Multi-Layer Based Data Aggregation Algorithm for Convergence Platform of IoT and Cloud Computing

M. Mamun-Ibn-Abdullah[a], M. Ali[b], M. Humayun Kabir[a]

[a]IoT Research and Innovation Lab, Dept. of Electrical and Electronic Engineering, Islamic University, Kushtia, Bangladesh

[b]Dept. of Electrical and Electronic Engineering, Rabindra Maitree University, Kushtia, Bangladesh

humayun@eee.iu.ac.bd

Abstract

Sensor Networks (SN) are deployed in smart domain to sense the environment which is essential to provide the services according to the users need. Hundreds or sometimes thousands of sensors are involved in sensor networks for monitoring the target phenomenon. Large scale of sensory data have to be handle by the sensor network which create several problems such as waste of sensors energy, data redundancy. To overcome these deficiencies one most practice solution is data aggregation which can effectively decrease the massive amount of data generated in SNs by lessening occurrence in the sensing data. The aim of this method is to lessen the massive use of data generated by surrounding nodes, thus saving network energy and providing valuable information for the end user. The effectiveness of any data aggregation technique is largely dependent on topology of the network. Among the various network topologies clustering is preferred as it provides better controllability, scalability and network maintenance phenomenon. In this research, a data aggregation technique is proposed based on Periodic Sensor Network (PSN) which achieved aggregation of data at two layers: the sensor nodes layer and the cluster head layer. In sensor node layer set similarity function is used for checking the redundant data for each sensor node whereas Euclidean distance function is utilized in cluster head layer for discarding the redundancy of data between different sensor nodes. This aggregation technique is implemented in smart home where sensor network is deployed to capture environment related information (temperature, moisture, light, H_2 level). Collected information is analyzed using ThinkSpeak cloud platform. For performance evaluation amount of aggregated data, number of pairs of redundant data, energy consumption, data latency, and data accuracy are analyzed and compared with the other state-of-art techniques. The result shows the important improvement of the performance of sensor networks.

Keywords: Data Aggregation Technique, Network Topology, Cluster Based Network, Periodic Sensor Network (PSN), Cloud Computing, IoT.

1. Introduction

The technological advancement of sensor networks composed of small and cost effective sensing devices make it possible to equip wireless radio transceivers for remote monitoring. The key advantage of wireless sensor node is it does not use wire infrastructure for power and network connections. These sensor nodes can be used to monitor the environment by collecting information from their surroundings utilizing a base station which serves as a data repository. This type of sensor network is called Wireless Sensor Network (WSN). There are many issues happen when the WSN deals with vast number of data such as enough energy consumption, creating data redundancy and so on.

Large scale of sensory data have to be handled by the WSN which create several problems such as waste of sensors energy for data acquisition, need large space for storage, and wide wireless link[1]. Data aggregation is more practical solution to overcome these deficiencies. Data aggregation is an effective way to reduce the large amount of data generated on sensor nodes by eliminating unnecessary sensing data. The aim of this method is to lessen the massive use of data generated by surrounding nodes, thus improving data latency as well as conserving network energy and providing more abstract information for the end user.

WSN can be of two types based on the network topology: flat network and cluster based network. Among them cluster based WSN possesses several benefits, better communication network, efficient topology management, reduce delays etc. [2]. Each node in WSN sends data to its neighboring node closed to sink in a multi-hop fashion. While collecting data closely placed nodes can experience the same data which causes lack of electric power. Such a technique cannot be considered as energy efficient. A proper aggregation technique can be introduced to make it energy efficient. In cluster based WSN each sensor node sends data to cluster head after that data is routed to sink through cluster head.

The cluster head can perform aggregation by receiving raw data before sending it to sink. In a cluster based WSN several points have to be settled down: i) the number of cluster, ii) the number of nodes in each cluster, iii) the number of cluster head which satisfied optimized performance parameter. In this research, a two-layer based data aggregation technique using cluster based Periodic Sensor Network (PSN) is proposed. At bottom layer called sensor layer which aggregates utilizing set similarity function used whereas Euclidean distance function is utilized in top layer called cluster head layer. The performance of the aggregation technique is analyzed based on the energy consumption, the data latency and accuracy.

