VaccineWatch : a monitoring system of vaccine messages from social media data

Somrak Numnark, Supawadee Ingsriswang and Duangdao Wichadakul
Information Systems Laboratory (ISL), Bioresources Technology Unit,
National Center for Genetic Engineering and Biotechnology (BIOTEC)
Klong Luang, Pathumthani 12120, Thailand

***Abstract*—To exploit social media data in vaccine-related areas, we proposed VaccineWatch, a monitoring system with visualizations and analytics of significant vaccine information from Twitter and RSS feeds. The system was designed and implemented as a web application with following distinguished features. First, it comes with graphical user interfaces that visualize perspectives of vaccine-related information mined from social media data. Second, it provides a set of filters allowing users to focus on their diseases, vaccines, countries, and/or companies of interest. Third, it includes the helper tools for the management of social media data collection and backend processes such as Twitter and RSS crawlers. The prototype of VaccineWatch is available at www.vacciknowlogy .org/VaccineWatch.**

Keywords—vaccine informatics, social media data mining and visualization, public health surveillance, vaccine monitoring system

I. INTRODUCTION

The increasingly large volume of social media data has drawn interest from public-health community for extracting various aspects of useful health-related information. Its nature of nearly real-time with highly contextual and networked data, when integrated appropriately, can generate geographically and timely information [1-3]. Salathé, M. et al., for example, used Twitter data to assess the vaccination sentiments towards the novel influenza A (H1N1) vaccine and suggested the likelihood of strongly increased disease outbreaks in the communities dominated by negative sentiments [4]. Culotta, A. analyzed the Twitter messages for the flu-related keywords and built models for forecasting future influenza rates [5] while Scanfeld, D. explored Twitter status updates about "antibiotic(s)," determined the categories, and derived the misuse or misunderstanding of antibiotics [6]. Sadilek, A. et al. built the model from Twitter data to study the spread of infectious diseases and the interactions between specific individuals in the progress of a contagion [7]. The same authors constructed a probabilistic model from Twitter posts, the social ties, and co-locations of individuals with other people and predicted if and when they will get sick [8].

With the successes of exploiting social media data in various public health for disease analyses and surveillances, here, we present VaccineWatch, a monitoring system of vaccine's news, alerts, announcements, events, and reports from social media data: Twitter and RSS feeds. The main objectives of the system are to 1) help people in vaccine-related areas discern significant information of vaccines available in the social media, and 2) produce the analytics and visualizations that help describe, forecast, and enhance policy planning, clinical trials, and timely reactions to disease warnings or outbreaks.

II. HOW TO USE

To enable the ease of use, the system provides graphical user interfaces that are organized into two sections: visualization and administration.

The visualization section consists of two main tabs: dashboard and display. The dashboard tab visualizes the information of the relationship between disease and areas (countries, cities) that have been reported. It is divided into two parts: (A) the table of disease list and (B) the map of reported diseases over the countries (Figure 1).

Yellow dots on the map mark the diseases that have been reported in an area. The number within the dots represents the number of diseases. Here, users can filter for specific diseases by choosing the disease names from the table of disease list or type in the disease name.

The display tab provides detailed information about the data collected in the system, which could be grouped by Twitter keywords, Twitter accounts, and RSS feeds (Figure 2(A)). Statistical data are presented in the form of word cloud and line chart (Figure 2 (B), (C)). Also users can filter data within a daily, weekly or monthly duration. The tagged data table (Figure 2(D)) shows the collected data with tagged terms, highlighted by different colors according to specific categories (Table I).

TABLE I. TAGGING COLOR BY CATEGORY

Tagger	Color
DISEASE	Purple
CITY	Green
COUNTRY	Nigger brown
COMPANY	Light brown
VACCINE	Blue

To demonstrate the system usage, the information in display tab can be used to answer the question, "What happened during the three months between 1 April – 30 June 2014? ."

