Data Mining and Visualization: Meteorological Parameters and Gas Concentration Use Case

© Yas A. Alsultanny

Arabian Gulf University
Manama, Kingdom of Bahrain

alsultanny@hotmail.com

Abstract. Knowledge extraction from big data is one of the important subjects now and in future. Mining in the big data needs many steps, which must be implemented very carefully. The final step in big data mining is visualizing the results or summarizing the results numerically. This paper aims to mining the big data recorded by environmental station. These stations are recording the concentrations of some gases and meteorological parameters. The 2D and 3D data visualization is used to evaluate the capability of visualization in determining the effect of meteorological parameters on some gases that caused pollution. The results showing the visualization is a very important tool, and visualization can be used in mining big data, by showing the concentrations of gases. The paper recommends using big data visualization periodically as an alarming tool for monitoring the levels of pollution gases concentration.

Keywords: metrological parameters, gases concentration, filtering; preprocessing, decision tree, meteorological parameters.

1 Introduction

Big Data Mining (BDM) and Data Visualization (DV) are two important hot topics in the field of knowledge discovery. The big data can be visualized and analyzed to extract knowledge. The visual analytical tools have steadily improved during the last years in order to work with big data. The data collected from different resources, such as the station for monitoring pollution gases. These stations usually have an hourly readings to measure concentrations of gases such as; ozone O_3, nitrogen dioxide NO_2, sulfur dioxide SO_2, carbon monoxide CO, carbon dioxide CO_2, particulate matter (PM_{10} and $PM_{2.5}$), moreover these stations have hourly readings for meteorological parameters such as; Temperature (Temp), Humidity (Hu), Wind Speed (WS), Wind Direction (WD), and Air Pressure (AP).

Big data is a term used to describe some of current directions in information technology, as a concept that take into consideration data analysis. The amount of data in the world is huge, and it grows in an annual basis of 50% of its original size [1]. It is important to note that most of the big data is unstructured data, where it is not organized and does not fit the usual databases [2]. Big data can be used as a useful tool to enhance decision making [3].

Data Mining is the technique to get useful knowledge out of databases; data mining requires pre-processing and analytic approach for finding the value. Data mining requires many operations such as data integration, data selection, and so on [4].

Visual analytic first defined by Tomas and Cook in 2005 [5] as; the science of analytical reasoning facility by interactive visual interface. Murray in 2013 [6] described Data Visualization as; "fortunately, we humans are intensely visual creatures. Few of us can detect patterns among rows of numbers, but even young children can interpret bar charts, extracting meaning from those numbers' visual representations. Visualizing data is the fastest way to communicate it to others".

Air pollution is important in our life; most of the pollutants in the air are a result of emissions from cars, trucks, buses, factories, refineries, and other sources. The objective of this paper is to highlight the aspects of Big Data miming to visualize air pollution concentrations and it is relative to meteorological parameters.

2 Literature Review

Big data rises with the huge growth of data. It refers to the storing, processing, and analyzing the vast amounts of data. Big data brings new challenges to visualization because of the speed, size and diversity of data. One of the most common definitions of big data is data that have volume, variety, and velocity [7-9]. The term "Big Data" is surrounded by a lot of advertising, where many software vendors claim to have the ability to handle big data with their products [10]. Innovations in hardware technology such as those in network bandwidth, memory, and storage technology have assisted the technology of Big Data. The new innovations coupled with the latent need to analyze the massive unstructured data that stimulated their development [11].

Data Mining is the field of discovering novel and potentially useful information from large amounts of data [12]. Data mining defined as the use of analytical tools to discover knowledge in a database. The analytical tools may include machine learning, statistics, artificial intelligence, and information visualization [13]. Data mining categorized into seven categories as Fayyad et al. in 1996 [14] stated. These categories are regression, clustering, summarization, dependency modeling, link analysis, and sequence analysis. Knowledge Discovery

Proceedings of the XIX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2017), Moscow, Russia, October 10-13, 2017

in Databases (KDD) is the processing steps used to extract useful information from large collections of data [15]. Data mining manly has two methods: classification is assigns items in a collection to target categories or classes, and clustering is a form of unstructured learning method. Decision trees are types of classifications such as: Reduced Error Pruning (REP) tree, K Nearest Neighbors (KNN), the J48 based on C4.5 algorithm, and M5P algorithm is an improvement of the Quinlan's M5 algorithm [16-20].

