http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 11 (10): 1477-1483, 2012
ISSN 1812-5638 / DOI: 10.3923/itj.2012.1477.1483

Data Dissemination in Wireless Sensor Networks with Clustering Method

Chao Gao, Xiaoya Hu, Bingwen Wang, Hongliang Gao and Wei Xiong
Department of Control Science and Engineering,
Huazhong University of Science and Technology, Wuhan, 430074, China

Abstract: In Wireless Sensor Networks (WSNs), since the network consists of low-cost sensor nodes with finite battery power, energy efficient is the primary consideration in algorithm design in order to prolong the network lifetime. Various schemes had been proposed to efficiently store and process sensed data. In this study, a data storage scheme called Cluster Based Data Storage (CBDS) is proposed to reduce the communication cost for transmitting data and to efficiently process data queries. The CBDS uses the clustering architecture to efficiently transmit and store data. To our knowledge, this is the first time using clustering method in data centric storage. The CBDS uses the virtual coordinate instead of geographic location information. So CBDS does not need to use the GPS or other technologies to locate, it also saves energy. Moreover, CBDS does well in handling hot spot problem. Analysis and simulation results show that the CBDS scheme outperforms the GTH-based DCS scheme.

Key words: Wireless sensor networks, data storage scheme; cluster protocol

INTRODUCTION

A Wireless Sensor Network (WSN) consists of a large number of distributed battery-operated and low-cost sensor nodes (Akyildiz *et al*., 2002; Yicka *et al*., 2008). WSN has low energy consumption, inexpensive, distributed and self-organizing features. And it brought greatly improved in the field of information awareness. Wireless Sensor Networks (WSNs) are widely used in civil and military fields (Idris *et al*., 2009; Poon *et al*., 2006; Ruan *et al*., 2010; Tovar *et al*., 2010) because of its superiority. Therefore, a large number of protocols and algorithms had been proposed for WSN from the data link on the bottom layer to the top application layer of the leveled-network model.

Since wireless sensor network is data centric, data storage and query is a hot research area in sensor networks. Data storage schemes in sensor networks focus on data storage, fusion and query. The most prominent feature of the wireless sensor network is sensor nodes are battery-operated. The first considered factor of proposed protocols in wireless sensor networks is energy efficient. So the purpose of data storage schemes in WSNs is to make nodes with limited energy and storage capacity used as effectively as possible.

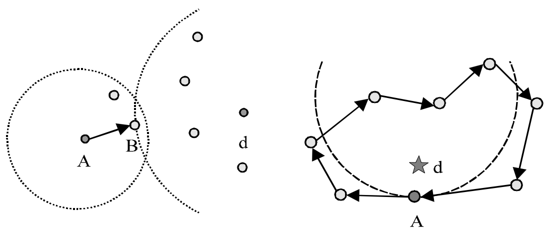
There are number of schemes had been proposed for efficiently storing and querying data. One widely mentioned method of them is the external storage-based (ES) data dissemination (Pottie and Kaiser, 2000; Yao *et al*., 2006). It depends on a centralized base station (sink) which connects the sensor network to the external network, usually with powerful data storage, processing and communication capabilities. In external-based storage, all of monitoring data collected by sensor nodes is sent to the sink for centralized storage. Although the data processing and query in the sink is very convenient and flexible, the ES schemes increase network traffic and nodes energy waste. In addition, since numbers of nodes send data through multi-hop to sink, the area near of the sink will become a hot zone (hotspot). As the network size increases and the number of nodes increase, this area will become the communication bottleneck, affecting the overall system stability and reliability. In order to avoid unnecessary network traffic, the local storage-based (LS) schemes, e.g., directed diffusion (Intanagonwiwat *et al*., 2003; Ye *et al*., 2005), use the opposite approach to ES schemes. The sensor nodes store monitoring data in local memory, data item sent to the sink only when node received queries. Since the sink does not know which node stores the interested data item, the LS schemes need a fast and efficient search algorithm to facilitate the sink to find the source node which holding the interest data. For example, in directed diffusion, a sink uses the flooding message algorithm to query data; the source node with the requested data receives the query message and then sends the data as response. So the LS scheme suitable for the applications in which data collected frequently but data query seldom. In addition, sensor nodes cannot

Corresponding Author: Xiaoya Hu, Department of Control Science and Engineering,
Huazhong University of Science and Technology, Wuhan, 430074, China

preserve the long-term historical detail data because of the limited storage capacity. Relevant data will be lost if the node failure or run out of power, system stability and robustness cannot be guaranteed.

