IADIS International Journal on Computer Science and Information Systems
Vol. 5, No.1, pp. 87-100
ISSN: 1646-3692

DISTRIBUTED PATTERN RECOGNITION FOR AIR QUALITY ANALYSIS IN SENSOR NETWORK SYSTEM

Yajie Ma
Information Science and Engineering College, Wuhan University of Science and Technology.

Yike Guo
Department of Computing, Imperial College London.

Moustafa Ghanem
Department of Computing, Imperial College London.

ABSTRACT

In this paper, we investigate the development of distributed real-time data mining algorithms for sensor network applications with a focus on air pollution monitoring applications. Our approach is based on a considering two-layer sensor network framework comprising mobile and stationary sensors. We present architectural abstractions for such a network that are suitable for conducting peer-to-peer data mining algorithms. We also develop and evaluate the performance of a real-time peer-to-peer data clustering algorithm for the identification of pollution hotspots and for analyzing the dispersion of pollution clouds. We conduct an experimental evaluation of our algorithms to compare the accuracy of our algorithms to centralized implementations and identify the associated tradeoffs.

KEYWORDS

Pattern recognition, Distributed clustering, Sensor Grid, Sensor Network, Air pollution.

1. INTRODUCTION

1.1 Sensor Networks for Urban Pollution Monitoring

Our motivations for this research are drawn from designing architectures and algorithms to support the analysis of data collected from large-scale sensor networks for environmental applications, and in particular those for air quality control and pollution monitoring [1–3]. We focus on newer systems such as that developed in the MESSAGE (Mobile Environmental Sensor System Across Grid Environments) project [4-7], which fully integrates existing static sensor systems and complementary data sources with the mobile environmental sensor system.

The sensor units used in a system such as MESSAGE can be generally deployed over wide geographic area to collect, process, and act upon large amounts of complex data. Each of the units is typically equipped with multiple sensing devices that measure different variables, and typically also has an on-board processing unit to perform local data processing and analysis operations as well as exchange data with other sensor units and base stations. The units can be fixed or mobile and connectivity between themselves and the base stations can proceed using a variety of network technologies and protocols including ad-hoc wireless networking methods and fixed networking methods. With the support of a two-layer network framework formed by the mobile sensors fixed by the roadside devices and stationary sensors carried by public vehicles, the MESSAGE provides a highly effective system for air pollution monitoring.

Irrespective of the actual technologies use, the key challenge that typically arises in the context of these applications is our ability to organize and process the large amounts of collected data effectively and efficiently in the network itself to minimize the reliance on centralized servers. In this paper, and based on our former work on the Discovery Net [8] and MoDisNet [9], we present architectural abstractions for such a network that are suitable for conducting peer-to-peer data mining algorithms within the MESSAGE system as well as in other environmental monitoring sensor networks. We use our abstractions to develop and evaluate the performance of a real-time peer-to-peer data clustering algorithm for the identification of pollution hotspots and for analyzing the dispersion of pollution clouds.

1.2 Paper Layout

The remainder of this paper is organized as follows. In Section 2 we present related and previous work on air pollution monitoring systems design. In section 3, we describe the MESSAGE system architecture and discuss how it can be adapted to meet the demands of distributed and peer-to-peer data mining. In section 4, we describe the implementation of our distributed clustering algorithm and discuss its performance properties. In Section 5, we describe the real-time pollution pattern recognition experiments to evaluate our approach.

2. PREVIOUS AND RELATED WORK

Under the current Environment Act of UK [10], most local authorities have air quality monitoring stations to provide environmental information to public daily via Internet. The

conventional approach to assessing pollution concentration levels is based on data collected from a network of permanent air quality monitoring stations. However, permanent monitoring stations are frequently situated so as to measure ambient background concentrations or at potential 'hotspot' locations and are usually several kilometers apart. For example, the Environmental Research Group (ERG) at King's College London is a leading provider of air quality information and research in the UK. In 1993, the group created the London Air Quality Network (LAQN) [11] in conjunction with the 33 London Boroughs and Regional Health Authorities. The group manages over 160 continuous monitoring sites, providing site management, quality assurance, local site operator support and reporting services including the dissemination of measurements via the Internet.

In the US, the CitySense research group [12] in Harvard University provides an urban scale sensor network testbed that is being developed by researchers at Harvard University and BBN Technologies. At the end of 2009, CitySense consisted of 100 wireless sensor notes deployed across Cambridge, MA. Each node consists of an embedded PC, network interface, and various sensors for monitoring weather conditions and air pollutants. CitySense is designed as an open testbed for researchers to evaluate wireless networking and sensor network applications in a large-scale urban setting.

