Big Smog Meets Web Science: Smog Disaster Analysis based on Social Media and Device Data on the Web

Jiaoyan Chen
College of Computer Science, Zhejiang University, Yuquan Campus, Yugu Road, Hangzhou, China
jiaoyanchen@zju.edu.cn

Huajun Chen*
College of Computer Science, Zhejiang University, Yuquan Campus, Yugu Road, Hangzhou, China
huajunsir@zju.edu.cn

Guozhou Zheng
College of Computer Science, Zhejiang University, Yuquan Campus, Yugu Road, Hangzhou, China
zzzgz@zju.edu.cn

Jeff Z. Pan
Department of Computing Science, The University of Aberdeen, Aberdeen, UK
jeff.z.pan@abdn.ac.uk

Honghan Wu
Department of Computing Science, The University of Aberdeen, Aberdeen, UK
honghan.wu@abdn.ac.uk

Ningyu Zhang
College of Computer Science, Zhejiang University, Yuquan Campus, Yugu Road, Hangzhou, China
zhangningyu@zju.edu.cn

ABSTRACT

Nowadays, people are increasingly concerned about smog disaster and the caused health hazard. However, the current methods for big smog analysis are usually based on the traditional lagging data sources or merely adopt physical environment observations, which limit the methods' accuracy and usability. The discipline of Web Science, the research fields of which include web of people and web of devices, provides real time web data as well as novel web data analysis approaches. In this paper, both social web data and device web data are proposed for smog disaster analysis. Firstly, we utilize social web data to define and calculate Individual Public Health Indexes (IPHIs) for smog caused health hazard quantification. Secondly, we integrate social web data and device web data to build standard health hazard rating reference and train smog-health models for health hazard prediction. Finally, we apply the rating reference and models to online and location-sensitive smog disaster monitoring, which can better guide people's behaviour and government's strategy design for disaster mitigation.

Categories and Subject Descriptors

H.2.8 [**Information Systems Applications**]: Database Applications—*data mining*

Keywords

Web Science; Big Smog; Health Hazard; Social Media; Device Data; Stream Data

*Corresponding author.

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author s site if the Material is used in electronic media.
WWW'14 Companion, April 7˜11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2576941.

1. INTRODUCTION

Nowadays, people are increasingly concerned about the environmental disasters and the health hazard caused by them. For example, many cities in China are suffering from deepening smog disaster, which has been widely reported to cause many diseases[8][5]. It is reported that the smog is related to a rise of 56% in lung cancer deaths in Beijing from 2001 to 2010 and an 8-year-old lung cancer patient in November 2013[11]. Therefore, it is very urgent and significant to analyze big smog and the caused health hazard for better strategies.

However, the existing methods for analysis of smog caused health hazard are always delayed as they usually adopt the traditional lagging data sources such as the daily records on hospital visits[2] and the mortality in the months after a smog disaster[1], all of which usually can't be timely available or cover a very large population. Meanwhile, the current smog monitoring usually only adopts the physical environment observation records from air quality stations and weather stations. In split of their widely deployment, they can only observe the physical environment statuses such as the concentration of $PM_{2.5}$ and the sky condition, which have low situation awareness.

As shown in some works like [15], social web can provide real-time and on-the-ground information for situation awareness enhancement during some natural disasters. Particularly, it has been proven that the health-related topics including ailment, treatment and symptom can be identified from microblogging service[9]. Moreover, the big cities that are struck by the smog disaster usually have a very large population of microblog users. Therefore, we can adopt the microblog data for quantitative analysis of smog caused health hazard.

Once we have quantified smog caused health hazard from social web data, it can further be integrated with device web data - physical environment observations. We can model the relationship between big smog and caused health hazard, and then build a standard reference for smog health hazard rating. Meanwhile, we can train models to predict the health hazard that may be caused through current physical envi-

ronment observations. Moreover, the integration of physical environmental observations and health hazard information can enhance the situation awareness of smog monitoring.

For the web of device, we utilize the open APIs to access the stream of observation records from 671 air quality stations[1] and 589 weather stations[2], whose rate is about 24,380 records per hour. For the web of people, we adopt the most popular Chinese microblog website named Weibo[3], which has more than 6 billion total users and updates about 4.2 million user tweets per hour. In the experiment, we collect millions of historical observation records and billions of historical user tweets from January to December in 2013.

