Information Systems 64 (2017) 281–291

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

Forecasting smog-related health hazard based on social media and physical sensor

Jiaoyan Chen [a], Huajun Chen [a,*], Zhaohui Wu [a], Daning Hu [b], Jeff Z. Pan [c]

[a] College of Computer Science, Zhejiang University, Hangzhou, China
[b] Department of Informatics, University of Zurich, Zurich, Switzerland
[c] Department of Computer Science, The University of Aberdeen, Aberdeen, UK

ARTICLE INFO

Available online 13 April 2016

Keywords:
Smog disaster
Health hazard
Social media
Urban data
Forecasting
Data mining

ABSTRACT

Smog disasters are becoming more and more frequent and may cause severe consequences on the environment and public health, especially in urban areas. Social media as a real-time urban data source has become an increasingly effective channel to observe people's reactions on smog-related health hazard. It can be used to capture possible smog-related public health disasters in its early stage. We then propose a predictive analytic approach that utilizes both social media and physical sensor data to forecast the next day smog-related health hazard. First, we model smog-related health hazards and smog severity through mining raw microblogging text and network information diffusion data. Second, we developed an artificial neural network (ANN)-based model to forecast smog-related health hazard with the current health hazard and smog severity observations. We evaluate the performance of the approach with other alternative machine learning methods. To the best of our knowledge, we are the first to integrate social media and physical sensor data for smog-related health hazard forecasting. The empirical findings can help researchers to better understand the non-linear relationships between the current smog observations and the next day health hazard. In addition, this forecasting approach can provide decision support for smog-related health hazard management through functions like early warning.

1. Introduction

Smog disasters are becoming more and more frequent and may cause severe consequences on the environment and public health in China. For example, in January 2013, smog had covered the capital of China, Beijing, for over 20 days. According to recent statistics [1], smog affects more than a quarter of the land and over 600 million people in China.

According to Virginia Hughes [2], smog is a health hazard that may adversely affect people's health. Sometimes it causes extreme and immediate public health emergency, like the one in 1952 London [3]. Therefore, it is necessary to develop a systematic approach to analyze, monitor and forecast smog-related health hazards in a timely manner.

On the other hand, social media as a real-time urban data source has become an increasingly important channel to observe events, trends and sentiment [4,5]. Negative comments on smog or complaints about smog-related health conditions from a small group of environment sensitive individuals can diffuse really fast on social media and cause much large scale of discussions and reactions. Therefore, social media with its network effects can be used to capture possible smog-related public health disasters in its early stage and provide warnings.

In the big data era, various technologies are developed to extract, process and analyze population-level social media

* Corresponding author.
E-mail address: huajunsir@zju.edu.cn (H. Chen).

http://dx.doi.org/10.1016/j.is.2016.03.011
0306-4379/

data, but few with the purpose of forecasting. Previous research [6] usually collected and analyzed such social media data for monitoring the impacts of nature environment on public health, but there is a lack of systematic approaches for forecasting smog-related health hazards with social media data.

Moreover, a variety of physical sensor platforms for monitoring smog status, including air quality stations, weather stations and earth observation satellites, are also widely deployed across China for both big cities and small towns [7], generating a huge amount of observational data about smog severity.

As Fig. 1 shows, we propose a predictive analytics approach that utilizes both social media and physical sensors for smog-related health hazard forecasting. It contains two major components: (1) modeling smog-related health hazards and smog severity with raw microblogging text and network information diffusion records and (2) forecasting the next day smog-related health hazards using an artificial neural network-based model.

To the best of our knowledge, our research is the first study to systematically model and analyze real-world social media and physical sensor data for smog-related health hazard forecasting. Firstly, this study can help researchers to better understand the non-linear relationships between current smog observations and the next day health hazard, in which physical sensors alone often fail to capture. Secondly, the proposed predictive analytics framework aims to provide decision support for smog-related health hazard management through functions like early warning for the coming smog-related public health emergency.

Moreover, we investigate the strengths of social media in smog-related health hazard forecasting. It can contribute more than physical sensors in forecasting the smog-related health hazards when the smog disasters are severe. Meanwhile, data about social observations' diffusion in social networks can further improve the forecasting accuracy.

2. Related work

2.1. Smog disaster and public health

On one hand, predictive analytics that are related to smog disasters or other kinds of air pollutions usually investigates the natural observations themselves without considering their related health hazard. Here are some examples. Merz et al. [8] conducted a time-series analysis of air monitoring data for the downtown Los Angeles station to detect the air pollution trends. Casado et al. [9] applied a series of geostatistics and visualization procedures to analyze hourly ozone measurements collected from 29 stations in the southeastern United States, which clearly confirmed the diurnal pattern of ozone fluctuations. Van et al. [10] investigated smog prediction problem in perspective of computational steering techniques which allow an optimal trade off between computation speed and prediction accuracy.

On the other hand, most studies that involve smog-related public health problems usually analyzed the impacts of smog on public health, but largely ignored real-time health hazard monitoring and forecasting. They mainly used objective indicators from physical sensors or statistics from hospitals. Pope and Dockery [11] conducted an extensive review on the research about health effects of particulate matter (PM) – the most harmful component in smog. They focused on the short-term and long-term PM exposure and its effects on mortality and some diseases. Recently, Hughes et al. [2] compared annual case numbers of chronic obstructive pulmonary disease (COPD) with smog trends in some cities to investigate the health effects of smog in past years.

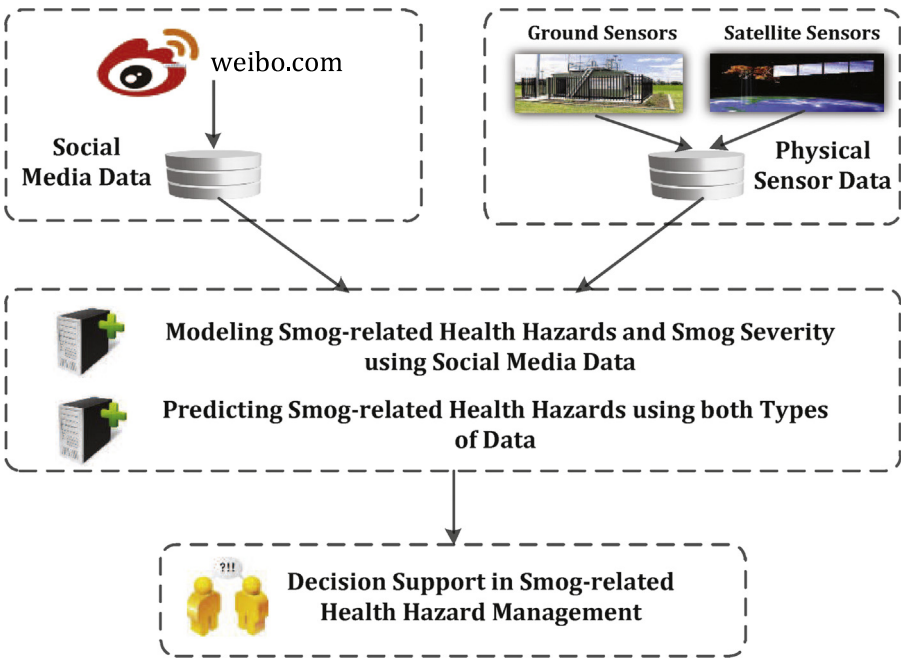

Fig. 1. Predicting smog-related health hazard with social media and physical sensor data.

