Data Correlation-Based Clustering Algorithm in Wireless Sensor Networks

Myungho Yeo[1], Dongmin Seo[2] and Jaesoo Yoo[1]
[1] Dept. of Information and Communication Engineering, Chungbuk National University
Cheongju, South Korea
[e-mail: mhyeo@netdb.cbnu.ac.kr, yjs@chungbuk.ac.kr]
[2] Dept. of Computer Science, Korea Advanced Institute of Science and Technology,
Daejeon, South Korea
[e-mail: dmseo@kaist.ac.kr]
*Corresponding author: Jaesoo Yoo

Received March 21, 2009; revised May 14, 2009; accepted May 30, 2009; published June 22, 2009

Abstract

Many types of sensor data exhibit strong correlation in both space and time. Both temporal and spatial suppressions provide opportunities for reducing the energy cost of sensor data collection. Unfortunately, existing clustering algorithms are difficult to utilize the spatial or temporal opportunities, because they just organize clusters based on the distribution of sensor nodes or the network topology but not on the correlation of sensor data. In this paper, we propose a novel clustering algorithm based on the correlation of sensor data. We modify the advertisement sub-phase and TDMA schedule scheme to organize clusters by adjacent sensor nodes which have similar readings. Also, we propose a spatio-temporal suppression scheme for our clustering algorithm. In order to show the superiority of our clustering algorithm, we compare it with the existing suppression algorithms in terms of the lifetime of the sensor network and the size of data which have been collected in the base station. As a result, our experimental results show that the size of data is reduced and the whole network lifetime is prolonged.

Keywords: Wireless sensor networks, clustering, correlation, energy-efficient, compression

This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korean government (MOST) (No. R01-2006-000-1080900) and the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (The Regional Research Universities Program/Chungbuk BIT (Research-Oriented University Consortium).

DOI: 10.3837/tiis.2009.03.007

1. Introduction

Recently, wireless sensor networks have found their way into a wide variety of applications and systems with vastly varying requirements and characteristics, including environmental monitoring, smart spaces, medical applications, and precision agriculture [1][2][3]. These networks are mainly to be used for the systematic gathering of useful information related to the surrounding environment (e.g. temperature, humidity, seismic and acoustic data, etc.)

In general, many types of sensor data exhibit strong correlation in both space and time. It has created new opportunities for data collection in a variety of environmental and industrial scenarios, where we expect data to be temporally and spatially correlated. Many researchers have investigated techniques for reducing the energy cost of collecting sensor data with these opportunities. Especially, [4] defines the term “suppression” to refer generally to query-independent techniques for reducing the cost of reporting changes in sensor data. The suppression techniques are divided into spatial and temporal suppressions respectively.

In the case of temporal suppression, a node does not transmit a value that is not changed since last reported [5]. The base station, in turn, assumes any unreported data remain unchanged. This scheme is effective when data seldom change. On the other hand, when a number of changes occur in a particular area, all nodes must report them to the base station. As a result, it causes a high cost.

In the case of spatial suppression, a node suppresses a value that is identical to those of its neighboring nodes. [6] suggests a more effective scheme based on averages, in which all nodes attempt to report their data at different time slots during a logical time step. Nodes overhear reports from their neighbors. When a node’s slot comes up, it first computes the average of all data overheard so far. If its value equals this average, its report is suppressed. Therefore, if a neighborhood contains all nodes with similar data, not all will be sent. To accurately derive the value of a node, the base station must know which neighbors were averaged when suppression was triggered. In [7], nodes are organized into clusters; all nodes in a cluster send their data to a cluster head node; suppression happens at the cluster head.

Clustering facilitates the distribution of control over the network [8][9]. Clustering saves energy and reduces network contention by enabling locality of communication. Nodes communicate their data over shorter distances to their respective cluster heads [10][11][12][13]. The cluster heads aggregate these data into smaller set of meaningful information, not of all nodes, but only the cluster heads need to communicate far distances to the base station. However, the cluster shapes are fixed based on a network topology and are not tailored to actual correlation patterns, because they just organize clusters based on the distribution of sensor nodes or the network topology but not the similarity of sensor data.

In this paper, we propose Data Correlation-based Clustering scheme (DCC), a novel clustering algorithm based on the similarity of sensor data without any assumptions such as high-correlated regions. DCC aims to improve suppression rate by two ways. One way is how to suppress sensor reading by a cluster head. Another way is how to organize the network topology. Simulation results show that DCC has better performance than the existing schemes in terms of the network lifetime and the compression of sensor data.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 states our motivation. Section 4 describes our proposed technique for a data correlation-based clustering and the enhanced suppression scheme. Section 5 discusses the simulation results. Finally, we discuss conclusions and future works in Section 6.

