeated for each of the entities so that a topic is identified for each of the entities. In brief, steps 815-825 use the owner’s website as a source for identifying the topic related to each of the entities.

In step 830, process 800 searches relevant online resources using each of the entities as keywords. In step 835, process 835 identifies the topic associated to each of the entities. The online resources may be selected based on the relevant topic of interest or industry. For example, the tourism board from a specified country may be used to identify topic for each of the entities which are relevant or e-commerce websites such as Amazon may be used to identify the topic (i.e. specific product range) for each of the segments (i.e.

models under the specific product range) by referring to the topic or category that are predefined in the websites.

In step 840, process 800 generates a topic-entity map based on the topics and entities determined. In particular, the entity that is used to determine the topic is being grouped to the topic. Process 800 ends after step 840. Table 1 below is an example of a topic-entity map. This topic-entity map generated below is based on using the name "Samsung" as the owner in step 510 in process 500. Further, as the number of products from Samsung is rather substantial, the topic-entity map below is a topic-entity map showing only subset of two of its products, i.e. phone and smart PC. As shown, the entities are grouped according to the topic. In other words, the topic "phone" is being determined based on the entities "galaxy s", "s2", "s3", "s4", "galaxy s iii", "galaxy sii", "samsung galaxy", and "samsung galaxy iv", and these entities are being grouped to topic "phone".

Topic	**Entities**
phone	**galaxy s**
	s2
	s3
	s4
	samsung galaxy
smart pc	**smart pc**
	samsung ativ s
	ativ smart pc
	samsung ativ

Table 1

Figure 9 illustrates a flow diagram of process 900 performed by the processor in server 110 or mobile device 115 in accordance with an embodiment of this invention. Process 900 is a process to further elaborate step 415 in process 400. Process 900 begins with step 905 by retrieving a list of users associated to the owner in the social media. Particularly, process 900 retrieve a list of users that are linked or associated to the owner. For example, if the social media is Twitter, process 900 will retrieve a list of followers that

follow the Twitter account of the owner. If the social media is Facebook, process 900 will retrieve a list of friends or users who “likes” the Facebook account or page of the owner.

In step 910, process 900 retrieves the contents which also include the profile information, for example, Facebook’s like and Twitter bio published by each of the users in the list of users after a certain cutoff date. One skilled in the art will recognise that the cutoff date at this step may be the same as that provided in step 520 in process 500. Alternatively, process 900 may request for another cutoff date without departing from the invention. Process 900 ends after step 910. The contents retrieved are stored in the memory of the server 110 or mobile device 115. Table 2 illustrates an example of the contents being stored in the memory of the server 110 or mobile device 115. In the table, the contents published by four users associated to the owner are retrieved and stored.

S/N	**Username**	**Date of content**	**Content**
1	User 1	2013-04-27	@Tony just in time for the s4
2	User 2	2013-04-26	I wanna go Sentosa, who wanna come??
3	User 2	2013-03-13	Is Samsung galaxy s4 really coming march 15th?
4	User 3	2013-04-13	I luv my ativ smart pc.
5	User 4	2013-04-1	price of s3 has dropped!
6	User 4	2013-03-18	tired and hungry
7	User 4	2013-03-15	Rumour from reliable source that s4 to be released next month

Table 2

Figure 10 illustrates a flow diagram of process 1000 performed by the processor in server 110 or mobile device 115 in accordance with an embodiment of this invention. Process 1000 is a process to further elaborate step 420 in process 400. Process 1000 begins with step 1005 by pre-processing the contents published by the users. The contents are pre-processed according to process 700 above.

In step 1010, process 1000 generates segment for each of the contents. The segments are generated by using the common words (such as conjunctions or determiners), stop words and punctuation as separator or delimiter. For example, if a content is “Samsung is holding a galaxy contest!”, three segments will be generated for the content as follows: (samsung) | (holding) | (galaxy contest). Process 1000 ends after step 1010. Table 3 illustrates the segments of each of the contents in table 2 after process 1000.