The Discovery Net [8] and MoDisNet [9] are both architectures for the analysis of air pollution data sets. The Discovery Net architecture is based on collecting data from statically located urban pollution monitoring sensors and uses workflow technique to analyze the distribution of pollutants centrally. The MoDisNet system is the mobile version of DiscoveryNet. Furthermore, the MoDisNet system enhanced the data analysis capability by managing the pollution data in a distributed style. Our work in this paper is based on both of the above systems. Both systems are designed for analyzing data collected at finer spatial and temporal granularity thus enabling the studying pollution hot spots change with time. They are also designed to allow the integration of the sensor data with all types of data, such as meteorological data. The inclusion of such data allows for studying pollution hot spots change and the factors affecting them in scientific experiments.

3. THE MESSAGE FRAMEWORK

3.1 The MESSAGE Data Collection Architecture

The key feature of the MESSAGE system [4-7] is to use a variety of vehicle fleets including buses, service vehicles, taxis and commercial vehicles as platform for environmental sensors. The system has been designed to work with a variety of sensors in three cities in the UK, London, Cambridge and New Castle. Figure 1(a) shows a high-level view of the data collection architecture of the MESSAGE project [6]. The approach is based on using a real-time data manager to collect sensor data from distributed sensors and to organize it into a number of data marts for integration with other data sources for access and analysis by various applications

The architecture is generic and can be used with any type of sensors. The sensors used in London are based DUVAS sensor units used in the system are similar to those used in previous projects are based on high-performance mobile UV spectroscopy sensors, which

measure SO2, NO, NO2, O3, NH3, and Benzene at ppb levels. The units also have the ability to send data using 3G/GPRS via the mobile phone network enabling greater autonomy and reliability in data transmission. In order to provide the data buffering locally on the device, they also support WiFi data transmission with a store and forward configuration. The DUVAS units are also equipped with on-board processing units providing them with an ability to perform on board data mining either on single units or collaboratively across multiple units.

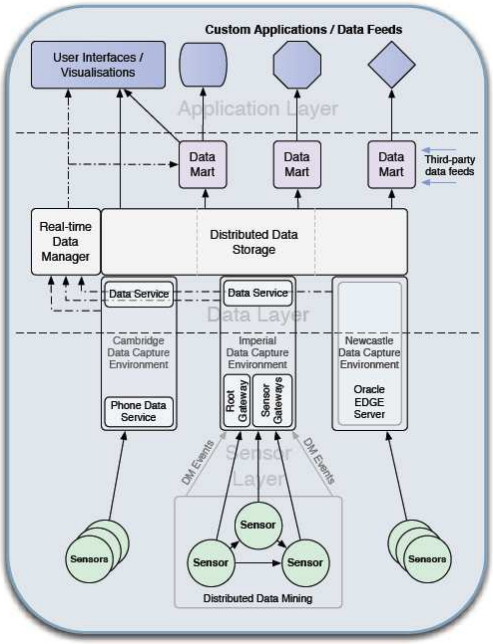


Figure 1(a). MESSAGE Architecture

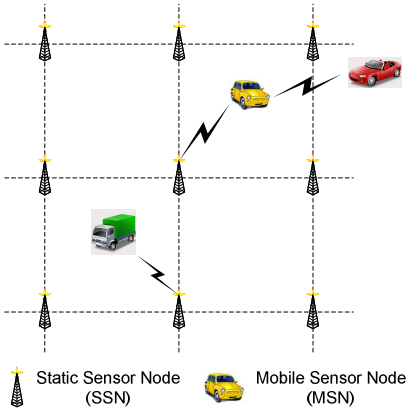


Figure 1(b). Two Layer network framework

The input data based on our former research [8] uses the air pollution data sampled from 140 sensors marked as the red dots (see Figure 4 and 5 in section 6) distributed in a typical urban area around the Tower Hamlets and Bromley areas in east London. There are some of the typical landmarks such as the main roads extending from A6 to L10 and M1 to K10, the hospitals around B5 and K4, the schools at B7, C8, D6, F10, G2, H8, K8 and L3, the train stations at D7 and L5, and a gas work between D2 and E1. 140 sensors collect data from 8:00 to 17:59 at 1-minute intervals to monitor the pollution volumes of NO, NO_2, SO_2 and Ozone. Then there are 600 data items for each node and totally 84000 data items for the whole network. Each data item is identified by a time stamp, a location, and a four-pollutant volume reading. Once sensor data is collected, data cleaning and preprocessing is necessary before further analysis and visualization can be performed. Most importantly, missing data must be performed using bounding data from the sensor, or also using data from nearby sensors at the same time. Interpolated data may be stored back to the original database, with provenance information including the algorithm used. Such pre-processing is standard, and has been conducted using the available MESSAGE component. The relatively high spatial density of sensors used also allows a detailed map of pollution in both space and time to be generated.