2. Related Work

2.1 Suppression Algorithms

There are several opportunities for reducing the cost of monitoring and reporting a group of sensor readings [4]. First, readings often change slowly over time and do not usually deviate from their expected values. For example, in industrial applications, a change in some monitored quantity may be a warning of impending failure and occurs rarely. Second, readings are often spatially correlated. In habitat monitoring, if one sound sensor detects a loud cry from an animal, its neighboring sensors likely hear the same sound. In general, many types of sensor readings exhibit strong correlation in both space and time. For example, if one sensor detects high soil moisture due to precipitation, sensors in neighboring areas with similar soil composition and elevation likely observe similar moisture readings. Furthermore, soil moisture usually stays at a saturated level during precipitation and tapers off to a normal level afterward. Many researches use the term suppression to refer generally to query independent techniques for reducing the cost of reporting changes in sensor values.

2.2 Clustering Algorithms

Many clustering algorithms have been proposed for sensor networks. In clustering algorithms, the groups of neighboring sensor nodes form clusters. In each cluster, one representative node called a cluster head collects sensor readings from its members and sends the collected data to the base station. In LEACH [13], the cluster head is periodically rotated among the sensor nodes to balance energy dissipation. Sensor nodes that are not elected as cluster heads choose their closest cluster head to join. LEACH assumes that any two nodes can communicate with each other directly. HEED [10] extends LEACH by incorporating communication range limits and cost information. In HEED, the initial probability for each sensor node to become a cluster head is dependent on its residual energy. Later on, sensor nodes that are not covered by any cluster heads double their probability of becoming a cluster head. This procedure iterates until all sensor nodes are covered by at least one head. Sensor nodes join cluster heads that have the lowest cost within their range. [14] proposed a simple correlation-based clustering algorithm considering non-uniform correlation distribution. The non-uniform correlation distribution is just defined as a set of Highly Correlated Regions (HCRs), which are predefined before deploying sensor nodes. [14] assumes that sensor nodes are correlated in the same HCR. Furthermore, the organization of clusters depends on the number of sensor nodes in each HCR. However, it is difficult to predefine HCRs and predict the correlation of sensor nodes without HCRs in practice.

3. Motivation

In [6], nodes are organized into clusters. All nodes in a cluster send their values to a cluster head node. Suppression happens at the cluster head and also en route. For example, suppose a set of clusters like **Fig. 1(a)**. The same color means the same sensor reading. C_A and C_B are organized by sensor nodes which sensed the same readings, respectively. Each cluster head suppresses some readings if they are identical to the cluster head's reading. According to this scheme, the readings of all the sensors may be suppressed by *A1* or *B1*. In practice, however, existing clustering algorithms organize clusters based on the locations of sensor nodes regardless of their readings. Thus, clusters may be organized easily by sensor nodes which have different values from their cluster heads like **Fig. 1(b)**. In this case, the clusters C_A and C_B

consist of *{B2, B3, B4}* and *{A5, A6}*, respectively. The readings differ from the cluster head node's reading. Therefore, their readings must be reported from each cluster head to the base station. In this case, opportunities to suppress across clusters are not exploited. Furthermore, the cluster shapes are fixed based on a network topology and are not tailored to actual correlation patterns. If a cluster head deals with only highly correlated cluster members, it could yield more efficient aggregation and smaller number of bits to transmit. Therefore, the problem is not only how to suppress sensor readings by a cluster head, but also how to organize the network topology in the orthogonal way.

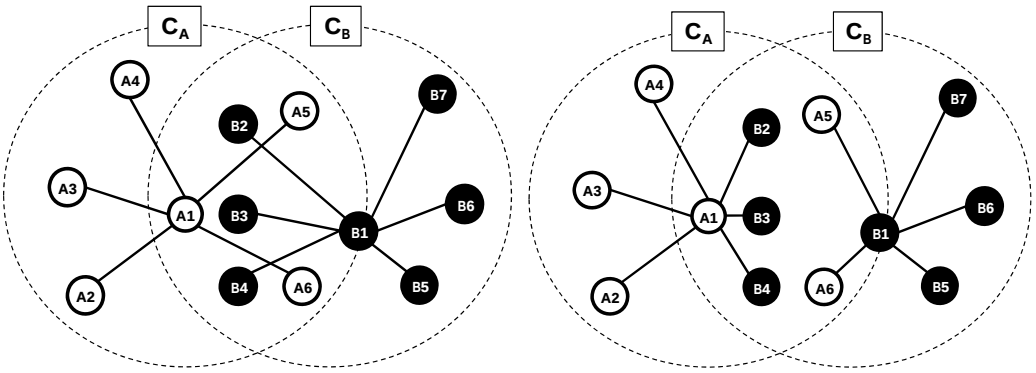


Fig. 1. Examples of clustering

4. Data Correlation-Based Clustering

In this section, we propose the DCC (Data Correlation-based Clustering) scheme. We define the term, *Data Correlation*, as similarity among sensor readings. Our hypothesis is that clustering based on data correlation will achieve better compression performance if compared with ordinary clustering based on locality of communication. DCC is an orthogonal approach to cluster-based schemes such as LEACH [13] and HEED [10]. To describe and evaluate our fashion simply, we use HEED that performs the single-hop communication. As shown in **Fig. 2**, a round in HEED includes two phases such as cluster formation and data gathering. After clustering, cluster heads broadcast a simple TDMA schedule to tell their members when their data are sent. Cluster heads gather the data of the members according to the TDMA schedule. Thus, we modify the advertisement sub-phase in the cluster formation phase and the TDMA schedule scheme of the procedure in the data gathering phase.

