Big data platform for air quality analysis and prediction

Yue Shan Chang
Dept. of CSIE
National Taipei University
New Taipei City, Taiwan
ysc@mail.ntpu.edu.tw

Kuan-Ming Lin
Dept. of CSIE
National Taipei University
New Taipei City, Taiwan
ra8822@gmail.com

Yi-Ting Tsai
Dept. of CSIE
National Taipei University
New Taipei City, Taiwan
smallxiami790612@gmail.com

Yu-Ren Zeng
Dept. of CSIE
National Taipei University
New Taipei City, Taiwan
gary2335854@gmail.com

Cheng-Xiang Hung
Dept. of CSIE
National Taipei University
New Taipei City, Taiwan
guitar10833@gmail.com

***Abstract*—With the advance of industry, air quality (AQ) is increasingly becoming worse. There are increasingly AQ monitors device have been deployed around country for monitoring air-quality all year long. To estimate and predict AQ, such as PM (particulate matter) 2.5, become an important issue for government to improve people's quality of life. As we can know, there are many factors can affect the AQ, such as traffic, factory exhaust emissions, weather, incineration of garbage, and so on. In most well-developed countries, these pollution sources are monitored for future environmental policy making. In this paper, we will propose a semantic ETL (Extract-Transform-Load) framework on cloud platform for AQ prediction. In the platform, we exploit ontology to concretize the relationship of PM 2.5 from various data sources and to merge those data with the same concept but different naming into the unified database. We implement the ETL framework on the cloud platform, which includes computing nodes and storage nodes. The computing nodes are used to execute data mining algorithms for predicting, and storage modes are used to store retrieved, preprocessed, and analyzed data. We utilize restful web service as the front end API to retrieve analyzed data, and finally we exploit browser to show the visualized result to demonstrate the estimation and prediction. It shows that the big data access framework on the cloud platform can work well for air quality analysis.**

Keywords—Air Quality, Big Data, Prediction, Cloud Environment,

I. SEMANTIC ETL FRAMEWORK

This section presents our proposed semantic ETL framework for accessing AQ required data, as shown in Fig. 1. The proposed framework is referred to [1]. For the extract layer, some software crawlers [2] have been implemented to automatically retrieve open data, such as weather, traffic, air quality, factory exhaust emission, and so on. The crawlers are periodically retrieving the data according to relevant open data generating speed. And then store the retrieved data into our database.

The core of the ETL framework is to transform retrieved data into required instance and format. In this layer, there are three parts: Data Preprocessing, Semantic Data Model, and Semantic Data Analytics, respectively. The Data Preprocessing is responsible for data normalizing and cleaning, removing duplication, filtering and grouping from the retrieved data. After the retrieved data is processed, we then exploit ontology methodology to construct the semantic data model. So that all processed data can be having meaningful relation between them. It is useful for integrating multimodal data sources into target domain data source. Therefore, users can easily inquire desired information based on the semantic model. This part includes ontology creation, mapping of data fields, and alignment of similar data. After data model is created, the data analytics is used for analyzing and mining air quality based on the semantic data model. The analyzing and mining algorithms can be implemented using some well-known algorithms in Apache Spark's[1] big data framework.

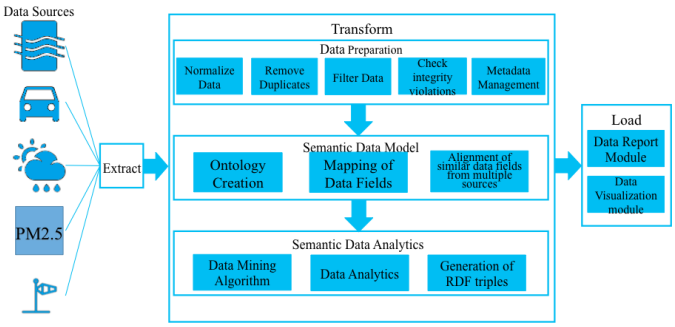


Fig. 1. Semantic ETL framework for AQ analysis

The third layer is Load Module that is used for reporting and visualizing the analytical result of AQ. In the layer, we can use restful web service as the API (Application Program Interface) to retrieve the analyzed result and show it onto a browser.

[1] https://spark.apache.org

978-1-5386-4959-6/18/$31.00 ©2018 IEEE

II. PROTOTYPING RESULT

In order to make comparison with retrieved open data, we refer the work [3] to build a LoRa-based PM2.5 monitor to collect local reading. In addition, because most of open data retrieved from government are formatted as CSV format, in order to easily retrieve, store, and preprocess these data, MongoDB is used as data repository to store all data and Hadoop file system with Spark framework is used as computing environment to accelerate system performance of data mining and analytics. We demonstrate an example to retrieve PM2.5 value of a certain point (latitude is 25.0129720000, longitude is 121.4586670000) from MongoDB, the query string is as following:

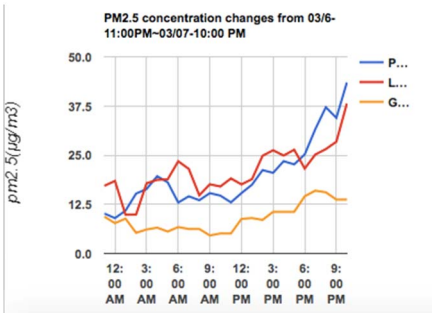
```
db.getCollection('aqi_map_pm25').find({"Lat" :
"25.0129720000", "Lng" :
"121.4586670000","Time":{"$gte":"2018-03-05
15:00"}}).sort({'$natural':-1})
```

The query results may have multiple records, following shows an example, that is JSON format. The result shows the query result of PM2.5 at the monitoring site located at the point (latitude is 25.0129720000, longitude is 121.4586670000).