The rest of the manuscript is decorated as follows: Section 2 explain the related works. Section 3 presents data aggregation techniques in Internet of Things (IoT). Section 4 describes proposed data aggregation technique. Section 5 evaluates the proposed data aggregation technique and finally, Section 6 concludes the research with mentioning future work.

2. Literature Review

Nowadays a number of WSN applications exist where the amount of the sensory data has exceeded several peta bytes yearly [4]. High energy consumption and complex data analysis issued in these applications. To overcome these problem researchers have highlighted to the data aggregation method in WSNs. As sensors, are generally operated by battery energy which is limited. In order to increasing network life time data aggregation is very important to operate in WSNs for decreasing the data transmission. Recently data aggregation has been well studied in WSNs. Mainly the performance of data aggregation technique depends on the network topology. In addition that, the researchers have put many network topologies for WSNs such as Tree [4,13,14], Cluster [5,8-12], Chain [6,15,16] or Structure Free [7,17,18] based topology.

Among the various network topologies clustering is preferred as it provides better controllability, scalability and network maintenance phenomenon. Moreover existing data aggregation techniques focus on CHs only. In this research, we propose a data aggregation technique based on cluster based Periodic Sensor Network (PSN) which achieved aggregation of data at two layers: firstly at the sensor nodes layer and secondly at the cluster head layer. In first layer set similarity function is used whereas Euclidean distance function is utilized in second layer. This aggregation technique is implemented in smart home where sensor network is deployed to capture environment related information (temperature, humidity, light, H_2 level). Collected information is analyzed using ThingSpeak cloud platform. The performance of the aggregation technique is evaluated based on the network energy consumption, the data latency and data accuracy.

3. Proposed Data Aggregation Technique

Data aggregation techniques used in IoT can be categorized in five groups: In-network, tree based, cluster based, centralized, and structure-less data aggregation. The aim of data aggregation method is to reduce the redundant data transmission for the expansion of life time energy in WSN. Cluster based network topology is focused here and it is divided into several clusters.

Clusters are formed on the basis of similarity of sensor nodes. For sending aggregated data to base station a cluster head is formed in each cluster. This cluster head can directly communicate with base station by using radio transmission. Sensed data goes to the destination through CHs by using cluster based sensor network topology. Sensor node collects data in periodically and transmitted from sensor node to CHs using single-hop communication.

Then, our proposed data aggregation technique works in two levels: the first one at the sensor nodes level, and the second at the CHs level.

Periodic Sensor Network (PSN) is defined as a wireless sensor network deployed on the purpose of periodic monitoring where periodic updates are sent to the sink from the PSN, based on the most recent information

$$M_i \bigcap_s M_j = \left\{ (y_i, y_j) \in M_i \times \frac{M_j}{link(y_i, y_j)} = 1 \right\}$$

sensed from the physical parameter. PSNs are typically arrays of sensor nodes interconnected using a radio communication network which allow their data to reach the sink. They are used for applications where certain conditions or processes need to be monitored constantly, such as the temperature in a conditioned space or pressure in a process pipeline. In this paper we consider PSN, where sensor nodes monitor a given phenomenon and send notifications and measurements back to the sink at each period p. In this case, we can notice the huge amount of data generated and sent to the sink.

Furthermore, a significant amount of redundant data is likely delivered, particularly in case of dense networks. Subsequently, data aggregation for periodic applications becomes a necessity in order to reduce the size of data and save energy. Data aggregation consists in eliminating the inherent redundancy in raw data collected from the sensor nodes, minimizing the number of transmissions to the sink and thus saving energy. In PSN, each period p is divided into τ equal time slots as follows: $p = [s_1, s_2, \ldots, s_\tau]$. At each slot s_j, each sensor S_i captures a new measure m_{ij}, then, it forms a vector of measures during the period p as follows: $M_i = [m_{i1}, m_{i2}, \ldots, m_{i\tau}]$.