Meanwhile, current studies investigating smog-related public health often adopt small data sets with limited population and time coverage. Motley et al. [12] measured only 66 volunteers' health status to acquire public health information during a smog disaster in Los Angeles. Chen et al. [13] used hospital visits to analyze health hazard trends under smog disasters, but the records only covered one hospital in Beijing during one high-smog period. Schwartz [14] investigated mortality caused by smog disasters, but the research was based on data records from some typical big smog disasters in some big cities.

2.2. Social media

Recently, social media including microblogging and social network services provided us real-time and large scale data sets related to public health. Lee et al. [6] and Culotta et al. [15] studied public health issues concerning with flu and cancer by analyzing Twitter messages. Paul et al. [16,17] mined topics of various ailments, symptoms and treatments from tweets with the Ailment Topic Aspect Model. Greene et al. [18] investigated several disease-specific information that was shared and exchanged on Facebook. Gardy et al. [19] acquired epidemiologic and genomic data through a social network for the research of tuberculosis.

Social media has also been applied to track the impacts of natural disasters as it will provide detailed information for situation awareness. Sakaki et al. [20,21] adopted Twitter data to detect and monitor disaster events including earthquake and typhoon in Japan with high probability and timeliness. Kongthon et al. [22] obtained up-to-date information about the disaster damage and the needs of the populace in 2011 Thai flood. Yin et al. [4] built an information system that utilized Twitter messages to enhance situation awareness during various crises and events, including natural disasters.

Although social media has been widely applied in investigating natural disasters and public health problems, there are very few studies that use social media to investigate smog disasters, not to mention smog-related health hazard. Mei et al. [23] was one of the earliest studies for smog disaster analysis with social media, but it aimed to infer the smog severity in the cities where no air quality stations were deployed, which was not related to smog-related health hazard. Our previous work [24] utilized Chinese tweets on Weibo to analyze the correlation between smog disasters and public health statuses, but did not study smog-related health hazard forecasting. Its prediction model simply approximated the health hazard with physical observations, which aimed at quantitatively analyzing the relationship and generating a standard for rating smog disasters' health hazard. The work presented in this paper actually extends our previous work [24] from historical relationship analysis to real-time forecasting. Our another work [25] in progress aims at forecasting smog disaster directly with different kinds of data including social media data. It will reinforce our study for decision making under smog disasters, but is quite different from the health hazard forecasting topic presented in this paper.

2.3. Artificial neural network

Artificial neural networks (ANNs) are computational models inspired by an animal's central nervous systems, and have been widely used in predictive analytics research. They are capable of learning complex non-linear discriminant functions [26], and can help solve public health problems like predicting active pulmonary tuberculosis [27] and Severe Acute Respiratory Syndromes (SARS) epidemic [28]. In our ongoing study [25], which applies different social observations and physical sensor observations to smog disaster forecasting, ANNs with single hidden layer or two hidden layers achieve a little higher performance than random forest and support vector machine. In recent years, ANNs are further extended with deep architectures. They have been proven to work very well for many complex prediction problems with big data in the fields like computer version, nature language processing and so on [29]. These theoretical properties and real world applications indicate that ANNs are able to approximate the relationship between the smog disaster and its health hazard.

There have been some state-of-the-art algorithms, such as back propagation (BP) [30], to train multiple layers ANNs for various regression problems. The recently proposed learning algorithm named extreme learning machine (ELM) [31] can train a single hidden layer feed-forward ANN at a high speed with high generalization performance. It can universally approximate any continuous target function and effectively solve many real-world regression problems, such as sales forecasting in fashion retailing [32]. According to some experiments [31], the single hidden layer ANN trained by ELM can achieve higher testing accuracy than some typical methods like support vector machine on many regression and classification benchmarks.

However, there is a lack of research which utilized prediction methods, such as ANNs, to fuse both social media data and physical sensor data for the forecasting of smog-related health hazard. This is mainly due to the lack of (1) systematic approaches for collecting, modeling and analyzing such information and (2) efficient prediction framework which can combine features from both social media and physical sensors.

3. A predictive analytics research framework

3.1. Overview

This research work addresses two challenges:

(1) Smog-related health hazard and smog severity modeling with social media.
(2) Smog-related health hazard forecasting using social media and physical sensor data.

We propose a predictive analytics framework, as shown in Fig. 2. It has three main parts. First, smog-related health hazard and smog severity are measured using raw social observations and social network diffusion data. Second, a health hazard prediction model is built using records of

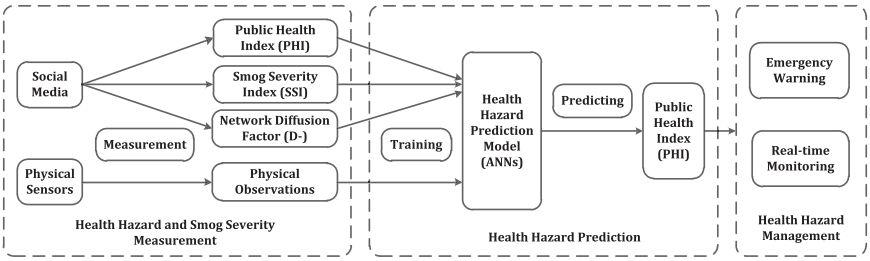

Fig. 2. The proposed predictive analytics framework for smog-related health hazard forecasting.

public health index, smog severity index, social network diffusion factor and physical observation, and is further utilized to forecast smog-related health hazards. Third, the forecasted and measured smog-related health hazards are applied to support decision making in smog-related health hazard management including real-time monitoring and emergency warning.

3.2. Smog-related health hazard and smog severity measurement

In our study, smog-related health hazards and smog severity are modeled as two indexes using social media information.