S/N	Username	Date of content	Content
1	User 1	2013-04-27	(just) | (time) | (s4)
2	User 2	2013-04-26	(wanna) | (Sentosa) | (wanna)
3	User 2	2013-03-13	(Samsung galaxy s4) | (coming march 15th)
4	User 3	2013-04-13	(luv) | (ativ smart pc)
5	User 4	2013-04-1	(price) | (s3) | (dropped)
6	User 4	2013-03-18	(tired) | (hungry)
7	User 4	2013-03-15	(Rumour) |(reliable source) | (s4) | (released next month)

Table 3

Figure 11 illustrates a flow diagram of process 1100 performed by the processor in server 110 or mobile device 115 in accordance with an embodiment of this invention. Process 1100 is a process to further elaborate steps 425 and 430 in process 400. Process 1100 begins with step 1105 by reading the generated segments for each of the contents published by the users. Each of the segments in each of the contents will be compared with each of the entities in the topic-entity map generated to determine a similarity score. The comparison is based on dice string similarity score using the following expression,

$$s = \frac{2n_t}{n_x + n_y}$$

where n_t is the number of characters found in both strings, n_x is the number of characters in string x and n_y is the number of characters in string y. For example, to calculate the similarity between “process” and “proceed”:

x = process bigram for x = {pr ro oc ce es ss}

y = proceed bigram for y = {pr ro oc ce ee ed}

Both x and y have 6 characters each and of which 4 of them are the same. Hence, the dice string similarity score is

$$s = \frac{2 \times 4}{6+6} = 0.667$$

In step 1115, contents having at least one segment with a similarity score above a certain threshold of 0.85 are assigned to the entities in the topic-entity map. This means that after comparing a segment of a content with one of the entities, the content will be assigned to the entity if the similarity score is above the certain threshold. It is possible that a content may be assigned to more than one entities in the topic-entity map. One skilled in the art will note that a higher threshold would return a more accurate match while a lower threshold would return a less accurate match and hence the exact value of the threshold is left as a design choice to those skilled in the art. Process 1100 ends after step 1115. Table 4 illustrates the contents of table 2 being assigned to the topic-entity map.

Topic	**Entities**	**Segment above threshold**	**Username**	**Date of content**	**Content**
phone	**galaxy s**				
	s2				
	s3	(s3)	User 4	2013-04-1	price of s3 has dropped!
	s4	(s4)	User 1	2013-04-27	@Tony just in time for the s4
		(Samsung galaxy s4)	User 2	2013-03-13	Is Samsung galaxy s4 really coming march 15th?
		(s4)	User 4	2013-03-	Rumour from reliable

				15	source that s4 to be released next month
	samsung galaxy	(Samsung galaxy s4)	User 2	2013-03-13	Is Samsung galaxy s4 really coming march 15th?
smart pc	**smart pc**	(ativ smart pc)	User 3	2013-04-13	I luv my ativ smart pc.
	samsung ativ s				
	ativ smart pc	(ativ smart pc)	User 3	2013-04-13	I luv my ativ smart pc.
	samsung ativ				

Table 4

Figure 12 illustrates a flow diagram of process 1200 performed by the processor in server 110 or mobile device 115 in accordance with an embodiment of this invention. Process 1200 is a process after step 430 in process 400 to determine the high value audience from the users. Process 1200 begins with step 1205 by reading the contents assigned or tagged to a topic in topic-entity map. Subsequently, process 1200 determines the users of each of the contents being tagged in the topic-entity map and retrieves all the contents published by the each of the determined users (including related or non-related contents) after the cutoff date. The contents are grouped according to the author of the contents i.e. the user who published the content. Table 5 illustrates the results from step 1205 for topic “phone”.