This section recalls the data aggregation scheme proposed and that will be enhanced in this paper. The method proposed in works in two phases, the first one at the nodes level called local aggregation and the second at the aggregator's level. At each period p each node sends its aggregated data set to its proper aggregator which subsequently aggregates all data sets coming from different sensor nodes and sends them to the sink.

A. First Layer: local aggregation. In this tier of aggregation, the idea is to identify similar data measurements captured by a sensor node i during a period p. In PSN, each sensor node i takes a new measurement y_{is} at regular time interval called slot s. It is likely that a sensor node takes the same (or very similar) measurements several times especially when s is too short. During a period p, each node forms a new set of captured measurements M_i and sends it to the aggregator. In order to reduce the size of the set M_i, a similarity function between measurements and a frequency of a measure are defined as follows:

Definition 1: (Similar Function): We define the similar function between two measurements as:

$$Similar\left(m_{i_j}, m_{i_k}\right) = \begin{cases} 1 & if \ \left\| m_{i_j} - m_{i_k} \right\| \leq \in, \\ 0 & otherwise \end{cases}$$

Where m_{ij} and $m_{ik} \in M_i$ and ε is a threshold determined by the application. Furthermore, two measures are similar if and only if their similar function is equal to 1.

Definition 2: (Measure's Weight, wgt (m_{ij})): The weight of a measurement m_{ij} is defined as the number of similar measures (according to the Similar function) to m_{ij} in the same vector M_i. Based on the notations defined above, we describe the local aggregation phase which is run by the nodes themselves at each period in the following manner: for each new captured measurement, a sensor node S_i searches for similarities of the new taken measurement. If a similar measurement is found, it deletes the new one and increments the corresponding weight by 1, else it adds the new measure to the set and initializes its weight to 1.1 After applying local aggregation, S_i will transform the initial vector of measures, M_i, to a set of measures, M_i, associated to their corresponding weights as follows: $M_i = \{(m_{i1}, wgt(m_{i1})), (m_{i2}, wgt(m_{i2})), \ldots, (m_{ik}, wgt(m_{ik}))\}$, where $k \leq \tau$.

B. Second Layer:

In this section, we propose a new method to search redundant data sets generated by the sensors using the distance functions. Distance functions are an important method that can find duplicated data sets by searching dissimilarities between these sets. Hence, a great number of distance functions have been proposed in the literature [22,23]. In this paper, we are interested in two distance functions that are widely used in various domains: Euclidean and Cosine distances. Let us consider two data sets M′ᵢ and M′ⱼ, generated by the sensor nodes S_i and S_j respectively, at the period p as follows:

$M_i^{'} = \left\{\left(m_{i_1}^{/}, wgt(m_{i_1}^{/})\right), \left(m_{i_2}^{/}, wgt(m_{i_2}^{/})\right), \ldots, \left(m_{i_1}^{/}, wgt(m_{i_1}^{/})\right)\right\}$ and $M_j^{'} = \left\{\left(m_{j_1}^{/}, wgt(m_{j_1}^{/})\right), \left(m_{j_2}^{/}, wgt(m_{j_2}^{/})\right), \ldots, \left(m_{j_{k_j}}^{/}, wgt\left(m_{j_{k_j}}^{/}\right)\right)\right\}$ where $|M_i'| = k_i$ and $|M_j'| = k_j$ Therefore, M_i' and M_j' are considered redundant if the calculated distance between them is less than a threshold (t_d) as follows:

$$Dist\left(M_i', M_j'\right) \leq t_d$$

In mathematics, the Euclidean distance is the ordinary distance, e.g. straight line distance, between two points, sensor objects. It is used in many applications and domains, such as computer vision and prevention of identity theft. Furthermore, the Euclidean distance is already used in WSN during the deployment phase in terms of sensors' localization and inter-sensors distance estimations. In this paper, we use the Euclidean distance on the data sets collected by sensors while adapting it to take into account the measures' weights. In general, the Euclidian distance (E_d) between two datasets M_i and M_j before applying the local aggregation, is given by:

$$E_d(M_i, M_j) = \sqrt{\sum_{k=1}^{\tau} \left(m_{i_k} - m_{j_k}\right)^2}$$

Where, $m_{i_k} \in M_i$ and $m_{j_k} \in M_j$, M_i and M_j are said to be redundant if E_d (M_i,M_j)≤t_d where t_d is a threshold determined by the application.

