ISBN 978-981-11-3671-9
Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)
Beijing, 25-27 June, 2017, pp. 584-589
doi: 10.18178/wcse.2017.06.101

Social Activities Assist Sensor Network through Multi-modal Information Fusion

Jia Wu [1], Zhe Yu [1, 2], Zhaocheng Su [1], Lingke Zhang [1, 2], Miao Li [1, 2] and Peng Miao [1, 2 +]

[1] School of Communication and Information Engineering, Shanghai University, Shanghai China
[2] Institute of Biomedical Engineering, Shanghai University, Shanghai China

Abstract. In China, environment issue has been attached greater importance, especially the management and protection in air quality. There are already sensor networks existing and locating in the major cities for PM2.5 monitoring. These sensor networks provide the real-time data of PM2.5 at different locations which are further used to publish and predict the future air quality. In this study, we propose a new frame work to fuse the multi-modal data sources, *i.e.* data stream from PM2.5 sensor networks and the real-time social activities from WeChat platform, for air quality assessment in real-time. Both semantic analysis and 2D interpolation methods are used in the multi-modal fusion procedure. As an application, we construct the cloud data server based on SinaAppEngine (SAE) and develop the service system using PHP and MySQL. Based on the new system, social activities were automatically transferred into quantitative evaluations of local air quality. Then using the GPS information of social activities, the evaluated air quality was fused with the sensor data and further interpolated on the map. The accuracy and robustness of our system were tested based on the data in the university campus. A one-month continuous monitoring was also performed and the detailed analysis demonstrated the monitoring performance of the new system. The tools and methods proposed in this study show more potentials in environmental monitoring, policy decision making and sociological research.

Keywords: social activities, sensor network, data fusion, PM2.5

1. Introduction

China is undergoing a period of rapid economic development, but there are many factors, such as coal power generation, industrial production, vehicle emissions and so on, leading to frequent domestic urban haze weather, which makes the index of PM2.5 become the hot topic domestically. Air pollution is not only an environmental problem, but also closely related to economic development, health care, haze management, so it is a complex social problem, and there are many difficulties in finding a solution.

At present, government and other organizations provide monitoring and prediction of index of PM2.5. Usually, sensor/equipment networks are constructed to measure the index of PM2.5 at different locations. Kinds of models are applied in the prediction of changes of air quality. However, current sensor networks cannot provide the monitoring of PM2.5 with full coverage on the map. People usually pay more attention to the air quality of their local area, including the living community, the work and study place, *etc*. For extreme events of air quality, the public needs to know the accurate changes of PM2.5 in real time, and they always give their self-evaluations of air quality in texts on the social platform like WeChat. Therefore, if the self-evaluated data from social network can be identified and matched to the sensor data, we can retrieve the detailed information of air quality at places without sensors [1].

This study combines PM2.5 sensor network, WeChat social platform and cloud server to achieve multi-mode data fusion of PM2.5 sensor data and social platform data. It can also provide the users with a full map

[+] Corresponding author. Tel.: +862166137233; fax: +862166137233.
E-mail address: pengmiao@shu.edu.cn.

coverage and bidirectional data feedback [2]. Using the proposed system, the real-time air quality map is dynamically created and updated with the users' feelings and comments in WeChat. The system provides a new tool and platform for the environmental monitoring. The social activities database also benefits the study of the social impact of air quality relating to policy decision and related research.

2. Multi-modal Data Fusion System for Environment Monitoring

The system architecture is shown in Fig.1. This system consists of sensor networks, cloud services and WeChat public App. Sensor networks acquire the quantitative measurements of environmental factors like PM2.5 concentrations, temperature, humidity, and other information. 4G/wifi/NBIoT networks provide the data acquisition and device control to form a robust sensor networks and feed the cloud server with real-time measuring data [3].

