Data Mining on Twitter for Improving Public Health and Safety

Alfred George Francisco Milan III
University of California, Santa Cruz

***Index Terms*—Epidemic intelligence, Twitter, API, social media, data collection, Normalization Cross-Correlation, United Kingdom, ILI rate, framework, Probe GPS Data, Self-Reporting Tweets, Traffic Congestion, Swine Flu, Flu Outbreaks, Geographical breakdown, National Surveillance Data, Accidents, Vehicles, Safety, Road Segments,interval, week, Gaussian, influenza, geo-code, accurate, matrix.**

Abstract

Data Mining on Social Media gives researchers, public officials, and corporations more insight on how their consumers and their constituents respond or interact. Data Mining is an efficient method to gather highly valuable source for data researchers, medical professionals, and product planners who can infer when they assess what their targeted audience wants and needs.

Within this paper, we would look into how data mining on social media contributes to computing traffic congestion by using social media data, with the supplement of GPS probe data, and how data mining can reveal how social media plays an active role in predicting flu outbreaks, specifically Swine Flu. Given how GPS probe and national surveillance data provide less information and take more time to get results from, we turn to social media, most notably Twitter, to predict traffic congestion and flu outbreaks by deconstructing tweets and using information from netizens' tweets. We partake in a normalization cross-correlation process to calculate the official surveillance data as means of comparing it with self-reported Twitter data. Thus, we conclude that Twitter, single-handedly, could be used as an early warning system to inform people about health precautions they need to undertake. We come to conclude that data mining on social data gave significantly bigger data than if not.

Intended conference of submission: 2010 8th IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops) http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=5462925

Kostkova, P., Szomszor, M., and St. Louis, C. 2014. swineflu: The use of twitter as an early warning and risk communication tool in the 2009 swine flu pandemic. ACM Trans. Manage. Inf. Syst. 5, 2, Article 8 (July 2014), 25 pages. DOI: http://dx.doi.org/10.1145/2597892

Senzhang Wang, Xiaoming Zhang, Jianping Cao, Lifang He, Leon Stenneth, Philip S. Yu, Zhoujun Li, and Zhiqiu Huang. 2017. Computing urban traffic congestions by incorporating sparse GPS probe data and social media data. ACM Trans. Inf. Syst. 35, 4, Article 40 (July 2017), 30 pages. DOI: http://dx.doi.org/10.1145/3057281

I. Introduction

Technology, specifically social media, has made society better than worse. Despite how technology has evolved to the point where people are dependent on social media, the said platform could alleviate societal problems. Given how infectious diseases are able to be transmitted easily, Twitter, single-handedly, could be used to increase public awareness of a fatal disease and warn the public of pre-precautions that they should undertake.

First, due to the return of old diseases, globalization, and emergence of unknown diseases, spreadable diseases are threatening public safety and health in the 21st century. Given that these diseases rapidly spread amongst the population, citizens need a swift distribution of awareness via early warning and response teams. With the emergence of social networking sites, such as Facebook and Twitter, the public may be more informed in a much more rapid way during the Swine Flu era.

More specifically, various types of social media have been identified by researchers as Epidemic Intelligence's potential hotbed of information. Epidemic Intelligence is a automated early detection system that informs health authorities about health threats and disease outbreaks, their verification, and investigation means to take the required measures to protect citizens. [1] Given how a user may inform their associated followers or friends about a given disease via status updates in real time, the general public now has the chance to be served under an improved early warning and response system.

Second, since news corporations used to exclusively control what health-related news should be disseminated to the public, it limits how much information is being released. However, the rise of social networking has made the individual their own reporter as it gives users the opportunity to *report* un-edited health information in a viral manner. Yet, given that the information is un-edited, the high quantity of information being released does not necessarily mean that such data has high quality.

Likewise, social media can be used as a detailed resource for Chicago public safety officials who lack sufficient information to address the problem of traffic congestion. Having Twitter as a platform for official traffic accounts, such as *Roadnow Chicago* and *Total Traffic LA*, allows researchers to get legitimate real-time data regarding road congestion, accidents, road constructions, and other traffic events. The whole city would have Twitter users as traffic sensors who actually act as pedestrians, drivers, and passengers that would report real time traffic conditions where applicable. Meanwhile, GPS data would lack the information that specific tweets would contain, given the displayed time and content the latter has.

Given how GPS probe data's low sampling frequency, it has not been typically enough to calculate a large arterial network's traffic conditions in full capacity. Thus, social media platform, more specifically Twitter, could be used to retrieve more specific and accurate data given how tweets often state the location and time a given traffic collision or incident has occurred. However, since one is only concerned about proving Twitter's capabilities in delivering accurate data for public health agencies, one would only examine how social media data is being used in the hybrid framework to find the city-wide traffic congestion estimation.

Thus, social media could be used as one type of source of Epidemic Intelligence since this type of source has the potential to change the efficiency of Epidemic Intelligence. However, Twitter is mainly used since its provided API allows anyone to openly access profile content and activity while Facebook does not. [2] As a result, Twitter would be the primary social networking platform that will be in use throughout the entire research due to its potential of accessing a rather large sample as means of monitoring [3] and forecast [4] the spread of diseases, such as the Swine Flu.

Furthermore, Twitter could also be used as a more effective option to collect traffic congestion data since probe data has proven itself to be obsolete in the presence of new technology.

II. Objectives

The research focuses on a multiple of things:

1) Classifying tweets, and analyzing demographics.
2) Identifying self-reporting tweets and correlate them with national surveillance datas.
3) Analyze term frequency
4) studying how social media data is extracted to compute hidden traffic congestion correlation patterns among the road segments.
5) Examining how different types of traffic related information including GPS probe data, traffic related tweets, social events, road features, Points of Interest (POIs),as well as patterns to create and study a hybrid model.