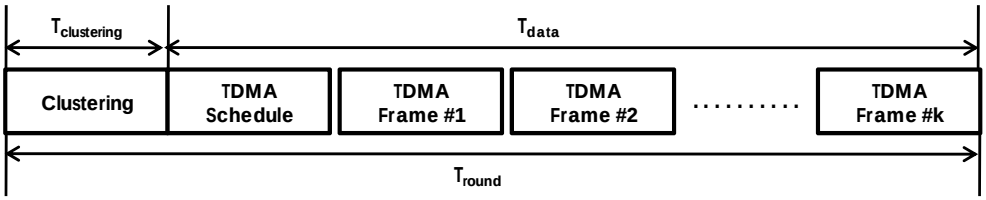


Fig. 2. Frame structure of HEED

4.1 Cluster Formation

In this paper, our major concern is how to organize clusters by adjacent sensor nodes which have similar readings. First, all sensor nodes first make mapping between the range of sensor data and advertisement period as shown in **Fig. 3**. This mapping is divided into *N advertisement-frames* and *N attribute-frames*. The number of frames, *N*, may be predefined. All sensor nodes can determine their advertisement frames without any centralized control as follows.

$$\left\lfloor \frac{V_{node}}{V_{\max} - V_{\min}} \times C_{attributeFrame} \right\rfloor \tag{1}$$

V_{node} is a current sensor reading in a current node. V_{max} and V_{min} are the maximum and minimum values in range of sensor reading, respectively. $C_{attributeFrame}$ is the number of attribute frames. All sensor nodes keep the sleep mode except for their advertisement frame. Thus, sensor nodes which are assigned in same advertisement frame can communicate together. In each advertisement frame, cluster heads are selected according to the probability of becoming a cluster head like HEED as follows.

$$CH_{prob} = C_{prob} \times \frac{E_{residual}}{E_{\max}} \tag{2}$$

E_{max} is an initial energy, $E_{residual}$ is a current residual energy in the node, and C_{prob} is an initial percentage of cluster heads among all n nodes. That is, a sensor node which maintains the highest residual energy broadcasts data in each advertisement frame. According to this fashion, sensor nodes which have similar sensor readings can be assigned into the same attribute frame. Furthermore, correlated sensor nodes communicate with each other in their advertisement frames during an advertisement period.

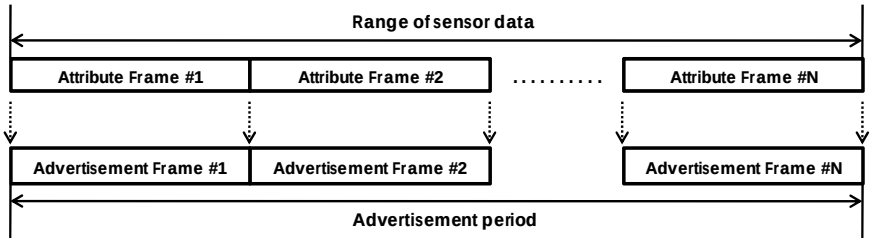
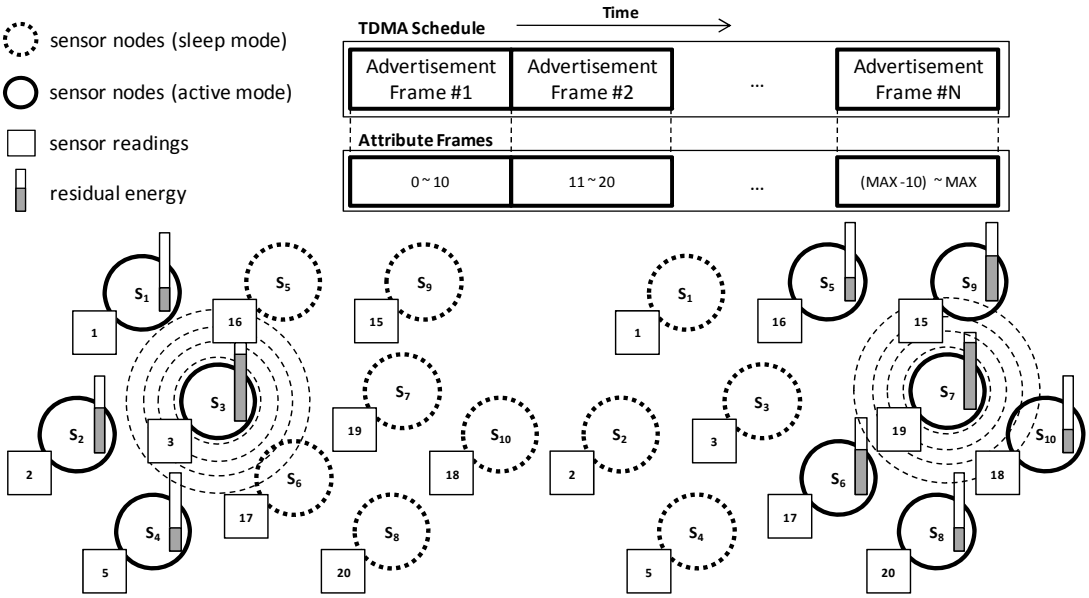


