AN OPTIMAL IOT ENABLED DATA PROCESSING APPROACH FOR EFFECTIVE DETECTION OF REMOTE AIR POLLUTION

Nivash J P\(^{1,a}\), L D Dhinesh Babu\(^{2,b*}\)

\(^1\)School of Information Technology and Engineering, VIT University, Tamil Nadu, India.
\(^2\)School of Information Technology and Engineering, VIT University, Tamil Nadu, India.

Email: jeevprabnivash@gmail.com

Received on 25-07-2016

Accepted on 20-08-2016

Abstract

In order to prevent air quality deterioration and human health consequences, globally the pollution makes up most important emergency. For these causes, in a local and on a regional scale mobile monitoring has increased in the last decade. By describing vital issues and effects, we have summed up the present condition in this area. Recent research says that massive data collected from various devices have to be processed efficiently. Many researchers have proposed ideas for data collection in a distributed way. Efficient data processing architecture is needed for massive data collected through IOT environment. Three-layer architecture was designed to deal the data processing in IOT. The data received from devices are initially stored in Data Assembly Layer (DAL). It is then passed to Data Separation Layer (DSL) for assuring the data efficiency. If the data is static, it is directly stored in the database by DSL else the Data Examining Layer (DEL) prepares the data to neighborhood structure. ODP (Optimal Data processing) model distributes the stress of data storage which reduces the workload of remote air pollution monitor.

Keywords: Remote Air Pollution Detector; Ambient Assisted Living; IOT for Air Pollution.

Introduction:

The air quality focus has become a major concern as a result of economic growth. To enhance public awareness the Air pollution Index (API) was introduced many years ago. API represents the quality of air in urban areas. In 1970 the United States Environmental Protection Agency presented the first API system. The dispersal process of air pollutants are based various parameters, such as carbon monoxide(CO), Sulfur dioxide(SO\(_2\)), Nitrogen Dioxide(NO\(_2\)), hydrocarbon(O\(_3\)), suspended particulates(PM\(_{10}\)), and so on(Lei & Wan, 2010).The toxicological studies were stimulated by air pollution...
disasters of common pollutants. Humans and animals face adverse effects by all identified pollutants (Cassell, 1969). The temporal and spatial structures of dust, haze, smoke, and other atmospheric elements can be depicted by sharp and finely detailed Satellite sensors (Husar, Hoijarvi, Falke, Robinson, & Percivall, 2008). The applications involving environmental monitoring must have expert environmental knowledge. Those applications must have reliable sensors, efficient communication and effective computer technologies (Martinez, Hart, & Ong, 2004). For forecasting and monitoring, Environment Observation and Forecasting System (EOFS) is considered as one of the large scale sensor network (Xu, 2002). The state of being certain may become bigger in a significant way if there are required sensors in the environment (Snidaro, Niu, Foresti, & Varshney, 2007). The massive data generated by sensors can be analyzed manually or by RAPD (Remote Air Pollution Detector). The pressure will be increased if massive data is sent to RAPD for central processing.

The main focus of this paper is data processing and reducing the workload of RAPD. The three important aspects of this paper are summarized below.

- Risk function is included in the ODP architecture through which the data analysis is quantified (Wang, Shao, Shu, Han, & Zhu, 2015).
- In ODP architecture filtering mechanism is used to obtain efficient real time input and to store static data in database.
- Variable Neighborhood search based Optimal Data Processing (ODP) algorithm is applied to analyze distributed data (Hansen & Mladenovi, 2001).

Related work:

In accordant with the surrounding environment Bamis et al proposed a self-configured system. By comprehensive records it is said that the data collected by sensors are accurate in this system (Bamis, Lymberopoulos, Teixeira, & Savvides, 2010). In the environment sensors may fail or provide incorrect data or incomplete data (McKerlich, Ives, & McGreal, 2013). By multi-layer verification model Mcnaull et al ensured the data accuracy (McNaull, Augusto, Mulvenna, & McCullagh, 2012).

Many authors have considered the accuracy analysis but massive data transmission and Remote Air Pollution Detector (RAPD) workload was not explored. By recognizing environmental factors Hristova et al states that more precise comprehensive data can be accumulated (Hristova, Bernardos, & Casar, 2008). Coronato et al suggested a system to discover abnormal behavior by knowing the environment. From regular unnatural behavior detection a different real-time
checkout process was adopted by this system (Coronato & De Pietro, 2013). Hossain et al designed a system to collect data comprehensively and for overall detection a distributed sensor environment was proposed in the environment (Hossain, Atrey, & Saddik, 2007). Winkley et al proposed a wireless sensor network based system for data collection. The data collected from the portable sensors are sent to a processor for classifier analysis (Winkley, Jiang, & Jiang, 2012). This system does not analyze environmental factors.