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

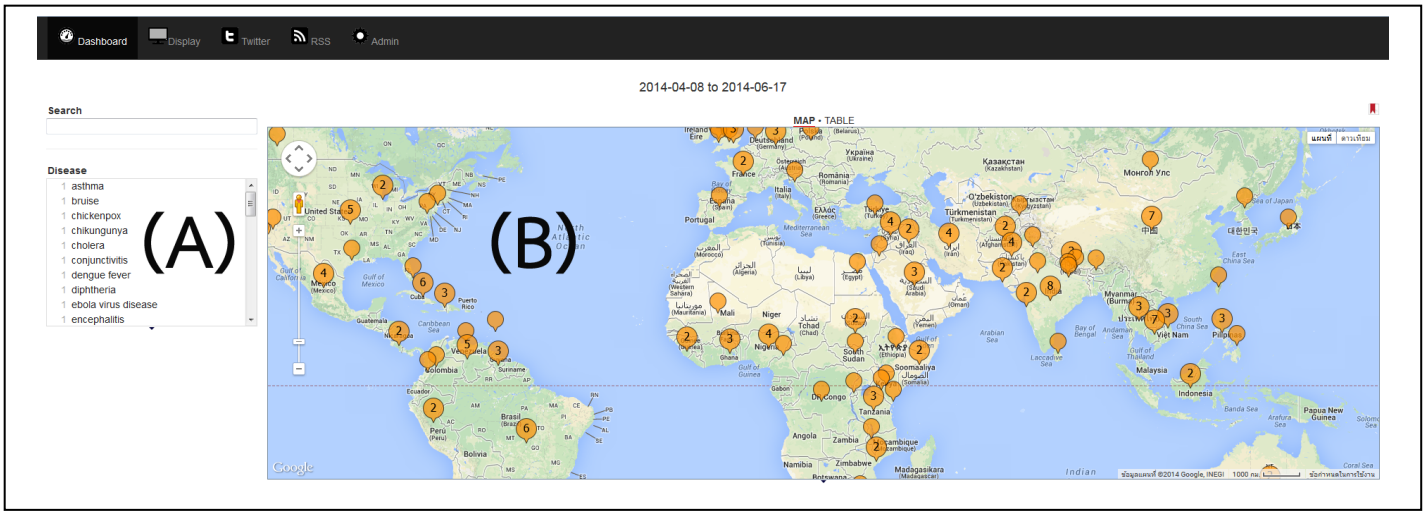


Fig. 1. Dashboard tab: (A) table of disease list and (B) map of reported diseases over the countries