We do so as means of dissecting tweets in order to harness the appropriate and acute information needed to check how social networking sites, specifically Twitter, are changing how data is being gathered.

III. The Role of Social Networks

A. Risk Communication and News Dissemination

Risk communication of diseases were then handled mainly by radio, television, the printed press to their audiences. Given technological advances in the 21st century, Social Networks and Twitter have now shown with no doubt that they play a significant part in the overall media distribution and the transmission of risk communication. [5] Furthermore, they have provided consumers with a platform to learn and report about diseases. During the Swine Flu Outbreak, Twitter had been recognized by researchers for its part in informing the masses about public health [6] and its active role as a new medium in the guise of a social network [7].

Likewise, GPS probe data is similar to the radio, television, and the press given the limited amount of precise information that GPS probe data could provide to researchers regarding traffic conditions. More importantly, the application of hashtags (styled as "#" with a term following it) could be analyzed to explore the idea of the presence of semantic identifiers as important markers in an upcoming, localized system. Especially due to the Twitter 130 character limit, new natural processing languages are needed. [8]

Inferring from the said information above, one can see the rising applications in the field of medicine. Despite its abuses, a recent study of antibiotic understanding on Twitter, [9] shows that medical information could be disseminated on social media. In particular, Twitter can be used to assess public knowledge (such as the misconception that one can stop taking antibiotics once the symptoms have disappeared or the assumption that antibiotics will treat one's cold) and therefore reveal a disparity in public understanding from medical information from qualified professionals.

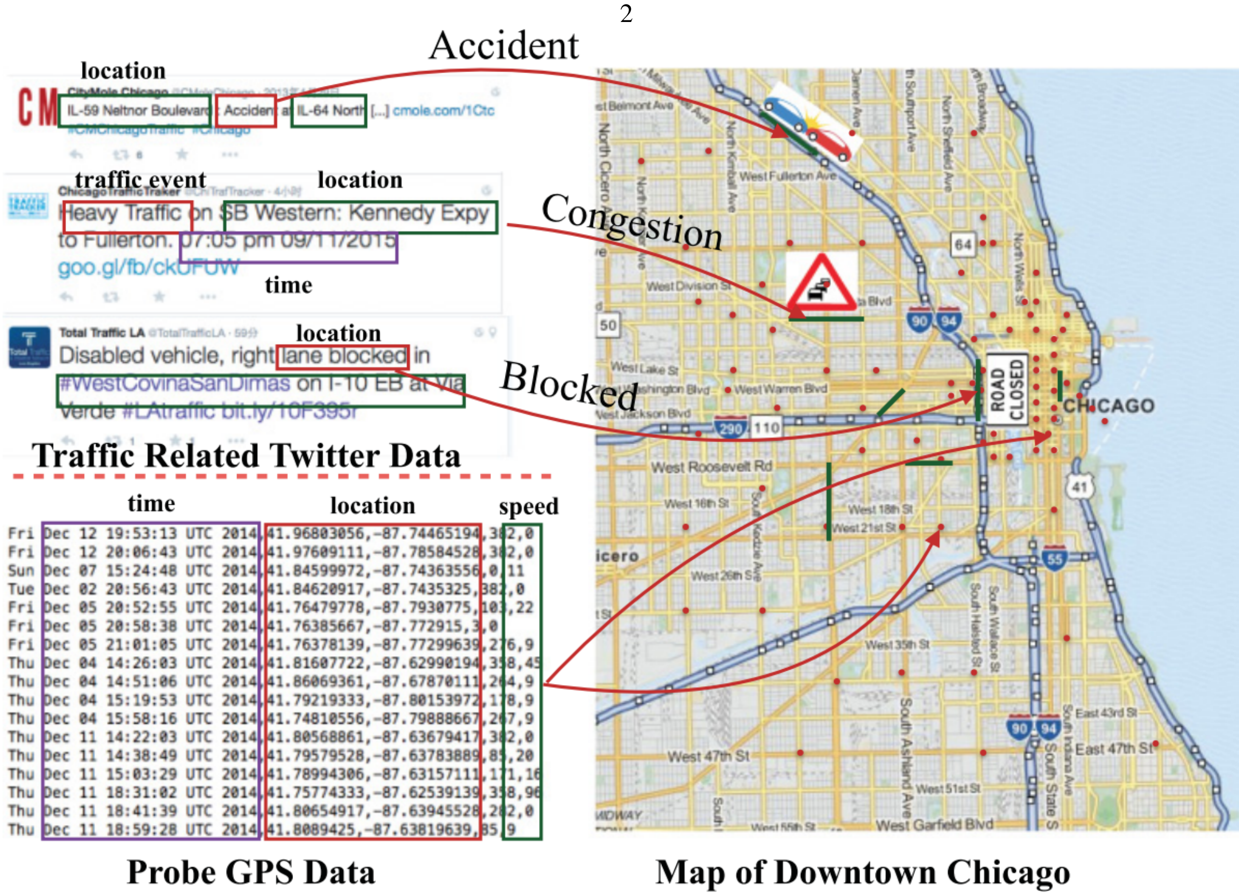


Fig. 1. Illustration of probe GPS data and social media data for traffic monitoring.