3.2 Two Layer Network Framework

Within the MESSAGE framework sensor units can be placed on vehicles (e.g. busses), thus allowing a fleet of moving vehicles to collect pollution data as the vehicles travel across a city. This enables collecting more data from more locations. Figure 1(b) shows the MESSAGE mobile sub-network formed by the Mobile Sensor Nodes (MSN in short) and the stationary sub-network organized by the Static Sensor Nodes (SSN in short). They sample the pollution data and execute the AD conversion to get the digital signals. According to the system requirements, the MSNs may pre-process the raw data (such as the noise reduction, local data cleaning and fusion, etc.) and then send these data to a nearest SSN. The SSNs take in charge of the data receiving, update, storage and exchange works.

4. DISTRIBUTED MINING ARCHITECTURE

In this paper we extend the MESSAGE architecture into a four-layer hierarchical architecture, to enable the distributed mining of the pollution data in P2P style within large scale mobile sensor networks. These extensions support the full scale analytical task ranging from dynamic real time mining of sensor data to the analysis of off-line data warehoused for historical analysis. The enables sensors equipped with sufficient computational capabilities to feed data to the warehouse as well as perform analysis tasks in collaboration with their peers when needed. In this section, we will first discuss the real-time sensor grid challenge, and then present the hierarchical e-Science grid architecture within MESSAGE with the explanation of each layer in detail. As with previous work [8], we use the term sensor grid rather than sensor network to highlight the need for processing and analyzing, rather than simply monitoring, large amounts of data.

4.1 Requirements for Distributed Sensor Data Mining

Enabling large-scale data mining within the MESSAGE architecture requires enabling distributed data mining to be conducted by the sensors themselves. This leads to the following four groups of requirements:

Peer-to-peer processing: Within a large-scale mobile sensor network architecture, the sensors themselves naturally form and communicate with each other as a P2P network. In order to satisfy the real-time analysis requirement, the sensors themselves will have to store part of the information and communicate with each other within a P2P network. The measurements from sensors, both mobile and static, will be filtered and processed using a set of specialized algorithmic processes, before being warehoused within a repository. The design and implementation of suitable P2P sensor network architecture will need to satisfy the real-time analysis requirements as well as the data storage/communication trade-offs. The sensors in such a system will need to be equipped with sufficient computational capabilities to participate in the grid environment and to feed data to the warehouse as well as perform analysis tasks and communicating with their peers.

Metadata representation: The sensor network is characterised by the heterogeneity and geographic distribution of the sensors. Within a sensor network, sensors can be located, accessed and integrated within a particular study. Not only is it essential to record the type of

pollutants measured (e.g. Benzene, SO2, NOx, etc) for each sensor, but also since sensors may be mobile it is essential to record the location of the sensor at each measurement time. Such information must be described at the semantic level thus allowing sensors published as sensor services to be discovered. Finally methods addressing the security and authentication issues relating to accessing and controlling the sensors must also be addressed.

Large data set storage and management: The sensor network is also characterised by sizes of data being collected and analysed. Moreover, different sensors typically provide measurements at different resolutions and scales and efficient data aggregation and data querying mechanisms must be provided, and such information must be described and published as sensor data services using standardised metadata techniques.

Real-time data analysis (stream mining): Historical data is well suited to large-scale analysis over multiple dimensions, but for dynamic queries over real-time sensor data streams, the data has to be taken directly from the sensors. These data points have little value for warehousing and also the real time mining querying cannot afford a "store and mining" model. A typical analytical work would involve the statistics at a certain location and about certain properties in that location. The sensor may not be able to offer this information on its own, due to its movement from the location, or due to inability to capture all relevant information pertaining to the query. Dynamically composed sensor network with a P2P communication model will support such information aggregation for these kinds of analytical queries. Streaming mining algorithms are also necessary to realize such a real-time analytical querying model.

4.2 Hierarchical Architecture

Figure 2 shows the modifications to the MESSAGE architecture to address the key requirements for distributed data mining through the middleware layer. The architecture comprises four layers:

Sensor Layer: Similar to the data capture architecture, this layer manages different types of sensors. Sensors within the environment are heterogeneous and may be mobile or static. Hence, the wireless connectivity can provide different access protocols to the IP backbone including WiFi (802.11.g), Zigbee (802.15.4), and WiMax (802.16). The sensors have the capability to sample one or more pollutants or other environmental properties such as noise or temperature. Sensors must have the capacity to buffer a reasonable amount of data in the event that connectivity is lost as well as to conduct local data analysis tasks. Then, sensors send their data packets to a known location. Ultimately, all data packets will be sent onto an IP network for transmission into the Data Layer of the architecture.