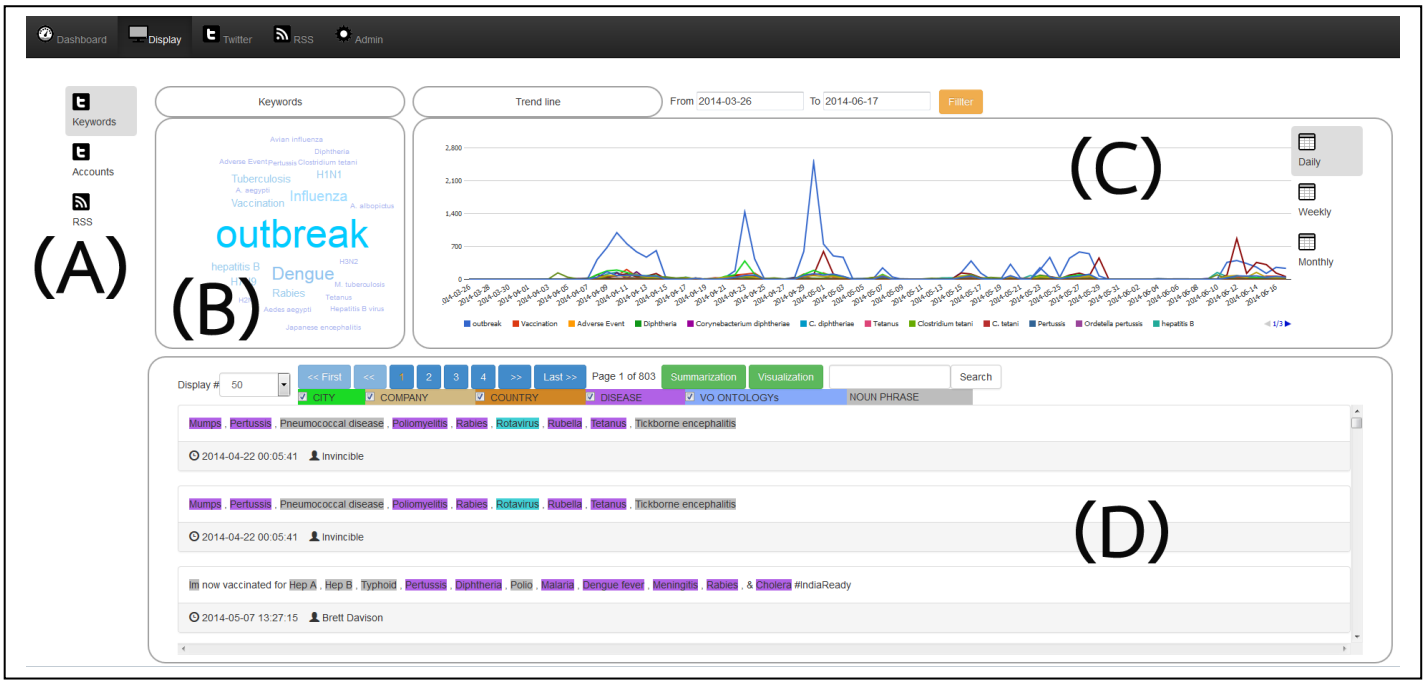


Fig. 2. Display tab: (A) grouped by Twitter keywords, Twitter accounts, and RSS feeds, (B) word cloud, (C) line chart, (D) tagged data table

