Supporting Temporal Analytics for Health-Related Events in Microblogs

Nattiya Kanhabua, Avaré Stewart,
Wolfgang Nejdl
L3S Research Center
Leibniz Universität, Hannover, Germany
{kanhabua, stewart, nejdl}@L3S.de

Sara Romano
Dipartimento di Informatica e Sistemistica
Federico II University, Naples, Italy
sara.romano@unina.it

ABSTRACT

Microblogging services, such as Twitter, are gaining interests as a means of sharing information in social networks. Numerous works have shown the potential of using Twitter messages (or *tweets*) in order to infer the existence and magnitude of real-world events. In the medical domain, there has been a surge in detecting public health related tweets for early warning so that a rapid response from health authorities can take place. In this paper, we present a *temporal analytics tool* for supporting a comparative, temporal analysis of disease outbreaks between Twitter and official sources, such as, World Health Organization (WHO) and ProMED-mail. We automatically extract and aggregate outbreak events from official outbreak reports in order to produce time series data used for the analysis. Our tool can support a correlation analysis and an understanding of the temporal developments of outbreak mentions in Twitter, based on comparisons with official sources.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Design

Keywords

Event Detection, Disease Outbreaks, Twitter, Time Series Analysis

1. INTRODUCTION

Twitter is a microblogging service that is gaining interests as a means for sharing real-world events ranging from a user's personal status to news reports. Given its nature and volume, Twitter messages (or *tweets*) are now seen as a valuable source for real-time Web applications, e.g., trend detection [2] and natural disaster detection [6]. In the medical domain, there has been a surge in detecting public health related tweets for Event-Based Epidemic Intelligence (e-EI). In general, health related tweets (e.g., user status updates or news) are commonly found in Twitter as, for example: (a) *"I have the mumps...am I alone?"*; (b) *"my baby girl has a Gastroenteritis so great!! Please do not give it to meee"*; (c) *"#Cholera breaks out in #Dadaab refugee camp in #Kenya http://t.co/...."*; (d) *"As many as 16 people have been found infected with Anthrax in Shahjadpur upazila of the Sirajganj district in Bangladesh"*. Such information can indicate the existence and magnitude of real-world health related events. Thus, Twitter can be considered as a collector of real-time information that could be used by health authorities as an additional information source for obtaining early warnings; thereby helping them to prevent and/or mitigate the public health threats.

Recent works have focused on validating the timeliness of Twitter by correlating tweets with real-world outbreak statistics, such as, *Influenza*-like-Illness rates [5] and detecting flu outbreaks [1, 3, 4]. In addition, the aforementioned works show the advantage of using Twitter for detecting real-world events focusing on *common and seasonal* diseases, such as, influenza or dengue fever, where only countries with a high density of Twitter users (e.g., United Sates, United Kingdom or Brazil) were subjects of the study. To the best of our knowledge, none of these previous work have focused on an temporal analysis of Twitter data for *general diseases* that are **not only seasonal, but also sporadic** diseases that occur in *low tweet-density areas* like Kenya or Bangladesh, as we will show in this work.

In this paper, we present a *temporal analytics tool* for supporting a temporal, retrospective analysis of infectious disease outbreaks in Twitter. The objective of this tool is to help *medical professionals* to: 1) analyze disease outbreaks with real-time data, and 2) compare the temporal development of an outbreak event mentioned in social media against official reports. In order that, we propose a method to automatically extract outbreak events from official health related reports from World Health Organization[1] (WHO) and ProMED-mail[2], both are denoted *external sources*. For a given real-world outbreak event, the tool provides the possibility to visualize and correlate the time series of Twitter and those extracted from external sources, and compare them using different granularities of time (*daily*, *weekly*, *monthly*) and location (*country*, *continent*, *latitude*, *worldwide*).

The rest of this paper is organized as follows. In Section 2 we provide the description of dataset used in this demo as well as the data models for an outbreak event and a tweet. In Section 3, we outline the system architecture. Finally, in Section 4, we describe the visualization and analysis of outbreaks in different data sources and we outline the proposed demonstration.