Topic: Phone

S/N	**Username**	**Date of content**	**content**	**Related**
1	User 1	2013-04-27	@Tony just in time for the s4 lor	Yes
2	User 2	2013-04-26	I wanna go Sentosa, who wanna come??	No
3	User 2	2013-03-13	Is Samsung galaxy s4 really coming	Yes

			march 15th?	
4	User 4	2013-04-1	price of s3 has dropped!	Yes
5	User 4	2013-03-18	tired and hungry	No
6	User 4	2013-03-15	Rumour from reliable source that s4 to be released next month	Yes

Table 5

In step 1210, process 1200 determines recency index for each of the user with the following expression,

Recency = invert((LR – R + 1)/max)

where LR is the latest post of the user, R is the latest post date of the topic and max is the time period between the earliest relevant content and the most recent content from the whole corpus where the corpus is all the content from the friends, fans and followers of the owner of the account in social media after the cutoff date. Invert value is needed as the more recent post should get the bigger value. The main purpose is to find out how recent has a user post relevant post regarding the topic. Based on information provided in table 5, the recency index for User 1, User 2 and User 4 can be derived as follows:

User 1:

Recency = invert((‘2013-04-27’-‘2013-04-27’+1)/(‘2013-04-27’-‘2013-03-13’))
= invert(1/45)
= 45

User 2:

Recency = invert((‘2013-04-26’-‘2013-03-13’’+1)/(‘2013-04-27’-‘2013-03-13’))
= invert((44+1)/45)
= 1

User 4:

Recency = invert((‘2013-04-1’-‘2013-04-1”+1)/(‘2013-04-27’-‘2013-03-13’))
= invert(1/45)
= 45

In step 1215, process 1200 determines the frequency index for each of the user with the following expression,

Frequency = Nr / total number of contents after cutoff date

Where Nr is the total number of contents related to the topic. The purpose is to find out the index for the total number of contents related to the topic in relation to all the contents published by the user after the cutoff date. This determines how interested each of the users is within the period of time. The assumption is, if the user is very interested, most of the content within the period of time will be relevant and hence the frequency will be higher. If there is only one content and the content is related to the topic, then the Frequency is one. Based on information provided in table 5, the Frequency index for User 1, User 2 and User 4 can be derived as follows:

User 1: Frequency = 1/6 = 0.167

User 2: Frequency = 1/6 = 0.167

User 4: Frequency = 2/6 = 0.333

In step 1220, the recency and frequency indexes for each of the users are each normalised to 100. For recency, the index is normalised based on value of max. For frequency, the index is normalised based on total number of contents. Since the frequency in step 1215 is based on the total number of contents, the frequency index only needs to be multiplied by 100. The normalised recency and frequency for User 1, User 2 and User 3 are derived as follows.

User 1: Recency = (45/45)*100 = 100

User 2: Recency = (1/45)*100 = 2.22

User 4: Recency = (45/45)*100 = 100

User 1: Frequency = (1/6)*100 = 16.67

User 2: Frequency = (1/6)*100 = 16.67

User 3:Frequency = (2/6)*100 = 33.33

In step 1225, the value index for each of the user is determined using the following expression,

Value index = (Recency + Frequency) / 2

Based on the normalised recency and frequency for User 1, User 2 and User 3 are derived above, the value index for User 1, User 2 and User 3 can be derived as follows.

User 1:Value index = (100 + 16.67)/2 = 25.00

User 2:Value index = (2.22 + 16.67)/2 = 9.45

User 4:Value index = (100 + 33.33)/2 = 66.67

In step 1230, the users are arranged in descending order according to the value index. The owner has an option to view the top N high-value audience or use this ranking to divide the group of audience to high-value, mid-value and low-value to determine the suitable action to engage with the audience. This approach provides a quantitative method to divide the audience and the focus can be upon this group of high-value audience who is highly likely to be interested in the topic based on the content shared. Process 1200 is repeated for each of the topics in the topic-entity map so that high value audiences can be identified for each of the topics. Process 1200 ends after step 1230. Based on the above example, the users are arranged as illustrated in table 6 below.

Topic: Phone

Username	Value index
User 4	66.67
User 1	25.00
User 2	9.45

Table 6

Process 1200 identifies high value audiences based on the topic level. One skilled in the art will recognise that process 1200 may be modified to identify high value audience based on the entity instead without departing from the invention.

Process 1200 may be modified to identify high value audience based on the entity in the following manner. Step 1205 may be modified to reading the contents assigned or tagged to an entity in topic-entity map and grouping the contents according to the author of the contents i.e. the user who published the content. Table 7 illustrates the results from the modified step 1205 for entity "s4".