B. The Role of Social Networks for Traffic Congestion

While tweets have arguable quality in terms of spreading verified medical information, tweets have an articulate quality of giving researchers exact information about traffic congestion. Since both general user and official traffic authority accounts' traffic-related tweets are gathered and run through, GPS probe data (which contain only *geographical coordinates, time, and vehicle speed)* and the Twitter data (which contain the *exact time, location and type of traffic incident)* are used to designate the appropriate road segments. In this paper, one would compare the GPS probe data and the Twitter data to see how much more precise information one can get from the latter given that Twitter provides more detailed real time data. As seen on Figure 1 below, probe GPS data and Traffic related Twitter data are different since the latter contains missing information, such as the type of incident, the specific location without the need for coordinates, and the direction of the traffic that the incident took place. [10]

Using past information, a spatio-temporal frequent pattern mining problem should be used because specific road segments are discovered to be most likely to cause inter-connected congestion. Thus, in order to efficiently discover all the frequent co-congestion patterns in downtown Chicago, we propose a search tree based method. As a result, we may be able to identify outliers in road networks while helping on estimating traffic congestion. To combine data from social events, road physical features, POI features, and weather information to improving obtaining the nonrecurring congestion, TCE_R, a coupled matrix and tensor factorization scheme, is proposed to collaboratively factorize the congestion matrix to low rank matrices with other matrices and tensors formed by such rich information. The said low matrices could then be multiplied in order to complete the sparse traffic congestion matrix, thus getting the traffic states.

However, due to the fact that applying a search tree based method and TCE_R, a coupled matrix and tensor factorization scheme, would be more applicable to traffic congestion and not flu outbreaks, the said topics would be set aside for now and one would instead focus on how traffic congestion data is being collected in this experiment. Thus, the said data could be used to improve the accuracy between self-reported tweets from general users and its correlation with official Twitter accounts.

Now that the role of traffic and swine flu related tweets have been discussed, the framework model should now be discussed as means of how social media data is being utilized within the said framework.

IV. Hybrid Framework

A. *Traffic Congestion*

The framework of our model of the proposed solution for traffic congestion in downtown Chicago is show in Figure 2. Social media data, GPS probe data, and the road network data could be seen from above as the types of framework data. For social media data, the event category, location, and time information from each tweet are gathered from identified traffic related tweets. [11] This type of data is estimated by the congestion matrix **Y**. For road network data, one can extract road physical features and POI features on each road segment, and construct the road feature matrix **X**.

Using the given information, one can construct the congestion matrix **Y**, and the event tensor **A**.

- The two dimensions of **Y** are road segment ID and time slot. $y_{ij} = 1$ means that a congestion occurred in the jth time of day in the road segment r_i.
- The three dimensions of **A** are road segment ID, event category, and time slot. A_{ijk} = 1 means that in the jth time slot of a day on the road segment r_i, an event with category k happens.

For probe GPS data: one can construct the congestion matrix **H** and the confidence matrix **Q** from having a road segment based on longitude and latitude mapped from each probe reading. This type of data is estimated by the congestion matrix **H**.

- With the probe data, the traffic state in the jth estimated time slot on r_i, the road segment, is represented by each entry h_{ij} of H. The reliability of h_ij, the estimated traffic state, is represented by **Q**'s every entry q_{ij}.
- Thus, we are able to know which road segments are more likely to be in congestion in some time intervals since prior knowledge is gained from constructing historical congestion probability matrices Y^h and H^h. A congestion correlation matrix **Z** is constructed since the certain road segments are very likely to co-occur congestion re discovered by a spatial-temporal frequent pattern mining method that we propose. This is believed to be possible since there is a large volume of historical traffic congestion information obtained from the two types of data sources. Thus, the probability of road segments r_i and r_j co-occurring congestion represents **Z**'s every entry z_{ij}.

This framework shows how social media data is used and its usage of GPS probe data is only used as a comparison to Twitter data at this time. In order to have a more comprehensive understanding of social media data and its extraction, one needs to take into account the classification and analysis of related tweets.

V. Categorization and Analysis of the Twitter Dataset

A. *Classification and Term Frequency Analysis*

Classifying tweets is quite important since one needs to identify which tweets are related to this research. But first, in order to investigate the use of Twitter for EI and risk communication, one needs to determine three things: self-reporting flu, tweets with a URL or link, and retweets. Twitter is used for three things: to share other Web resources (general users often use url-shortening services to adhere to the site's 140-character limit), to retweet a tweet as means of spreading the original writer's tweet, and to check for phrases within tweets. Overall, tweets with URLs/links take up 65% of all tweets that contain the term *flu* while retweets take up from 1% to 3% from May 2009 to December 2009, since retweeting was gaining in popularity at the time.

One would also examine the daily term frequency as means of monitoring the data's behavior. Figure 4 shows that only one term (*p*andemic, which was used in 31% of tweets containing the term flu on 11/06/09) caused a significant event of tweets. As one can infer, the usage of one word in a viral manner could start an important event.

B. *Demographic Analysis*

Having analyzed the elementary characteristics of tweets, one can move on to global users who posted swine-flu-related tweets. However, one is only looking at the locations of those said tweets since most of the data would include the location of the twitter users. For instance, local outbreaks should be identified by epidemic intelligence applications. Due to the lack of information on global coordinates and the absence of geo-tagged data, finding the location field of the Twitter user's profile would be the next viable method. Using as a comparison to the Twitter data, official health