Fig. 3. Frame structure of the advertisement sub-phase

Fig. 4 shows an example of the clustering formation. Assume that there are 10 sensor nodes which have different sensor readings and residual energies. Because *{S_1, S_2, S_3, S_4}* have sensor readings between *0* and *10*, they are assigned into the *attribute frame #1*. Others which have sensor readings between *11* and *20* are assigned into the *attribute frame #2*. Thus, at the *advertisement frame #1*, *{S_1, S_2, S_3, S_4}* change their mode into the active mode, listen to any advertisement message. And then S_3 which has the most residual energy broadcasts an advertisement message to activated sensor nodes. One cluster is organized by *{S_1, S_2, S_3, S_4}*. Then, others maintain the sleep mode until their advertisement frame and overhearing is prevented. At the *advertisement frame #2*, *{S_1, S_2, S_3, S_4}* become the sleep mode, while others become the active mode. Another cluster is organized from them. Finally, two clusters are

organized by sensor nodes which have similar sensor readings. One cluster is $\{S_1, S_2, S_3, S_4\}$ and the other is $\{S_5, S_6, S_7, S_8, S_9, S_{10}\}$.



(a) $T=\{Advertisement\ Frame\ \#1\}$ (b) $T=\{Advertisement\ Frame\ \#2\}$

Fig. 4. An example of the clustering formation

4.2 Data Gathering

During the cluster formation, DCC guarantees independent communication among sensor nodes with different advertisement frames. It leads to organize clusters based on data correlation. To guarantee independent communication in each cluster in the data gathering phase, DCC exploits a similar approach. Each attribute frame is mapped into each TDMA frame as shown in **Fig. 5**. According to the corresponding TDMA frame, sensor nodes can be awaked and communicate with their cluster members.

Fig. 6 shows an example of the data gathering. At the *TDMA frame #1*, $\{S_1, S_2, S_3, S_4\}$ are activated in the *cluster 1*. The cluster head S_3 collects sensor readings from its members. At the *TDMA frame #2*, $\{S_5, S_6, S_7, S_8, S_9, S_{10}\}$ are activated in the *cluster 2*. The cluster head S_7 collects sensor readings from its members.

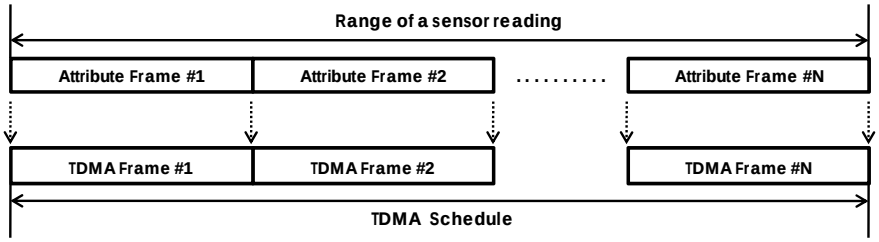
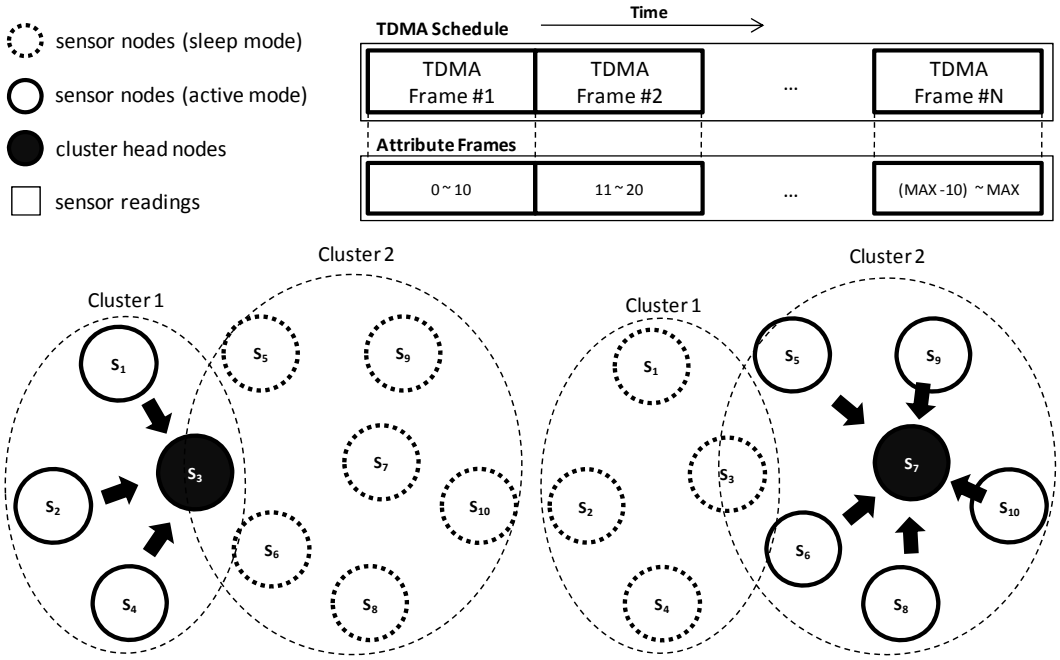