Further more, in order to support online smog disaster monitoring and deal with massive web data streams, we propose a big streaming data processing framework, which relies upon the distributed streaming data processing system named storm[7] and big data warehouse called Hive[14].

In summary, our work has the following contributions:

(i) Social web data is utilized to define and calculate Individual Public Health Indexes (IPHIs) for the quantization of smog caused health hazard.

(ii) Social web data and device web data is integrated to build standard reference for health hazard rating and train smog-health models for health hazard prediction.

(iii) A big streaming data processing framework for smog disaster knowledge discovery is proposed to support smog disaster mitigation.

The reminder of the paper is organized as follows. Section 2 overviews the work. Section 4 introduces web data processing. Section 4 describes smog disaster knowledge discovery. Section 5 presents smog disaster monitoring. Section 6 displays the results and evaluation. Section 7 concludes the paper and discusses the future work.

2. OVERVIEW

2.1 Preliminaries

Definition 1 (Individual Public Health Index): we define six Individual Public Health Indexes (IPHIs) to quantify the health hazard from different ailments (Throat, Cold, Respiration), symptoms (Cough, Ail) and treatments (Mask). The IPHIs and their corresponding text description terms - words and phrases, are listed in Table 1. It references the works of [9][10], but is further modified for Chinese language and smog health hazard. Each IPHI is calculated with the frequencies of all its corresponding terms:

$$iphi = f(term_1, term_2, \cdots, term_m)$$

, where $term_i$ represents the frequency of ith term and m is the number of terms.

[1]http://pm25.in/
[2]http://openweathermap.org/
[3]http://weibo.com/

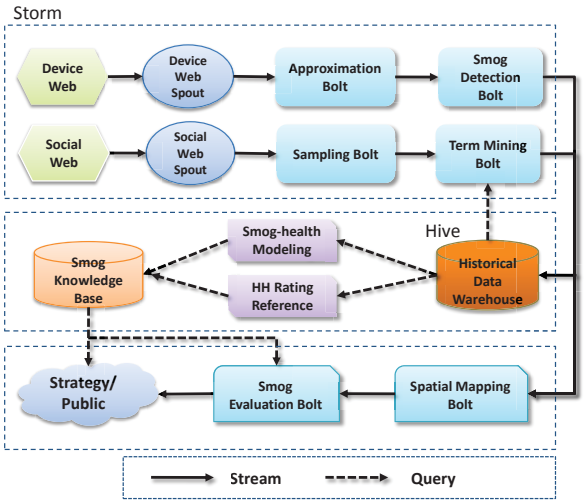

Figure 1: Big streaming data processing framework for big smog analysis. HH: Health Hazard

IPHI	Words or Phrases
Cough (C)	cough
Mask (M)	wear/take/prepare/buy mask
Throat (T)	sphagitis, amygdalitis, throat/tonsil dry/hurt/itch/inflammation
Cold (CD)	cold, sneeze, cough, snot, nose dry/hurt/itch
Respiration (R)	rhinitis, lung ailment, phlegm, breath hard, trachitis, pneumonia
Ail (A)	ill, visit hospital/doctor, medicine, fever

Table 1: The words and phrases for the IPHIs.

2.2 Framework Overview

Figure 1 shows our big streaming data processing framework, which consists of three parts: web data processing, smog knowledge discovery and smog disaster monitoring.

Web Data Processing: In this part (shown in the first row in Figure 1), we adopt the distributed real time computation system named Storm[7] and design a processing topology to process the online web data. In detail, the topology contains several connected spouts and bolts: two parallel spouts for the access of social web data and device web data, two successive bolts for tweet sampling and term mining - calculate the IPHIs for quantified health hazard information, and two successive bolts for air quality approximation[16] and smog disaster detection. As the framework can continuously process real time web data, it can further support online smog disaster monitoring.

Smog Knowledge Discovery: In this part (shown in the second row in Figure 1), we firstly build a Hive based data warehouse to store and index the historical smog data records, which include the observation records from device web and mined health hazard information from social web. Based on the historical social web data and device web data in the warehouse, we discover advanced smog knowledge: build standard reference to rate the health hazard according to the current IPHI records; train the smog-health model with artificial neural network to model the relationship between the physical environment observations and the IPHIs. The discovered smog knowledge is maintained in a knowledge base, and can be further applied to the evaluation of

a smog disaster and provided to the public for in-depth understanding of big smog.