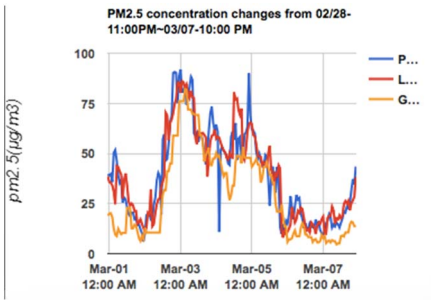
```
{
   "_id" : ObjectId("5a9d1d4423baca0a66ac48a9"),
   "SiteName" : "板橋",
   "Time" : "2018-03-05 18:00",
   "Lat" : "25.0129720000",
   "Lng" : "121.4586670000",
   "PM25" : "25"
}
```

After data retrieved from governmental and nongovernmental open data of AQI (Air Quality Index), these data are necessary to be cleaned and filtered by preprocessing module, and then store into backend database. After preprocessing, the original data can be showed in web browser for observing. Therefore, we can make a comparison between all retrieved data, as shown in Fig. 2. Fig. 2(a) shows three PM2.5 value today collected from our PM2.5 sensing device, nongovernmental open data and governmental open data. The blue curve shows local reading collected by our deployed LoRa[2]-based PM2.5 sensor, the red curve is the reading of nongovernmental data, and the orange curve is the reading of governmental data. Both nongovernmental and governmental data are inferred from three nearest neighborhood monitor sites by IDW (Inverse Distance Weighting) [4]. Fig. 2(b) shows one week's comparison. From the figure we can see that the government's data show smaller results. There may be two main reasons. One is that the government's data is more accurate and has better accuracy. Second, government data are collected from relatively distant sensing points and then calculated through IDW calculations. Therefore, the value may decrease. Based on the platform, a long-term analysis and prediction can be made.

In the work, we exploit various statistical and machine learning methods to construct the prediction model for predicting the PM2.5 value. First a decision tree, which is a tree-like graph model of decisions, is used to predict next possible value. We run the experiments on Spark[3] environment, which is a well-known big data processing framework, and exploiting spark MLlib to execute the classification and prediction. We use the real PM 2.5 data collected from nongovernmental data at San Shia District, New Taipei city as training and testing data. In addition, we also collected weather-related data from Central Weather Bureau[4] as the attributes of classification, such as temperature, wind direction, wind speed, rainfall, traffic flow, humidity, atmospheric pressure, time, and so on. The accuracy of one-hour prediction to predict next hour value is around 81.5%. In addition, we also conduct experiments run on TensorFlow[5] framework using Recurrent Neural Networks (RNN) with long short-term memory model [5] to predict next hour PM2.5 value. The average RMSE (Root-Mean-Square Error) is around 8.6. Because this is a work in progress, we believe that these two kinds of forecast results will be improved after we adjust the parameters.



(a)



(b)

Fig. 2. Big Data platform for computing and storing AQ data

III. CONCLUSION AND FUTURE WORK

We have proposed a semantic ETL (Extract-Transform-Load) framework on cloud platform for AQ analysis and prediction. We exploit ontology to concretize the relationship of PM 2.5 from various data sources and to merge those data with the same concept but different naming into the unified database. The computing nodes are used to execute data mining algorithms for predicting AQ, and storage modes are used to store retrieved, preprocessed, and analyzed data. We exploit

[2] https://www.lora-alliance.org

[3] https://spark.apache.org

[4] https://www.cwb.gov.tw/V7/

[5] https://www.tensorflow.org

browser to show the visualized result to demonstrate the estimation and prediction. It shows that the big data access framework on the cloud platform can work well for air quality analysis. In the future, we will conduct more experiments using a variety of classification methods and othes DNN model to imrove the prediction accuracy.

ACKNOWLEDGMENT

This work was partially supported by Ministry of Science and Technology of Taiwan, Republic of China under Grant No. MOST 106-3114-M-305 -001 –A and by National Taipei University under Grant No. 106-NTPU_A-H&E-143-001 and 107-NTPU_A-H&E-143-001.

REFERENCES

[1] Srividya K. Bansal, Sebastian Kagemann, “Integrating Big Data: A Semantic Extract- Transform-Load Framework,” Computer, Vol. 48, No. 3, pp. 42-50, Mar. 2015

[2] Gautam Pant, Filippo Menczer, “MySpiders: Evolve Your Own Intelligent Web Crawlers,” Autonomous Agents and Multi-Agent Systems, Vol. 5, No. 2, pp 221–229, June 2002.

[3] Yung-Sheng Lin, Yu-Hsiang Chang, Yue-Shan Chang, “Constructing PM2.5 Map Based on Mobile PM2.5 Sensor and Cloud Platform,” 2016 IEEE International Conference on Computer and Information Technology (CIT), 8-10 Dec. 2016, pp. 702-707.

[4] Feng-Wen Chen, Chen-Wuing Liu, “Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan,” Paddy and Water Environment, September 2012, Vol. 10, No. 3, pp 209–222.

[5] Hakkani-Tür, D., Tur, G., Celikyilmaz, A., Chen, Y., Gao, J., Deng, L., Wang, Y. (2016) Multi-Domain Joint Semantic Frame Parsing Using Bi-Directional RNN-LSTM. Proc. Interspeech 2016, 715-719.