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.
ACM 978-1-4503-1156-4/12/10.

[1] http://www.who.int
[2] http://www.promedmail.org/

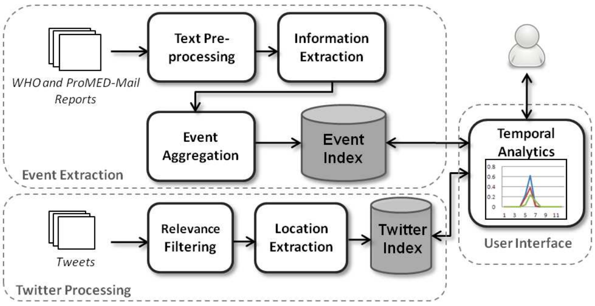


Figure 1: System architecture.

2. DATASET AND MODELS

We used Twitter data that was collected for one year period during the 1st of January 2011 and the 31st of December 2011. In the same period, we collected about 3,000 official outbreak reports publicly available from the external resources (WHO and ProMED-mail). Twitter data was collected using a list of English keywords (supplied by medical domain experts) that are related to medical conditions or disease names, and symptoms. Finally, our Twitter collection consists of over 112 millions of tweets.

An event corresponds to a real-world outbreak event defined as a quadruple: $e = (v, m, l, t_e)$ described by four attributes that provide information on *who* (victim v) was infected by *what* (disease or medical condition m), *where* (location l) and *when* (time t_e). A key aspect of our event model is the extraction of temporal expressions. We distinguish between the two temporal aspects associated with a disease outbreak e: 1) t_p or the publication time of a document d reporting about e, and 2) t_e or the time of the outbreak, which is the time period that the outbreak has actually taken place. In this case, t_e can be determined by the time mentioned in d, or temporal expressions. Further, temporal expressions can be explicit, implicit or relative. Examples of explicit temporal expressions are “May 25, 2012” or “June 17, 2011” that can be mapped directly to dates months, or years on the Gregorian calendar. An implicit temporal expression is an imprecise time point or interval, e.g., “Independence Day 2011” that can be mapped to “July 04, 2011”. Examples of relative temporal expressions are “yesterday”, “last week” or “one month ago”.

Thus, in our event model, if an ongoing outbreak started **last week** and it is first reported in the news of **today**, then the time of the outbreak event is the temporal expression (**last week**), and not its publication time (**today**). Note, that is in contrast to the aforementioned works, which use the publication time of documents as proxy for the occurrence of an event. Finally, we represent a tweet as a triple: $tw = (w, l, t_p)$ where w is the contents or texts of the tweet, l is the location of the tweet, and t_p is the publication time of the tweet.

3. SYSTEM ARCHITECTURE

An overview of our system is depicted in Figure 1. Our system consists of three main steps: 1) outbreak event extraction (*Event Extraction*), 2) Twitter filtering and location extraction (*Twitter Processing*), and 3) visualization and correlation of outbreak-related data (*User Interface*).

3.1 Event Extraction

The *Event Extraction* module automatically extracts outbreak events from official outbreak reports in a pipeline fashion. The stages of the pipeline consist of: 1) *Text Preprocessing* for tokenization, sentence extraction, part-of-speech tagging, and 2) *Information Extraction* for named entity recognition and temporal expression extraction. In order that, a series of language processing tools were employed including OpenNLP[3] (for tokenization, sentence extraction and part-of-speech tagging), OpenCalais[4] (for named entity recognition) and HeidelTime [7] (for temporal expression extraction). We assume that an event is described as a sentence containing mentions of a disease name and a location. The time of an event is determined as a temporal expression mentioned in the same sentence or its surrounding context sentences. The results from this step are a set of events associated to a given named entity and occur in particular time and place (denoted as the **event profiles** of a disease). In addition, each outbreak is associated with the number of victim/suspected cases. The extracted outbreak events were manually verified with the input of domain experts in order to filter out irrelevant reports, such as, those containing updates or discussion about the characteristics of a disease. Finally, the extracted outbreak events will be used as ground truth for further analysis.