Middleware Layer: This is the core layer of the sensor network architecture. This layer covers the functions of execution management, distributed data mining activity and sensor registry and control. As the MESSAGE system is designed primarily to support the distributed analysis of the pollution data, the distributed data mining entity is designed for on-line examination of the monitored area that is down to the level of streets and buildings. The execution management, which is critical for the system performance, is the core service of this layer as well as the whole network architecture. It enables virtual organization management, resource management and load balancing, etc.

Data Layer: As with the Data capture architecture, the Data Layer handles the capture of streams of data packets coming from a large number of mobile environmental sensors and

ensures the efficient, reliable handling of this data and its insertion into the real-time data store. This data must then be transported to the data store and inserted in the database in an efficient manner. Since there is the potential for the volume of data to be significant, the key to this layer of the architecture is the efficient management of data. Data must be stored in a real-time database that provides a schema that allows for efficient storage of large quantities of data. Database oriented queuing systems meet the desired scalability and performance characteristics and deliver sophisticated business management capabilities. Querying of the data should also be optimized but the database is not designed for large numbers of real-time queries. Instead, data will be batch queried at off-peak times and fed into Data Marts.

Application Layer: Similar to the data capture architecture, the Application Layer retrieves information from the Real-time Database Layer, specifically the real-time data store, and uses this information as the input to applications. There are a number of Application Groups, including traffic monitoring and control, on-line or off-line data mining, public data query and visualization tools, etc. The user-defined service makes the system extensible so that the users of the system can take advantage of new services that become available. As many application groups require distinct sets of information, the data access and storage are designed to be a utility service to aid common data access task (e.g. remote ODBC database access), and the storage service allows data that has been accessed to be store locally.

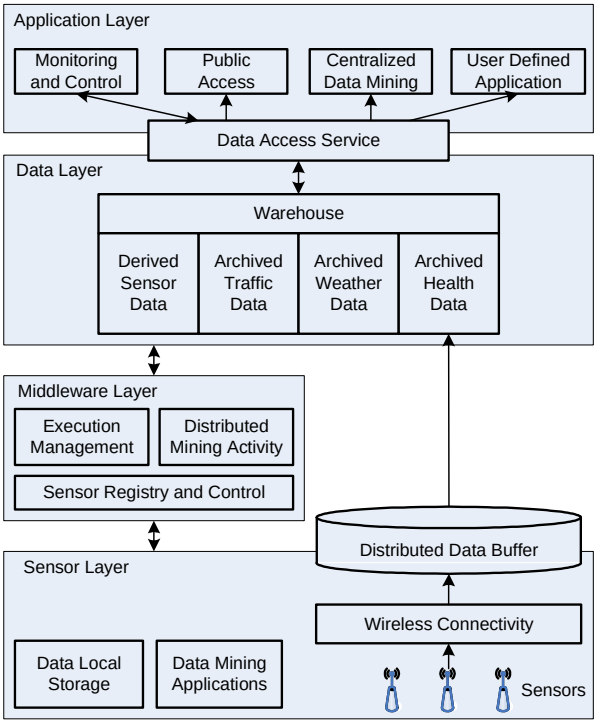


Figure 2. P2P sensor grid architecture

5. DISTRIBUTED CLUSTERING ALGORITHM

Data mining for pollution monitoring in sensor networks in urban environment faces several challenges. First, the methods of data collection and pre-processing highly rely on the complexity of the environment. For example, the distribution and features of pollution data are correlated to inter-relationships between the environment, geography, topography, weather and climate and the pollution source, which may guide the design of the data mining algorithms. Also, the mobility of the sensor nodes increases the complexity of sensor data collection and analysis [13, 14]. Second, resource-constrained (power, computation or communication), distributed and noisy nature of sensor networks presents challenges for storing the historical data in each sensor, even for storing the summary/pattern from the historical data [15]. Third, sensor data come in time-ordered streams over network, which makes traditional centralized mining techniques inapplicable. As the result, the real-time distributed data mining (DDM in short) schemes are significantly demanded in such scenario. Considering the pattern recognition application, in this section, we introduce a peer-to-peer clustering algorithm as well as the performance analysis.

5.1 P2P Clustering Algorithm

To realize the DDM algorithm with the capability to provide the information exchange in P2P style, a P2P clustering algorithm is designed to find out the pollution patterns in the urban environment according to the sampled air pollutants' volumes.

After investigating the clustering algorithm and its results in MoDisNet, we found that, because the clusters of the pollutants are not always in nonconvex shapes, the *K*-means algorithm doesn't apply well in such a scenario. Hence in this paper, we design a hierarchical clustering algorithm based on DBSCAN in [16]. In comparison with the algorithm in [16], our algorithm has the following characteristics:

1. Nodes only require local synchronization at any time, which is better suited to a dynamic environment.
2. Nodes only need to communicate with their immediate neighbors, which reduces the communication complex.
3. Data are inherently distributed in all the nodes, which makes the algorithm be widely used in large, complex systems.

