Modeling Flu Trends with Real-Time Geo-tagged Twitter Data Streams

Jaime Chon, Ross Raymond, Haiyan Wang, and Feng Wang(✉)

School of Mathematical and Natural Sciences,
Arizona State University, Tempe, USA
fwang25@asu.edu

Abstract. The rich data generated and read by millions of users on social media tells what is happening in the real world in a rapid and accurate fashion. In recent years many researchers have explored real-time streaming data from Twitter for a broad range of applications, including predicting stock markets and public health trend. In this paper we design, implement, and evaluate a prototype system to collect and analyze influenza statuses over different geographical locations with real-time tweet streams. To evaluate the accuracy of the influenza estimation based on tweet streams, we correlate the results with official statistics from Center for Disease Control and Prevention (CDC). Our preliminary results have demonstrated that real-time tweet streams capture the dynamics of influenza at national level, and could potentially serve as an early warning system of influenza epidemics or flu trends.

Keywords: Influenza · Mathematical modeling · Geo-tagged twitter stream

1 Introduction

Recent years have witnessed tremendous growth of online social media such as Facebook and Twitter which offer people innovative platforms for sharing news stories, exchanging information and reporting latest statuses such as flu and epidemics. Although government agencies such as CDC (Centers for Disease Control and Prevention) of United States regularly report official and aggregated statistics on the trends of influenza or outbreaks such as SARS and Ebloa, these statistics often fail to reflect the latest development and progress since there is delay caused by manual data collections and complicated reporting processes.

The rising popularity of social media has led people to share their flu statuses and symptoms online, thus allowing an alternative channel to collect, analyze and monitor the latest trends of influenza development. Towards building a flu-surveillance system and studying whether Twitter data can be used as a robust indicator of influenza, this paper designs, implements, and evaluates a prototype system which automatically collects, analyzes and models geo-based flu tweets from real-time Twitter data streams for characterizing and modeling flu trends.

K. Xu and J. Zhu (Eds.): WASA 2015, LNCS 9204, pp. 60–69, 2015.
DOI: 10.1007/978-3-319-21837-3_7

Specifically, we extract flu tweets from real-time data streams and tag each tweet with geographical locations based on three information sources: (i) the geographical locations of the user who tweeted the message, (ii) the physical location where the user sent the tweet and enabled their geographical location tracking in the Twitter App, or (iii) the geographical location mentioned in the content of the tweets. The availability of geo-tagged flu tweets allows us to characterize and model flu trends at different geographical locations in real-time, which could serve as early warning signals before CDC releases official statistics, typically a few days or weeks later.

To verify the relevance of flu trends modeled by our system, we correlate geo-tagged flu tweets with the reported flu cases releasee from CDC official statistics. Our experimental results reveal a strong temporal correlation between these two metrics at coarse-grained geographical levels such as countries, but show little correlations between these metrics at fine-grained geographical levels such as cities, states or regions. The strong correlation between flu tweets and CDC statistics at national level demonstrates the potential application of our system for providing early prediction and warning of flu trends.

The contributions of this paper are two-fold as follows:

- We explore geo-tagged flu tweets from real-time Twitter data streams to characterize, analyze and model the trends and statuses of influenza and epidemic over different geographical locations.
- We develop a prototype framework for geo-based Twitter flu data analysis and modeling, correlate flu tweets with CDC statistics on the reported flu cases, and demonstrate the potential application of our system for providing early prediction and warning of flu trends.

The remainder of this paper is organized as follows. Section 2 describes the architecture of the real-time system we have developed for collecting, processing, analyzing and modeling influenza statuses with tweet data streams. In Sect. 3, we present the preliminary results with real-world tweets and characterize the correlation between flu tweets and reported flu cases at CDC. Section 4 describes related work of modeling influenza with social media data, while Sect. 5 concludes this paper and outlines our future work.

2 System Framework for Tweet Data Collection, Processing and Analysis, and Mathematical Modeling

Figure 1 illustrates our proposed framework to collect, analysis, and modeling Twitter data with mathematical models.