4. Experimental results and Evaluation

We have developed a control and monitoring system using Raspberry Pi in smart home. Different types of sensors (motion, temperature, humidity, light, H_2) are deployed in the environment to collect information. Collected information's are stored in ThingSpeak cloud platform for remote monitoring. Proposed data aggregation algorithm is implemented in this smart home to evaluate its performance and Figure 1. shows test bed of proposed data aggregation technique.

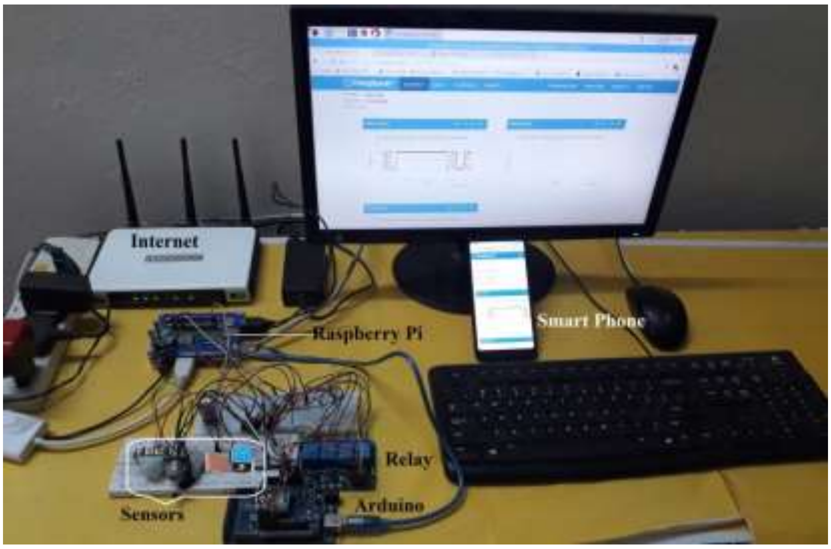


Figure 1. Test-bed of proposed data aggregation technique.

Throughout this evaluation work we have taken different values for different parameters. Table 1 shows the details about different parameters.

Table 1. Performance Measure Parameter.

Parameters Name	Description	Values
Threshold (ε)	Defined in Similar function	0.03, 0.05,0.07,0.1
Period(τ)	The number of sensor measurement taken by each sensor node during a period	200,500,1000
Distance Threshold(t_d)	The distance threshold	0.35,0.4,0.45,0.5

During the aggregation, in first layer using similar function each sensor node searches the similarity between measures captured at each period and assigns for each measure its weight. Therefore, the result of the aggregation in this phase depends on the chosen threshold ε, the number of the collected measures in period τ and the changes in the monitored condition.

Figure. 2 shows the percentage of remaining data, or aggregated data, which will be sent to the CH, with and without applying the aggregation in first layer at the sensors layer. At each period, the amount of data collected by each sensor is reduced at least by 78% (and up to 96%) after applying the aggregation phase.

Otherwise, the sensor node sends all the collected data, e.g.100%, without applying the aggregation phase. Therefore, our technique can successfully eliminate redundant measures at each period and reduce the amount of data sent to the CH. We can also observe that, with the local aggregation phase, data redundancy among data increases when τ or ε increases. This is because the similar function will find more similar measures to be eliminated in each period.