Definition 1. Public Health Index (PHI) is the sum of total relative frequencies of smog-related health hazard phrases in the current tweets. D-PHI is an enhanced public health index that includes consideration of diffusion in social networks.

Definition 2. Smog Severity Index (SSI) is the weighted sum of total relative frequencies of smog severity phrases in the current tweets. D-SSI is an enhanced smog severity index that includes consideration of diffusion in social networks.

Calculation of the two indexes includes five steps. First, both smog-related health hazard phrases and smog severity phrases are extracted. Smog-related health hazard phrases are those that are commonly used in Weibo (a Chinese social media site that is similar to Twitter) to complain about health problems that may be caused by smog disasters. According to some smog-related medical studies, smog disasters usually cause nose, eyes and throat irritation as well as heart and respiratory diseases in the short term [2,13]. We collect 200 Chinese phrases that are commonly used to complain about these health problems. Table 1(a) presents their English counterparts – some may represent multiple Chinese phrases with the same meaning.

Smog severity phrases are those that are commonly used in tweets to describe the current condition of a smog disaster. Based on the study of smog-related tweets, we collect 160 common Chinese phrases and define a severity order for each of them, as shown in Table 1(b).

Second, raw tweets with tags of time and location are gathered from Weibo. We partition a city into small grids, and then collect the current tweets of each grid area continuously. In detail, the program calls some APIs that enable us to acquire raw tweets (not forwarded) posted in a specified circular area defined by one position and one radius. The program ensures that the grid area is totally covered by the circular area.

Third, daily relative frequency *rf* of each phrase is calculated:

$$rf(p) = af(p, T_C) \times idf(p, T_H)$$
$$\begin{cases} af(p, T_C) = \frac{\sum_{d \in T_C} f(p,d)}{|T_C|} \\ idf(p, T_H) = \log \frac{|T_H|}{|\{d \in T_H : p \in d\}|}, \end{cases} \quad (1)$$

where T_H and T_C represent historical and current tweet sets respectively, p represents a phrase, d represents a tweet, $f(p, d)$ represents the frequency of phrase p in tweet d, $af(p, T_C)$ represents the average frequency of phrase p in the current tweet set T_C, $idf(p, T_H)$ represents the inversed document frequency of the tweets with phrase p in the historical tweet set T_H. The logarithm function is to scale up the fraction of rare tweets. The above algorithm is derived from the typical tf-idf algorithm [33]. The difference lies in the replacement of the largest word frequency in current tweet set with the size of current tweet set, which aims at eliminating the influence of other heat phrases on Weibo.

Fourth, PHI and SSI are calculated with the relative frequencies of all the phrases:

$$\begin{cases} PHI = \sum_{p \in P_1} rf(p) \\ SSI = \sum_{p \in P_2} rf(p) \times order(p), \end{cases} \quad (2)$$

where P_1 stands for the set of smog-related health hazard phrases shown in Table 1(a), P_2 stands for the set of smog severity phrases shown in Table 1(b) and $order(p)$ stands for a phrase's severity order. The calculation of SSI is weighted, because most Chinese tweets about smog itself come from experts such as the local environment agency or people who have a good knowledge of smog disaster. Usually, they use a fixed word set to describe the severity, and subjective severity level that the words' reflect is unified. In contrast, the health status tweets are mostly posted by common people. The severity level of one description word may vary from people to people. Using severity words for weighting may not reflect the peoples' subjective idea about severity, which is what we really want in this study.

Fifth, social network diffusion is considered to calculate D-PHI and D-SSI. On Weibo, any of people's actions, including retweet and like, indicates an agreement to the original tweet. Therefore, one record of such action is regarded as a duplicated raw tweet, based on which we calculate the network

Table 1
The phrases used in Weibo about smog-related health hazard and smog severity.

(a) Smog-related health hazard phrases

Type	Phrases
Nose	*Sneeze, runny nose, stuffy nose, nose disease, nose itches, nose irritation*
Eye	*Eye disease, eye pain, eye itching, dry eyes, eye irritation*
Throat	*Throat is dry, throat disease, tonsillar disease, cough*
Respiratory	*Have phlegm, pulmonary disease, asthma, bronchial disease, breathing difficulty, respiratory disease*
Heart	*Irregular heart rhythms, heart disease, heart hearts, high blood pressure*
Others	*Wear a mask*

(b) Smog severity phrases

Order	Phrases
1	*Have a little smog, a little bad air quality, air is slightly polluted, a little dusty sky*
2	*Have smog, bad air quality, air is polluted, dusty sky, high AQI, high* $PM_{2.5}$
3	*Have a severe smog, smog outbreak, air is severely polluted, extremely bad air quality, very dusty sky, extremely high AQI, extremely high* $PM_{2.5}$
−1	*There is no smog, air quality is good, sky is clear, the smog has gone*
−2	*Air quality is very good, sky is very clear*

diffusion-based average frequency:

$$daf(p,T_C)=\frac{\sum_{d\in T_C}(f(p,d)\times(g(d)+1))}{|T_C|+\sum_{d\in T_C}g(d)}, \tag{3}$$

where $g(d)$ represents a tweet's total number of retweet and like. Once *daf* is calculated, we use it to replace the average frequency *af* in Formula 1 to calculate the relative frequency *rf*, and further compute the value of D-PHI and D-SSI according to Formula (2).

3.3. Smog-related health hazard forecasting

As Fig. 3 shows, we develop an ANN-based prediction model to forecast the next day smog-related health hazard (PHI record) with the inputs including the current and the past air quality observations, meteorology observations and social observations.

3.3.1. Features

The inputs of the prediction model contain four kinds of features. The first kind of features (F_a) is extracted from air quality observations, including both air pollution concentrations (CO, NO_2, SO_2, O_3, $PM_{2.5}$ and PM_{10}) and air quality index (AQI) which comprehensively evaluates the air quality. The second kind of features (F_m) is extracted from records of various meteorological elements, including humidity, cloud value, pressure, temperature and wind speed, all of which have been proven to affect smog disasters greatly. For example, high wind speed and low cloud value usually make smog pollution to decrease in the next day. The third kind of features (F_s) comes from records of smog severity index (SSI) and social network diffusion incorporated smog severity index (D-SSI), both of them represent people's opinions and observations on the current and future smog disasters. The fourth kind of features (F_h) uses the current and recent PHI and D-PHI records, which enables the model to take the auto-correlation factor of the time-series data into consideration.