Based on a network Antonio et al proposed a comprehensive architecture for monitoring. However workload will be increased if the massive data is sent for analyzing (Jara, Zamora-Izquierdo, & Skarmeta, 2013). Through behavior cognition and sensor synchronization Zhou et al attempted to construct smart house (Zhou, Jiao, Chen, & Zhang, 2011). In the area of analysis and data transmission there are several unsolved issues (Mulvenna et al., 2011). In Ambient Assisted Living (AAL) majority of data processing algorithms are based on FIFO. Here data wait in a queue (Davis, Kollmann, Pollex, & Slomka, 2011). When there is an increase in data conventional data processing algorithms will not meet the requirements.

In accordance with the above description this paper concerns about data assembling, data separation and data examining. This architecture may reduce the network load and processing pressure of RAPD.

**Data Processing Overview:**

First-In First-Out (FIFO) based algorithms are considered as the conventional data processing algorithms so far. FIFO cannot be used when the data increases. The ODP is mainly concerned on assembling the data, separating the data, and examining.

The ODP (Optimal Data Processing) architecture handles the pressure of RAPD and decreases the network load. The Variable Neighborhood Search algorithm based solution is used for optimization problems (Hansen & Mladenovi, 2001). The whole algorithm concentrates on two parts: building alterable neighborhoods consistently and finding a favorable solution locally. The local search calls a procedure get the most out of solution and rebuild the neighborhood as stated by the solution. The computation complexity is increased and this procedure is time consuming. ODP is an algorithm based on Variable Neighborhood Search (VNS) architecture. Instead of local search a random point is chosen as the local best solution by IDP and according to the random point IDP is rebuilt. RAPD is ensured by the simplification of computation complexity.
In ODP architecture the solution space $\zeta = \{\xi_1, \xi_2, \xi_3, ..., \xi_n\}$ where beginning solution $\xi$ is acquired.

If $\forall \xi^* \in \zeta$ fulfills $f(\xi^*) \geq f(\xi)$, then the old optimal solution $\xi$ is replaced by new global optimal solution $\xi^*$ where $f(\xi)$ indicates utility function. In the same manner neighborhood structure of global optimal solution $\xi$ is $N(\xi)$. $N(\xi)$ is a subset of $\zeta(N(\xi) \in \zeta)$ like $\xi \in \zeta$. $N_k$ is indicated as $k$ specific neighborhood in a $k$ partitioned neighborhoods of a neighborhood structure.

These $k$ distinguished neighborhoods are indicated as $N_k(\xi)$ especially when $\xi$ is capable global optimal solution.

Through serialized set of matrix transformations $N_k$ produces different neighborhood structures. Unlike $\xi^*$, $\xi'$ is indicated as the local optimal solution ($\xi' \in \zeta$).

Another $\xi$ cannot meet with $f(\xi') < f(\xi)$. In ODP, $\xi'$ is generated randomly $N_k(\xi)$. The value of $k$ is returned, when $\xi'$ satisfying $f(\xi') > f(\xi)$ and according to $\xi'$ the algorithm rebuilds the neighborhood structure. Else, it shifts to next neighborhood(Wang et al., 2015). For location routing RVNS based algorithms are adopted(Hansen & Mladenovi, 2001).

**Optimal Data Processing Architecture (ODP):**

IDP is separated into three layers Data Assembly Layer (DAL), Data Separation Layer (DSL), Data Examining Layer (DEL). DAL receives and stores data. The data stored is processed in DSL. VDP is adopted in DEL for processing data through neighborhood structure. Fig.1 represents the ODP architecture which is three-layer structure. The remote sensing detected data is stored in DAL. These sensors collect the levels of carbon monoxide, sulfur dioxide, and nitrogen dioxide and send to local server.

The data from sensors are then stored in a buffer and send through DSL. In DSL through filter and classifier data is separated by two parts according to the risk function input. The static data part is stored in database and data stream part is stored in a relevant data set.

The abnormal data is discovered is discovered in DEL by ODP. According to the levels of data, it is set in different neighborhoods.