Fig. 5. Frame structure of TDMA schedule

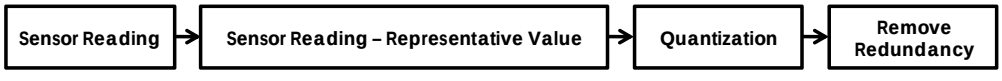


(a) T={ TDMA Frame #1} (b) T={ TDMA Frame #2}

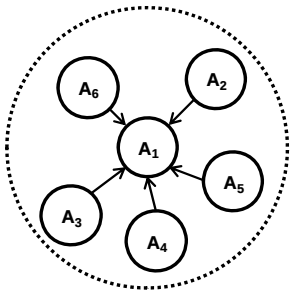
Fig. 6. An example of the data gathering

4.3 Enhanced Suppression Scheme

We also propose a spatio-temporal suppression scheme for our clustering algorithm. In the case of temporal suppression, when sensor data are transmitted to the cluster head, nodes do not transmit sensor data if their data are not changed since last reported. For example, at the current round, N_1 does not transmit its data to the cluster head because its data equal the collected data at the next round. Then, the cluster head assumes any unreported data remain unchanged. **Fig. 7(a)** shows the procedure of spatial suppression in DCC. In the case of spatial suppression, it is performed at the cluster head. When clusters are being created, a cluster consists of nodes which already have the correlated sensor readings. The cluster head also computes the difference between sensor readings and a representative value. The representative value is a maximum or minimum value in each attribute-frame. The cluster head also performs quantization to achieve more compactness. Because the existing algorithms organize clusters regardless of the correlation of sensor readings, they must use long bit codes for the quantization of sensor readings. If the cluster head has redundant data, they are removed except for each node's identification. For example, we consider a sensor network to collect sensor readings like **Fig. 7(b)**. The cluster consists of nodes which are allocated to an identical channel with attribute values 10 ~ 20 as shown in **Fig. 7(b)**. Cluster head A_1 computes the difference and quantization of collected data from $A_1 \sim A_6$. Therefore, A_1 remove redundant data except for each node's identification and transmits *{0x0000, {*A_1*,* A_3*,* A_5*}}* to the base station.



(a) Suppression procedure



	A_1	A_2	A_3	A_4	A_5	A_6
Sensor Reading	10	15	10	13	10	11
Difference (RV: 10)	+0	+5	+0	+3	+0	+1
Quantization	0000	0101	0000	0011	0000	0001
Remove Redundancy	{0x0000, {A_1, A_3, A_5}}, {0x0101, {A_2}}, {0x0011, {A_4}}, {0x0001, {A_6}}					

(b) Example of suppression scheme

Fig. 7. Enhanced suppression scheme

5. Performance Evaluation

We simulated our proposed algorithm and the existing clustering algorithms in various environments using Java Network Simulator (JNS) [15]. **Table 1** shows the values of performance evaluation parameters. To explain our scheme simply, we use the following typical energy dissipation model. It does not depend on applications and distances. It can be extended easily for long distance transmission in many applications.

$$E_{Rx} = E_{elec} \times k \tag{3}$$

$$E_{Tx} = E_{elec} \times k + e_{amp} \times k \times d^2 \tag{4}$$

E_{Rx} and E_{Tx} mean energy consumptions for receiving and transmitting data, respectively. The radio dissipates $E_{elec} = 50nJ/bit$ to run the transmitter or receiver circuitry and $e_{amp} = 100pJ/bit/m^2$ to run the transmission amplifier. k is the number of transmitted bits and d is the distance between the sender and the receiver. To show the superiority of our proposed algorithm, we compare it with varieties of LEACH-c which are adapted for temporal or spatio-temporal suppressions in terms of network traffic and network lifetime.