With all these features, we need to conduct feature selection. On one hand, we should filter out some unimportant kinds of observations such as some specific air pollutants and meteorology elements, as they may be not very predictive for the next day PHI or may be quite correlated with some other inputs. For example, we find that O_3 does not improve the prediction much when it is inputted with the other air pollutions. On the other hand, we should find out important records in time line for each observation. For example, we find that only the current records instead of all the recent records are important for wind speed and wind direction. In contrary, for D-PHI and PHI, both the current records and the records in the past 6 h are useful. In selecting features, a view independent subset searching method is adopted. It does not search all the subsets of the whole feature set ($F_a+F_m+F_s+F_h$), but finds out proper features from each kind of features individually, which reduce the complexity from $2^{|F_a|+|F_m|+|F_s|+|F_h|}$ to $2^{|F_a|}+2^{|F_m|}+2^{|F_s|}+2^{|F_h|}$. It is reasonable because in our application one kind of features represents one independent view to observe the forecasting target, which means two features from different views will not be highly correlated. Meanwhile, we further decrease the search spacing through testing records of each observation in time line from the current to the past. If we find the record at time $t-i$ is not important, the records before that time point will also be regarded as unimportant.

3.3.2. Model

The ANN-based prediction model is built with a method that searches for the optimum feed-forward ANN structure. It contains three components: parameters adjusting, training and testing. We use both ELM algorithm [31] and BP algorithm [30] for training, and adopt a typical cross-validation strategy which partitions the whole sample set into m complementary subsets in testing.

We present the model building procedure with the case of ELM which only uses one hidden node layer. First, initial activation function G, hidden node number L and regular parameter c are set. Second, input weights $\mathbf{a_i}$ and bias b_i of each hidden node are randomly generated. Namely, the input $\mathbf{x}\in\mathbf{R}^d$ are mapped into a random feature space in each hidden node:

$$h_i(\mathbf{x})=G(\mathbf{a_i},b_i,\mathbf{x}). \tag{4}$$

Third, output weights $\boldsymbol{\beta}$ of all the hidden nodes are calculated through regular linear solution. Namely, the algorithm minimizes the following objection function with small residual error and output weight norm.

$$L_{LEM}=\frac{1}{2}\|\boldsymbol{\beta}\|+\frac{c}{2}\|\mathbf{T}-\mathbf{H}\boldsymbol{\beta}\|, \tag{5}$$

where $\|\cdot\|$ denotes the Frobenius norm, $\mathbf{H}$ is hidden layer output matrix:

$$\mathbf{H}=\begin{bmatrix}\mathbf{h}(\mathbf{x_1})\\ \vdots\\ \mathbf{h}(\mathbf{x_N})\end{bmatrix}=\begin{bmatrix}h_1(\mathbf{x_1}) & \cdots & h_L(\mathbf{x_1})\\ \vdots & \vdots & \vdots\\ h_1(\mathbf{x_N}) & \cdots & h_L(\mathbf{x_N})\end{bmatrix}$$

and **T** is training data target matrix:

$$\mathbf{T} = \begin{bmatrix} \mathbf{t}_1^T \\ \vdots \\ \mathbf{t}_N^T \end{bmatrix} = \begin{bmatrix} t_{11} & \cdots & t_{1\,m} \\ \vdots & \vdots & \vdots \\ t_{N1} & \cdots & t_{Nm} \end{bmatrix}$$

where N is the number of training samples and m is the output dimension. Fourth, the trained model is tested. Fifth, another setting of parameters G, L and c is adopted and goes to the second step, or the program stops if all the parameter settings have been traversed. The ANN that achieves the highest testing accuracy is adopted. The optimum hidden node number L is found by incrementally searching with a stopping condition when the testing accuracy begins to decrease. The case for BP algorithm is quite similar with an additional parameter – number of hidden node layers, but without regular parameter c.

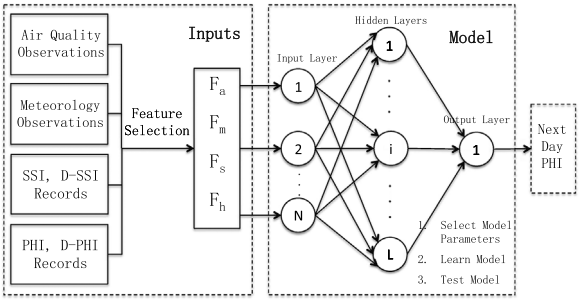

Fig. 3. The model and its inputs for smog-related health hazard forecasting.

4. Evaluation results

In evaluation, we analyze the advantages of incorporating both social media features and physical sensor features, display the improvement brought by utilizing network information diffusion and compare the forecasting accuracies of ANNs and other regression methods. We also present some forecasting results for Beijing and Shanghai when they are attacked by big smog disasters.

4.1. Data and experiments

We use physical sensor data and social media data in 8 cities (Beijing, Shanghai, Shijiazhuang, Tianjin, Nanjing, Hangzhou, Guangzhou and Wuhan) for experiments. Both data cover more than one year from May 2013 to November 2014. The former contains about 592 million hourly records about air quality and weather, while the latter contains about 315 million tweets with their retweet and like records. Meanwhile, the tweet number exceeds 10,000 for most days in both Beijing and Shanghai.

In evaluation, the records observed in the current and previous days are used as inputs. The next daily day PHI record which quantifies the smog-related health hazard is forecasted and compared against the observed records. We use PHI instead of D-PHI because information diffusion on the social network will non-uniformly magnify the observation thus making the index less objective and harder to forecast. For social media features, PHI/D-PHI and SSI/D-SSI records are calculated daily and the records of the current and past days