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

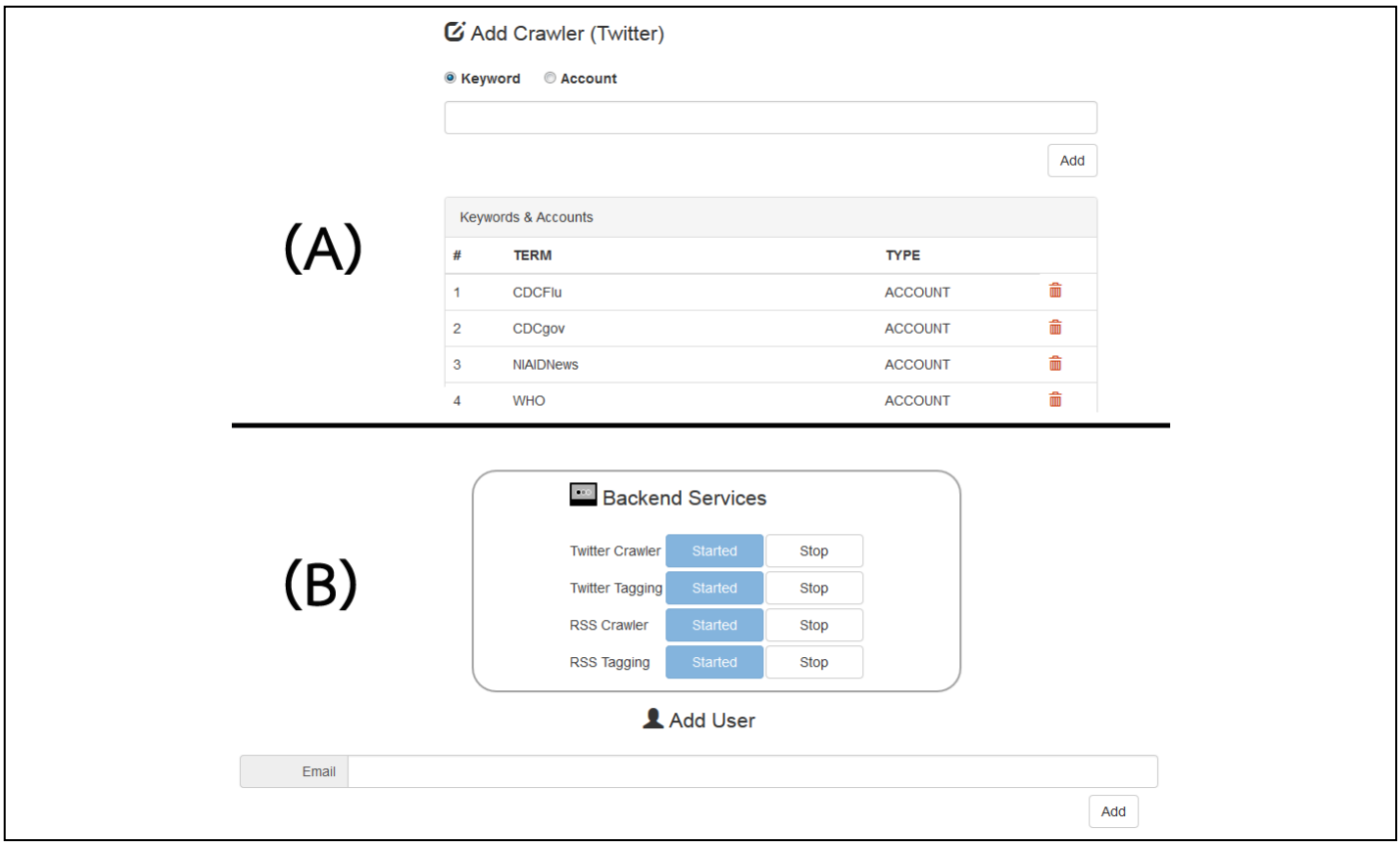


Fig. 3. Administration pages: (A) data source management, (B) backend process management