Smog Disaster Monitoring: In this part (shown in the third row in Figure 1), online and location-sensitive smog monitoring with health hazard information is proposed through appending two successive bolts to the streaming data processing topology. One is the spatial mapping bolt, which maps the physical observation records and mined health hazard information of the smog disaster area. The other is the disaster evaluation bolt, through which we apply the current IPHIs to the standard reference to rate the health hazard, and apply the physical environment observation records to the smog-health models for health hazard prediction. We combine both outdoor observation records and the mined health hazard information as detailed smog evaluation, which is provided to the public and the strategy maker.

3. WEB DATA PROCESSING

3.1 Sampling

In our work, data stream of social web is processed window by window on Storm cluster. For each window of data records - tweets, they are continuously processed by different bolts. In the sampling bolt, we firstly group and filter the tweets according to spatial area to support location-sensitive smog monitoring. Secondly, we filter out the tweets that are unrelated to the public health topics, because the size of total tweets in one window can still be vary large and they may cause too much unnecessary computing for the following bolts.

For the tweets from mobile device, e.g., smartphone, their locations can usually be directly determined by the tags of latitude and longitude. However, it has been proven that less than 1% of the current tweets on Twitter have accurate geographic positions[6], and the situation is similar to Weibo. Luckily, we find that almost all the Weibo users are marked with a registered location - a city's secondary administrative district entered when the user registers, and the real geographic position of a tweet is usually consistent with the user's marked city. Accordingly, we directly use the registered city as the tweet's geographic position. This kind of tweet can't be used in location-sensitive monitoring as it needs tweets with accurate locations, but they can be applied to smog knowledge discovery.

According to wide reports[11] and historical case studies[3], big smog usually causes outbreak of various ailments, symptoms and treatments, and the words and phrases (terms) listed in Table 1 cover almost all of them. In filtering, we leave the tweets that contain any of the listed terms, and then we get the current health topic related tweet set: T_C. Although the terms in Table 1 may be uncomplete and ignoring the tweets without any of the terms may lose some information, the extracted information is basicaly sufficient to evaluate the public health hazard status.

In order to be suitable for Chinese tweets on Weibo and ensure high accuracy, each term's synonyms in Chinese language are found and used for filtering. In real word, a tweet that contains any word or phrase in Table 1 may not only represent a human status of ailment, treatment or symptom, but also some other things, e.g. an advertisement of medicine. Therefore, we test our filtering strategy with a collection of historical tweets before applying it to our approach, and some basic testing results are briefly evaluated in Section 6.1. In detail, we collect a tweet set filtered by a term, check whether each filtered tweet represents a status of ailment, treatment or symptom indicated by the term, and then calculate its filtering accuracies: c.

3.2 Term Mining

Once we have current tweet set T_C and historical tweet set T_H (query the historical data warehouse), we can mine the public health information and quantify the different aspects of health hazard with IPHIs. One of the mostly popular approaches for mining health related topic from tweet text is word frequency analysis, whose effectiveness has been displayed in some works[13]

Firstly, all the Chinese texts of the tweets are segmented through an existing java based lightweight Chinese segmentation system named ik-analyzer[4]. Meanwhile, we add the terms (words and phrases listed in Table 1) to the corpus so that the word segmentation system can recognize the terms and regard each term as one word.

Secondly, we calculate the relevance weight of each term using the strategy of tf-idf[12]:

$$w_{term} = tf(term, T_C) \times idf(term, T_H)$$
$$\begin{cases} tf(term, T_C) = \frac{f(term, T_C)}{max\{f(w, T_C), \forall w \in d\}} \\ idf(term, T_H) = log\frac{|T_H|}{|\{d \in T_H : term \in d\}|} \end{cases} \quad (1)$$

where tf is the function to calculate the frequency of the term in T_C, idf refers to the calculation of inverse document frequencies in T_H, d represents a tweet, and $|T_H|$ counts the tweet number. The higher value of w_{term} means a higher frequency in current tweets and a lower frequency in historical tweets, which further indicates that the status of ailment, symptom or treatment indicated by the term is more common in the current tweets than in the historical tweets.

Finally, we calculate the value of each IPHI with the relevance weights of its corresponding terms:

$$iphi = \sum_{i=1}^{i=m} w_i \times c_i \quad (2)$$

where w_i is the relevance weight of ith term, c_i is the filtering accuracy of the term and m is the number of terms. We multiply the relevance weight of the term with the term's filtering accuracy because this will eliminate the part except for the status of ailment, symptom or treatment. Then we sum the products because we currently regard that the terms equally represent the statuses of ailment, symptom and treatment, which can be further enhanced in our future work through adding weight to each item.