At last the data which is abnormal is sent to RAPD. The oldest data from the buffer is transferred to database by a mechanism which cleans the buffer. There should be a balanced data load in both static and dynamic datasets. Load balancing techniques were discussed to a great degree in heterogeneous and homogeneous environments. Static and Dynamic were the basic load balancing techniques (Dhinesh Babu & Venkata Krishna, 2013).
Data Assembly Layer (DAL):

In DAL the sensor architecture of remote air pollution detector is deployed. After collecting the data it is sent for processing. This is a local server which can communicate with RAPD. RAPD has a control over this local server and communicates when necessary. The sensor network which collects geospatial data is called geosensor network (Botts, Reed, Percivall, & Davidson, n.d.). These sensors are placed in real world. The measured data collected by geosensor networks are converted to digitalized values and gathered in DAL. The obtaining storing order of data and is shown in Fig.2. The recent data are sent at regular time intervals to local server. The data re-arrangement is needed since there is difference in obtaining order and collecting order, so timestamp is used. The collecting order and timestamping order needs to be same.

The main job of DAL is to align data and place in respective slot. It is done according to the timestamp.

A buffer is formed by many sections and a section is formed by similar slots. The buffer is similar to a 2D array which has two-division structure. 1st division is dividing buffer into many sections. Every section shares the work in parallel. In the 2nd division, slots with dissimilarities are divided. The obtaining order of data and receiver order are not same in local server. The received data is aligned according to timestamp. The data will be stored only if the timestamp is same. Each sensor has a space allocated in every slot. Once all the sensors store their data, it is sent to DSL.
Data Separation Layer (DSL):

The relevant data is stored in buffer through DSL. Primary sensors and Auxiliary sensors collect data in the form of VariableMatrix and ParameterMatrix (Wang et al., 2015). In an ideal situation only accurate data appears. As a result of network conditions some inaccurate data may mix with the input. The accuracy is judged by adding a filter for these data. As a reference point a relevant dataset is brought in. Primary and Auxiliary sensors collect large quantity of historical data and stored in relevant dataset. It is periodically updated by RAPD. According to air pollution status the data in DSL can be divided into stream data and static data. A Classifier is designed in DSL which includes library of stream data feature and static data feature. The stream data and static data can be separated by comparing feature library. The static data is transferred directly from buffer to database by Classifier and their spaces in buffer are also released. Later the data from primary sensors are divided into many levels.

Data Examining Layer (DEL):

As a composite function, Risk function reflects serious environment changes or pollution status at a particular region. Entirely different risk function should be acquired to determine level of gases in air pollution detector. According to parameters from auxiliary sensors the data from primary sensors are judged. The risk function can be planned based on environmental criteria. According to the present global solution ODP will build a neighborhood structure. The buffer is separated into particular neighborhood according to present solution. The data process interruption is lesser compared to
other algorithms like FIFO (First In First Out) and LRU (Least Recently Used). Taking rarely used data to database can serve space to latest data.

**Functional Evaluation:**

In Least Recently Used (LRU) buffer can be considered as stack. If the new value is not in buffer it is placed on the top moving other values down and least recently used value is removed from buffer. Nevertheless if new value is found in buffer, it is laid at the top. Position 1 is top of the stack in buffer and position $\beta$ is at the bottom. The position $i$ is most recently used item $i = 1, ..., \beta$. Let $\phi(\chi) = \Pr(X = \chi)$ where $\chi = (\chi_1, ..., \chi_\beta) \in \eta$ the set of buffer tenancy(Towsley, 1990).

$$\eta = \{ \chi: \sum_{i=1}^{\beta} 1(\chi_i = k) \leq D_k, 1 \leq k \leq K, \sum_{k=1}^{K} \sum_{i=1}^{\beta} 1(\chi_i = k) = B \}.$$

These probabilities meet the expectations of

$$\phi(\chi) = \sum_{i=1}^{\beta} \phi(\chi_2, ..., \chi_{\beta} l) \alpha_{\chi_l} + \sum_{i=1}^{\beta} \phi(\chi_2, ..., \chi_{\beta} l) \alpha_{\chi_l} l \in \{ \chi: \sum_{i=2}^{\beta} 1(\chi_i = k) < W_k \} (W_{\chi_1} - \sum_{i=2}^{\beta} 1(\chi_i = \chi_1) - \psi_{\chi,\chi_1}) / W_{\chi_1}$$

Here $\psi_{\chi,\chi_1} = 1$ if $\chi = \chi_1$ and 0 otherwise.