2.1 Twitter Data Collection

Twitter provides two mechanisms for programmatic access to their data encoded in JSON format: REST API and Stream API. The REST API provides access to user profile and follower data. This data allows the generation of a user following

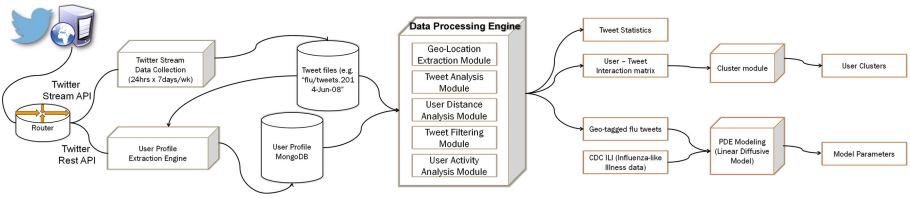

Fig. 1. Twitter data collection, Analysis, and Modeling Framework

topology, which captures who follows whom in Twitter and is critical to study the influence between Twitter users. The Streaming API continuously delivers real-time stream of Tweets matching a given search query over a persistent HTTP connection. The stream API offers two endpoints: Sample and Filter. The Sample endpoint delivers a small random sample (typically 1 %) of all public statuses (Tweets). The Filter endpoint delivers all Tweets that match a given query, which can include keywords, locations, and users, up to 1 % of all public Tweets. The aggregate Tweet information can be used as an indicator of what happens in real life.

We collect both Twitter user and Tweets information. Since it is hard to have a complete view of the whole Twitter user topology due to the large scale and dynamics of Twitter network, we start our collection with tweets associated with a specific topic or event, such as flu, Ebola, earthquake, etc., then collect the user profiles of users actively interact with these tweets and further build the topology of these active users.

Twitter Stream Data Collection module handles establishing and maintaining the connection with Twitter servers to retrieve tweets based on chosen keywords and/or tweets with a particular user as a source. Tweets are saved in JSON format in flat files with collection dates as file name. JSON format is chosen to eliminate encoding/decoding errors, efficiency, and flexibility. Flat file is chosen to remove the overhead of a traditional database;

User Profile Extraction module handles the collection of user data, including user profile and user's follower ids for every user observed in the Tweet data stream. The fetched data are stored in a MongoDB database [1], which is a document-based non-SQL database which provides fast and scalable storage. MongoDB is chosen due to its ability to handle the complex random reads needed by our user processing algorithms.

2.2 Twitter Data Processing

Data processing can be further divided into data cleaning and data analysis.

Tweet Filtering Module carries out the cleaning of the raw tweet stream from Twitter. The main functions of the module are: filtering and handling

of messages (notifications), removing extraneous fields from Tweets and user profiles, and reordering user's Tweets.

The raw Tweet stream contains not only Tweets matching the given query, but additional messages (notifications) from Twitter. These messages need to be filtered from the Tweet Stream. Two examples of notifications are: a tweet matching the query has been deleted, or the backlog for the stream is filling up. The latter example occurs when the tweet processing algorithm is not fast enough to process the tweets as fast as Twitter sends them.

We also remove all extraneous fields for each tweet to conserve disk space. This filter can be safely applied since a tweet contains a number of fields not relevant to our processing needs, the removed fields are only relevant in the case of displaying a tweet to a user. For example, the tweet will contain a url for a CSS file containing custom formatting information for a particular tweet.

The last component of this module reorders the tweets on disk to ensure tweets are in order of their timestamp. This needs to be done because Twitter doesn't guarantee that tweets will arrive in order, users are in different time zones, and most importantly all of the tweet processing algorithms depend on processing tweets in order for the simplicity/efficiency of the algorithms.

Geo-location Extraction Module. Based on [2], 84.2 % of twitter users have specified location in their twitter profiles and 10.3 % of twitter users have geo-location enabled Tweet. However, there are still challenges as follows: only a very small percentage of Twitter users add gps information to their tweets,; a significant number of users attempt to thwart automated systems by using bogus locations in their profile or by using valid locations with non-standard spelling or characters ("a" would be replace by "@"); all geocoding services have api limits that would be easily reached, and currently all geocoding services rely on input being as close to a location as possible and not on random text that may contain a location. We implement our module based on the "carmen" library [7] for geocoding tweets. This library provides us with a framework to resolve Tweet locations, and a small database of known locations.The included database contains names of states, abbreviations, cities, and common misspellings. We have expanded the database to include a lot more of the previously mentioned entities, as well as zip codes, airport codes, monuments, etc. We add enhancement to the Carmen library by adding a new resolver to process the Tweet text. The four fields we use to geo-tag a tweet are coordinates, place, profile.location, and text.

Tweet Analysis Module. Besides generating statistics of collected tweets, a major functionality of this module is to discover the retweet relationship, which is a mapping from a source (original) tweet to all its recursive retweets/replies. Discovering which tweets are retweet, reply, or contain the same content as the source tweet is important since when the flu tweet count is calculated, only source tweets instead of all tweets mentioning flu are considered. This is because for source tweets with tens of thousands of retweets/replies, most of the retweets are

just simple "take care" or "get well" which does not reflect whether the person who retweets the source tweet has flu symptoms or not. The flu cases are majorly captured in the source tweets.