4. SMOG KNOWLEDGE DISCOVERY

4.1 Health Hazard Rating

To support health hazard evaluation of a smog disaster, we build a standard reference for health hazard rating. Firstly, we validate a regular rule that is widely reported[5][8]: *high AQI*[5] *usually causes outbreak of some ailments.* We validate the regular rule through a set of historical IPHI and AQI records, which cover the whole cycle of a smog disaster, as shown in Section 6.2. With the above regular rule,

[4]http://code.google.com/p/ik-analyzer/

[5]AQI (Air Quality Index) is a number to evaluate how polluted the air is.

we propose our method to quantify the health hazard rating standard.

Secondly, we quantify the correlation between each IPHI and AQI: calculate the correlation coefficient with historical records of many typical smog disasters. Then a most correlated IPHI ($i\hat{p}hi$), which has the highest correlation coefficient, is adopted. For $i\hat{p}hi$, we can collect a large sample set: $\left\{(i\hat{p}hi_i, aqi_i)|i = 1, 2, \cdots, l\right\}$, and use the exponential function (h) to fit the samples:

$$i\hat{p}hi = g(aqi) = a_1 \times e^{a_2 \times aqi} + a_3 \times e^{a_4 \times aqi} \quad (3)$$

where a_1, a_2, a_3 and a_4 are the parameters of the exponential function.

Finally, we can build the standard health hazard rating reference with exponential fitting function g and the current AQI based smog disaster rating table. Assume there is a smog disaster rate with the AQI range of $[\alpha_i, \alpha_{i+1})$, we can calculate the $i\hat{p}hi$ range of corresponding health hazard rate: $[\delta_i, \delta_{i+1})$:

$$\begin{cases} \delta_i = g(\alpha_i) \\ \delta_{i+1} = g(\alpha_{i+1}) \end{cases} \quad (4)$$

Then we can define a mapping function $\Delta(i\hat{p}hi)$ as the standard reference for health hazard rating:

$$r = \Delta(i\hat{p}hi) = \begin{cases} 1, & i\hat{p}hi < \delta_1 \\ 2, & \delta_1 \le i\hat{p}hi < \delta_2 \\ 3, & \delta_2 \le i\hat{p}hi < \delta_3 \\ \cdots & \\ n, & \delta_{n-1} \le i\hat{p}hi \end{cases} \quad (5)$$

where r represents the health hazard rating.

4.2 Smog-health Modeling and Prediction

In this part, we try to model the relationship between environment observations and IPHIs for health hazard prediction. In fact, it is the further enhancement of the standard health hazard rating reference. More physical environment factors except for AQI (aqi) are taken into consideration: humidity (h) and wind speed (ws), both of which are regarded as important short-term meteorological factors that may influence the health hazard caused by smog disaster. Meanwhile, models are trained for all the positively correlated IPHIs $\{iphi_1, iphi_2, \cdots, iphi_m\}$ instead of the most correlated one $i\hat{p}hi$.

Because of the complexity of the relationship between IPHI and multiple environmental variables, some simple fitting methods may be incapable to accurately model the relationship. In our approach, we adopt one typical Artificial Neural Network - Single Layer Feedforward Neural Network (SLFN), as well as a fast and accurate training algorithm named Extreme Learning Machine (ELM)[4]. Meanwhile, we use an incremental method[4] to optimize the hidden node number of SLFN, so as to achieve the optimal approximator. The procedure of training smog-health models is shown in Algorithm 1.

Once the smog-health models have been trained, we can adopt them for health hazard prediction. One typical case is predicting the whole day IPHIs that may be caused by a smog disaster through the currently observed the weather records and AQI records, which can further be applied in smog disaster monitoring.