The algorithm runs in each SSN (MSN only takes in charge of collecting data and sending data to a closest SSN). In order to describe this algorithm, we give some definitions first (suppose the total numbers of SSN is n ($n > 0$)).

- SSN_i: a SSN node with the identity i ($i = 0, \ldots, n\text{-}1$);
- S_i: an Information Exchange Node Set (IENS) of SSN_i, which is a set of some of the SSNs that can exchange information with SSN_i;
- CS: candidate cluster centre set. Each element in CS is a cluster centre;
- $C^l_{i,j}$: the cluster center of jth ($j \geq 0$) cluster that is computed in SSN_i in lth recursion ($l \geq 0$), $C^l_{i,j} \in CS$;
- $Num_{i,j}$: the number of members (data points) belongs to jth cluster in SSN_i;
- $E(X, Y)$: the Euclidian distance of data X and Y;
- D: a pre-defined distance threshold;
- δ: a pre-defined offset threshold.

The algorithm proceeds as follows.

1. Generates S_i and local data set. Node SSN_i receives data from MSNs as local data and chooses a certain number of SSNs as S_i in term of a random algorithm (the detail of the random algorithm is beyond the scope of this article).
2. Generates *CS*. This process is described by the following pseudo code:

SSN_i chooses a data item j from its local data set into *CS* as $C^0_{i,j}$;
for each other data item k in the local data set of SSN_i
    for each data item $m \in CS$
        if $E(k, m)>D$
            put k into *CS as* $C^0_{i,k}$;

---

3. Distributes data. For each candidate cluster centre $C^0_{i,j} \in CS$ and a data item Y, if $E(C^0_{i,j}, Y) < D$, then distribute Y into the cluster. Thus each local cluster of SSN_i can be described as $(C^0_{i,j}, Num_{i,j})$
4. Updates *CS*. Node SSN_i exchanges local data description with all the nodes in S_i. After SSN_i receives all the data descriptions it wants, it checks to see if two cluster centres $C^0_{i,j}$, $C^0_{i,k}$ satisfy $E(C^0_{i,j}, C^0_{i,k}) < 2D$, then it combines these two clusters and updates the cluster centre as $C^1_{i,j}$.
5. Compares $C^0_{i,j}$ and $C^1_{i,j}$. Computes the offset between $C^1_{i,j}$ and $C^0_{i,j}$. If the offset $\leq \delta$, then the algorithm finishes; otherwise SSN_i replaces $C^0_{i,j}$ with $C^1_{i,j}$, and go to step 3.

5.2 Clustering Accuracy Analysis

The evaluation of the accuracy of the algorithm aims to investigate in what degree our P2P clustering algorithm can assign the data items into the correct clusters in comparison with the centralized algorithm. To do so, we design an experimental environment for data exchange and algorithm execution. The network topology of the simulation is shown in Figure 3. We use 18 sensor nodes, including 12 SSN nodes from node 0 to node 11 and 6 MSN nodes from node 12 to node 17. Data are sampled at each MSN and sent to a nearest SSN. The air pollution data we use is consisted of the volumes of four pollutants NO, NO_2, SO_2, and O_3 sampled at 1-minute intervals in urban environment from 8:00 to 17:59 within a day collected from 6 MSNs (as described in Section 3.1). Then, the total number of data items in the dataset is 3600. Data can be sent and received in bi-directions along the edges.

The comparison of the average clustering accuracy of the centralized and distributed clustering algorithms is shown in Table 1. For the centralized clustering algorithm, we suppose node 8 be the sink (central point for data processing), which means every other MSN sends the data to node 8. And the classic DBSCAN algorithm is running in node 8 for centralized clustering. For the accuracy measurement, let X^i denote the dataset at node i. Let $L^i_{km}(x)$ and $L^i(x)$ denote the labels (cluster membership) of sample x ($x \in X^i$) at node i under centralized DBSCAN algorithm and our distributed clustering algorithm respectively. We define the Average Percentage Membership Match (*APMM*) as

$$APMM = \frac{1}{n}\sum_{i=1}^{n}\frac{|\{x \in X^i : L^i(x) = L^i_{km}(x)\}|}{|X^i|} \times 100\% \quad (1)$$

Where n is the total number of SSNs.

For the distributed clustering algorithm, we vary the number of nodes in the Information Exchange Node Set (IENS) of each SSN from 1 to 10. Let D = 10 and δ = 1. Data are randomly assigned to each SSN. Table 1 shows the *APMM* results.