There are three ways that a tweet can be retweeted. User clicks the "retweet" button, or "reply" button, or directly retweets a Tweet by typing RT at the beginning of a Tweet to indicate they are re-posting someone else's content.

To identify the source tweets and count the number of retweets, we go through each tweet and check if it is a reply (checks if the "in_reply_to_status_id" has a value) or a retweet (checks the "retweeted" field and if the pattern "RT @" occurs within the text), then increments a counter if the tweet belongs to an already identified source tweet. If the tweet is not categorized as reply or a retweet, the tweet id and text is stored as a source tweet.

In the case where the "retweeted" field is missing and the tweet contains the "RT @¡user name¿" pattern in the text, the algorithm will employ various techniques to compare the text to the text of already discovered tweets. One key item to note is that the retweet pattern can occur several times. The algorithm will loop through each "RT @" pattern from the outermost to the innermost and check if the user name belongs to any observed tweet. If this check passes, the algorithm will now attempt to match the text to the already observed texts. This means taking into account truncated text, as well as user added messages in the beginning and in the end of the text.

User Activity Analysis Module is used to measure a user's level of engagement by extracting the interaction between users and tweets and produces the user-tweet interaction matrix. The matrix records which users are involved in which source tweets and can be used to further cluster users into groups based on their tweeting behavior and cluster tweets into groups based on the users that are interested in them.

User Distance Analysis Module is used to measure how far a user is from the source by calculating the distance metrics for all the users actively participating in the discussion of a specific topic. Our definition of distance is a function of user activity, user profile, and their distance to the source user. The definition can be extended to add graph metrics such as the k-shell value of each user. User distance is an input to the PDE modeling module which can characterize and predict the spreading of a specific topic.

2.3 Mathematical Modeling

After data is collected, cleaned, and analyzed, the last component in our prototype is the mathematical modeling. This component includes two modules: (1) User/Tweet Clustering. We apply existing clustering algorithms such as spectrum clustering, k-mean algorithm to cluster users and tweets. We can verify our clusters using the collected user and tweet information. (2) PDE model design. PDE models are used to describe temporal and spatial diffusion of flu related topics and predict flu trend in real life.

3 Preliminary Results

3.1 Statistics of Collected Tweets

We have collected raw Twitter data covering several dimensions and categories as described below: (1) Any tweet that contains the keyword *flu*. This is one of the most common illnesses that CDC tracks closely. (2) Tweets containing the keyword *Ebola*. Investigating this dataset can help identify the spreading pattern for an actual outbreak of an unexpected disease, therefore increases the accuracy and speed of the predictive model in identifying new outbreaks. (3) Tweets related to the Malaysian Airlines flight disappearance. Tracking an event with high media coverage gives insight into the unique communication fluctuations due to such event.

Table 1 gives a brief summary of the scale of the collected data.

Table 1. Summary of collected data

Category	Amount
Total Size of Data	150GB
Total Number of Tweets	121,556,931
Total Number of Source Tweets	40,518,977
Total Number of Unique Users	19,083,164
Data Collection Start	October 11, 2013
Data Collection End	March 17, 2015

3.2 Correlation Between Twitter Flu Tweet Trend and CDC Reported ILI Case Trend

To verify the relevance of flu trends modeled by our system, we correlate geo-tagged flu tweets with the reported flu cases released from CDC official statistics. We adopt the flu data collected between January 3 and March 26, 2014, which is a subset of all the collected flu data that align with the flu season.

Figure 2 shows the number of weekly new flu tweets in Twitter and the number of weekly reported ILI cases provided by CDC. It shows strong linear correlation between the lines. In order to measure the linear correlation, we adopt Pearson's product-moment correlation coefficient $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$, where $cov(X, Y)$ is the covariance between variablex X and Y, and σ_X is the standard deviation of X. The result shows that the correlation coefficient between Twitter weekly new flu tweets count and the newly reported ILI is as high as 0.9297.