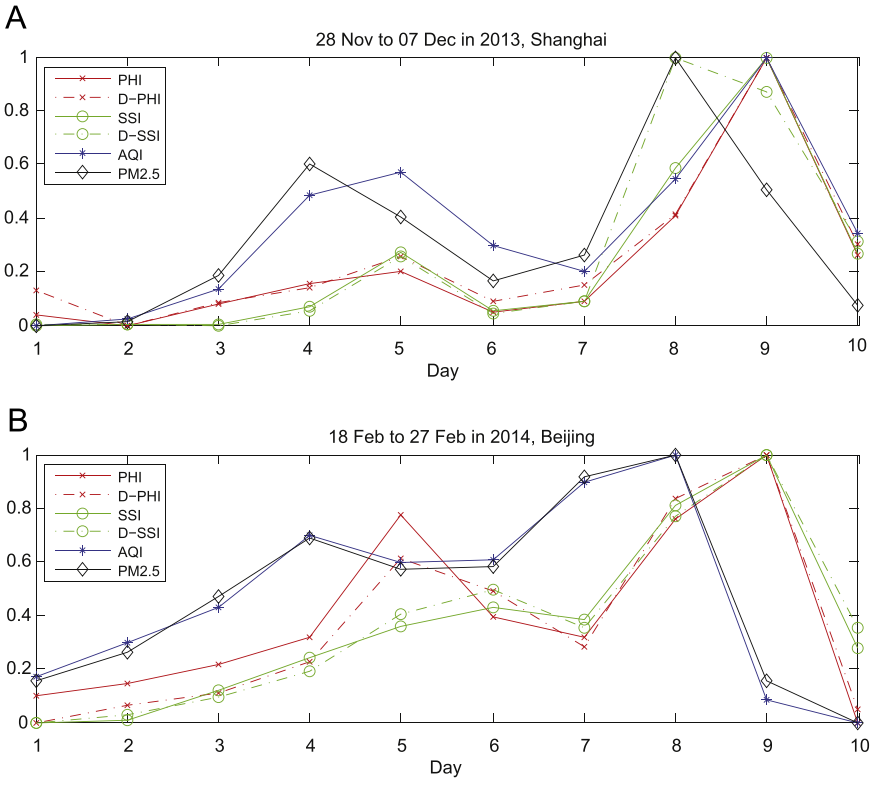

Fig. 4. Normalized daily PHI, D-PHI, SSI, D-SSI, AQI and $PM_{2.5}$ records in Shanghai and Beijing.

are used, while for physical sensor features, the records are observed hourly and the records before 24 o'clock are used. Days without enough tweets, which are caused by Weibo API limitations and network failures in data collecting, are discarded to ensure that PHI/D-PHI and SSI/D-SSI records are calculated based on a high population. Finally, 3240 samples are generated according to the data processing described in the paper. They are partitioned into a training set and a testing set with a cutting time.

Through the model building method described in Section 3.3.2, we get optimized ANN structures for different evaluations with different features. For ELM, which is a training algorithm for single hidden layer ANN, the optimized hidden node number ranges from 35 to 45, while for BP, the optimized ANN structure contains 2 hidden layers with 12–15 nodes in each layer. We evaluate the effect of using more hidden node layers, but find that it does not improve the generalization performance as our sample set is not very large. Meanwhile, two classic SVM regression methods, nu-SVR and epsilon-SVR provided by LIBSVM [34], as well as random forest regression method provided by sklearn, are also applied for comparison. Each experiment is conducted with multiple pairs of training and testing sample sets (partitioned with different cutting times) and each test is repeated multiple times. The average of the results is finally adopted for evaluation.

4.2. Correlations between the features and the next day PHI

We conduct some data analysis to evaluate the correlations between our features and the forecasting target – the next day PHI. It includes two parts: visual comparisons and statistical tests. Fig. 4 displays the records of PHI and some of the considered features during two big smog disasters in Shanghai and Beijing. In the figure, the trends of SSI, D-SSI and D-PHI are relatively consistent with that of PHI, and the latest $PM_{2.5}$ record is usually consistent with the next day PHI records. The latest AQI record may either be consistent with the current day PHI or the next day PHI. Meanwhile, we calculate the correlation coefficients between the next day PHI and the current day SSI, D-SSI, D-PHI, $PM_{2.5}$ and AQI with the data in all 8 cities (3240 samples). The coefficients are 0.434, 0.452, 0.481, 0.387 and 0.399, and their corresponding p values are all less than 0.042 in two-sided confidence testing. The above analysis indicates that SSI, D-SSI, D-PHI, AQI and $PM_{2.5}$ are quite correlative with PHI, and are indicators to forecast the next day PHI.

The results in our previous study [35] can further confirm the above conclusion about the correlation. According to [35], both meteorology elements like wind speed and air pollutants like O_3 highly correlate with $PM_{2.5}$ and are important short-term factors of smog disasters. They can indirectly influence the next day smog-related health hazard. Briefly, the next day PHI is correlated with our social media and physical sensor features, and it is reasonable to utilize them for smog-related health hazard forecasting.

4.3. Comparison between social media and physical sensor features

First, we compare the testing accuracies of the health hazard prediction model using different kinds of features. As shown in Table 2, the model using both physical sensor and social media features (P+S) has much lower root-mean-square error (RMSE) than that using either physical sensor features (P) or social media features (S). On average, the RMSE of P+S is about 20% lower than P and about 25% lower than S when single hidden layer ANNs and ELM are adopted for training. The comparison result is similar when multiple hidden layers ANNs and BP are adopted.

Second, we investigate the advantages of physical sensor features and social media features, which helps explain why their integration can predict more accurately. All the tested samples are classified into four categories according to smog severity which is evaluated by AQI here. Average RMSEs and relative errors are recalculated for each category, as shown in Figs. 5 and 6.

From the figures, we can find out that for the days that are not seriously polluted (AQI < 200), physical sensor features can achieve lower RMSE and relative error than social media features, while for the days that are severely polluted (AQI ≥ 300), social media features perform better. This result is consistent with our common sense, as people post much more tweets in those severely polluted days, which provide more positive samples. Actually, according to the statistics,

Table 2
Testing accuracies (RMSEs) of the health hazard prediction model using different training methods and features. ELM-ANN and BP-ANN represent one hidden layer ANNs with ELM and multiple hidden layers ANNs with BP. P and S represent physical sensor features and social media features, while the prefix D- means considering network diffusion simultaneously.

City	ELM-ANN				BP-ANN				nu-SVR				epsilon-SVR				Random Forest			
	P	S	PS	PDS	P	S	PS	PDS	P	S	PS	PDS	P	S	PS	PDS	P	S	PS	PDS
Beijing	.085	.096	**.065**	.065	.086	.093	**.065**	.068	.074	.099	.069	.068	.086	.100	.080	.074	.095	.099	.078	.077
Tianjin	.091	.107	**.075**	.070	.089	.105	.077	.072	.081	.110	.078	.075	.085	.110	.081	.073	.092	.109	.082	.079
Shijiazhuang	.091	.105	.072	.081	.089	.103	**.064**	.084	.093	.112	.081	.086	.084	.113	.078	.094	.088	. 110	.077	.075
Shanghai	.092	.096	**.068**	.063	.091	.098	.069	.060	.089	.107	.078	.071	.081	.108	.075	.078	.089	.099	.070	.066
Hangzhou	.102	.117	**.087**	.074	.102	.118	.105	.075	.122	.120	.108	.078	.119	.118	.112	.080	.105	.103	.092	.081
Nanjing	.104	.086	**.077**	.063	.103	.086	.081	.074	.094	.087	.085	.075	.097	.084	.083	.079	.092	.095	.079	.073
Wuhan	.101	.125	.084	.074	.100	.122	.085	.076	.102	.121	.086	.077	.104	.125	.088	.077	.099	. 119	**.083**	.079
Guangzhou	.099	.096	.088	.079	.096	.100	**.087**	.079	.118	.125	.102	.087	.112	.125	.106	.080	.101	.105	**.087**	.080
Average	.096	.103	**.077**	.071	.095	.103	. 079	.073	.097	.110	.086	.077	. 096	.110	.088	.079	. 095	.104	.081	.076

Underline means the best PDS item in each line, and bold font means the best PS item in each line.