"To visualize" has two meanings. "To form a mental image of something" refers to a cognitive, internal aspect whereas "to make something visible to the eye" refers to an external, perceptual role [21]. Visualization is any kind of technique to present information [22-23]. Data visualization refers to any graphic representation that can examine or communicate the data in any discipline [24]. The 3D visualization is gradually becoming the main trend in many fields including population gases and meteorological parameters [25].

3 Data Visualization

This study proposes a visualization method to represent graphically air pollution big data, to be an efficient method for knowledge discovery. This visual methodology is useful for people who are working in field of air pollution to have an efficient readability and accuracy of data analysis. Data visualization is the use of computer for visual representations of data. It aims at helping decision maker to detect effectively into big data. Data visualization is an efficient and intuitively accessible approach to identify patterns in large and diverse data sets.

Gases and metrological parameters visualizations can have two goals: Explanatory and Exploratory. Gases and metrological parameters data are usually recorded by automatic stations at regular time intervals. Metrological data is typically multivariate that often consists of many dimensions. Air pollution is a major concern in any city through the world. The visualization technique is used to aid visual analysis of the air pollution problem, followed by metrological data for knowledge discovery.

There are many steps must be taken in order to prepare data for visualization, these steps are shown in Figure 1. The steps are: stations sensors adjustment, data recording, data filtering, data preprocessing, normalization, aggregation, and visualization.

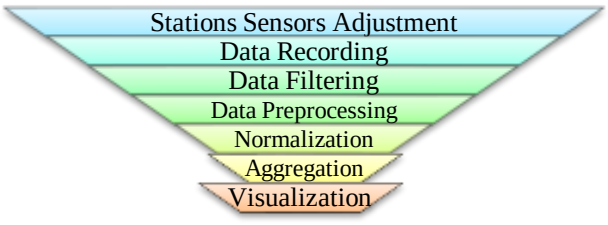

Figure 1 Big data acquisition and utilization

4 Data Collection and Analysis

The data available for this paper were collected from Arabian Gulf countries from one station in state of Kuwait; it was hourly time series data for eleven years, after data filtering and preprocessing, the data for one year 2015 was analyzed in this paper. The data represented on an hourly averaged reading, where the yearly readings for each gas or parameter must be 8,760 (24 hr*365 day), but the real readings after filtering and processing are 8,630, with 130 (1.5%) missed reading. The Rapidminer version 7.5 was used for processing and visualization the data of this paper.

Figure 2 shows the effect of temperature on the concentration of the five gases (O_3, NO_2, CO, CO_2, and SO_2) and PM_{10}. The figure visualizes the data distribution by using two-dimensional diagrams; the temperature has an opposite effect on O_3 and NO_2. The concentration of O_3 increased directly during the hottest hours, when the temperature was above 40ºC. While the temperature had a reverse effect on NO_2, the concentration of this gas became lower during the hottest hours, and its concentration was in its lightest levels, when the temperature was less than 10Cº. The effect of temperature on CO and CO_2 is very limited and this is clear from the figure, this indicates the temperature has no effect on these two gases. The hottest hours have a direct effect on SO_2 and PM_{10}, their concentrations usually increased during summer and especially in the hottest hours of a day.

Figure 3 shows the effect of humidity on the five gases and PM_{10}. The humidity has a reverse effect on O_3 and NO_2, their concentrations are increased with lower concentration of humidity, moreover the concentrations of CO, CO_2, and SO_2 increased with lower percentage of humidity. The PM_{10} concentration significantly reduced, when the humidity percentage was higher than 70%. These results are true, because the highest percentages of humidity, reducing the five gases and PM_{10} disperse.

Figure 4 shows the three dimensions scatter diagrams to visualize the effect of both temperature and humidity at the same time on the five gases and PM10. The figure shows again most of the readings of O3 are concentrated in the region of hottest temperature and low percentage of humidity. The concentrations of NO2 increased at the lowest temperature and humidity. For CO, CO_2, SO_2, and PM_{10} their readings are concentrated in the region of hottest temperature and low percentage of humidity.

A decision tree is a predictive model [26]. It was implemented in this paper to predicate PM_{10}, which is measured in part per million (ppm), by stating the effect of temperature and wind speed. To implement the decision tree the PM_{10}, temperature, and humidity were classified into: 0=0-50, 1=51-150, 2=151-400, 3=401-700, 4=701-1000, 5=1001-1500, 6=1501-2500, 7=2501 and more. The temperature in centigram degree (Cº) classified into: 0=0-6, 1=7-11, 2=12-16, 3=17-21, 4=22-26, 5=27-35,6=36-46, 7=47 and more. The wind speed meter per second (m/s) classified into: 0=0-2, 1=3-5, 2=6-8, 3=9-12, 4=13 and more. The decision rules of the decision tree to predicate PM_{10}, as an example by using temperature and wind speed-readings are as follows.