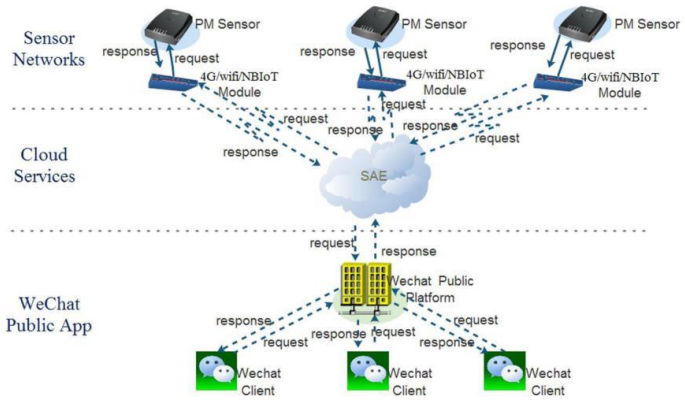

Fig. 1: System architecture.

The cloud platform service is to store and process data from the sensor networks and social platform. It outputs a fully-covered PM2.5 map to the WeChat public App. The WeChat public App gathers the public participant's description of local air quality and send all the text data to the cloud service in real-time. Using the semantic analysis and 2D interpolation method, both types of data are fused and finally visualizes and interacts with public activities.

3. Sensor Networks

In this study, we developed the sensor networks with both abilities of PM2.5 measurement and wireless communication. The network is composed of PM2.5 measuring module, wireless communication module and power management module. Fig.2 is the hardware design of the sensor. The power management module uses 18V-25W polysilicon solar panels, which can store the power in the battery to solve the outdoor distribution of the sensor. The acquisition of sensor data is controlled by the STM32 and the processing algorithm is programmed into the microcontroller before its working. The self-organizing network module can automatically switch WIFI mode or 4G mode or NBIoT mode based on the network environment. This design of wireless communications provides a robust access of bidirectional communications including data upload to and control request from the cloud database in real time [4].

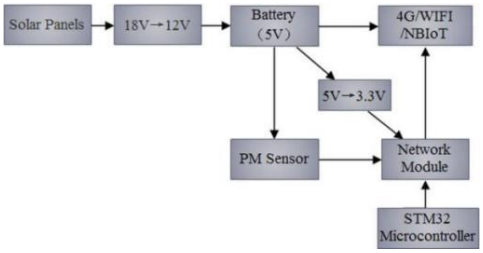

Fig. 2: The hardware design of sensor.

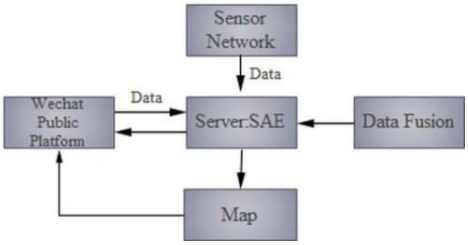

Fig. 3: System design of cloud server.

4. WeChat Public App Design

We have developed a WeChat public App as a main social platform to gather the text from individuals who express the feeling and concerns of local air quality. Each text data also contains the GPS location information. All the users who feed the system their text data can see the dynamic map of current air quality in the WeChat public App. The App also connected with the cloud service through the communication modes provided by the mobile phone.

In this study, we use PHP to develop the public App. The main development includes: (1) set the URL and Token (for authentication), (2) set the URL points to the cloud server for storing the collected data, and (3) the public platform server requires access to the developer computer to enable normal reception and transmission of data.

For the security consideration, the connection request of the WeChat platform is set to use the HTTP GET parameter, and there are three parameters: signature, timestamp, and random number. The signature is the character string after the timestamp, nonce and token SHA1 encrypted. After receiving the public platform server request, the cloud server uses the same encryption algorithm to sign and compare the signature of the public platform server to exclude the malicious third party connection.

When the WeChat clients send messages to our public App, the text messages will be send to the cloud server in the XML format, including sender's ID, message type, requesting information, text data, GPS data and so on. The text data are further processed using the semantic analysis algorithm and fused into the current data map. If there is valid requesting information from the individual user, after the server parses the XML, it can read the request, get the corresponding data from the local MySQL database, and response it back to the individual user through the public App in XML format [5]. The individual user can receive the corresponding data or information based on their request.