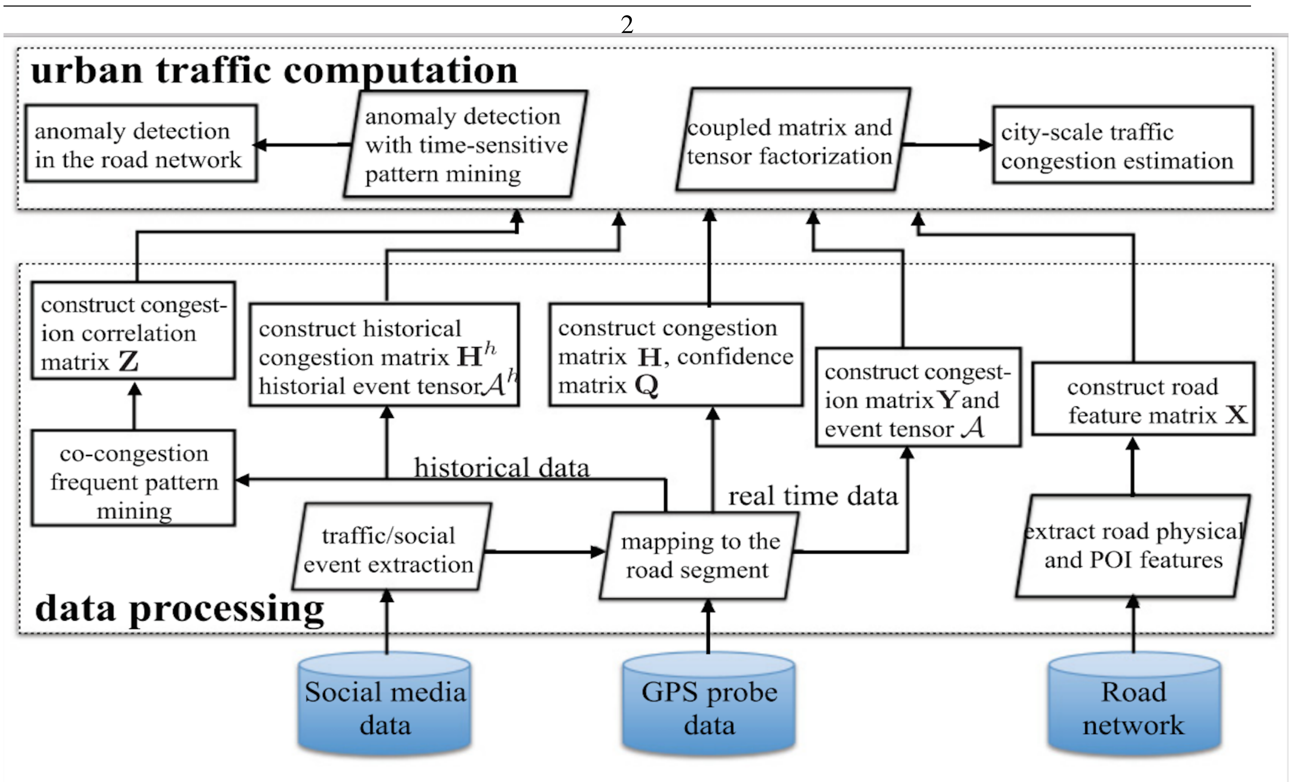


Fig. 2. Framework of our model

agencies' surveillance data in the United Kingdom must be examined in order to get a sample of 140,770 users whose locations are used as a basis of Twitter users' demographics. As shown in Figure 3, that method is used since this would be needed to analyze text processing over a series of tweets. [12] Also as seen on Figure 3, the U.S., U.K., and Canada has the biggest representation combined with 60%, 22%, 6%, respectively.

Thus, one needs to decipher the Geonames [1] country code from a location string by matching the location string to a Wikipedia entry, extracting coordinates, and matching coordinates to Geonames code.

First, this is especially useful if the tweet is from a town/city that bears the same name as another place. For example, Ontario, California and Ontario, Canada are geographically different. Second, geographic coordinates are excavated by querying from Dbpedia[2], the Semantic Web resource, once a potential matching Wikipedia article is retrieved. Finally, a Geonames service is used to retrieve the country code by using the geographical coordinates used in the previous step. Afterwards, we need to compare such demographics and their related tweets to national surveillance data.

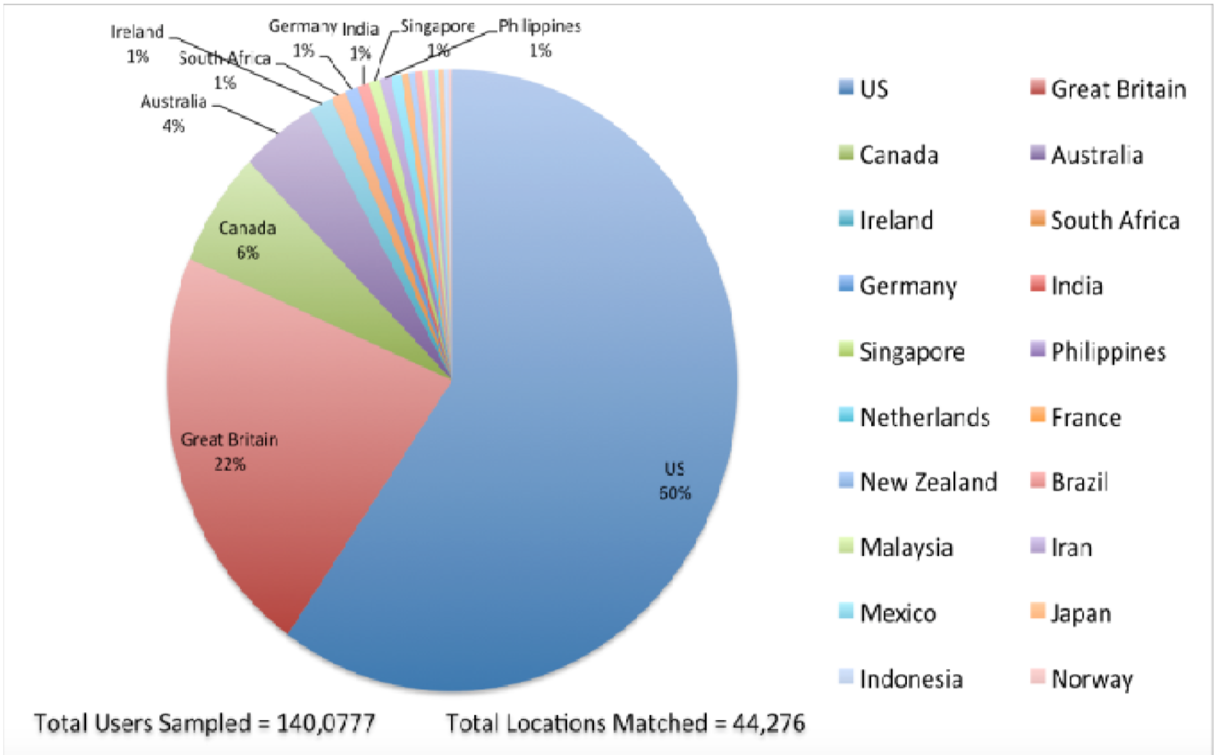
[1] http://www.geonames.org/
[2] http://dbpedia.org/