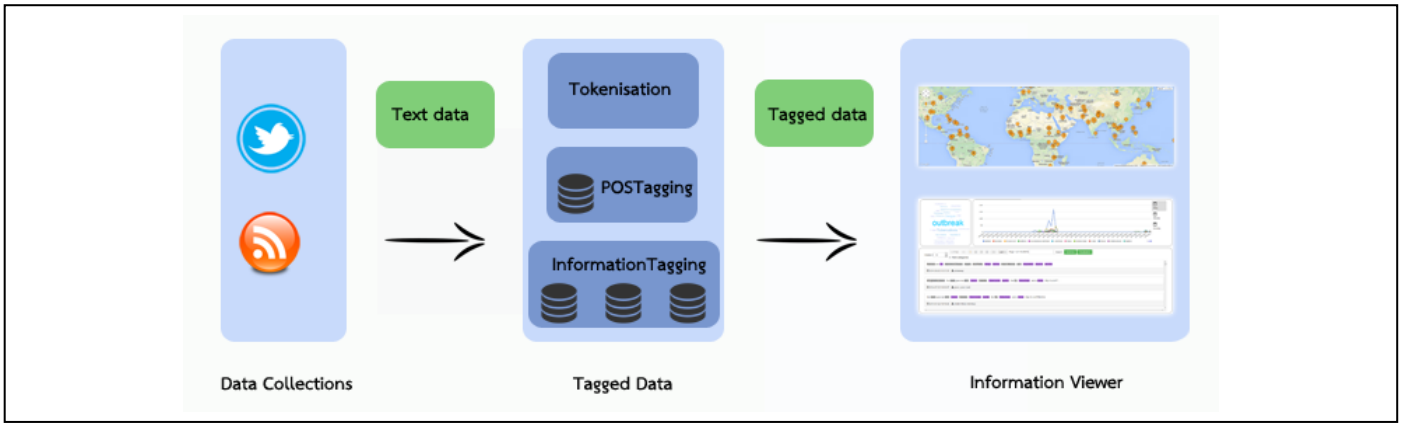


Fig. 4. System overview

Starting from the display tab focusing on data grouped by Twitter keywords, users can select a specific time period by choosing the start and end dates on the filter area. The word cloud and the line graph represent the amount of data stored during the specified duration. Each data series in the line chart represents the number of tweets collected from Twitter using a specific keyword. The bigger word in the word cloud, the more tweets it appears in. Within the specified three months, the top ten Twitter keywords were outbreak (42.7%), Dengue (11.4%), Influenza (9.0%), Rabies (8.3%), Tuberculosis (6.7%), Vaccination (6.1%), hepatitis B (5.1%), H7N9 (5.0%), H1N1 (2.8%), and Pertussis (2.3%) counted from 24,912 non-redundant tagged messages.

The highlighted terms with different colors in the tagged data table help users capture terms of interest in the collected tweets. Users may explore the detailed messages for a specific keyword of interest using the search input textbox at the tagged table data area. By searching for "dengue fever," 1,509 related messages remained. The click on "Summary" button will open the new window showing

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

the top 50 tagged terms, vaccine alerts, and announcements of companies identified from the search result related to "denguefever". The click on the "Show co-occurring terms," will open a new page of network graph implemented using Cytoscape Web [9]. The right mouse click on an edge between two terms (nodes) will link to the list of messages containing both terms. Here, it shows that during the three month, "dengue fever" was getting high attentions in Brazil.

Likewise, clicking on "Visualization" button will open the dashboard page to display the map of diseases to cities/countries extracted from the search result. The detailed number of diseases and their co-occurring areas extracted from the messages will be analyzed and shown under the TABLE menu. In this example, the analysis result contains (1) six diseases (dengue fever, leishmaniasis, measles, pneumonia, rabies, and tuberculosis) and 23 countries/cities were related to "dengue fever", and (2) dengue fever and measles are topmost reported in Brazil (688 messages).

The administration section is organized into two parts. The first part (Figure 3(A)) is used to add/remove Twitter keywords, Twitter accounts, and RSS URLs as sources of data collection process. This interface enables flexibility in gathering data of interest by the administrator. The second part (Figure 3(B)) is used to manage the backend processes including Twitter Crawler, Twitter Tagging, RSS Crawler and RSS Tagging.

III. IMPLEMENTATION

Figure 4 shows the overall architecture of VaccineWatch system with three core processes: (A) data collection, (B) data tagging, and (C) information viewer. These processes are described below.

A. Data collection

Data collection is a preprocessing step that collects social media data from Twitter and RSS, and records them into MySQL database. Types of collected data include: (1) Twitter by keywords (i.e., outbreak, vaccination, Influenza, H7N9, H5N1, Japanese encephalitis, etc), (2) Twitter by accounts (i.e., WHO, CDCgov, NIAIDNews, etc.), and (3) RSS by URLs (i.e., http://www.who.int/feeds/entity/csr/don/en/rss.xml). The system integrates Twitter4J [10] for gathering Twitter data and ROME [11] for retrieving RSS data. As of the three months, a total of 172,182 social media records were stored in the database (Table II). Figure 5 shows an example of Twitter data collected in the system.

TABLE II. DATA COLLECTED IN SYSTEM

Collected data types	Number of collected data
Twitter by keywords	167,818
Twitter by accounts	4,317
RSS by URLs	47

Recommended vaccines for travel to Cambodia : Diphtheria , Tetanus , Hepatitis A , Hepatitis B , Japanese B Encephalitis , Polio , Rabies

Fig. 5. Sample social media data (using "Diphtheria" to query Twitter by keyword)

B. Data Tagging

Data tagging, the process of assigning semantics to social media data, consists of three steps: 1) tokenization, 2) parts of speech (POS) tagging, and 3) information tagging.

1) Tokenization step separates a social media data as a sequence of words called tokens. Figure 6 shows the tokens of the sample data generated from this step.