5. Cloud Server

5.1. Data Processing on the Cloud Server

Based on Sina cloud server (Figure 3), we create programs and databases using the shared MySQL. The data processing on the cloud server mainly includes four parts: (1) The measurement data from sensor networks and the text data from the social network are saved into the SAE database in real-time. (2) Semantic analysis of the text data about air quality are done and quantitative evaluations are obtained. (3) Both kinds of data are fused on the map and 2D interpolation is performed to form the full map information. (4) WeChat client request of full map information should be responded in real-time.

To simplify the calculation of air quality map, we level the quantitative data from both sensor network and social network into six grades based on the "Ambient Air Quality Index (AQI) Technical Regulation (Trial)" (HJ633-2012). AQI is divided into six grades and converted to PM2.5 concentrations, as shown in Table 1. Therefore, a colour map can be constructed as a demonstration of air quality map. So, in the response of individual request from the WeChat user, only the changed values will be updated. Furthermore, in the 2D interpolation, the nearest neighbour strategy is applied to reduce the calculation load. The individual WeChat user can also select the single sensor and view the measured index curve of the PM2.5 data by clicking on the sensor tag on the map.

5.2. Semantic Analysis of the Text Data

Fig.4 shows the semantic analysis procedure to obtain the quantitative air quality evaluation based on the text data from the WeChat public platform. In this processing, feature extraction is performed using the conditional random field CRF model to access two sets of feature data. After that, the key semantic class in the sentence is identified and match to the key semantic class. The we calculate the similarity between the key semantic class and the keyword in the dictionary, select the keyword with the similarity to replace the keyword semantic class, and match the environment descriptor keyword with the PM2.5 concentration values [6].

In our method, CRF graph model of the chain structure is used. The observation sequence is expressed as $W = (w_1, w_2, ...w_n)$ and sequence of the marker (status) $Y = (y_1, y_2, ...y_n)$, which are defined as follows:

$$P_\lambda(Y|W) = \frac{1}{Z(W)} \exp\left(\sum_{t \in T} \sum_{k} \lambda_k f_k(y_{t-1}, y_t, W, t)\right) \tag{1}$$

where f_k is the characteristic function, λ_k is the eigenvalues of the corresponding eigenfunction, t is the marker, and $Z(w)$ is the normalization factor, which makes the probability distribute between 0 and 1.

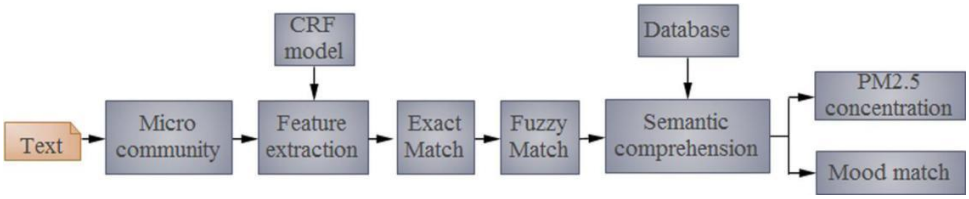

Fig. 4: Diagram of semantic analysis system.

CRF model parameter estimation is usually done using the L-BFGS algorithm. It decodes process of CRF, solves the unknown marked string, and obtains the maximum probability Y^* of joint in the string, namely:

$$Y^* = \arg\max_Y P(Y|W) \tag{2}$$

For the linear chain CRF, the Viterbi method with a training procedure can be applied. The CRF training data set was constructed using both sensor measurements and social text data. In this study, volunteers are required to publish the contents covered as many as possible common spoken language and included a variety of fields used in the system. Then independent observers mark the training data with comparable PM2.5 values. Finally, we established a common language dictionary to map the text descriptions to the PM2.5 value which contains the feature templates. After the feature model is trained and completed, it can be used for processing other text data from the social network.