VI. Data Collection

A. *Official Traffic Authority Twitter Accounts versus Self-Reporting Tweets*

If one wants to collect data on a smaller scale (such as a city), one has to take into account, for example, the Official Twitter Accounts that Traffic Authorities and Departments have, as well as self-reporting tweets from the masses. For the Official Traffic Twitter Accounts, 10 Twitter Accounts have been identified to specialize in dispensing traffic information in the greater Chicago area. By doing so, one can retrieve a tweet such as *"Heavy Traffic on NB Western: Fullerton to Kennedy Expy. 06:15 pm 02/13/2015"* that clearly shows the necessary information needed to process the road segment, traffic event category, and time information.

For the General Sensor User, we retrieve a sample population of 10,000 Chicago-ans who use Twitter and have collectively tweeted more than 30 million times. Within these tweets, one identifies two things: the type of traffic event that occurred by identifying words such



2

Fig. 3. Geographical breakdown of matched Twitter data.

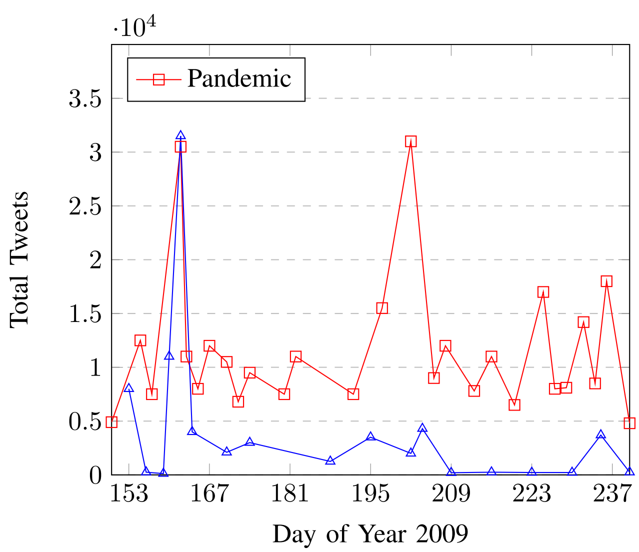


Fig. 4. Overall Twitter activity versus the tweets containing the term "pandemic"

as *"collision," "jam," "road construction," "slow/heavy traffic"*, and geo-coding tweets to identify the affected road segments. Figure 5 shows how one would extract the social event type from tweets and geo-code them.One can use the said tweet *"Heavy Traffic on NB Western: Fullerton to Kennedy Expy. 06:15 pm 02/13/2015"* and identify a given number of streets. One should also map general users' tweets to the corresponding road segments since only less than 2 million out of the 32.5 million total tweets are geo-tagged.

To check whether the previous two things accurately selects a tweets that reports a traffic incident, a sample of 1,000 tweets is selected and verified manually. As a result, about only 38 out of 1,000 false tweets, thus proving that the said two things effectively selected related tweets.[13] Using one of the official Twitter Accounts, *Chicago Events*, one can see in their tweet, *"Concert added: sch.mp/adPEt - RT@themizzi The Mizzerables has a show on 05/24/2015 at 08:00 PM @ Beat Kitchen in Chicago IL."* the event type *Concert*, time 05/24/2015 at 08:00 PM, and location *Beat Kitchen in Chicago IL.*.[14] Such tweets are easy to extract the said information from. By monitoring social events within the Chicago area, one can predict the intensity of traffic due to those social events by using a two-dimensional Gaussian model to measure the impact intensity of the event on the nearby road segments based on their