A substitute expression for $E(Y_k)$ using different approaches was derived by Flajolet et al which manipulates the properties of LRU(P.Flajolet, Gardy, 1992).

An item from partition $\kappa$ is placed at position $i$ and it is denoted by $\lambda_\kappa(i)$. If it is placed in position $i$ it takes 1 and 0 otherwise, $\lambda_\kappa(i) = 1(X_i = \kappa)$. Here $\lambda_\kappa = \sum_{i=1}^{\beta} \lambda_\kappa(i), \kappa = 1, ..., K$. Let $\rho_\kappa(i) = \Pr[X_\kappa(i) = 1], 1 \leq \kappa \leq K, 1 \leq i \leq Y$.

For the last entry from partition $\kappa$ the probability is $\rho_\kappa(1)$ and given by $\rho_\kappa(1) = \alpha_\kappa$. Let the average of number of items from partition $i$ can be represented by $y_k(i)$ which is present in $i$ positions of buffer. This is showed as

$$y_k(i) = \sum_{l=1}^{t} E[X_l(i)] = \sum_{l=1}^{t} \rho_\kappa(l), i = 1, 2, ..., Y, \tag{1}$$

And

$$\rho_\kappa(i) = r_\kappa(i-1) = \left( \frac{\alpha_\kappa(1-y_k(i-1))}{r(i-1)} \right)^+, i = 1, 2, ..., Y - 1; \kappa = 1, ..., K, \tag{2}$$

Where
Buffer hit probability can be calculated using recursions of above two equations.

Let us see FIFO buffer replacement now. In FIFO the buffers can be considered as queue with Position Y as head and Position 1 as tail. The buffer stays unchanged if a request item is in buffer. The request item is kept in position 1 if it is not in the queue and other items are moved within the buffer. The oldest item i.e., queue head is took out from the buffer.

The stationary probability $\phi(\chi)$ behavior is described in the following equations.

$$\phi(\chi) = \phi(\chi) \sum_{i=1}^{Y} \chi_i/W_{\chi_i} + \sum_{i=2}^{B} \phi(\chi_2, ..., \chi_Y, i)\alpha_{\chi_i} \in \left\{ \kappa: \sum_{i=2}^{B} 1(\chi_i = k) < W_{\chi_k} \right\} \frac{W_{\chi_k} - \sum_{i=2}^{Y} 1(\chi_i - \chi_1) - \psi_{\chi_i, 1}}{D_{\chi_1}},$$

$\chi \in \eta$

When a request is done an item is moved out of buffer of R being the probability. This is exactly alike to the probability.

$$R = \sum_{K=1}^{K} \alpha_{\kappa} \left( 1 - \frac{V[\lambda_{\kappa}]}{W_{\kappa}} \right)$$

From a partition $\kappa$ an item occupying any position is the probability of $V[\lambda_{\kappa}]$/Y. An item from partition $\kappa$ is added in the probability of $\alpha_{\kappa}(1 - V[\lambda_{\kappa}]/W_{\kappa})$. Comparing these two probabilities gives

$$\alpha_{\kappa}(1 - V[\lambda_{\kappa}]/W_{\kappa}) = \frac{W[\lambda_{\kappa}]}{Y} R.$$  

(Towsley, 1990) manipulated with some algebra and get

$$V[\lambda_{\kappa}] = \frac{W_{\kappa}}{1 + \frac{RW_{\kappa}}{a_{\kappa}Y}}$$

$V[\lambda_{\kappa}]$ is got by solving equations (3) and (5) in an iterative manner. According to Towsley et al the following algorithm can be better suited.

**step 1:** Assign $R := 1$; $\lambda$ sum := 0;

**step 2:** Iterate while ($|\lambda$ sum $- Y| > \psi$)

$$V[\lambda_{\kappa}] := \frac{W_{\kappa}}{1 + \frac{RW_{\kappa}}{a_{\kappa}Y}}; \quad \kappa = 1,2, ..., K$$

$$\lambda$$ sum := $\sum_{K=1}^{K} V[\lambda_{\kappa}]$

$$R := R \times \lambda$$ sum / $Y$;