Table 1: PM2.5 Grades and Categories

AQI	PM2.5（ug/m3）	Level	Category	Color
0-50	0-35	one	excellent	green
51-100	36-75	two	good	yellow
101-150	76-115	three	mild contamination	orange
151-200	116-150	four	moderately polluted	red
201-300	151-250	five	severe pollution	purple
>300	>250	six	serious pollution	brown red

The above procedure runs on the cloud server to extract the characteristics of the text data as the first step. Then the well-trained CRF model is used for air quality evaluation by locating the key semantic class in the sentence. Sometimes, there is no direct match between the text and semantic class, then the Dice similarity is used to find a fuzzy match. The Dice similarity is defined as follows:

$$Sim(A,B)_{Dice} = \frac{2.|A \cap B|}{|A| + |B|} \tag{3}$$

where A, B denote two vocabulary set, the symbol || and $\cap$ means the set's length and intersection of sets respectively. When the maximum value of $Sim(A,B)_{Dice}$ is found, the semantic class fuzzy matching is completed.

5.3. Missing Data Processing

In practice, it is necessary to solve the problem that the sensor network may provide incomplete data set because of the wireless communication and out of power. In this study, we use the cubic B-spline interpolation method to obtain the missing data [7]. The cubic B-spline interpolation method has a second order continuous derivative, and its curve has second order smoothness. Usually, the error of the cubic B-spline interpolation method is small enough for our applications. In the cubic B-spline interpolation, the permutation point $a = x_0 < x_1 < \ldots < x_n = b$ and its function $f(x)$ corresponding to the value of y_i are given on interval $[a,b]$. The function $S(x)$ satisfies: (1) $S(x)$ is a cubic polynomial in each

subinterval $[x_j, x_{j+1}](j=0,1,2,...,n-1)$. (2) $S(x)$ is the cubic derivative of the node $x_0 < x_1 < ... < x_n$, at each inner node $x_j(j=0,1,2,...,n-1)$ with a continuous derivative up to the second order, namely, $S(x) \in C^2[a,b]$. If $S(x) = y_i(j=0,1,2,...,n-1)$ and satisfies (1), (2), then $S(x)$ is the cubic spline function.

6. Experiments and Results

6.1. Accuracy of Sensor Network

Compare the data from our sensor network located in our university campus with the data of the Shanghai downtown area, there are consistent errors between 6% and 9% (Table 2). This bias came from the distance between the two measuring locations [8].

Table 2: Measured Data from Two Sensor Networks

Time	Downtown area/(μg/m^3)	University campus/(μg/m^3)	Errors
2016/4/12 14:00	120	109.41	-8%
2016/4/19 01:00	89	80.22	-9%
2016/5/20 20:00	96	89.64	-6%
2016/5/22 14:00	147	136.22	-9%
2016/5/22 15:00	152	142.91	-6%

6.2. Monitoring the University Campus Air Quality

Regional monitoring experiment in Shanghai University Baoshan campus. There are 100 volunteers in the Baoshan campus of Shanghai University sending their texts everyday through our WeChat public App for one month. The texts are about weather conditions and mood in the micro-community. The data of first week were used to establish a feature model and match the PM2.5 concentration value. The last three weeks, the feature model was used to obtain the evaluations of air quality.