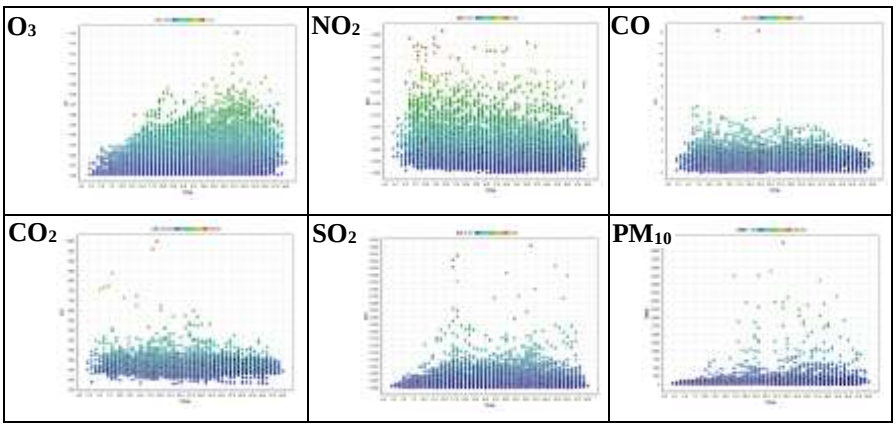

Figure 2 Effect of temperature on the five gases and PM_{10}

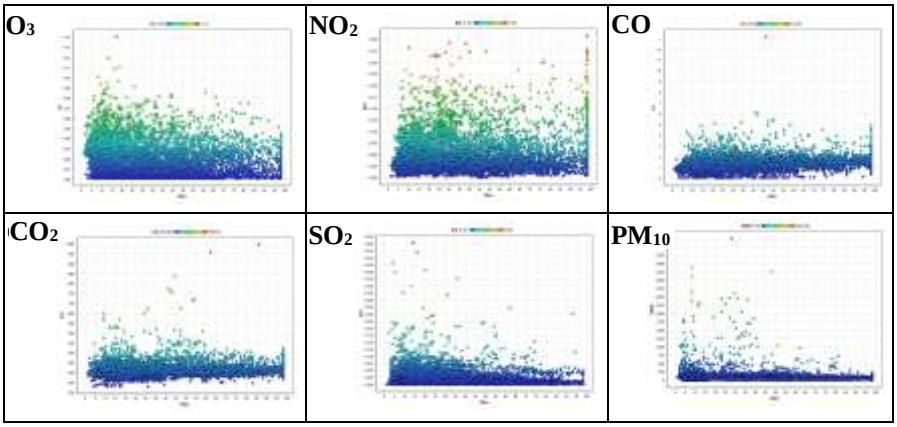

Figure 3 Effect of humidity on the five gases and PM_{10}

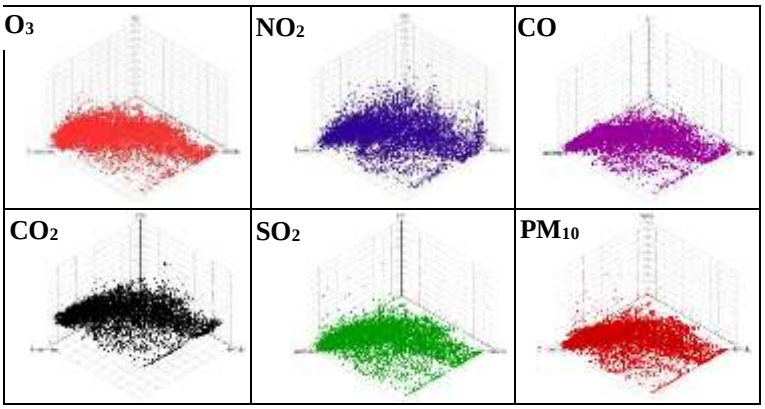

Figure 4 Effect of temperature and humidity on the five gases and PM_{10}

It shows when wind speed between 6-12m/s and temperature 22=35 C°, the PM_{10} will be between 151-400 ppm.

Tree

WS > 3.500: 2 {1=0, 0=0, 2=2, 3=2, 7=0, 4=0, 5=1, 6=0}
WS ≤ 3.500
| WS > 2.500
| | Temp > 4.500: 2 {1=15, 0=0, 2=28, 3=8, 7=0, 4=4, 5=4, 6=5}
| | Temp ≤ 4.500: 1 {1=41, 0=11, 2=16, 3=0, 7=2, 4=0, 5=2, 6=3}
| WS ≤ 2.500: 1 {1=4802, 0=1472, 2=634, 3=77, 7=3, 4=23, 5=16, 6=10}

5 Conclusion

The problems of storing and analysis of big data are facing all the organization through the world, especially the environmental organizations taking interest in monitoring pollution gases. These organizations have one or more online reading stations installed near industrial cities and oil refinery stations.