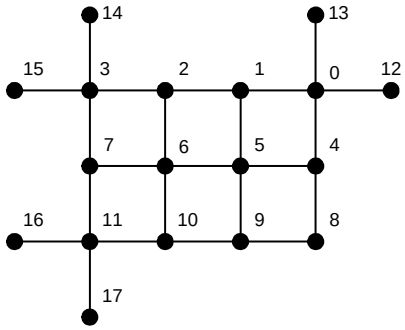


Figure 3. The network topology of the simulation.

Table 1. Centralized Clustering vs. Distributed Clustering (*APMM* results)

IENS	1	2	3	4	5	6	7	8	9	10
APMM	86.3%	91.2%	92.67%	93.46	93.55%	93.74%	93.93%	94.23%	94.59%	94.97%

From Table 1 we can see that, when the number of nodes in IENS is no less than 2, in other words, when each SSN exchanges data with at least two other SSNs, the *APMM* exceeds 91%. When the number of nodes in IENS is no less than 4, the *APMM* exceeds 93%. The results are achieved under the condition of assigning the data to each SSN randomly. In reality, if the patterns of the dataset are various in different locations, the *APMM* may be lower than the results in Table 1. In such situations, a good scheme of how to choose the nodes to construct the IENS would be very important.

6. EXPERIMENTAL ANALYSIS OF PATTERN RECOGNITION

6.1 Clustering Accuracy Analysis

The pollution hotspots identification uses the air pollution data to find out the distribution of some key pollution locations within the research area. Our former work in Discovery Net can only classify the pollution data into several pollution levels, such as high or low, but can't tell us the distribution of different pollutants in different locations and their contributions to the pollution levels. To improve the data analysis capability, in this data analysis experiment, we use the distributed clustering algorithm to cluster the pollutants into groups which can recognize different pollution patterns. From the experimental results of Discovery Net, we pickup all the high pollution level locations in the research area at 15:30 and 17:00

respectively to check the contribution of different pollutants (NO, NO_2, SO_2 and Ozone) to the pollution levels. The results are shown in Figure 4.

In this figure, different clusters/patterns correspond to different colors, which reveal the relationship between the combination and volumes of different pollutants. According to the clustering result, we use red color to denote the pattern of high volume of NO_2 and Ozone whilst low volume of NO; blue color features the pattern of high volume of SO_2 and Ozone; yellow color only contains high volume of SO_2. From the figures we can see, at 15:30 the hotspots are located at the schools (which are highlighted by circles and almost all featured by high volume of NO_2 and Ozone) and the gas work (which is highlighted by square and featured by high volume of SO_2). At 17:00, the hospitals (highlighted by the ellipses) and the gas work all contribute to the pollutant of SO_2.

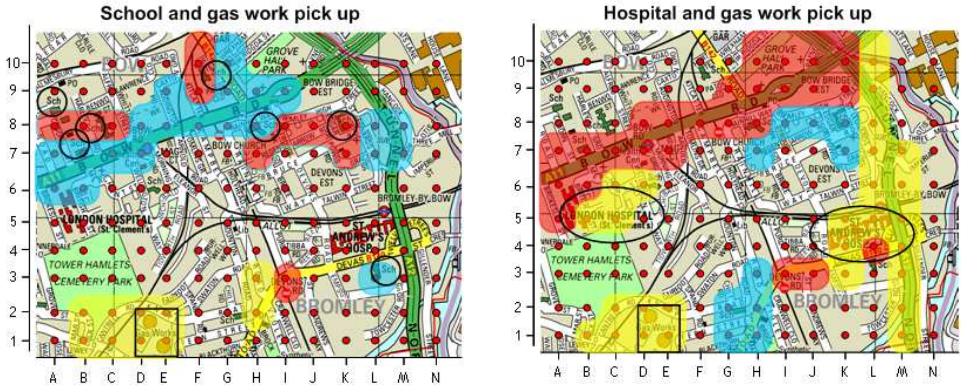


Figure 4. Pollution hotspots identification.

Another kind of hotspot located on the main roads. However, they present different patterns at different time on different roads. Main road A6-L10 is covered in blue at 15:30 while red at 17:00. There are two reasons for this circumstance. First, the road transport sector is the major source of NO_X emissions and the solid fuel and petroleum products are two main contributors of SO_2. Second, NO_2 and Ozone are all formed through a series of the photochemical reactions featuring NO, CO, hydrocarbons and PM. Generating NO_2 and Ozone needs to take a period of time. This is why the density of NO_2 is always high on the main road whereas Ozone at 17:00 is higher than that at 15:30. Another interesting fact is that, at 17:00 main roads A6-L10 and M1- K10 show different pollution patterns. From the figure we can see, the pollution pattern on M1-K10 is very similar to the patterns at the gas work and hospital areas, but not similar to the pattern on the other main road. We investigated this area and found that, a brook flows along this area in the near east and a factory area locates on the opposite side of the brook which is beyond the scope of this map. This can explain why the pollution patterns are different on these two main roads.