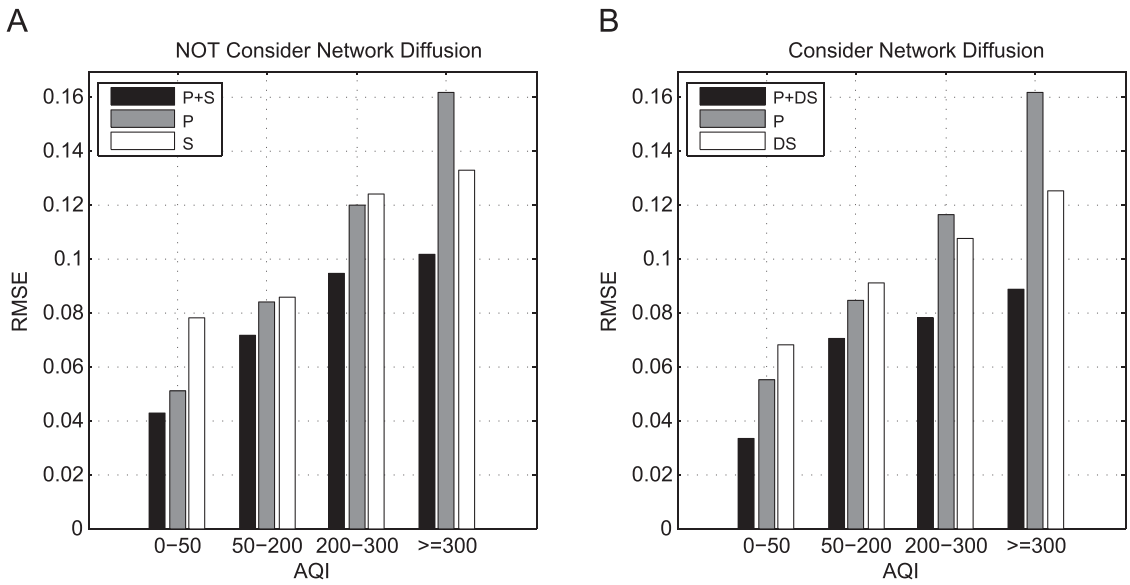

Fig. 5. RMSEs under different AQI ranges with physical sensor features (P) and social media features (S/DS).

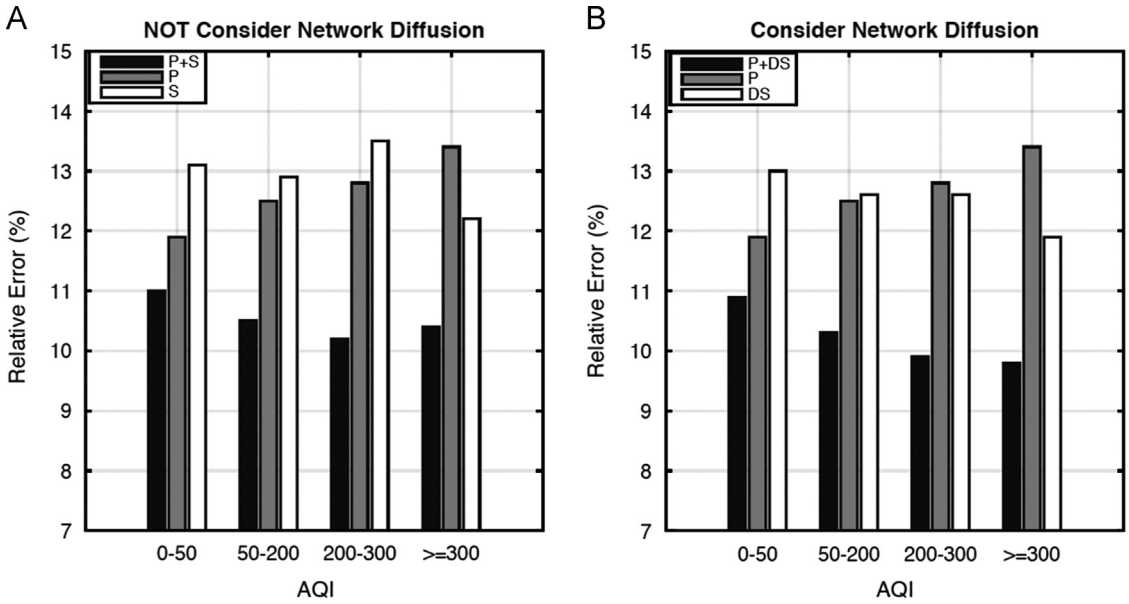

Fig. 6. Relative errors under different AQI ranges with physical sensor features (P) and social media features (S/DS).

people posted 14.5% more tweets about smog disasters or smog-related health hazard in days when AQI exceeds 200 than in days when AQI is less than 200. Additionally, we can also find that social network diffusion can further improve the accuracy of social media features for polluted days. It will be discussed in details in the next subsection.

4.4. Network diffusion factor

In our predictive analytics framework, social network diffusion is considered to further improve prediction accuracy. As shown in Table 2, the RMSEs with the diffusion factor considered (P+DS) are smaller than those without the diffusion factor considered (P+S) for most cities. Actually, the decrement of average RMSE brought by the diffusion factor ranges from 7.6% to 10.5% when different training approaches are adopted.

Figs. 5 and 6 help explain why social network diffusion can improve prediction accuracy. First, the RMSEs of DS are less than those of S in the AQI ranges of 0–50, 200–300 and ≥ 300, and the relative errors of DS are less than those of S in all the four AQI ranges. Second, in both figures, when compared with P, DS outperforms P in two AQI ranges (200–300 and ≥ 300), while S outperforms P in only one AQI range (≥ 300). This may be because considering retweets and likes enlarges the signal of severe pollution and health hazard, thus

reducing the noise in learning the nonlinear relationships. Especially, the number of retweets and likes to smog-related tweets becomes larger in extreme weather days when it is either severely polluted or has good air quality. It is confirmed by our statistical analysis to the retweet and like records. When compared with the + when AQI ranges from 50 to 200, the days when AQI exceeds 300 have 38% more highly retweeted or liked tweets (more than 40 retweets or likes), and the days when AQI is less than 50 have 12% more highly retweeted or liked tweets.