Table 1: Simulation parameters

Parameters	Values
The size of sensor networks	*100m* × *100m*
Communication range	*10m ~ 50m*
Location of the base station	*(x=50, y= 50)*
Initial energy	*0.1J*
The number of sensor nodes	*100 ~ 400*
The size of a sensor reading	*8byte*
The size of a join/advertisement message	*32byte*
The size of a sensor ID	*24byte*
The number of attribute frames	*10~50*

5.1 The Number of Sensor Nodes

Fig. 7 and **Fig. 8** show the network lifetime and the network traffic as the number of sensor nodes is changed, respectively. We assume that the network lifetime is a time to detect a dead node at first. The network traffic is an average data transmission rate per a node during a round. LEACH-c sends all readings to the base station. These readings can be suppressed by a temporal or spatio-temporal suppression technique. *T* denotes LEACH-c algorithm with a temporal suppression scheme from [7]. *ST* denotes LEACH-c algorithm with both a spatial suppression scheme and a temporal suppression scheme from [5] and [7].

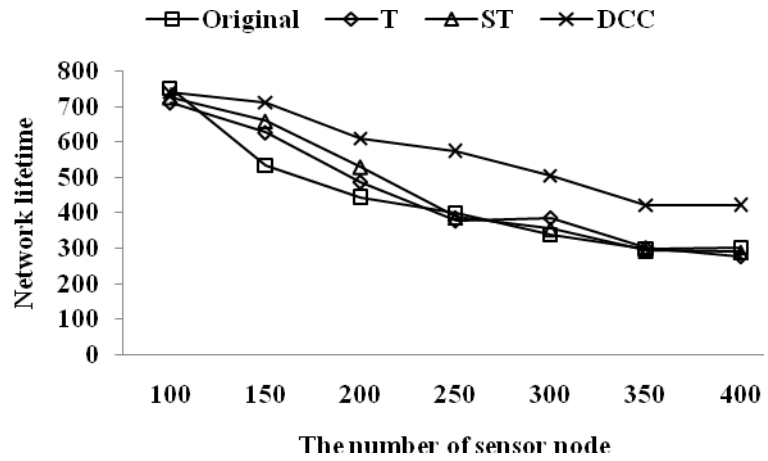


Fig. 7. Network lifetime

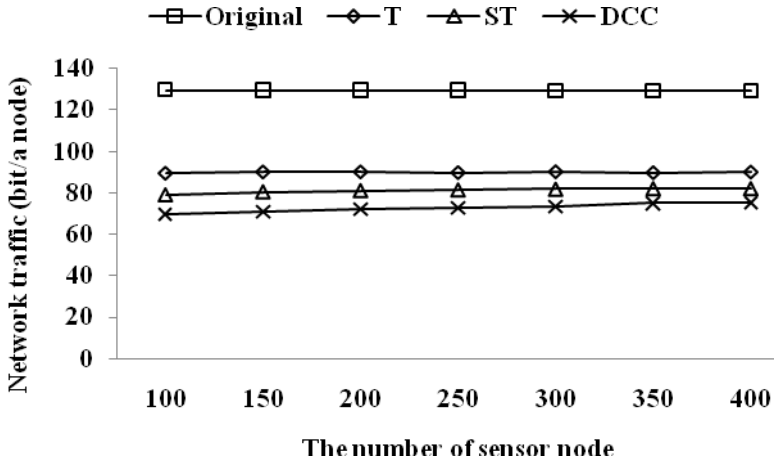


Fig. 8. Network traffic (bit/a node)

We assume that all sensor nodes are distributed uniformly over a whole network. We fixed the size of a sensing field to *100m* × *100m* and changed the number of sensor nodes from 100 to 400. It clearly helps to reduce the amount of transmitted data to the base station and prolong the lifetime of sensor networks. However, the performance gaps of the spatio-temporal suppression and the temporal suppression are not prominent as shown in **Fig. 7** and **Fig. 8**. In

other words, the spatial suppression is not effective to reduce the size of data. The reason is that each cluster consists of various nodes which have different data. DCC extends the network lifetime due to the improvement of suppression rate by increasing the correlation of sensor readings. In the result, simulation results shows that DCC achieves better performance than LEACH-c with the spatio-temporal technique as shown in **Fig.7** and **Fig. 8**.

5.2 The Number of Attribute Frames

Fig. 9 and **Fig. 10** show the network lifetime and the network traffic as the number of attribute frames is changed, respectively. We assume that 200 sensor nodes are deployed uniformly and communicate with each other inside *20m*. We changed the number of attribute frames from 10 to 50. As the number of attribute frames grows, the attribute range becomes narrow. It seems like that organizing more clusters saves more energy and reduces network contention by enabling the communication locality. Although the network traffic is decreased a little bit as the number of attribute frames increases, the network lifetime is shortened. It is because both the number of clusters and the number of sensor nodes which must transmit readings to the base station directly are increased. In the result, the whole communication cost is increased and the network lifetime becomes shortened.