6.2 Pollution Clouds Dispersion Analysis

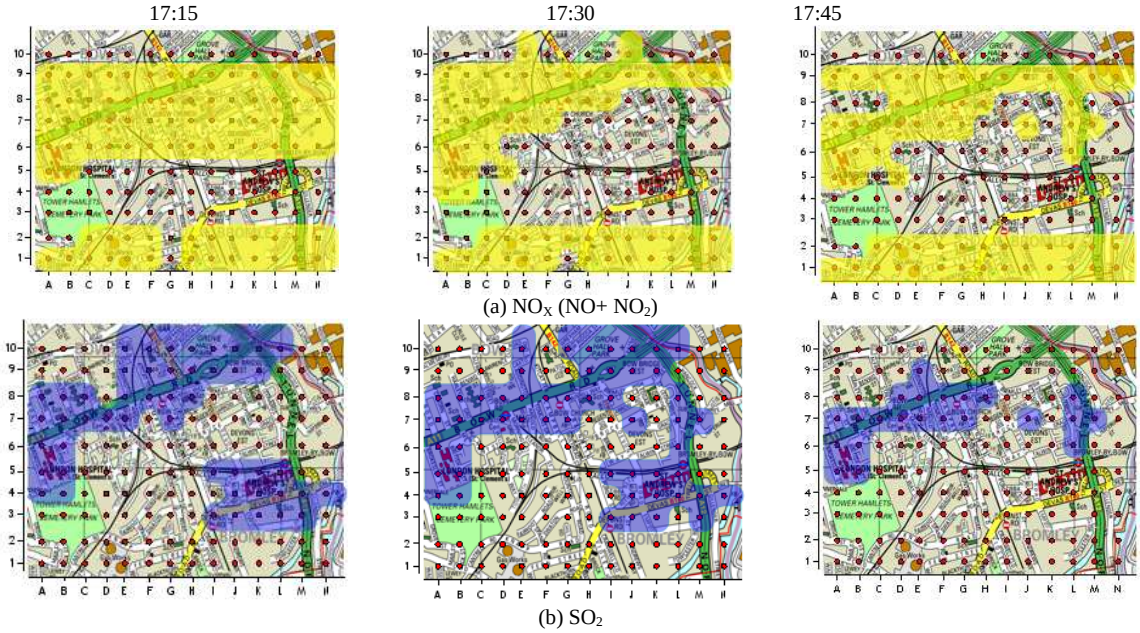
In this experiment, we investigate the dispersion of different pollution clouds to see their movements and changes. We pick up the pollutants of NO_X (NO+ NO_2) and SO_2 respectively

and calculate the pollution clouds of them at the time points of 17:15, 17:30 and 17:45. The results are shown in Figure 5(a) and (b).

According to the environmental reports of the UK, it is always the worst pollution distribution time period within a day after 17:00. The road transport sector contributes more than 50% to the total emission of NO_X, especially in urban areas. At the meantime, the factories are another emission source of the nitride pollutants. Besides the major source of SO_2 generated by the solid fuel and the petroleum products from the transport emission, some other locations such as the hospitals contribute some kind of pollutants, including the sulphide and nitride. These features are well illustrated by Figure 4.

In Figure 5(a), the main road A6-L10 and its circumjacent areas are severely covered by high volume of NO_X. The same situation appears in the area from A1 to N2 which includes a gas work (between D2 and E1), side roads (A1 to J2), factories and parking lots (K1 to L2). And we can notice that the dispersion of the NO_X clouds fades as the time goes by, especially around the main road area. However, the NO_X clouds will stay for a long time in A1 to N2 area.

The dispersion of SO_2 cloud in Figure 5(b), however, shows different feature. The cloud mainly covers the main roads, as well as two hospitals (around B5 and K4). In comparison with the result at 17:15, the SO_2 cloud blooms at 17:30, which lays almost over all the two main roads and hospitals. However, it fades quickly at 17:45 and uncovers a lot of areas, especially the main road M1-K10 and hospital K4. This status may due to the different environmental conditions in this area (the dispersion of SO_2 depends on a lot of factors such as the temperature, wind direction, humidity and air pressure, etc.). Besides, it also can be attributed to the existence of the brook in the near east – SO_2 can be absorbed into water to form sulphurous acid very easily, which decreases the volume of SO_2 in the air whereas increases the pollution of the water.



(a) NO_X ($NO+ NO_2$)

(b) SO_2

Figure 5. Pollution clouds dispersion of NO_X and SO_2.