Using the 2D and 3D scatter diagram to visualize the data reading is one of the important tools. That can be used by decision makers to explore the concentration

of pollutant gases and effect of meteorological parameters, by using these types of visualization the decision makers can take their decision in stopping or reducing the working hours of the factories or refinery stations that cause the major pollution.

We recommend each factory of refinery, using the same methods of visualizing the pollution gases to take their decision to stop their factory of refinery station or reducing the hours of working hours, when the temperature rises to more that 45C°.

References

[1] Gantz, J., Reinsel, D.: Extracting value from chaos. IDC IVIEW. https://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

[2] Lohr, S.: The age of big data. http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html

[3] Shumway R.: One solution for air pollution: big data. http://www.deseretnews.com/article/865617771/One-solution-for-air-pollution-Big-data.html

[4] Han, J., Kamber, M., Jian, P.: Data mining: concepts and techniques. Elsevier Inc., (2012)

[5] Thomas, J., Cook, J.: Illuminating the path: the research and development agenda for visual analytics. National Visualization and Analytics Center (2005)

[6] Murray, S.: Interactive data visualization for the web. O'Reilly Media, Inc. (2013)

[7] http://www.sas.com/en_us/home.html

[8] De Mauro, A., Greco, M., Grimaldi, M.: Grimaldi formal definition of big data based on its essential features. Journal of Library Review, vol. 65, no. 3, pp. 122–135 (2016)

[9] Dion, M., AbdelMalik, P., Mawudeku, A.: Big data and the Global Public Health Intelligence Network (GPHIN). vol. 41, pp. 209-219 (2015)

[10] Heudecker, N., Beyer, A., Laney, D., Cantara, M., White, A., Edjlali, R., McIntyre, A.: Predicts 2014: big data. gartner insight. Gartner Research, Stanford, Connecticut (2013)

[11] Bhagattjee, B.: Emergence and taxonomy of big data as a service. Working Paper CISL# 2014-06. Massachusetts Institute of Technology (2014)

[12] Cheng, S., Liu, Shi, Y., Jin, Y., Li, B.: Evolutionary computation and big data: key challenges and future directions. Proceedings of the First International Conference on Data Mining and Big Data, Bali. Indonesia, pp 3-14, June 25-30 (2016)

[13] Redpath, R.: A comparative study of visualization techniques for data mining. MSc thesis. School of Computer Science and Software Engineering, Monash University, Australia (2000)

[14] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. American Association for Artificial Intelligence, pp. 37-54 (1996)

[15] Frawley, J. Piatetsky-Shapiro, G., Matheus, J.: Knowledge discovery in databases: an overview; knowledge discovery in databases. AAAI Press/ The MIT Press, Menlo Park, California, USA (1991)

[16] Tan, P., Steinbach, M., Kumar, V.: Introduction to data mining. Pearson Addison Wesley (2006)

[17] Witten, I., Frank, E., Hall, M., Pal, C.:Data mining: practical machine learning tools and techniques. Elsevier Inc., 4th Edition (2017)

[18] Kantardzic, M.: Data mining: concepts, models, methods, and algorithms. Johm Wiley and Sons Inc., 2nd Edition (2011)

[19] Masethe, M., Masethe, H.: Prediction of work integrated learning placement using data mining algorithms. Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA, vol I, WCECS 2014, 22-24 October (2014)

[20] Neeb, H., Kurrus, C.: Distributed K-nearest neighbors. https://stanford.edu/~rezab/classes/cme323/S16/projects_reports/neeb_kurrus.pdf

[21] Oxford English Dictionary, Visualization. Oxford University Press (2009)

[22] Chen, C., Hardle, W., Unwin, A.: Handbook of data visualization. Springer (2008)

[23] Keim, A., Mansmann, J., Thomas, S., Ziegler, H.: Visual analytics: scope and challenges. Berlin, Heidelberg, Springer-Verlag (2008)

[24] Few, S.: Now you see it: simple visualization techniques for quantitative analysis. Analytics Press, Oakland (2009)

[25] NESSI.: Big data a new world of opportunities. White Paper (2012)

[26] Rokach, L., Maimon, O.: Data mining with decision trees: theory and applications. World Scientific Publishing (2008)