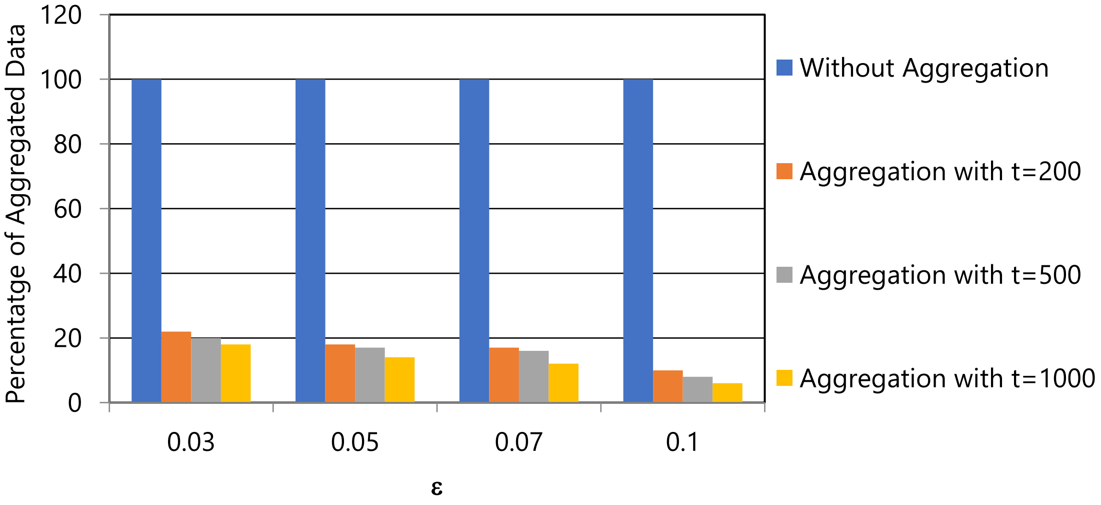
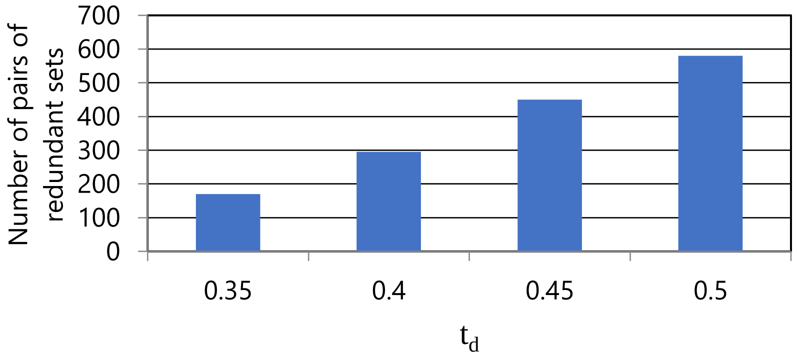


Figure 2. Percentage of data after applying aggregation technique

When receiving all the sets from its member nodes at the end of each period, CH applies the second aggregation level in order to find all pairs of redundant sets.

Figure 3 shows the number of pairs of redundant sets obtained at each period when applying Euclidean distance technique in CH layer

Figure 3. Number of pairs of redundant sets at each period τ=500, ε=0.07.



In this section, our objective is to study the energy consumption at the sensor nodes and CH levels. In sensor networks, energy consumption is highly dependent on the amount of data sent and received. Figure 4 shows the energy consumption comparison with and without applying the local aggregation phase by each sensor node and when varying τ and ε. Since the local aggregation significantly reduces the redundancy among data collected by the sensor node, it allows it to proportionally save its energy when transmitting its data to the CH at each period. It is important to notice that our technique can save from 75% up to 95% of the energy of a sensor node.

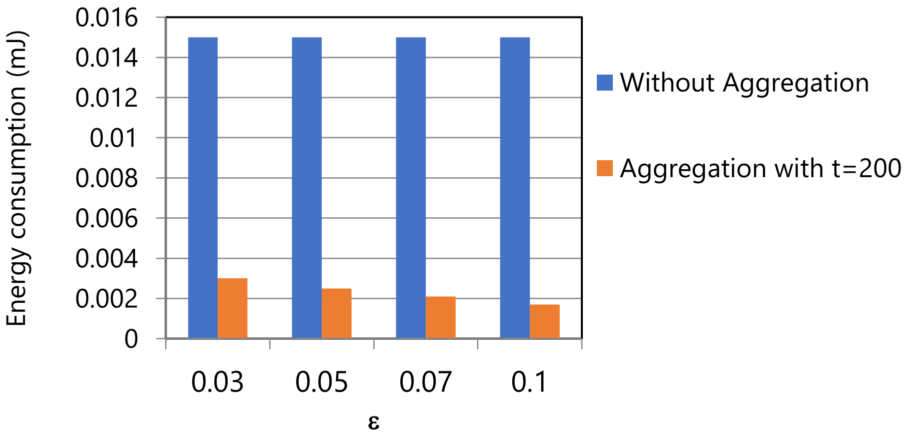


Figure 4: Energy consumption in each sensor node.