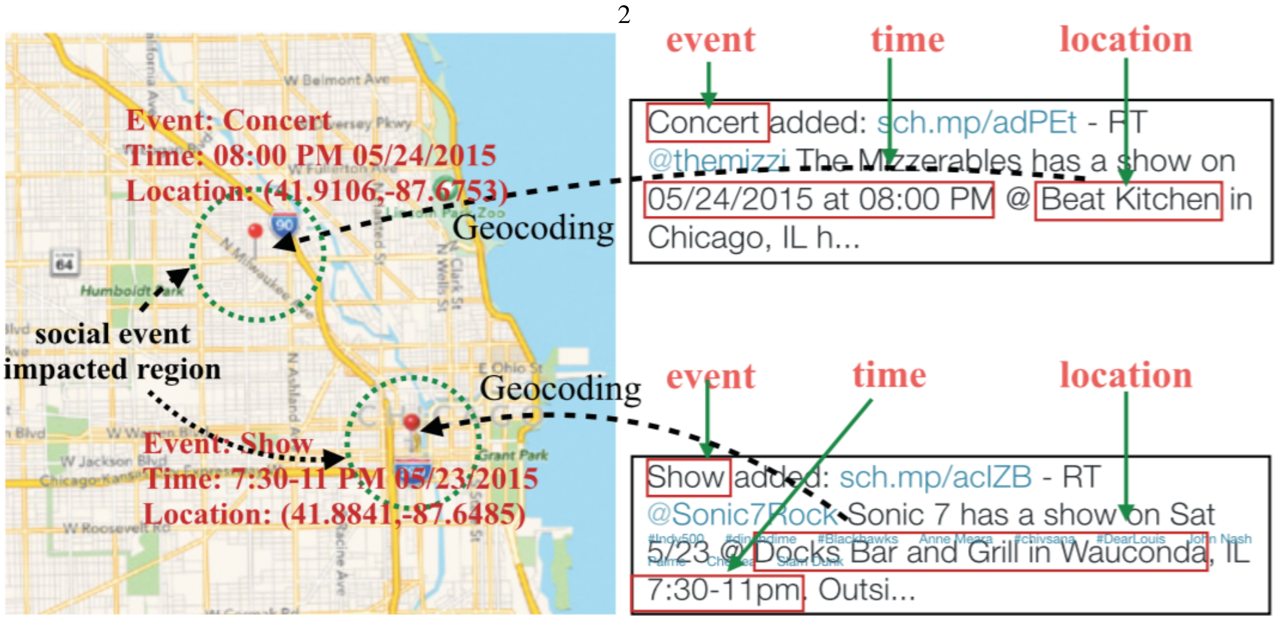


Fig. 5. Social events extraction and geocoding.

Euclidean distance:

$$I(r_i, se_j) = f(loc_{r_i}; loc_{se_j}, \sum_{se_j}) \tag{1}$$

$$= \frac{1}{2\pi|\sum_{se_j}^{1/2}|} \times e^{-1/2(loc_{r_i} - loc_{se_j})T}$$

where r_i is the road segment, se_j is the social event, loc_{r_i} is the location of r_i , and loc_{se_j} is the location of se_j.

As one can tell from Equation (1), as the distance between $loc_{r_i} - loc_{se_j}$ increases, the social event se_j's effect on road segment r_i also increases. As shown by Figure 1, Probe Data findings in the Chicago area on December 2014 primarily contain latitude, longitude, time, speed, and heading. However, we have to correctly map the said probe data with corresponding road segments due to lack of road segment information by computing the geographical gap between said road segment locations and pick the segment with the shortest geographical gap to said probe data.

Using a TCE_R Coupled Tensor and Matrix Factorization Scheme would be necessary to retrieving such rich information for estimating traffic congestion but not to retrieve flu related tweets. However, if the research were to gauge the outbreak of a given serious flu outbreak or illness in a city scale, such method would be applied.

B. *Self-Reporting Tweets versus U.K. and U.S. National Surveillance Data*

Having looked into how one would use Twitter content information on a Gaussian equation, one would look into the signal from self-reporting tweets from users who reported having the disease. Since the case numbers (which are reliably confirmed only by laboratory results) matter less than the signal change for EI early warning systems, one is unable to decipher the amount of the self-reported users that actually have the disease. Figure 6 shows the Twitter self-reporting signal for the U.K. Meanwhile, figure 7 shows a similar approach to correlate U.S. users self-reporting the disease against U.S. official surveillance data taken from the CDC Web site.

As a potential platform for early warning detection, one tests Twitters accuracy by collecting official surveillance data from the U.S. and U.K. Health Protection Agency (HPA), which provides weekly reports on the influenza-like illness (ILI) consultation rate for the United Kingdom [Public Health England 2013]. Meanwhile, we compare the said data with the percentage of self-reporting flu tweets that one calculates from the number of individuals that self-report the flu for each day in the investigation period without taking into account global trends in Twitter activity (e.g., spam, increased retweeting, and increased posting of links). Figure 4 exemplifies the strong correlation between the percentage of Twitter activity reporting flu and HPA ILI consultation rate for England and Wales. As one can tell, a sharp peak in Twitter activity occurs at the same time (around week 28, 6/07/2009) as the rapid increase in the number of HPA consultations.