We further divide the tweet counts by regions and investigate the correlation between regional tweet counts and regional CDC regional ILI cases to investigate whether Twitter tweet data can be used to indicate the flu trend as the level of region. Figure 3 [6] illustrates the 10 regions defined by CDC. Figure 4 shows no obvious correlation. For example, during week 2 of our data collection, region 5 has the highest number of tweet flu count while region 6 has the highest number

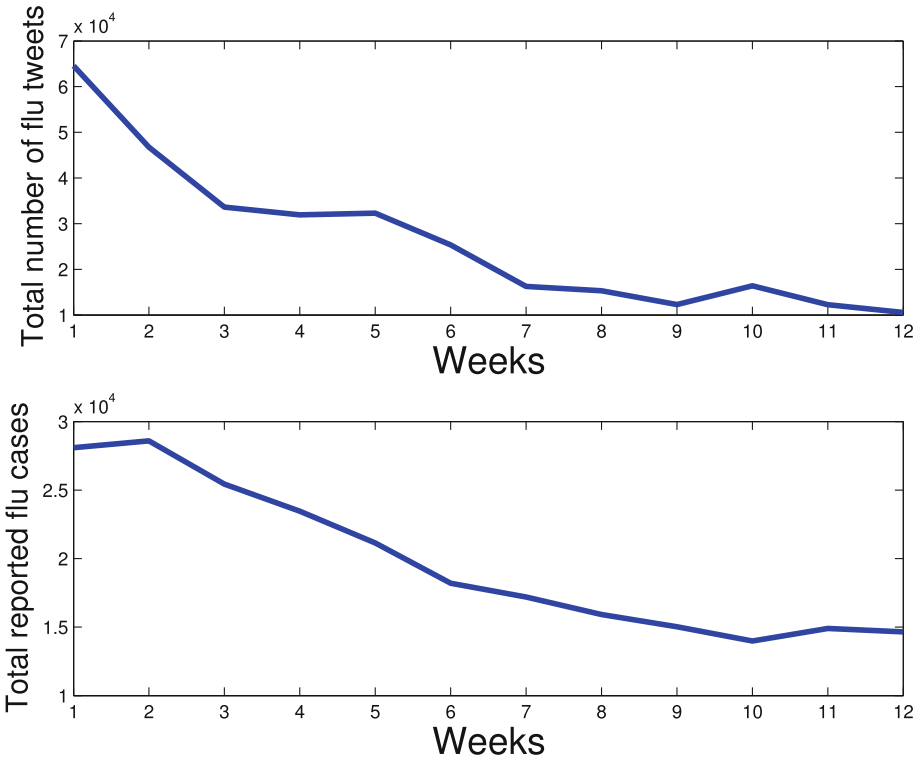

Fig. 2. Twitter flu tweet trend vs. CDC reported ILI case trend

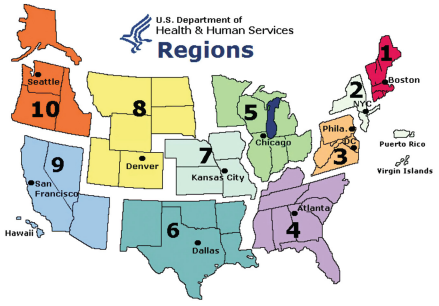

Fig. 3. CDC region

of CDC ILI cases. The lack of correlation between twitter flu count and CDC ILI cases at regional level might be caused by noises in the tweet text and needs further investigation.

In summary, our preliminary results reveal a strong temporal correlation between these two metrics at coarse-grained geographical levels such as countries, but show little correlations between these metrics at fine-grained geographical

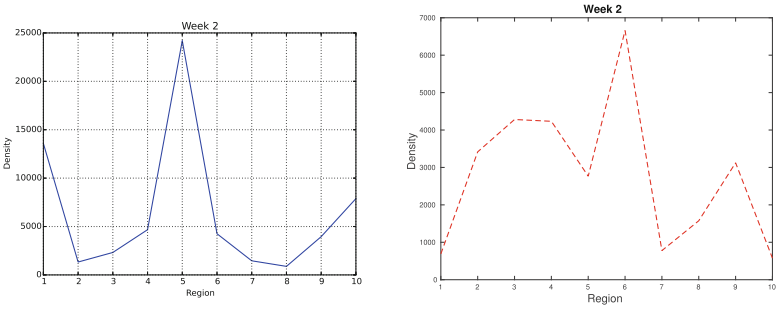

Fig. 4. Regional Twitter flu tweet trend vs. CDC reported regional ILI case trend

levels such as cities, states or regions. The strong correlation between flu tweets and CDC statistics at national level demonstrates the potential application of our system for providing early prediction and warning of flu trends.