3.2 Twitter Processing

The *Twitter Processing* module identifies relevant tweets that are matched with an outbreak event. In our case, a tweet will be matched with an outbreak event if its medical condition and location are equal to those of the outbreak event. In this module, the *Relevance Filtering* stage identifies irrelevant tweets and discards them from the dataset. The reason for filtering is that the contents of tweets are highly ambiguous and noisy. More precisely, a disease name mentioned in a tweet can refer to the context **not relevant** to an outbreak. Examples of irrelevant tweets for epidemic analysis are: (a) *“A two hour train journey, Love In the Time of Cholera.”* or (b) *“I liked a @YouTube video http://youtu.be/... a Metallica, Megadeth, & Anthrax - Helpless”.* Both tweets mention an infectious diseases, namely: Cholera and Anthrax, but their meanings are related to literature and music respectively. Additionally, tweets about vaccine, marketing campaigns, are considered irrelevant.

We distinguish relevant tweets from irrelevant ones using positive- and negative keywords as features. *Positive keywords* associated to diseases are pathogen (e.g., Streptococcus pyogenes) and symptoms (e.g., sore throat, fever, bright red tongue with a strawberry appearance, rash, bumps, itchy, and red streaks). We collect *negative keywords* associated to diseases from two freely-available resources: 1) MedISys[5] providing a list of negative keywords created by medical experts, and 2) Urban Dictionary[6], a Web-based dictionary of slang, ethnic culture words or phrases. In this case, the *Relevance Filtering* module consists of a classifier trained with features commonly used in a medical domain [5] in order to eliminate non-relevant tweets from the collection. The *Location Extraction* step aims to identify the location infor-

[3] http://opennlp.apache.org/

[4] http://www.opencalais.com/

[5] http://medusa.jrc.it/medisys/homeedition/en/home.html

[6] http://www.urbandictionary.com/

mation of a tweet. We identify tweets' locations in three different ways ordered by importance: 1) text-contained location, 2) geolocation information (latitude and longitude), and 3) user profile's registered location. Note that, we do not consider a tweet language (provided as an attribute by the Twitter API) in determining location information due to less accuracy. Consequently, a location associated to each tweet will be normalized into four different granularities of geographic concept hierarchy, namely: country, continent, latitude and worldwide.

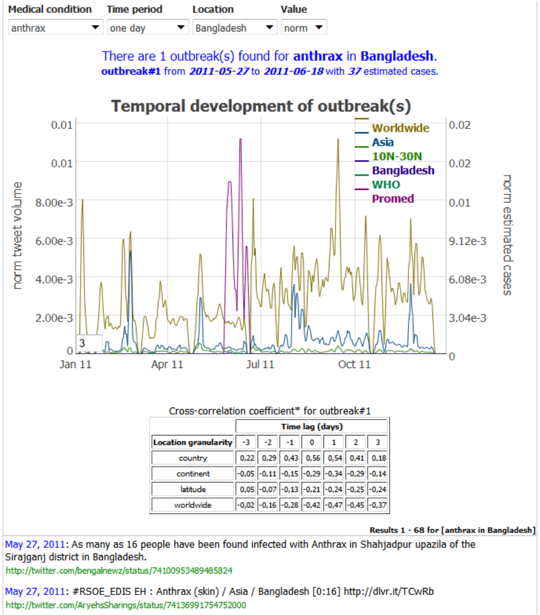
The intuition of using different geographic granularities is that public attention might depend on the geographic distance from an outbreak event. For example, people might talk or share their opinions about an ongoing outbreak in a neighboring country because they are concerned that the outbreak can spread into their country. For this reason, we also consider a *continent-level* location, *latitude-level* location and *world-level* location, in addition to a *country-level* location. Since the location mention could be a city or a geolocation information, all recognized location will be resolved with the Yahoo! PlaceFinder API[7]. Relevant tweets with location information are then indexed for further analysis.