Algorithm 1 Algorithm of smog-health model training

Require: IPHIs $\{iphi_1, iphi_2, \cdots, iphi_m\}$, time spans ts, cities cs, residual error ϵ and hidden node number N_0

```
1) Query warehouse with ts and cs, get result R
2) For each iphi_i:
3)   Generate training/testing sample tuples tr and te
4)   Initialize the model model_i
5)   While ||E|| > ε and N < N_0:
6)     Increase hidden node number N
7)     Train the added nodes with tr
8)     Test a_i with te, get residual error E
9)   EndWhile
10)  Output model_i
11) EndFor
```

5. SMOG DISASTER MONITORING

This part describes online and location-sensitive smog monitoring with physical environment observations, current health hazard rating and predicted health hazard information (IPHIs). It is a typical application that can be built on the stream processing framework and the discovered smog knowledge - standard reference for health hazard rating and smog-health models.

As shown in Figure 1, the application can be built as part of the stream processing topology, so the stream of web data can be continuously processed to support online smog monitoring, whose real-time depends on the size of time window of the stream processing system. Meanwhile, we add approximation bolt to approximate the air quality of all spatial areas with the method proposed in [16], which can help overcome the problem of insufficient air quality stations and support location-sensitive smog monitoring.

In monitoring, the smog detection bolt firstly detects the smog disaster through the timely environmental records such as AQI and sky condition. Meanwhile, the tweets of the disaster area or city are sampled and mined for health hazard information (IPHIs). Then both environmental records and mined records of IPHIs of the disaster area are passed to spatial mapping bolt for combination. We further calculate the health hazard rating through applying the records of Cold IPHI ($i\hat{p}hi$) to the standard reference $i\hat{p}hi$, and predict the whole day health hazard through the smog-health models.

Finally, all the information of a smog disaster, including the current outdoor physical environment, current health hazard rating and the predicted whole day health hazard (IPHIs) are provided to guide the behaviour of the public or the design of strategies.

6. RESULTS AND EVALUATION

6.1 Sampling Result

When we group and filter the tweets according to spatial area, we use the registered location for those tweets that have no latitude or longitude. To verify the feasibility, we randomly collected 100,000 geographic position marked tweets, and the marked geographic position is compared with the user's registered city. The result is shown in Fig. 3(A), which indicates that 81% of the tweets have consistent registered location and marked geographic position, and 5% have no registered location.

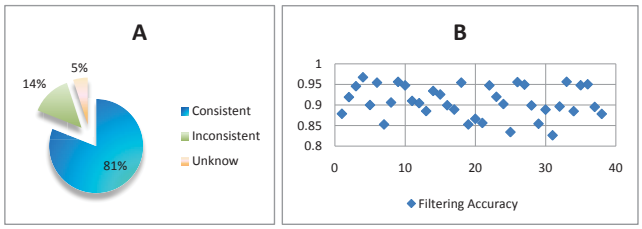

Figure 2: A: Comparison between tweet position and user register city; B: Accuracies of tweet filtering with 38 terms

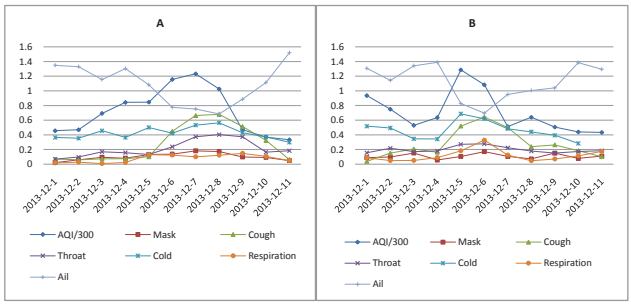

Figure 3: AQI and IPHIs from 12-01-2013 to 12-01-2013. A: Hangzhou. B: Shanghai

We can further calculate the accuracy of tweet filtering with the following formula.

$$acc = p_1 + (1 - p_1) \times \frac{p_2}{p_2 + p_3}, \quad (6)$$

where p_1 is the percentage of tweets with geographic location, p_2 is the percentage of tweets that have consistent user registered city and geographic position and p_3 is the percentage of tweets that have no user registered city. The final calculated result is 85.41%, which can basically ensure the feasibility of our approach.

Except for geographic location, tweets are also filtered by all the Chinese synonyms of the 38 terms listed in Table 1. For each term, we randomly select a number of detected tweets and manually judge whether the tweets represent the right statuses of ailment, treatment or symptom indicated by the term. In Figure 2, we list the filtering accuracies of all terms - 36 out of 38 terms can achieve accuracies higher than 87%, and 5 of them even achieve accuracies higher than 95%. Accordingly, we conclude that the proposed terms for each IPHI and the simple filtering approach have high soundness for tweet sampling.