In this section, we compare the execution time required for the three data aggregation methods used in our technique when varying t_d, ε and τ respectively as shown in Figure 5.

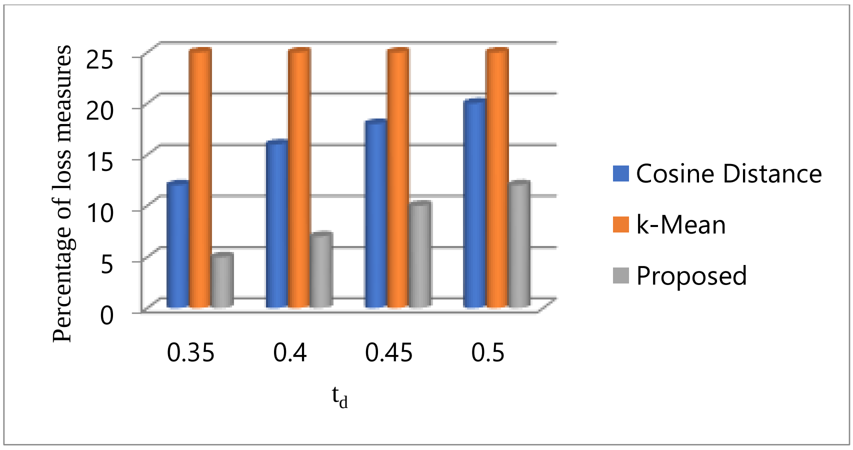


Figure 5. Execution time at the CH with τ=500 and ε=0.07.

Data accuracy is an important factor in WSNs which represents the measure “loss rate”. In our simulation, data accuracy has been evaluated based on the percentage of loss measures at the CH; in other words, we divided the number of measures taken by the sensor nodes whose values (or similar values) do not reach the sink, after applying each aggregation method, over the whole measures collected by the sensors at each period. Figure. 6 shows the results of data accuracy for the data aggregation functions used in our technique for different values of t_d, ε and τ. We can notice that the results for data accuracy, for the proposed method is up to 91.5%, Cosine distance is up to 83.5%, and k-Mean algorithm up to 75%.

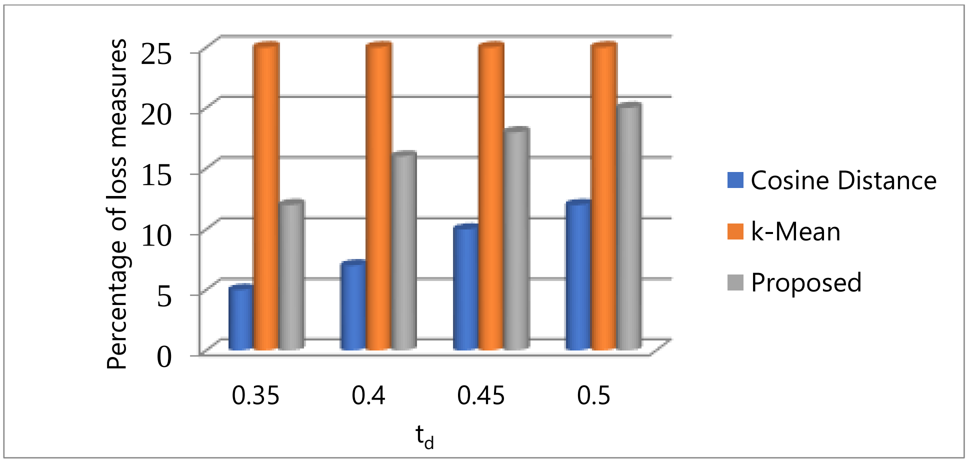


Figure 6: Data accuracy with τ=500 and ε=0.07.