3.3 User Interface

The *User Interface* component allows a user to visualize and analyze the temporal development of an outbreak event in Twitter by comparing with the outbreak information automatically extracted from official sources. The data are represented as an interactive, zoomable plots of time series that can be easily explored by the user. The charts are implemented by employing the Dygraphs JavaScript visualization library[8]. Given a time series graph, it is possible to explore and display values on mouse over allowing the user to analyze time series data in a particular time period. In addition, we also present the cross-correlation coefficient (CCF), which is a statistical method to estimate how variables are related at different time lags. This measurement can be interpreted as the similarity between two time series in volume, with consideration of time shifts. That is, the CCF value indicates whether there is a correlating trend in Twitter wrt. a real-world outbreak event. Figure 2 demonstrates the temporal analysis of the anthrax outbreak occurred in Bangladesh in 2011. The interface allows the user to filter the time series visualization considering different granularities of *time* and *location*. Moreover, a moving average parameter (in days) can be adjusted by the user. In Figure 2, the Twitter time series corresponding to all location granularities are shown, given a one-day time granularity and a smoothing parameter of three days. Below of the graph, the cross correlation results of +/-3 days of time lags are also displayed.

4. DEMONSTRATION PLAN

In this demonstration, we will show the system interfaces for comparing the global status of outbreaks in Twitter with the outbreak information automatically extracted from the external resource. In addition, we will show how to visualize an outbreak event on the interactive timeline, by the selection of a medical condition (disease name) and location and with the possibility to set different time granularity (daily, weekly or monthly) and location granularity (country, continent, latitude and worldwide). Moreover we will show how it is possible to visualize Twitter messages and outbreaks reports posted in a particular time period and how to perform cross correlation analysis between the time series corresponding to Twitter trend and real-world outbreaks events. During the demonstration we will restrict the visualization and analysis to 20 outbreaks from the list of 80 available in our web application in order to allow us to examine, together with the audience, all the possible settings on the data within a controlled space showing in depth all the possible events visualizations and analysis. The web application we will show during the demonstration is available at `http://meco.l3s.uni-hannover.de:8081/timemed/`.



Medical condition: anthrax
Time period: one day
Location: Bangladesh
Value: norm

There are 1 outbreak(s) found for **anthrax** in **Bangladesh**.
outbreak#1 from ***2011-05-27*** to ***2011-06-18*** with ***37*** estimated cases.

Cross-correlation coefficient* for outbreak#1

	Time lag (days)						
Location granularity	-3	-2	-1	0	1	2	3
country	0,22	0,29	0,43	0,56	0,54	0,41	0,18
continent	-0,05	-0,11	-0,15	-0,29	-0,34	-0,29	-0,14
latitude	0,05	-0,07	-0,13	-0,21	-0,24	-0,25	-0,24
worldwide	-0,02	-0,16	-0,28	-0,42	-0,47	-0,45	-0,37

Results 1 - 68 for [anthrax in Bangladesh]

May 27, 2011: As many as 16 people have been found infected with Anthrax in Shahjadpur upazila of the Sirajganj district in Bangladesh.
http://twitter.com/bengalnewz/status/74100953489485824

May 27, 2011: #RSOE_EDIS EH : Anthrax (skin) / Asia / Bangladesh [0:16] http://dlvr.it/TCwRb
http://twitter.com/AryehsSharings/status/74136991754752000

Figure 2: Temporal development of the 2011 anthrax outbreak in Bangladesh.

5. REFERENCES

[1] E. Aramaki, S. Maskawa, and M. Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of EMNLP'2011*, 2011.

[2] M. Cataldi, L. Di Caro, and C. Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of MDMKDD'2010*, 2010.

[3] A. Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics (SOMA'2010)*, 2010.

[4] V. Lampos and N. Cristianini. Nowcasting events from the social web with statistical learning. *ACM TIST*, 3, 2011.

[5] M. J. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. In *Proceedings of ICWSM'2011*, 2011.

[6] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of WWW'2010*, 2010.

[7] J. Strötgen and M. Gertz. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 2010.

[7] http://developer.yahoo.com/geo/placefinder/

[8] http://dygraphs.com/