4.5. Comparison between ANNs and other methods

The accuracies of the health hazard prediction models using ANNs, nu-SVR, epsilon-SVR and random forest, are shown in Table 2. We can find that two ANNs' methods outperform the SVM regression methods and the random forest regression method in forecasting the next day PHI, and the single hidden layer ANNs trained by ELM achieve slightly higher prediction accuracy than the multiple hidden layers ANNs trained by BP.

In detail, when only physical sensor features or social sensor features are inputted (P columns and S columns in Table 2), ANNs (ELM-ANN and BP-ANN) achieve very similar performance as the other three methods, especially the random forest regression method. However, when both kinds of features are jointly inputted (P+S columns in Table 2), ELM-ANN outperforms nu-SVR and epsilon-SVR for all eight cities, and outperforms Random Forest for six cities. The average RMSE is about 10.5% smaller than that of nu-SVR, 12.5% smaller than that of epsilon-SVR and 5.2% smaller than that of Random Forest. Meanwhile, when the network diffusion is considered (P+DS columns in Table 2), we can get similar comparison results. On the other hand, we can find that another ANN model trained by BP achieves similar accuracy as that trained by ELM, which confirms that ANNs are suitable to in our application of forecasting smog-related health hazard.

4.6. Real world case studies

According to the above evaluation, our forecasting approach with single hidden layer ANNs, ELM algorithm and both types of features achieves the highest forecasting accuracy. We adopt such settings for the evaluation of our approach in real world cases. In this part, the forecasting performance during two big smog disasters in Beijing and Shanghai are presented. The forecasted PHI records and the measured PHI records during the two big smog disasters are shown in Fig. 7. The trends of the forecasted PHI (P+S) is basically consistent with that of the measured value (Target), and the consistency becomes even higher when the network diffusion is considered (P+DS). The results indicate that this method can indeed work for real-world situations.

5. Conclusion and future work

In this study, we propose a predictive analytics framework for smog-related health hazard forecasting using information from both social media and physical sensors, which is helpful for smog analysis but not investigated. In this framework, we first propose a new method for smog-related health hazard measurement based on individuals' smog and health related comments, as well as their diffusions on social media. Next, we develop a prediction model that utilizes ANNs to learn the non-linear relationships between the current physical and social smog observations and the next day smog-related health hazard.

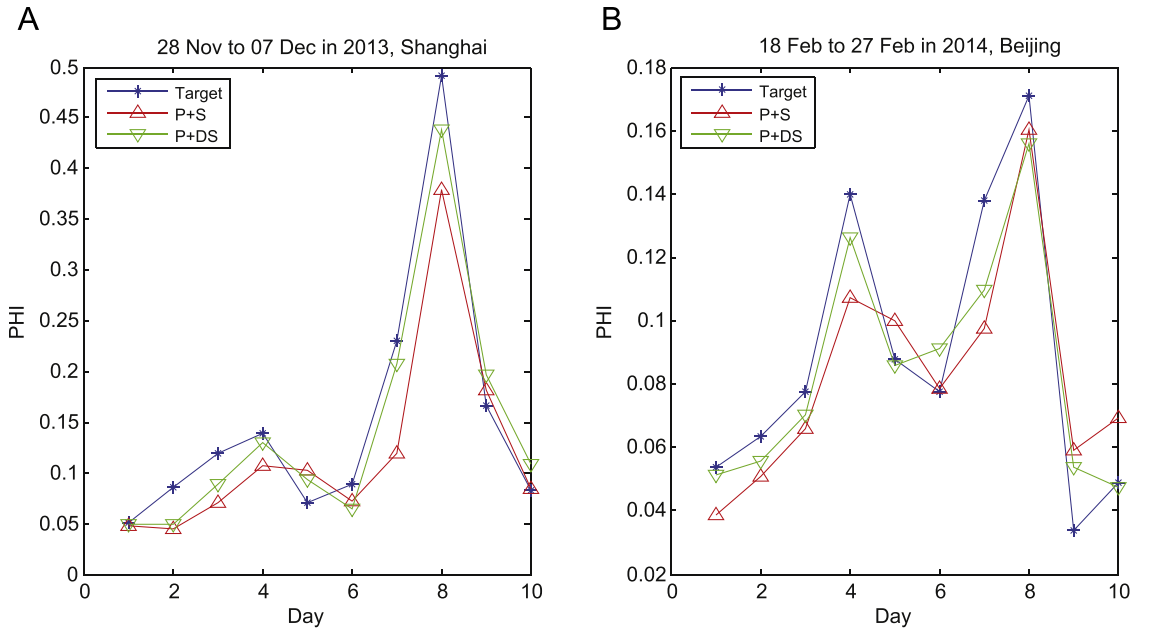

Fig. 7. Results of smog-related health hazard (PHI) forecasting for Beijing and Shanghai.

The evaluation results indicate that the performance of ANN together with both social media and physical sensor features is the best among all candidates that we used in the experiments. We also find that social media features provide more predictive information than physical sensor features under the situations when the smog disaster is severe. Moreover, such benefit from social media data will be enlarged if we further consider information diffusion on social network.

The study also contains some limitations which should be studied on in the future work. One major limitation lies in using social media data for health hazard observations as each of its steps may bring in some errors. For example, we use keywords to find health-related or smog-related tweets, but actually people may use the same keywords in different contexts to mean different things. The percentage of correct tweets after filtering by a keyword mostly ranges from 85% to 95% [24]. We can find some other public health information such as hospital visit records as supplements.

Meanwhile, the adaptivity of the approach for real world circumstances will also be considered in our future work. On one hand, some visual analytics [36] functions will be added into our on going demo system. Through presenting some similar historical circumstances or forecasting results by different features, the system can provide more information for flexible decision making. On the other hand, a new prediction model that utilizes the data from consistent historical circumstances by understanding the underlying semantic of the data is being investigated.

Acknowledgments

This work is funded by projects of NSFC61070156, YB20-13120143 of Huawei and Fundamental Research Funds for the Central Universities, and LY13F020005 of NSF of Zhejiang.