6.2 Health Hazard Rating Reference

The standard reference for health hazard rating is based on a basic regular rule: *high AQI usually causes outbreak of some ailments.* In Figure 2(A), we display daily records of AQI and IPHIs of Hangzhou and Shanghai, both of which are attacked by big smog in the beginning of December in 2013. The figure can basically show the positive correlation between AQI and each IPHI except for Ail. The IPHI of Ail represents too general ailments and is more likely to be influenced by many other meteorological factors such as temperature.

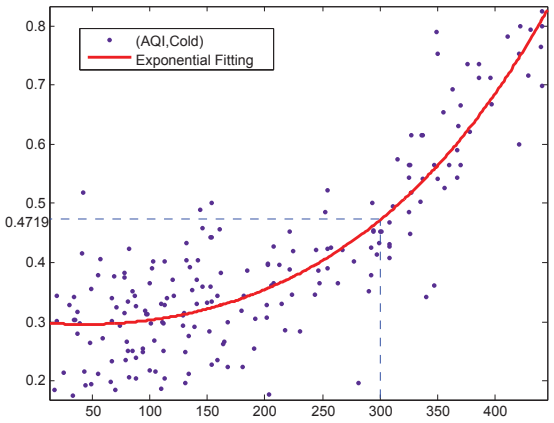

Figure 4: Exponential fitting between AQI and IPHI (Cold).

IPHI(Cold)	Rate	Color	Description
$[0, 0.2961)$	1	Green	Good
$[0.2961, 0.3030)$	2	Yellow	Moderate
$[0.3030, 0.3219)$	3	Orange	Unhealthy for sensitive group
$[0.3219, 0.3546)$	4	Red	Unhealthy
$[0.3546, 0.4719)$	5	Purple	Very unhealthy
$[0.4719, +\infty)$	6	Gray	Hazardous

Table 2: Standard reference for health hazard rating

We further choose 16 typical smog disasters in four large Chinese cities - Beijing, Shanghai, Tianjing and Hangzhou, and calculate the correlation coefficients of the six IPHIs: Cough (0.5673), Mask (0.5423), Throat (0.5142), Cold (0.7868), Respiration (0.3927) and Ail (-0.6776). As Cold is the mostly correlated IPHI with AQI, it is adopted as the IPHI ($\hat{iphi}$) for health hazard rating.

Once we have chosen $\hat{iphi}$, we can calculate the parameters a_1, a_2, a_3 and a_4 of the exponential function g with historical samples. Similar to the above correlation analysis, we prepare the samples from the records of the 20 typical smog disasters, and the sample number is 200. The result of exponential fitting is shown in Figure 4. The parameters of the fitting function is calculated: $(a_1, a_2, a_3, a_4) = (0.2057, -0.003033, 0.09418, 0.004729)$, and RMSE of the fitting is 0.07625.

With the value of all the parameters in g, we can calculate the boundary values of $\hat{iphi}$ for each health hazard rate, and work out the function $\Delta(\hat{iphi})$. Then we define standard reference for health hazard rating, as shown in Table 2. We regard that the smog disaster is hazardous when the IPHI of Cold is equal or higher than 0.4719. In Figure 4, this evaluation criterion is compared with the AQI-based evaluation criterion: the former judges through the horizontal blue line while the latter judges through the vertical blue line.

6.3 Smog-health Model

The standard health hazard rating reference models the relationship between the IPHI of Cold and AQI. However, some other IPHIs are also important to the caused health hazard, and the smog disaster does not include AQI records, but also some other environment factors. Smog-health models trained in this part are actually the enhancement of the standard health hazard rating reference.

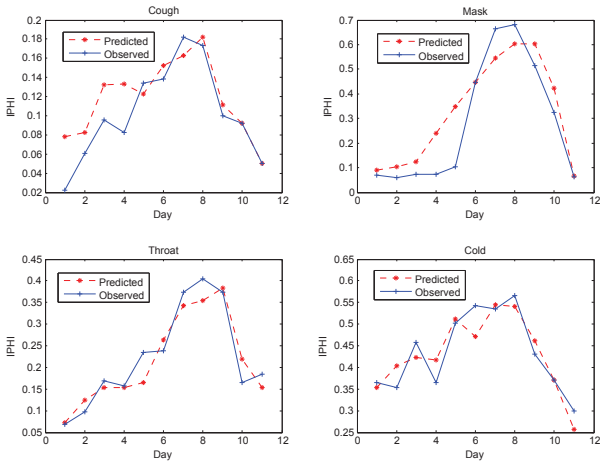

Figure 5: Predicted results of four IPHIs.