To summarize this section, Table 2 shows the flexibility of each method regarding energy consumption, data latency and accuracy, and complexity of the method at the CHs.

Table 2. Comparison between different Aggregation methods.

Method	Energy Consumption Conserving	Data Latency	Data Accuracy
Cosine Distance	Good	Medium	Medium
k-means Algorithm [24]	Good	Very good	Low
Proposed Method (Set similarity function with Euclidean Distance)	Very good	Medium	Good

From the table it shows explicitly that the proposed method can conserve best energy consumption among the three methods. For data latency its performance is medium. Although k-means algorithm provides very good data latency but its data accuracy is low.

5. Conclusion

The main goal of data aggregation algorithms is to collect and aggregate data in an energy efficient way to improve network lifetime. In the clustering classification, the cluster head decision is a major challenge. If the network as a whole is taken, then the power consumption can be optimized by rotating its cluster head inside a separate cluster. An important concern in the energy conservation network, especially as the energy conservation of the cluster head in the cluster-tree network has to be higher because of the different operations, which they control over the network. In this research, we propose a data aggregation technique based on Periodic Sensor Network (PSN) which achieved aggregation of data at two layers. Firstly at the sensor nodes layer, and secondly at the cluster head layer. In first layer set similarity function is used whereas Euclidean distance function is utilized in second layer. This aggregation technique is implemented in smart home where sensor network is deployed to capture environment related information (temperature, moisture, light, H_2 level). Collected information is analyzed using ThingSpeak cloud platform. The performance of the aggregation strategy is analyzed based on the energy consumption, the data latency and accuracy. The result shows how these methods can significantly improve the performance of sensor networks. In future we can implement machine learning techniques to select the different parameters: Threshold (ε), Period (τ), Distance Threshold (t_d)

Acknowledgement: This research work supported by University Grants Commission (UGC) of Bangladesh, in Financial Year 2018-2019 through Islamic University, Kushtia-7003, Bangladesh.

References

1. Beom-Su Kim, Ki-Il Kim, Babar Shah, Francis Chow and Kyong Hoon Kim, “Wireless Sensor Networks for Big Data Systems,” Sensors 2019, 19, 1565.

2. Asim Zeb, A. K. M. Muzahidul Islam, Mahdi Zareei, Ishtiak Al Mamoon, Nafees Mansoor, Sabariah Baharun, Yoshiaki Katayama, and Shozo Komaki “Clustering Analysis in Wireless Sensor Networks: The Ambit of Performance Metrics and Schemes Taxonomy,” International Journal of Distributed Sensor Networks, 2016, pp.1-24.

3. T. Zhu, S. Cheng, Z. Cai, and J. Li, “Critical Data Points Retrieving Method for Big Sensory Data in Wireless Sensor Networks,” EURASIP Journal on Wireless Communications and Networking, Vol.2016, No.1, pp.1–14, 2016.

4. C. Wang, L. Xing, V. M. Vokkarane, and Y. Sun, “Reliability of Wireless Sensor Networks with Tree Topology,” International Journal of Performability Engineering, Vol. 8, No.2, pp.213–216, 2012.

5. K. R. Bhakare, R. Krishna, and S. Bhakare, “An Energy-Efficient Grid based Clustering Topology for a Wireless Sensor Network,” International Journal of Computer Applications, Vol.39, No.14, 2012.

6. H. A. Marhoon, M. Mahmuddin, and S. A. Nor, “Chain-based Routing Protocols in Wireless Sensor Networks: A survey,” ARPN Journal of Engineering and Applied Sciences, Vol. 10, No. 3, pp. 1389–1398, 201.

7. C. Chao, and T. Hsiao, “Design of Structure-Free and Energy-Balanced Data Aggregation in Wireless Sensor Networks,” Journal of Network and Computer Applications, Vol.17, pp.229–239, 2014.

8. X. Kui, J. Wang, S. Zhang, and J. Cao, “Energy balanced Clustering Data Collection based on Dominating set in Wireless Sensor Networks,” Ad Hoc and Sensor Wireless Networks Journal, Vol.24, No.3-4, pp.199–217, 2015.