7. SUMMARY AND CONCLUSIONS

In this paper, we introduce the network framework and sensor unit design for an air pollution monitoring system. Based on the discussion of real-time sensor grid challenge, we present a hierarchical sensor grid architecture, which is featured by the four-layer structure and can provide a platform for different wireless access protocols. The experiments of air pollution analysis based on distributed P2P clustering algorithm, which investigates the distribution of pollution hotspots and the dispersion of pollution clouds. The experimental results are useful for the government and local authorities to reduce the impact of road traffic on the environment and individuals.

We are currently extending the application case studies to monitor PM_{10} and finer detection (e.g. $PM_{2.5}$). As addressing global warming becomes more important, there are increasing requirements for greenhouse gas emission monitoring and reduction. Information on greenhouse gases is therefore also needed for long term monitoring purposes with similar linkages to traffic and weather data to understand the contribution of traffic to environmental conditions.

ACKNOWLEDGEMENTS

The project Mobile Environmental Sensing System Across a Grid Environment (MESSAGE), Grant No. EP/E002102/1, was funded jointly by the Engineering and Physical Sciences Research Council (EPSRC) and the Department for Transport. The project also has the support of nineteen non-academic organisations from public sector transport operations, commercial equipment providers, systems integrators and technology suppliers. More information is available from the web site www.message-project.org.The authors are grateful for the support of colleagues at Imperial College London and all colleagues within the MESSAGE consortium. In particular we acknowledge the help of Dr. Jeremy Cohen and Dr. Robin North who have worked together with us in MESSAGE project and gave valuable suggestions on grid technology and pollution analysis.

REFERENCES

1. A. Vaseashta, G. Gallios, M. Vaclavikova, et al, 2007. Nanostructures in Environmental Pollutuion Detection, Monitoring, And Remediation. *In Science and Technology of Advanced Materials*, Vol. 8, Issues 1-2, pp. 47-59.
2. M. Ibrahim, E. H. Nassar, 2007. Executive Environmental Information System (ExecEIS). *In Journal of Applied Sciences Research*, Vol. 3, No.2, pp. 123-129.
3. N. Kularatna, B.H. Sudantha, 2008. An Environmental Air Pollution Monitoring System Based on the IEEE 1451 Standard for Low Cost Requirements. *In IEEE Sensors Journal*, Vol. 8, Issue 4, pp. 415-422.
4. MESSAGE: Mobile Environmental Sensing System Across Grid Environments. http://www.message-project.org.
5. J. Cohen, R. J. North, S. Wilkins, J. Darlington, Y. Guo, N. Hoose, Y. Ma, J. W. Polak. Creating the MESASAGE infrastructure. *Traffic Engineering and Control*, December 2009.

6. J. Cohen, B. Fuchs, R. North, N. Hoose, J. Darlington and J. Polak. Computational Grid-based Data Management and Analysis for a Mobile Sensor Network Deployment. *16th Intelligent Transport Systems (ITS) World Congress*. Stockholm, Sweden. September 2009.
7. J. Cohen, R. North, S. Fayer, J. Darlington and J. Polak. An e-Science Infrastructure for the On-demand Management, Analysis and Visualisation of Environmental Data. UK e-Science All Hands Meeting 2009, Oxford, UK. December 2009.
8. M. Richards, M. Ghanem, M. Osmond, et al, 2006. Grid-based Analysis of Air Pollution Data. *In Ecological Modelling*, Vol.194, Issue 1-3, pp. 274-286.
9. Y. Ma, M. Richards, M. Ghanem, et al, 2008. Air Pollution Monitoring and Mining Based on Sensor Grid in London. *In Sensors*, Vol. 8, pp. 3601-3623.
10. Environment Act 1995. http://www.opsi.gov.uk/acts/acts1995/ukpga_19950025_en_1.
11. London Air Quality Network. http://www.londonair.org.uk/
12. http://www.citysense.net/
13. M.J. Franklin. Challenges in Ubiquitous Data Management, 2001. *In Lecture Notes In Computer Science*. Vol.2000, pp. 24-31.
14. F. Perich, A. Joshi, T. Finin, et al, 2004. On Data Management in Pervasive Computing Environments. *In IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 5, pp. 621-634.
15. Y. Diao, D. Ganesan, G. Mathur, et al, 2007. Rethinking Data Management for Storage-centric Sensor Networks. *Proceedings of The Third Biennial Conference on Innovative Data Systems Research*, Asilomar, CA, USA, pp. 22-31.
16. M. Ester, H.-P. Kriegel, J. Sander, et al, 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of The 2nd International Conference on Knowledge Discovery and Data Mining*, Oregon, Portland, pp. 226-231.