References

[1] China will establish network to monitor smog's health effects, Southern Weekly, ⟨http://www.infzm.com/content/95493⟩, October 2013 (in Chinese).
[2] V. Hughes, Public health: where there's smoke, Nature 489 (7417) (2012) S18–S20.
[3] D.L. Davis, A look back at the London smog of 1952 and the half century since, Environ. Health Perspect. 110 (12) (2002) A734–A735.
[4] J. Yin, A. Lampert, M. Cameron, B. Robinson, R. Power, Using social media to enhance emergency situation awareness, IEEE Intell. Syst. 27 (6) (2012) 52–59.
[5] S. Vieweg, A.L. Hughes, K. Starbird, 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, ACM, New York, NY, USA, 2010, pp. 1079–1088.
[6] K. Lee, A. Agrawal, A.N. Choudhary, Real-time disease surveillance using Twitter data: demonstration on flu and cancer, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13, ACM, New York, NY, USA, 2013, pp. 1474–1477.
[7] Y. Yuan, S. Liu, R. Castro, X. Pan, PM2.5 monitoring and mitigation in the cities of China, Environ. Sci. Technol. 46 (7) (2012) 3627–3628.
[8] P.H. Merz, L.J. Painter, P.R. Ryason, Aerometric data analysis time series analysis and forecast and an atmospheric smog diagram, Atmos. Environ. 6 (5) (1972) 319–342. (1967).
[9] L.S. Casado, S. Rouhani, C.A. Cardelino, A.J. Ferrier, Geostatistical analysis and visualization of hourly ozone data, Atmos. Environ. 28 (12) (1994) 2105–2118.
[10] R. Van Liere, J.J. Van Wijk, Steering smog prediction, in: High-Performance Computing and Networking, Springer, London, 1997, pp. 241–252.
[11] C.A. Pope III, D.W. Dockery, Health effects of fine particulate air pollution: lines that connect, J. Air Waste Manag. Assoc. 56 (6) (2006) 709–742.
[12] H.L. Motley, R.H. Smart, C.I. Leftwich, Effect of polluted Los Angeles air (smog) on lung volume measurements, J. Am. Med. Assoc. 171 (11) (1959) 1469–1477.
[13] R. Chen, Z. Zhao, H. Kan, Heavy smog and hospital visits in Beijing, China, Am. J. Respir. Crit. Care Med. 188 (9) (2013) 1170–1171.
[14] J. Schwartz, Air pollution and daily mortality: a review and meta analysis, Environ. Res. 64 (1) (1994) 36–52.
[15] A. Culotta, Towards detecting influenza epidemics by analyzing Twitter messages, in: Proceedings of the First Workshop on Social Media Analytics, SOMA '10, ACM, New York, NY, USA, 2010, pp. 115–122.
[16] M.J. Paul, M. Dredze, You are what you Tweet: analyzing Twitter for public health, in: Fifth International AAAI Conference on Weblogs and Social Media (ICWSM), AAAI Publications, Barcelona, Spain, 2011, pp. 265–272.
[17] M.J. Paul, M. Dredze, A model for mining public health topics from Twitter, Technical Report, Johns Hopkins University, 2011.
[18] J.A. Greene, N.K. Choudhry, E. Kilabuk, W.H. Shrank, Online social networking by patients with diabetes: a qualitative evaluation of communication with Facebook, J. Gen. Intern. Med. 26 (3) (2011) 287–292.
[19] J.L. Gardy, J.C. Johnston, S.J.H. Sui, V.J. Cook, L. Shah, E. Brodkin, S. Rempel, R. Moore, Y. Zhao, R. Holt, et al., Whole-genome sequencing and social-network analysis of a tuberculosis outbreak, New Engl. J. Med. 364 (8) (2011) 730–739.
[20] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes Twitter users: real-time event detection by social sensors, in: Proceedings of the 19th International Conference on World Wide Web, WWW '10, ACM, New York, NY, USA, 2010, pp. 851–860.
[21] T. Sakaki, M. Okazaki, Y. Matsuo, Tweet analysis for real-time event detection and earthquake reporting system development, IEEE Trans. Knowl. Data Eng. 25 (4) (2013) 919–931.
[22] A. Kongthon, C. Haruechaiyasak, J. Pailai, S. Kongyoung, The role of Twitter during a natural disaster: case study of 2011 Thai flood, in: 2012 Proceedings of PICMET '12 on Technology Management for Emerging Technologies (PICMET), IEEE, 2012, pp. 2227–2232.
[23] S. Mei, H. Li, J. Fan, X. Zhu, C.R. Dyer, Inferring air pollution by sniffing social media, in: 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2014, pp. 534–539.
[24] J. Chen, H. Chen, G. Zheng, J.Z. Pan, H. Wu, N. Zhang, Big smog meets web science: smog disaster analysis based on social media and device data on the web, in: Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion, WWW Companion '14, ACM, New York, NY, USA, 2014, pp. 505–510.
[25] Y. Zhou, J. Chen, H. Chen, Observing social web for smog disaster forecasting, in: Proceedings of the 2015 ACM Conference on Web Science, WebSci '15, in press.
[26] A. Landi, P. Piaggi, M. Laurino, D. Menicucci, Artificial neural networks for nonlinear regression and classification, in: 2010 10th International Conference on Intelligent Systems Design and Applications (ISDA), 2010, pp. 115–120.
[27] A.A. El-Solh, C.-B. Hsiao, S. Goodnough, J. Serghani, B.J. Grant, Predicting active pulmonary tuberculosis using an artificial neural network, Chest J. 116 (4) (1999) 968–973.
[28] Y. Bai, Z. Jin, Prediction of SARS epidemic by BP neural networks with online prediction strategy, Chaos, Solitons Fract. 26 (2) (2005) 559–569.
[29] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Netw. 61 (2015) 85–117.
[30] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Cognit. Model. 5 (1988) 3.
[31] G.-B. Huang, L. Chen, C.-K. Siew, Universal approximation using incremental constructive feedforward networks with random hidden nodes, IEEE Trans. Neural Netw. 17 (4) (2006) 879–892.
[32] Z.-L. Sun, T.-M. Choi, K.-F. Au, Y. Yu, Sales forecasting using extreme learning machine with applications in fashion retailing, Decis. Support Syst. 46 (1) (2008) 411–419.
[33] J. Ramos, Using tf-idf to determine word relevance in document queries, in: Proceedings of the First Instructional Conference on Machine Learning, 2003.

[34] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. (TIST) 2 (2011) 27:1–27:27.
[35] J. Chen, H. Chen, J.Z. Pan, M. Wu, N. Zhang, G. Zheng, When big data meets big smog: a big spatio-temporal data framework for China severe smog analysis, in: Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, BigSpatial '13, ACM, New York, NY, USA, 2013, pp. 13–22.
[36] P.C. Wong, J. Thomas, Visual analytics, IEEE Comput. Graph. Appl. 5 (2004) 20–21.