Recommended vaccines for travel to Cambodia : Diphtheria , Tetanus , Hepatitis A , Hepatitis B , Japanese B Encephalitis , Polio , Rabies

Fig. 6. Sample result of tokenization step

2) Parts of Speech (POS) tagging step is used as a part of multi-word named entities (MWNEs) identification. The system uses an open source Apache OpenNLP [12] for recognizing English parts of speech and then combines adjacent nouns together. For example, the "Hepatitis" and following "A" will be combined to "Hepatitis A". The input of this process is tokens generated from tokenization step and the output of this step is tokens with English part-of-speech tags (Figure 7).

<VBN>Recommended</VBN> <NNS>vaccines</NNS> <IN>for</IN> <NN>travel</NN> <TO>to</TO> <NNP>Combodia</NNP> <COLON>:</COLON> <NNP>Diptheria</NNP> <COMMA>,</COMMA> <NNP>Tetanus</NNP> <COMMA>,</COMMA> <NNP>Hepatitis</NNP> <NNP>A</NNP> <COMMA>,</COMMA> <NNP>Hepatitle</NNP> <NNP>B</NNP> <COMMA>,</COMMA> <NNP>Japanese</NNP> <NNP>B</NNP> <NNP>Encephalitis</NNP> <COMMA>,</COMMA> <NNP>Polio</NNP> <COMMA>,</COMMA> <NNP>Rabies</NNP>

Fig. 7. Sample result of POS tagging step : (1) VBN is "Verb, past participle", (2) NNS is "Noun, plural", (3) IN is "Preposition or subordinating conjunction", (4) NN is "Noun, singular or mass", (5) TO is "To", (6) NNP is "Proper noun, singular", (7) COLON is "Colon", (8) COMMA is "Comma"

Figure 8 depicts the POS tagging result of the tokenized sample data. The blue color represents nouns generated from OpenNLP or noun phrases generated from MWNEs identification.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

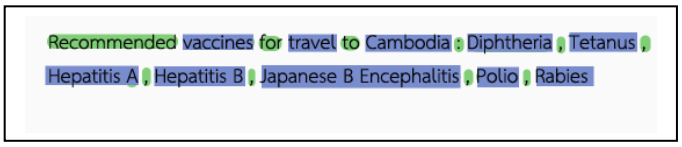


Fig. 8. Sample result of POS Tagging step

3) Information Tagging process

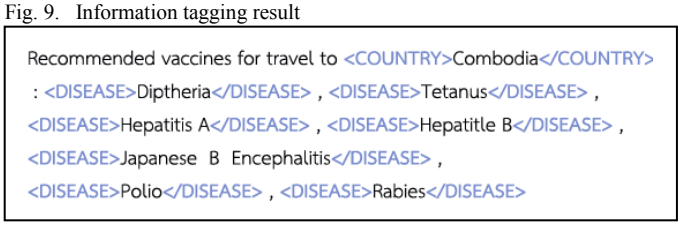
Vaccine-related information are tagged by dictionary-based approach. The VaccineWatch's vocabulary sets include: diseases, cities, countries, vaccine companies, and Vaccine Ontology (VO) [13]. Each set was built with a specific Entities Tagger (Table III). The input of this step is nouns or noun phrases generated from the previous step. The entities taggers were then sequentially used to tag the input using their corporated dictionary.

TABLE III. ENTITIES TAGGER

Entities Tagger	Source of vocabulary	Number of terms
Diseases	http://dbpedia.org	5606
Cities	http://dbpedia.org	20229
Countries	http://dbpedia.org	2835
Companies	http://www.cdc.gov/vaccines/about/terms/USVaccines.html	52
Vaccines	http://www.violinet.org/vaccineontology/	4046

The results of the information tagging process using inputs generated from previous step are shown in Figure 9. From the three months, the total number of the successfully tagged messages is 40,669, then, the number of distinct diseases, cities, countries, companies and terms in vaccine ontology were 206, 207, 154, 3 and 376 respectively.

Fig. 9. Information tagging result



C. Information Viewer

The marked-up text is stored in the database and can be queried via the visualization section previously described. The graphical user interfaces were implemented using SIMILE Widgets [14] and Google Charts [15].