We firstly choose the positively correlated IPHIs - Cough, Mask, Throat and Cold, whose correlation coefficients are higher than 0.5. Both Respiration and Ail are ignored as they are less correlated to AQI. Then we select 30 typical smog disasters (ts and cs) for samples ($|tr| = 250, |te| = 50$), set predefined residual error $\epsilon = 0.05$ and hidden node number $N_0 = 50$, and finally train the corresponding smog-health models ($model_C, model_M, model_T, model_{CD}$) according to Algorithm 1.

Once the models are trained, we adopt an additional smog disaster (11 samples) for prediction and evaluation: predict the whole day IPHIs with the AQI, humidity and wind speed in the morning and compare the predicted values with the values mined from social web data. The result is shown in Figure 5, and for each IPHI, RMSE is calculated: Cough (0.0277), Mask (0.1085), Throat (0.0353) and Code (0.0374). We can see that the predicted curves can basically fit the trends of the observed curves, and the residual errors of the most days are acceptable.

7. CONCLUSION AND FUTURE WORK

In this paper, we utilize social web data for smog caused health hazard qualification, and integrate the health hazard data with the device data to build standard health hazard rating reference and train smog-heath models for health hazard prediction. Once we have discovered the smog knowledge, we apply it to online and location-sensitive smog disaster monitoring, which can better guide people's behaviour and government's strategy design for smog disaster mitigation. Meanwhile, a big streaming data processing framework is proposed to deal with massive web streams.

In the future, we will firstly include another important physical sensor data - remote sensing image, which can detect smog disaster in large area. Secondly, we will implement and evaluate the smog disaster monitoring application, and further provide it to the public through web API and smartphone application. Last but not least, we should further evaluate the standard health hazard rating reference with more real world health hazard data, e.g., hospital visit records.

8. ACKNOWLEDGMENTS

This work is funded by LY13F020005 of NSF of Zhejiang, NSFC61070156, YB2013120143 of Huawei and Fundamental Research Funds for the Central Universities.

9. REFERENCES

[1] M. L. Bell, D. L. Davis, and T. Fletcher. A retrospective assessment of mortality from the london smog episode of 1952: the role of influenza and pollution. *Environmental Health Perspectives*, 112(1):6, 2004.

[2] R. Chen, Z. Zhao, and H. Kan. Heavy smog and hospital visits in beijing, china. *American journal of respiratory and critical care medicine*, 188(9):1170–1171, 2013.

[3] D. L. Davis. A look back at the london smog of 1952 and the half century since. *Environmental Health Perspectives*, 110(12):A734, 2002.

[4] G.-B. Huang, L. Chen, and C.-K. Siew. Universal approximation using incremental constructive feedforward networks with random hidden nodes. *Neural Networks, IEEE Transactions on*, 17(4):879–892, 2006.

[5] V. Hughes. Public health: Where there's smoke. *Nature*, 489(7417):S18–S20, 2012.

[6] J. Lingad, S. Karimi, and J. Yin. Location extraction from disaster-related microblogs. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 1017–1020, 2013.

[7] N. Marz. Storm: Distributed and fault-tolerant realtime computation, 2012.

[8] M. Patience. Beijing smog: When growth trumps life in china. *BBC News Magazine*, 2013.

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

[10] M. J. Paul and M. Dredze. A model for mining public health topics from twitter. *HEALTH*, 11:16–6, 2012.

[11] Z. Pinghui. Smog blamed as girl, 8, becomes youngest lung cancer patient. *South China Morning Post*, 2013.

[12] J. Ramos. Using tf-idf to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*, 2003.

[13] D. Scanfeld, V. Scanfeld, and E. L. Larson. Dissemination of health information through social networks: Twitter and antibiotics. *American Journal of Infection Control*, 38(3):182 – 188, 2010.

[14] A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, and R. Murthy. Hive: A warehousing solution over a map-reduce framework. *Proc. VLDB Endow.*, 2(2):1626–1629, Aug. 2009.

[15] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 1079–1088, New York, NY, USA, 2010. ACM.

[16] Y. Zheng, F. Liu, and H.-P. Hsieh. U-air: When urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1436–1444, New York, NY, USA, 2013. ACM.