**step 3:** $h_{\kappa} := \frac{V[\lambda_{\kappa}]}{W_{\kappa}}, \kappa = 1,2, ..., K$
In this section the performances are compared with FIFO, LRU (Dan & Towsley, 1990) and Optimal data processing model through simulation. In ODP by reorganizing the neighborhood data can be dealt. Abnormal data can be found in this case. In this simulation City Pulse Traffic Data streams are taken. To discover dissimilar environmental factors various sensors are deployed. The emission density of carbon monoxide (CO), Sulfur dioxide (SO₂), and Nitrogen Dioxide (NO₂) was represented in Fig. 3, 5, and 7 respectively. The emission source was got from United States Environmental Protection Agency. In Fig. 4, 6, and 8 the data process interruption was compared with FIFO, LRU, and ODP. It is noted that process interruption is decreased when the buffer size is increased. The process interruption is less in ODP when compared with FIFO and LRU.

**Fig. 3 Carbon Monoxide Emission**

![Carbon Monoxide Emission](image)

**Fig. 4 Carbon Monoxide Data Process**

![Carbon Monoxide Data Process](image)
Fig. 5 Sulfur Dioxide Emission

Fig. 6 Sulfur Dioxide Data Process Interruption

Fig. 7 Nitrogen Dioxide Emission Density
Table.1 is formed from the Clean Air Act 1970 in the Federal Register, vol. 36, part 2, p. 8187, 30th April, 1971, and some details are from pages 27 and 28. It represents the air quality standard. Some Air monitoring instruments are intensive and takes longer period for reading independently(Bibbero, 1971). Rapid measuring devices are essential for warning pollutants of complex concentrations.

**Table.1 Air pollution Effects:**

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>Concentration</th>
<th>Item Affected</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbon monoxide</td>
<td>10-15ppm</td>
<td>Health</td>
<td>8 hours</td>
</tr>
<tr>
<td></td>
<td>8-14ppm</td>
<td>Health</td>
<td>week</td>
</tr>
<tr>
<td>Sulfur Dioxide</td>
<td>0.11ppm</td>
<td>Health</td>
<td>3-4 days</td>
</tr>
<tr>
<td></td>
<td>0.04ppm</td>
<td>Health</td>
<td>mean</td>
</tr>
<tr>
<td></td>
<td>0.03ppm</td>
<td>Vegetation</td>
<td>mean</td>
</tr>
<tr>
<td>Photochemical oxidants</td>
<td>0.5ppm</td>
<td>Vegetation</td>
<td>4 hours</td>
</tr>
<tr>
<td></td>
<td>0.03-0.3ppm</td>
<td>Health</td>
<td>1 hour</td>
</tr>
<tr>
<td></td>
<td>0.1ppm</td>
<td>Health (Eye irritation) peak</td>
<td></td>
</tr>
<tr>
<td>Particulates</td>
<td>300µg/m³</td>
<td>Health</td>
<td>24 hours</td>
</tr>
<tr>
<td></td>
<td>80µg/m³</td>
<td>Health</td>
<td>mean</td>
</tr>
<tr>
<td></td>
<td>60µg/m³</td>
<td>Materials</td>
<td>mean</td>
</tr>
</tbody>
</table>

**Fig.8 Nitrogen Dioxide Data Process Interruption**

Table.1 is formed from the Clean Air Act 1970 in the Federal Register, vol. 36, part 2, p. 8187, 30th April, 1971, and some details are from pages 27 and 28. It represents the air quality standard. Some Air monitoring instruments are intensive and takes longer period for reading independently(Bibbero, 1971). Rapid measuring devices are essential for warning pollutants of complex concentrations.
Conclusion:
In this paper, major issues and challenges of massive data processing in IOT environment were analyzed. Three-layered Optimal Data Processing model was designed to solve the data processing issues in IOT environment. The Variable Neighborhood Search algorithm based solution is used for optimization problems. The data process interruption is analyzed and compared with ODP, FIFO and LRU. The simulation results show that data process interruption decreases in FIFO and LRU when the buffer size is increased. Here we have used static dataset for comparing the data process interruption. The Air Pollution effects and issues were discussed with carbon monoxide(\(\text{CO}\)), Sulfur dioxide(\(\text{SO}_2\)), and Nitrogen Dioxide(\(\text{NO}_2\)) pollutants. ODP model can be used for dynamic datasets like Health care monitoring, Environmental sensing, Forest fire detection, Landslide detection, Water quality monitoring, Natural disaster prevention, and so on.

References:
18. http://doi.org/10.1109/jsac.2013.sup.0513005


**Corresponding Author:**

Nivash J P*,

**Email:** jeevprabnivash@gmail.com