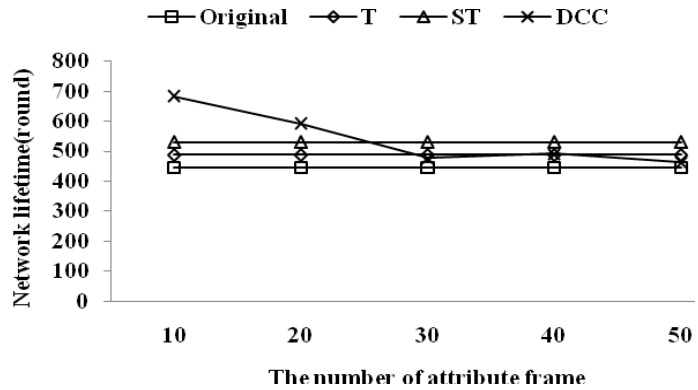


Fig. 9. Network lifetime

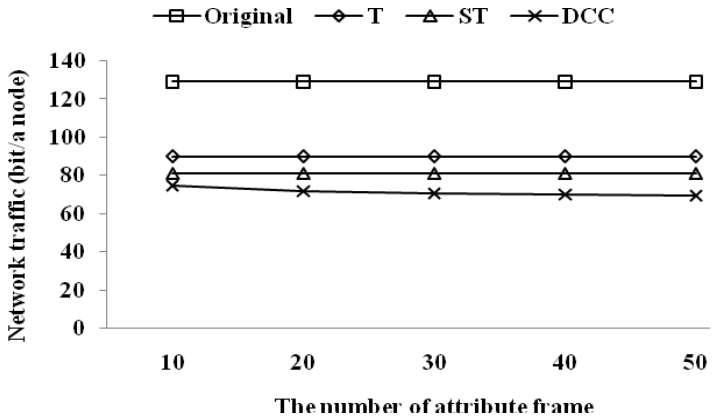


Fig. 10. Network traffic (bit/a node)

5.3 Communication Range

Both **Fig. 11** and **Fig. 12** show the performance of DCC according to the communication range. We assume that 200 sensor nodes are deployed uniformly. The number of attribute frames is fixed to 20 in DCC. We change the communication range from *10m* to *50m*. When the communication range is increased, its head may get more chances to suppress sensor readings which are received from members. However, cluster heads consume more energy to collect sensor readings from members because each cluster head contains more sensor nodes and the distance between the cluster head and members is increased.

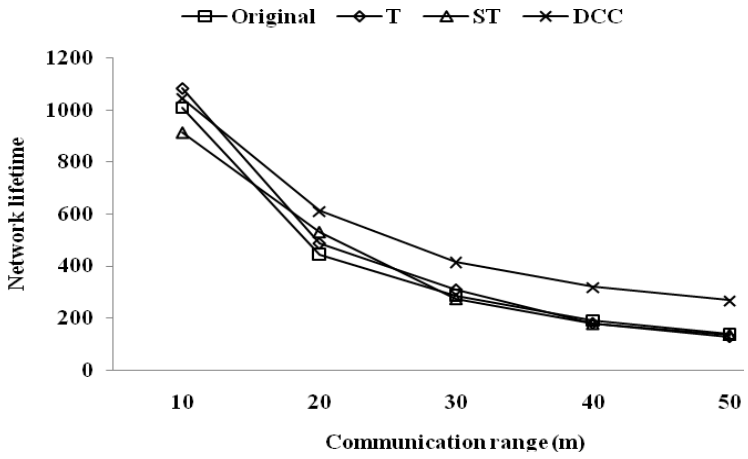


Fig. 11. Network lifetime

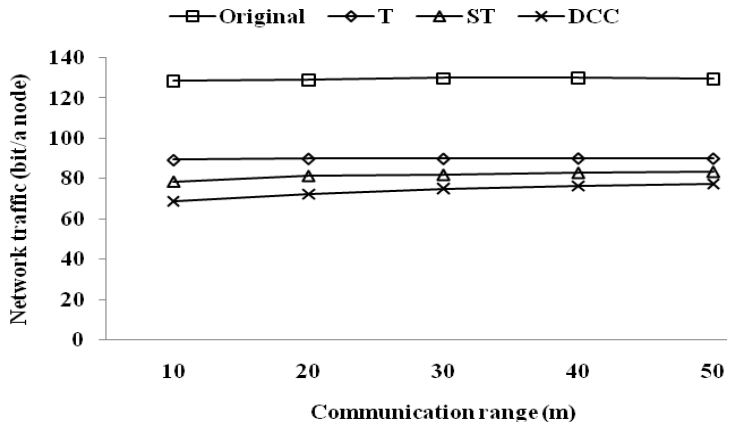


Fig. 12. Network traffic (bit/a node)

6. Conclusions

In this paper, we have proposed a new clustering algorithm based on data correlation. The proposed algorithm significantly improves suppression rates by increasing the correlation of

sensor data between cluster members and their cluster head. In addition, we have proposed a spatio-temporal suppression scheme to reduce the network traffic. We have evaluated the performance through various experiments on the networks with 200 nodes. In the result, the size of reported data which have been collected in the base station was reduced and the whole network lifetime was prolonged. As a result, we have shown that the proposed algorithm conserves energy and extends the lifetime of sensor networks. In the future, we will extend the proposed algorithm so that it supports a load balancing scheme.