Entity: s4

S/N	Username	Date of content	content
1	User 1	2013-04-27	@Tony just in time for the s4 lor
2	User 2	2013-04-26	I wanna go Sentosa, who wanna come??
3	User 2	2013-03-13	Is Samsung galaxy s4 really coming march 15th?
4	User 4	2013-04-1	price of s3 has dropped!
5	User 4	2013-03-18	tired and hungry
6	User 4	2013-03-15	Rumour from reliable source that s4 to be released next month

Table 7

Step 1210 may be modified determine the recency index for each of the user with the following expression, Recency = invert((LR – R + 1)/max), where LR is the latest post of the user and R is the latest post date of the entity and max is the time period between the earliest relevant content and the most recent content from the whole corpus where the corpus is all the content from the friends, fans and followers of the owner of the account in social media after the cutoff date. Based on information provided in table 7, the recency index for User 1, User 2 and User 4 can be derived as follows:

User 1:

Recency = invert((‘2013-04-27’-‘2013-04-27’+1)/(‘2013-04-27’-‘2013-03-13’))

= invert(1/45)

= 45

User 2:

Recency = invert((‘2013-04-26’-‘2013-03-13’+1)/(‘2013-04-27’-‘2013-03-13’))

= invert((44+1)/45)

= 1

User 4:

Recency = invert((‘2013-04-1’-‘2013-03-15’+1)/(‘2013-04-27’-‘2013-03-13’))

= invert(18/45)

= 2.5

Step 1215 may be modified to determine the frequency index for each of the user with the following expression, Frequency = Nr / total number of contents after cutoff date, where Nr is the total number of contents related to the entity. Based on information provided in table 7, the Frequency index for User 1, User 2 and User 4 can be derived as follows:

User 1: Frequency = 1/6 = 0.167

User 2: Frequency = 1/6 = 0.167

User 4: Frequency = 1/6 = 0.167

Step 1220 remains the same and recency and frequency indexes for each of the users are each normalised. The normalised recency and frequency for User 1, User 2 and User 3 are derived as follows.

User 1: Recency = (45/45)*100 = 100

User 2: Recency = (1/45)*100 = 2.22

User 4: Recency = (2.5/45)*100 = 5.55

User 1: Frequency = (1/6)*100 = 16.67

User 2: Frequency = (1/6)*100 = 16.67

User 3: Frequency = (1/6)*100 = 16.67

Step 1225 remains the same and the value index for each of the user is determined using the following expression, Value index = (Recency + Frequency) / 2. Based on the normalised recency and frequency for User 1, User 2 and User 3 are derived above, the value index for User 1, User 2 and User 3 can be derived as follows.

User 1:Value index = (100 + 16.67)/2 = 25.00

User 2:Value index = (2.22 + 16.67)/2 = 9.45

User 4:Value index = (5.55 + 16.67)/2 = 11.11

Step 1230 remains the same and the users are arranged in descending order according to the value index. This modified process is repeated for each of the entities in the topic-entity map so that high value audiences can be identified for each of the entities. Based on the above example, the users are arranged as illustrated in table 8 below for entity "s4".

Entity: s4

Username	Value index
User 1	25.00
User 4	11.11
User 2	9.45

Table 8

As the processor in mobile device 115 may not be able to handle the processes as described above efficiently, one skilled in the art will recognise that the mobile device 115 may be modified to execute the processes described above via server 110. This means that mobile device 115 may include processes that instruct server 110 to execute the processes as described above remotely and thereafter return the results to the mobile device 115.

The above is a description of embodiments of a system and method in accordance with this invention. It is envisioned that others can and will design alternative methods and systems as set forth in the following claims based upon this disclosure.