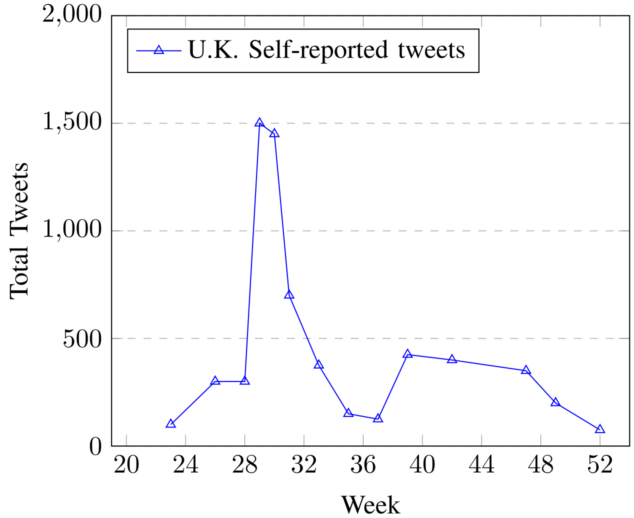


Fig. 6. Self-reporting tweets from U.K. Users

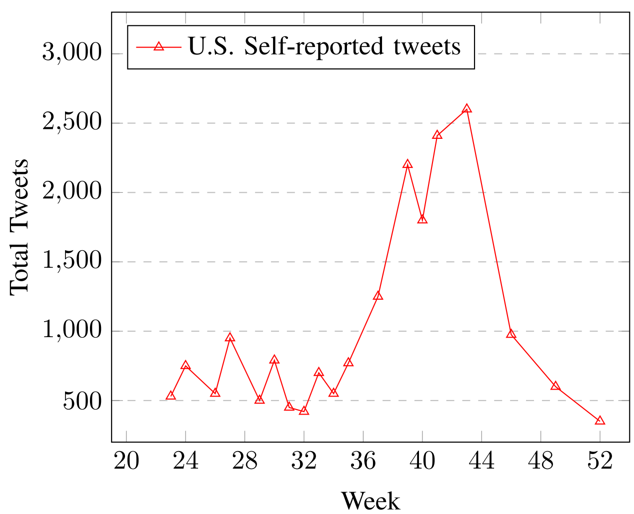


Fig. 7. Self-reporting tweets from U.S. Users

VII. Experiment: Using Normalized Cross-Correlation

Having looked at what one is comparing the Twitter data to, one needs to find the correlation between the Twitter and official surveillance data sources, and how close the spikes are in a given time period for the U.K. and U.S. The self-reporting tweets from U.S. and U.K. users are shown figures 6 and 7. Yet, one calculates the same correlation using the normalized cross-correlation ratio between various signals from Twitter, and the official RCGP U.K. surveillance data (collated by HPA) and the US from the CDC ILINet data to validate the correlation between the two separately. We also collect a weekly aggregation of Twitter data for the said ratio.

Eq. (2) gives the normalized cross-correlation function, where x(t) is the total number of tweets during week 1, y(t − i) is the number of reported cases according to the HPA and CDC during week (t − i) as shown below.

$$r = \frac{\sum_t (x(t) - x) \times (y(t-i) - y)}{\sqrt{\sum_t (x(t) - x)^2 \times \sum_t (y(t-i) - y)^2}} \qquad (2)$$

The term r is calculated across all self-reporting flu tweets that have links, and those that are retweets for values of i between -4 and 4. The various values of r for weekly offsets between i = 4 and i = 4 are displayed at figure 10. The similarity of two signals against a moving time lag is computed to get the said correlation ratio. Thus, r's values:

- for $i = 0$, exemplifies how much the two data signals are interconnected
- for $i = -1$, exemplifies how much the second signal is dependent on the first one.

During the interval between -1 and 1:

- when $r = -1$, signifies opposite signs despite having similar shapes.
- when $r = 0$, signifies that a correlation does not exist.
- when $r = 1$, signifies that the correlation between the two signals are strong and, hence, have near-identical shapes.

VIII. Results

As one can see, figure 8 and 9 show the strong correlation between self-reporting Twitter activity and corresponding national surveillance data. Thus, one can see that the normalization process used is effective in identifying self-reported flu tweets among other Twitter activity. Furthermore, one may see from figures 10 and 11 the weekly offset of both the U.K. and the U.S. and their near-symmetric shapes. In terms of gauging widespread flu outbreaks, public officials may be able to use self-reported Twitter data as a reliable complement to official surveillance data.

As an early warning system, Twitter has proved rather as a more effective system of collecting much needed public health data since the said social networking site can predict a week of flu-related Twitter activity while official surveillance data would take two or three weeks

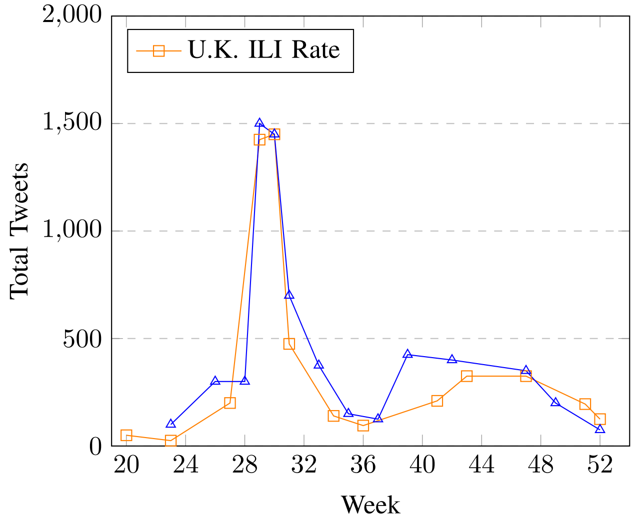


Fig. 8. U.K.'s ILI rates vs. Self-Reported Tweets.

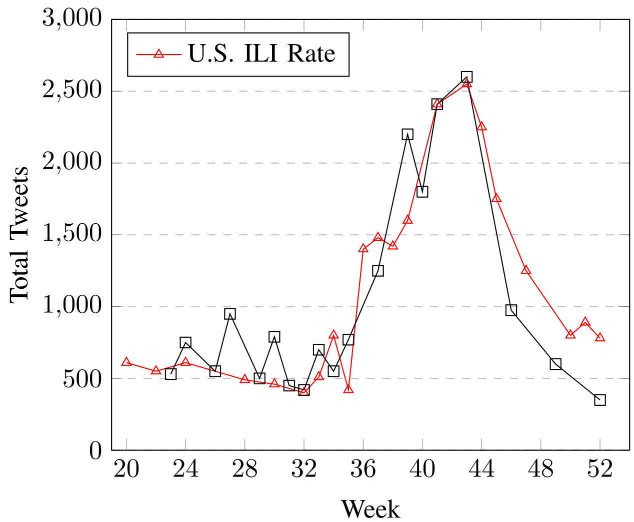


Fig. 9. U.S.'s ILI rates vs. Self-Reported Tweets.

later to obtain the same exact data. Furthermore, by using social event extraction and geo-coding tweets, we are able to further ensure the accuracy of self-reported tweets. From doing so, there is only 3.8% error among a sample of 1,000 tweets, thus showing that Twitter is a rather reliable source of information, in terms of data mining for traffic congestion and flu outbreaks.