References

[1] D. Estrin, L. Girod, G. Pottie and M. Srivastava, "Instrumenting the World with Wireless Sensor Networks," In *Proceedings of International Conference Acoustics, Speech, and Signal Processing*, Vol.4, pp. 2033-2036, May 2001.

[2] G. J. Pottie and W. J. Kaiser, "Wireless Integrated Network Sensors," In *Proceedings of Comm. ACM*, Vol.43, No.5, pp. 51-58, May 2000.

[3] I.F. Akyildiz, W. Su, Y. Sankarasubramanism and E. Cayirci, "A Survey on Sensor Networks," In *Proceedings of IEEE Communications Magazine*, Vol.40, No.8, Aug. 2002.

[4] A. Silberstein, R. Braynard and J. Yang. "Constraint Chaining: On Energy-Efficient Continuous Monitoring in Sensor Networks," In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 157-168, Jun. 2006.

[5] X. Meng, L. Li, T. Nandagopal and S. Lu, "Event contour: An efficient and robust mechanism for tasks in sen-sor networks," In *Proceedings of Technical report*, pp. 1-13, 2004.

[6] S. Pattem, B. Krishnamachari and R. Govindan, "The impact of spatial correlation on routing with compression in wireless sensor networks," In *Proceedings of International Conference on Information Processing in Sensor Networks*, pp. 28-35, Apr. 2004.

[7] M. Sharaf, J. Beaver, A. Labrinidis and P. Chryanthis, "Tina: A scheme for temporal coherency-aware in-network aggregation," In *Proceedings of the 2003 ACM Workshop on Data Engineering for Wireless and mobile Access*, pp.67-76, Sept 2003.

[8] O. Younis, M. Krunz and S. Ramasubramanian, "Node Clustering in Wireless Sensor Networks: Recent Developments and Deployment Challenges," *IEEE Networks*, Vol.20, No.3, pp. 20-25, Jun 2006.

[9] S. Basagni, "Distributed Clustering Algorithm for Ad Hoc Networks," In *Proceedings of International Symposium Parallel Architectures, algorithms, and Networks*, pp. 310-315, 1999.

[10] O. Younis and S. Fahmy, "Distributed clustering in adhoc sensor networks: A hybrid, energy-efficient approach," In *Proceedings of IEEE INFOCOM*, pp. 366-379, Mar 2004.

[11] S. Banerjee and S. Khuller, "A Clustering Scheme for Hierarchical Control in Multi-hop Wireless," In *Proceedings of IEEE INFOCOM*, pp. 1-10, Apr. 2001.

[12] J. Kamimura, N. Wakamiya and M. Murata, "Distributed Clustering Method for Energy-Efficient Data Gathering in Sensor Networks," In *Proceedings of the 1st IEEE Communications Society Conference (SECON 2004)*, Oct 2004.

[13] W. R. Heinzelman, A. Chandrakasan and H. Balakrishnan, "Energy- Efficient Communication Protocols for Wireless Microsensor networks," In *Proceedings of the Hawaii International Conference on System Sciences*, Vol.8, pp. 3005–3014, Jan 2000.

[14] D. Maeda, H. Uehara and M. Yokotama "Efficient Clustering Scheme Considering Non-uniform Correlation Distribution for Ubiquitous Sensor Networks," *IEICE Transactions on Fundamentals*, Vol. E90-A, No.7, pp. 1344-1352, Jul 2007.

[15] JNS: Java Network Simulator, http://jns.sourceforge.net/

Myungho Yeo received the B.S. and M.S. in Information and Communication Engineering from Chungbuk National University, Korea in 2004 and 2006, respectively. He is currently working towards Ph.D. degree on Department of Information and Communication Engineering from Chungbuk National University, Korea. His main research interests include main-memory database systems and wireless sensor networks.

Dongmin Seo received the B.S., M.S. and Ph.D. in Department of Information and Communication Engineering from Chungbuk National University, Korea in 2002, 2004 and 2008, respectively. He is now the postdoctoral of Korea Advanced Institute of Science and Technology, Korea. His main research interests include MOD (Moving-Objects Database) systems and XML database systems.

Jaesoo Yoo received the B.S. in Computer Engineering from Chunbuk National University, Korea in 1989 and also received M.S. and Ph.D. in Computer Science from Korea Advanced Institute of Science and Technology, Korea in 1991 and 1995. He is now a professor in Department of Information and Communication Engineering, Chungbuk National University, Korea. His main research interests include database systems, sensor data management, location based services, distributed computing and storage management systems.