What is claimed is:

1. A system for determining high value audience from a social media comprising:

a processing unit;

instructions for directing said processing unit to:

retrieving information created by an owner of an account in said social media, said information being content being published by said owner,

generating a topic-entity map based on said information retrieved, wherein said topic-entity map includes topics and entities, said entities being determined based on said information retrieved from said account of said owner, and each of said topics is determined based on the determined entities;

retrieving contents published by users in said social media, said users being associated with said account of said owner,

extracting relevant contents published by said users in said social media,

determining a similarity score for each content published by respective users against each of said entities in said topic-entity map,

assigning each of said contents to one of said entity in said topic-entity map in response to said similarity score of said contents being above a certain threshold value,

determining users for each of said contents being grouped to one of said topic in said topic entity map;

retrieving all contents retrieved for each of said determined users;

determining a recency index for each of said determined users;

determining a frequency index for each of said determined users;

normalising said recency and said frequency of each of said users;

determining a value index for each of said users based on an average of said normalized recency and frequency indexes of respective said users; and

arranging said users according to said value index from highest to lowest, and a media readable by said processing unit that stores said instructions.

2. The system according to claim 1 wherein said instruction to retrieving information created by said owner of an account in said media further includes instructions for directing said processing unit to:

receiving user information from said owner, wherein said user information includes username and name of social media;

receiving a cutoff date from said owner; and

extracting all information from said cutoff date for said username in said name of social media.

3. The system according to claim 2 wherein said instruction to generating a topic-entity map based on said information retrieved further includes instructions for directing said processing unit to:

pre-processing said contents being published by said username in said name of social media;

segmenting each of said contents;

arranging segments based on frequency; and

determining top N segments as entities.

4. The system according to claim 3 wherein said instruction to pre-processing said contents being published by said username in said name of social media further includes instructions for directing said processing unit to:

removing symbol in said contents; and

normalising exaggerative words.

5. The system according to claim 3 wherein said instruction to segmenting each of said contents further includes instructions for directing said processing unit to:

dividing said content into segments, wherein conjunction, determiners, stop words and punctuation in said content are being used to separate said content into said segments.

6. The system according to claim 3 wherein said instruction to generating a topic-entity map based on said information retrieved further includes instructions for directing said processing unit to:

requesting a website from said owner;

retrieving data from said website in response to receiving a website from said owner;

performing a similarity search between each of said entities and said data retrieved;

determining a topic for each of said entities based on said data retrieved.

7. The system according to claim 3 wherein said instruction to generating a topic-entity map based on said information retrieved further includes instructions for directing said processing unit to:

searching relevant online resources using each of said entities as keywords; and

determining a topic based on a category in which said entity is being classified in a search result derived from said step of searching relevant online resources.

8. The system according to claim 7 wherein said instruction to generating a topic-entity map based on said information retrieved further includes instructions for directing said processing unit to:

generating said topic entity map by grouping said entity that determines said topic to said topic.

9. The system according to claim 7 or claim 8 wherein said instruction to retrieving contents published by users in said social media further includes instructions for directing said processing unit to:

retrieving a list of users associated to said owner;

retrieving contents published by said users after said cutoff date.

10. The system according to claim 9 wherein said instruction to extracting relevant contents published by said users in said social media further includes instructions for directing said processing unit to:

pre-processing each of said contents being published by said users; and

generating segments for each of said contents.

11. The system according to claim 10 wherein said instruction to pre-processing each of said contents being published by said users further includes instructions for directing said processing unit to:

removing symbol in said contents; and

normalising exaggerative words.

12. The system according to claim 10 wherein said instruction to generating segments for each of said contents further includes instructions for directing said processing unit to:

dividing said content into segments, wherein conjunction, determiners, stop words and punctuation in said content are being used to separate said content into said segments.

13. The system according to claim 10 wherein said instruction to determining said similarity score for each content published by respective users against each of said entities in said topic-entity map further includes instructions for directing said processing unit to:

determining said similarity score for each of said segments against each of said entities in said topic-entity map.

14. The system according to claim 1 wherein said recency index for each of said determined users is determined with the following expression,

Recency = invert((LR – R + 1)/max)

where LR is the latest post date of said determined user, R is the latest post date of the topic by said determined user and max is the time period between the earliest relevant content and the most recent relevant content.

15. The system according to claim 1 wherein said frequency index for each of said determined users is determined with the following expression

Frequency = Nr / total number of contents after cutoff date

where Nr is the total number of contents related to said topic and total number of contents after cutoff date is the total number of contents published by all of said determined users.