IX. Conclusion

Since Epidemic Intelligence, global public health surveillance, and population monitoring are not reliant on new technology to gather information, the time period

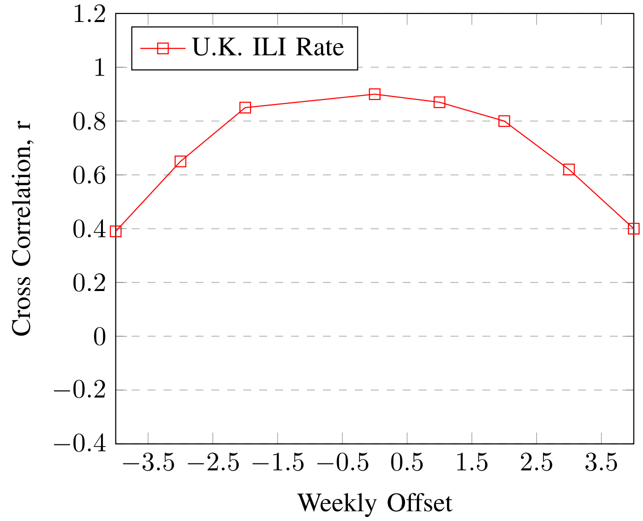


Fig. 10. The cross-correlation plot between Twitter and the ILI reporting in the U.K.

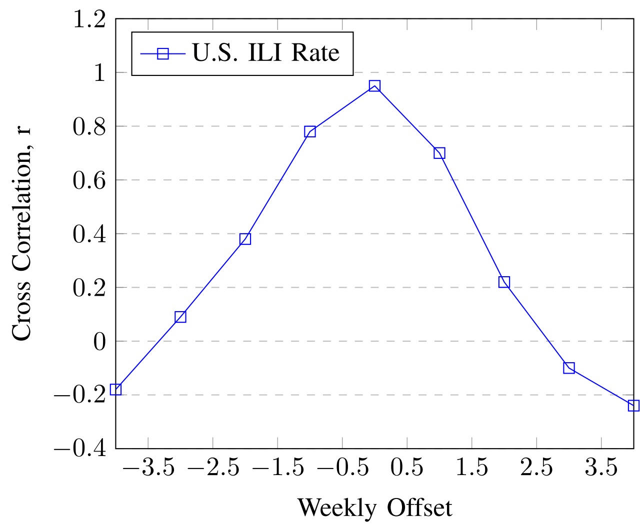


Fig. 11. The cross-correlation plot between Twitter and the ILI reporting in the U.S.

of getting information would be longer and the amount of information given would be substantially less. In this paper, we applied part of a novel framework to effectively integrate GPS probe data and social media data to more accurately compute urban traffic congestion, and, therefore, flu outbreaks. By using a social networking site, specifically Twitter, to gather data weeks before other primitive sources do, public health authorities would be more convinced to use Twitter as a complement to their various data sources. By collaborating with the HPA and CDC, as well as taking verified official Twitter

Accounts that specialize in traffic congestion, we are able to see how accurate Twitter can predict various data as an early warning system.

X. Future Work

Given how African-Americans and Latinos, and millennials have proven to be the most active Twitter users, we may exploit those social groups' social media activity in future work as means of improving the data collected. Furthermore, Twitter's journalistic merit should be further proven via a content analysis. Also, perhaps, we may look more into the TCE_R scheme in the future to further assess its potential in gathering the correct information and verify the effectiveness, and robustness of the search tree based method that was touched on earlier in the paper.

References

[1] Coulombier D. Kaiser R. Paquet, C. and M Ciotti. Epidemic intelligence: A new framework for strengthening disease surveillance in europe. Eurosurveill, 2006.

[2] D Williams. Api overview.

[3] de Bie T. Lampos, V. and N Cristianini. Flu detector tracking epidemics on twitter. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2010.

[4] V. Lampos and N. 2010 Cristianini. Tracking the flu pandemic by monitoring the social web. IAPR, 2010.

[5] B Duncan. How the media reported the first day of the pandemic h1n1 2009: Results of eu-wide media analysis. Eurosurveill, 2009.

[6] M. J. Paul and M Dredze. You and what you tweet: Analyzing twitter or public health. AAAI, 2011.

[7] Changhyun L. Hosung P. Kwak, H. and S Moon. What is twitter, a social network or a news media? International Conference on World Wide Web, 2010.

[8] Fuhry D. Demir E. Ferhatosmanoglu H. Sriram, B. and M Demirbas. Short text classification in twitter to improve information filtering. ACM, 2010.

[9] Scanfeld V. Scanfeld, D. and E. L Larson. Dissemination of health information through social networks: Twitter and antibiotics. 2010.

[10] Anto Satriyo Nugroho Sri Krisna Endarnoto, Sonny Pradipta and James Purnama. Traffic condition information extraction and visualization from social media twitter for android mobile application. ACM, 2011.

[11] Freddy Lecue Elizabeth M. Daly and Veli Bicer. Westland row why so slow?: Fusing social media and linked data sources for understanding real-time traffic conditions. ACM, 2013.

[12] Kostkova P. Szomszor, M. and E. 2010 de Quincey. swineflu: Twitter predicts swine flu outbreak in 2009. International Conference on Electronic Healthcare, 2009.

[13] Leon Stenneth Philip S. Yu Senzhang Wang, Lifang He and Zhoujun Li. Citywide traffic congestion estimation with social media. ACM, 2015.

[14] Leon Stenneth Philip S. Yu Senzhang Wang, Lifang He and Zhoujun Li. Citywide traffic congestion estimation with social media. ACM, 2015.