IV. DISCUSSION

To retrieve and synthesize vaccine-related information from social media data, we developed a system with three main processes :(1) data collection, (2) data tagging and (3) information viewer.

In data collection process, due to the limited number of returned messages (180) for a fixed time interval (15 minutes) of Twitter API [16], some data may be lost if there are a lot of tweets during the API waiting period.

The VaccineWatch used the open source software, running as background processes, to collect social media data from Twitter and RSS feeds. To make the system flexible and easy to use, we also provided the administration pages to configure and control these software.

In data tagging process, the effect of sequential tagging can cause errors. For example, "New Delhi" may refer to a city name or disease name. If "New Delhi" refers to the disease name and the sequential tagging starting from city name, the results will be incorrect. We plan to add the grammar and contextual analyses to make a more accurate labeling. In addition, we plan to work on the authenticity of colleted data. Another problem we found is the coverage of the vocabularies used by the system, to improve that, we can iteratively investigate the tagged messages, manually curate and extend the corresponding vocabulary sets.

In information viewer, the data available include the geographical map between disease and areas, the top 50 tagged terms, the vaccine alerts and announcements. As part of the ongoing work, we are extending the system with additional analyses and visualizations, the interactive timeline, and advanced search options.

V. CONCLUSION

The VaccineWatch monitoring system is presented to help identify and extract vaccine-related information from social media data. This software is built on top of multiple pieces of software [10,11,12,13,14,15,16] and comes with graphical user interfaces (GUIs) for ease of use. It provides various visualizations that help users capture both spatial and temporal information between vaccines, diseases, countries/cities, and companies, together with the top 50 tagged terms, messages related to vaccine and disease alerts, and company announcements. The flexible management of data sources and backend processes provides users the extensible and customizable system. With these features, the VaccineWatch is a very useful web application that can be used as a vaccine and disease surveillance system.

ACKNOWLEDGMENT

This work was supported by the National Vaccine Institute (NVI), Ministry of Public Health (MOPH), Thailand.

REFERENCES

[1] Brownstein JS, Freifeld CC, Madoff LC: Digital Disease Detection — Harnessing the Web for Public Health Surveillance. *N Engl J Med* 2009, 360(21):2153-2157.

[2] Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, Campbell EM, Cattuto C, Khandelwal S, Mabry PL *et al*: Digital Epidemiology. *PLoS Comput Biol* 2012, 8(7):e1002616.

[3] Dredze M: How Social Media Will Change Public Health. *Intelligent Systems, IEEE* 2012, 27(4):81-84.

[4] Salathé M, Khandelwal S: Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control. *PLoS Comput Biol* 2011, 7(10):e1002199.

[5] Culotta A: Detecting influenza outbreaks by analyzing Twitter messages. *arXiv:10074748* 2010.

978-1-4799-7294-4/14/$31.00 ©2014 IEEE

[6] Scanfeld D, Scanfeld V, Larson EL: Dissemination of health information through social networks: twitter and antibiotics. *Am J Infect Control* 2010, 38(3):182-188.

[7] Sadilek A, Kautz HA, Silenzio V: Modeling Spread of Disease from Social Interactions. *ICWSM* 2012.

[8] Sadilek A, Kautz HA, Silenzio V: Predicting Disease Transmission from Geo-Tagged Micro-Blog Data. *AAAI* 2012.

[9] Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD: Cytoscape Web: an interactive web-based network browser. *Bioinformatics* 2010, 26(18):2347-2348.

[10] Twitter4J [http://www.twitter4j.org]

[11] ROME [http://rometools.github.io/rome/]

[12] Apache OpenNLP [http://opennlp.apache.org/]

[13] Vaccine Ontology (VO) [http://www.violinet.org/ vaccineontology/]

[14] SIMILE Widgets [http://www.simile-widgets.org/]

[15] Google Charts [https://developers.google.com/ chart/]

[16] Twitter API [https://dev.twitter.com/docs/rate-limiting/1.1]

978-1-4799-7294-4/14/$31.00 ©2014 IEEE