16. A method for determining high value audience from a social media comprising:

retrieving information created by an owner of an account in said social media, said information being content being published by said username in said name of social media

generating a topic-entity map based on said information retrieved, wherein said topic-entity map includes topics and entities, said entities being determined based on said information retrieved from said account of said owner, and each of said topics is determined based on the determined entities;

retrieving contents published by users in said social media, said users being associated with said account of said owner,

extracting relevant contents published by said users in said social media,

determining a similarity score for each content published by respective users against each of said entities in said topic-entity map,

assigning each of said contents to one of said entity in said topic-entity map in response to said similarity score of one of said contents being above a certain threshold value,

determining users for each of said contents being grouped to one of said topic in said topic entity map;

retrieving all contents retrieved for each of said determined users;

determining a recency index for each of said determined users;

determining a frequency index for each of said determined users;

normalising said recency and said frequency of each of said users;

determining a value index for each of said users based on an average of said normalized recency and frequency indexes of respective said users; and

arranging said users according to said value index from highest to lowest.

17. The method according to claim 15 wherein said step of retrieving information created by said owner of an account in said media further includes:

receiving user information from said owner, wherein said user information includes username and name of social media;

receiving a cutoff date from said owner; and

extracting all information from said cutoff date for said username in said name of social media.

18. The method according to claim 17 wherein said step to generating a topic-entity map based on said information retrieved further includes:

pre-processing said contents being published by said username in said name of social media;

segmenting each of said contents;

arranging segments based on frequency; and

determining top N segments as entities.

19. The method according to claim 18 wherein said step to pre-processing said contents being published by said username in said name of social media further includes:

removing symbol and platform-specific term in said contents; and

normalising exaggerative words.

20. The method according to claim 18 wherein said step to segmenting each of said contents further includes:

dividing said content into segments, wherein conjunction, determiners, stop words and punctuation in said content are being used to separate said content into said segments.

21. The method according to claim 18 wherein said step to generating a topic-entity map based on said information retrieved further includes:

requesting a website from said owner;

retrieving data from said website in response to receiving a website from said owner;

performing a similarity search between each of said entities and said data retrieved;

determining a topic for each of said entities based on said data retrieved.

22. The method according to claim 18 wherein said step to generating a topic-entity map based on said information retrieved further includes:

searching relevant online resources using each of said entities as keywords; and

determining a topic based on a category in which said entity is being classified in a search result derived from said step of searching relevant online resources.

23. The method according to claim 22 wherein said step to generating a topic-entity map based on said information retrieved further includes:

generating said topic entity map by grouping said entity that determines said topic to said topic.

24. The method according to claim 22 or claim 23 wherein said step to retrieving contents published by users in said social media further includes:

retrieving a list of users associated to said owner;

retrieving contents published by said users after said cutoff date.

25. The method according to claim 24 wherein said step to extracting relevant contents published by said users in said social media further includes:

pre-processing each of said contents being published by said users; and

generating segments for each of said contents.

26. The method according to claim 25 wherein said step to pre-processing each of said contents being published by said users further includes:

removing symbol and platform-specific term in said contents; and

normalising exaggerative words.

27. The method according to claim 25 wherein said step to generating segments for each of said contents further includes:

dividing said content into segments, wherein conjunction, determiners, stop words and punctuation in said content are being used to separate said content into said segments.

28. The method according to claim 25 wherein said step to determining said similarity score for each content published by respective users against each of said entities in said topic-entity map further includes:

determining said similarity score for each of said segments against each of said entities in said topic-entity map.

29. The method according to claim 14 wherein said recency index for each of said determined users is determined with the following expression,

Recency = invert((LR – R + 1)/max)

where LR is the latest post date of said determined user, R is the latest post date of the topic by said determined user and max is the time period between the earliest relevant content and the most recent relevant content.

30. The method according to claim 14 wherein said frequency index for each of said determined users is determined with the following expression

Frequency = Nr / total number of contents after cutoff date

where Nr is the total number of contents related to said topic and total number of contents after cutoff date is the total number of contents published by all of said determined users.