9. P. Zou and Y. Liu, "A Data-Aggregation Scheme for WSN based on Optimal Weight Allocation," Journal of Networks, Vol.9, No.1, pp.100–107, 2014.

10. M. Shanmukhi and O. Ramanaiah, "Cluster-based Comb-Needle Model for Energy-Efficient Data Aggregation in Wireless Sensor Networks," Applications and Innovations in Mobile Computing (AIMoC), pp. 42–47, 2015.

11. T. Du, Z. Qu, Q. Guo, and S. Qu, "A High Efficient and Real Time Data Aggregation Scheme for WSNs," International Journal of Distributed Sensor Networks, Vol.2015, No.2015, pp.11, 2015.

12. H. Harb, A. Makhoul, M. Medlej, and R. Couturier, "An Aggregation and Transmission Protocol for Conserving Energy in Periodic Sensor Networks," 24th IEEE International Conference Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2015, pp.134–139.

13. A. Norouzi, F. S. Babamir, and Z.Orman, "A Tree based Data Aggregation Scheme for Wireless Sensor Networks using GA," Wireless Sensor Network, Vol. 4, No. 8, pp. 191–196, 2012.

14. Y. Lu, I. Comsa, P. Kuonen, and B. Hirsbrunner, "Dynamic Data Aggregation Protocol based on Multiple Objective Tree in Wireless Sensor Networks," Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), IEEE, pp.1–7, 2015.

15. Y.-K. Chiang, N.-C.Wang, and C.-H. Hsieh, "A Cycle-based Data Aggregation Scheme for Grid-based Wireless Sensor Networks," Sensors, Vol.14, No.5, pp.8447–8464, 2014.

16. N. Javaid, M. R. Jafri, Z. A. Khan, N. Alrajeh, M. Imran, and A. Vasilakos, "Chain-based Communication in Cylindrical Underwater Wireless Sensor Networks," Sensors, Vol.15, pp.3625–3649, 2015.

17. J. Luo and J. Cai, "A Dynamic Virtual Force-Based Data Aggregation Algorithm for Wireless Sensor Networks," International Journal of Distributed Sensor Networks, Vol.2015, No.2015, pp.7, 2015.

18. M. K. Al-Azzawi, J. Luo, and R. Li, "Virtual Cluster Model in Clustered Wireless Sensor Network Using Cuckoo Inspired Metaheuristic Algorithm," International Journal of Hybrid Information Technology, Vol.8, No.4, pp.133–146, 2015.

19. H. Natarajan and S. Selvaraj, "A Fuzzy based Predictive Cluster Head Selection Scheme for Wireless Sensor Networks," In Proc. of the 8th International Conference on Sensing Technology, pp.560–567, 2014.

20. J. M. Bahi, A. Makhoul, and M. Medlej, "An Optimized In-network Aggregation Scheme for Data Collection in Periodic Sensor Networks," Ad-hoc, Mobile, and Wireless Networks: 11th International Conference, ADHOC-NOW 2012, Belgrade, Serbia, July 9-11, 2012. Proceedings, pp.153–166, 2012.

21. D. Kumar, "Performance Analysis of Energy Efficient Clustering Protocols for Maximising Lifetime of Wireless Sensor Networks," IET Wireless Sensor System, Vol.4, No.1, pp.9–16, 2014.

22. M. Friedmana, M. Lastb, Y. Makoverb, and A. Kandelc, "Anomaly Detection in Web Documents using Crisp and Fuzzy-based Cosine Clustering Methodology," Information Sciences, Vol.177, pp.467–475, 2007.

23. J. Ye, "Cosine Similarity Measures for Intuitionistic Fuzzy Sets and Their Applications," Mathematical and Computer Modelling, Vol.53, No.12, pp. 91–97, 2011.

24. H. Harb, A. Makhoul, and R. Couturier, "An Enhanced k-means and Anova-based Clustering Approach for Similarity Aggregation in Underwater Wireless Sensor Networks," IEEE Sensors Journal, Vol.15, No.10, pp. 5483–5493, 2015.