4 Related Work

Many works have used Twitter and other types of social media data to assess and categorize the kind of information for measuring the spread of a disease in a population. For example, [8] looked at three web based biosecurity intelligence systems they pointed out the importance of social media, namely Twitter, recognizing how fast the information is passed and also seeing that many issues or messages were not altered through other means. Reference [9] suggested a complementary infoveillance approach during the 2009 H1N1 pandemic, using Twitter. They used content and sentiment analysis to 2 million tweets containing the keywords swine flu, swineflu, or H1N1. For this, they created a range of queries related to different content categories, and showed that the results of these queries correlated well with the results of manual coding, which suggests that near real time content and sentiment analysis could be achieved, allowing monitoring of large amounts of textual data over time. Reference [10] collected tweets matching a set of 15 previously chosen search terms including flu, vaccine, tamiflu, and h1n1 and applied content analysis and regression models to measure and monitor public concern and levels of disease during the H1N1 pandemic in the United States. Using a regression model trained on 1 million influenza related tweets and using the incidence rates reported by the Centers for Disease Control (CDC) as reference, the authors reported errors ranging from 0.04 % to 0.93 % for the estimation of influenza like illness levels. Reference [11] also analyzed cholera related tweets published during the first 100 days of the 2010 Haitian cholera outbreak. For this, all tweets published in this period and containing the word cholera or the hashtag #cholera were considered, and these data were compared to data from two sources: Health Map, an automated surveillance platform, and the Haitian Ministry of Public Health (MSPP). They showed good correlation between Twitter and Health Map data, and showed a good correlation (0.83)

between Twitter and MSPP data in the initial period of the outbreak, although this value decreased to 0.25 when the complete 100 days period was considered. Reference [12] applied SVM machine learning techniques to twitter messages to predict influenza rates in Japan. Reference [13] analyzed Twitter messages using regression models, in the United Kingdom and the United States respectively, obtaining correlation rates of approximately 0.95. More recently, [14] proposed a tool for real-time disease surveillance using Twitter data, showing daily activity of the disease and corresponding symptoms automatically mined from the text, as well as geographical incidence.

5 Conclusions and Future Work

Twitter, a popular online social media with hundreds of millions of users, provides a simple and robust platform for users to post their latest statuses including influenza symptoms. Rather than waiting for CDC to release the official statistics on the aggregated flu trends, this popular online social media offers an alternative channel to continuously monitor, collect and model influenza trends via flu tweets from real-time Twitter data streams. In this paper, we develop a prototype system to automatically collect, analyze and model flu trends via Twitter data streams. More importantly, we explore the geographical locations from user profiles, tweet location feature that attaches the current user location to a tweet, and the geographical information in the content of the tweets to tag flu tweets with coarse-grained and fine-grained locations. These geo-tagged flu tweets provide an accurate view of the latest flu trends at different regions. In our experiments, we correlate geo-tagged flu tweets with CDC statistics on the reported flu cases, and find the potential application of our system for providing early prediction and warning of flu trends. Our future work lies in developing more accurate geo-tagging mechanism and extending this framework to characterize and model flu trends from national, regional, and state level.

Acknowledgments. This project is supported by NSF grant CNS #1218212.

References

1. https://www.mongodb.org
2. http://www.beevolve.com
3. https://dev.twitter.com/
4. http://www.json.org/
5. https://www.openstreetmap.org/
6. http://www.cdc.gov/flu/weekly/regions2008-2009/hhssenusmap.htm
7. Dredze, M., Paul, M., Bergsma, S., Tran, H.: Carmen: a twitter geolocation system with applications to public health. In: AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI) (2012)
8. Lyon, A., Nunn, M., Grossel, G., Burgman, M.: Comparison of web-based biosecurity intelligence systems: biocaster, epispider and healthmap. Transboundary Emerg. Dis. **59**(3), 223–232 (2012)

9. Chew, C., Eysenbach, G.: Pandemics in the age of twitter: content analysis of tweets during the 2009 H1N1 outbreak. PLoS ONE **5**(11), e14118 (2012)
10. Signorini, A., Segre, A.M., Polgreen, P.M.: The use of twitter to track levels of disease activity and public concern in the u.s. during the influenza a h1n1 pandemic. PLoS ONE **6**(5), e19467 (2011)
11. Chunara, R., Andrews, J.R., Brownstein, J.S.: Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. Am. J. Trop. Med. Hyg. **86**(1), 39–45 (2012)
12. Aramaki, E., Maskawa, S., Morita, M.: Twitter catches the flu: detecting influenza epidemics using Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing Association for Computational Linguistics, pp. 1568–1576 (2011)
13. Lampos, V., Cristianini, N.: Tracking the flu pandemic by monitoring the social web. In: Proceedings of the 2nd International Workshop on Cognitive Information Processing (CIP), pp. 411–416 (2010)
14. Lee, K., Agrawal, A., Choudhary, A.: Real-time disease surveillance using twitter data: demonstration on flu and cancer. In: Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 11–14 August 2013 Chicago, IL, pp. 1474–1477. ACM (2013)