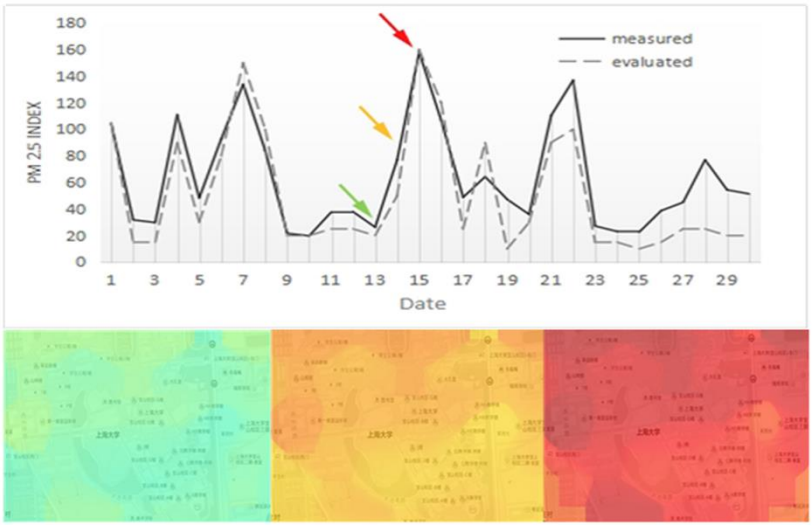

Fig. 5: (a) PM2.5 measured (solid curve) and evaluation (dotted curve) results based on the sensor network and semantic analysis in December. (b, c, d) The merged maps of three days selected (green, yellow and red arrows in (a)).

Figure 5 (a) showed the measured data of the PM2.5 index from one sensor (located in Xiangying Building) in December. The corresponding evaluation results based on the semantic analysis of text data was also shown in this figure. Compared with the measured data, we found that the semantic evaluation of the PM2.5 index provided similar results as the sensor network monitoring. But sometimes the keywords posted by text data can be affected by extreme PM2.5 event. For instance, December 20, 2016, there was prediction alarm from the news that the country ushered in the most serious haze. But the actual air quality is very good. We found that there were many "good" and other similar keywords, after semantic analysis of the text. Therefore, the evaluation of PM2.5 index from the social network is lower than the measured data from the sensor network. December 26, Shanghai issued a blue low temperature alarm, and then semantic analysis

gave the PM2.5 index smaller. December 19 news released that heavy haze spread 12 provinces and cities, that day the semantic analysis gave the PM2.5 index significantly higher than the actual situation. Figure 5 (b, c, d) showed three merged maps corresponding to the days selected in December data (excellent, good and moderately polluted air quality respectively).

7. Discussions and Conclusions

In this study, social networks for the first time work together with the sensor networks to provide a new environmental monitoring strategy. Based on the cloud server and WeChat platform, the multi-modal data fusion of the air quality measurement and the text data from social network are merged to form a full cover map of air quality. Our study provides a new platform for investigation the interactions between environmental monitoring and social community response. The tools and methods developed here show more potentials in environmental monitoring, policy decision making and sociological research. Different data analysis and prediction methods can be included into our system to provide more useful information.

8. Acknowledgements

This study is supported by Shanghai Science and Technology Committee (STCSM) (16511105502).

9. References

[1] T. Huy N, L. Ketsiri, and M. Nicole. A tool for public PM 2.5-concentration advisory based on mobile measurements. *Journal of Environmental Protection* 2012, 3(12):1671-1688.

[2] Y. Li, Q. Chen, H. Zhao, L. Wang and R. Tao. Variations in PM10, PM2.5 and PM1.0 in an urban area of the Sichuan Basin and their relation to meteorological factors. *Atmosphere* 2015, 6(1):150-163.

[3] Z. Yang, L. Zhu, H. Ding, and Z. Guan. A Priority-based Parallel Schedule Polling MAC for Wireless Sensor Networks. *Journal of Communications* 2016, 11(8):792-797.

[4] S. M. Oteafy, and H. S. Hassanein. Resource Re-use in Wireless Sensor Networks: Realizing a Synergetic Internet of Things. *Journal of Communications* 2012,7(7):484-493.

[5] T. Hoffman. Probabilistic latent semantic analysis. *Proceedings of the fifteenth conference on uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc.1999, 25(4):289--296.

[6] T. Hoffman. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning* 2001, 42 (1-2): 177-196.

[7] P.W. Foltz. Latent semantic analysis for text-based research. *Behavior research methods, instruments, & computers* 1996, 28(2):197-202.

[8] H. Wang, Y. Zhuang, Y. Wang, Y. Su, H. Yuan, G. Zhuang, Z. Hao. Long-term monitoring and source apportionment of PM2.5/PM10 in Beijing, China. *Journal of environmental sciences* 2008, 20(11):1323-1327.