For efficiently storage and querying events (Ratnasamy *et al*., 2003; Shenker *et al*., 2003) had proposed Data-Centric Storage (DCS) scheme for wireless sensor networks. This scheme uses Geographic Hash Table (GHT) to find the storage node (called home node or rendezvous node) for each type of event. DCS sensor networks store events by event names. Each sensor reading (event) is mapped to a sensor node or some of sensor nodes by the geographic hashing function based on the values of the event's name. Therefore, all events with the same value are stored at the same node. A data consumer that is interested in an event of a specific name will use the name to identify the rendezvous node and can retrieve the event directly from that node without flooding query messages. In order to identify the rendezvous node of an event, the event's name is hashed to a geographic location with a hash function and the node closest to the hashed location is the rendezvous node. The event is routed to the rendezvous node from the original sensor node according to Greedy Perimeter Stateless Routing (GPSR) which is a classic geographical routing protocol.

In GPSR and other geographical routing protocol, nodes are assumed to know their geographical positions. Most of the geographical routing protocols rely on the GPS technology which consumes high energy. The GPSR has two key algorithms (or called mode), named *Greedy algorithm* and *perimeter algorithm* as shown in Fig. 1. Greedy algorithm is the default forwarding algorithm of GPSR protocol. This algorithm greedily delivers a data packet to its destination location by forwarding the packet to a neighbor node whose position is closest to the location. GPSR switches to use perimeter algorithm when a routed packet encounters a node whose neighbors are farther away to the destination location than itself. Perimeter algorithm is used until it finds a node that is closer to the destination location than the entry node at which perimeter forwarding mode starts. Please note that GPSR switches to perimeter algorithm not only when packets encounter network voids caused by network's deployment of non-uniform, it can also switches to the routing mode because of empty destination locations where there are no nodes residing. A packet when reaching to a node closest to an empty destination location will be forwarded using perimeter algorithm until it loops at that closest node. In DCS sensor networks, destination locations are produced by hashing event names. As the result, destination locations of packets are usually empty and perimeter walks occur in almost every event insertion and query. As pointed out by Seada *et al*. (2004), geographic face routing algorithms including GPSR's perimeter mode perform poorly, especially the overhead of perimeter walks is significant in practical situations.



Greedy forwarding mode Perimeter forwarding mode

Fig. 1: The GPSR routing protocol

The CBDS scheme which this paper proposed uses the clustering architecture. There are a lot of previous works on clustering protocol. LEACH (Heinzelman *et al*., 2002) is a classic clustering algorithm in wireless sensor networks. It is the earliest and most famous clustering protocol in WSNs. Randomized rotation of cluster heads is proposed to evenly distribute the energy consumption among all of the sensor nodes in the networks. The LEACH algorithm's time unit is called round. Each round consists of two phase. The first is set-up phase when the clusters are organized, followed by a steady-state phase when sensor nodes can collect and transmit data. The duration of the steady phase is longer than the duration of the set-up phase so as to minimize the network overhead. The cluster head selection algorithm of LEACH is defective, because it does not consider the residual energy. HEED protocol (Younis and Fahmy, 2004), the cluster head node election took into account the residual energy and node distribution around the neighbor nodes. CTPEDCA (Wang *et al*., 2011) protocol proposed a cluster-based and tree-based power efficient data collection and aggregation method for WSNs. PEGASIS (Lindsey and Raghavendra, 2002) proposed a near optimal chain-based protocol that is an improvement over LEACH.

In this study, a novel data storage method, called Cluster Based Data Storage (CBDS) scheme is proposed to solve the problems mentioned above. The CBDS can reduce the communication cost for transmitting data and efficiently process data queries. This scheme uses a clustering architecture to efficiently transmit and store data. In most of other data storage method, the nodes need to know their location (via GPS and other methods). But in CBDS algorithm, location information is not required, it used the virtual coordinate instead of geographic location information. So CBDS did not need to use the GPS or other technologies to locate, it also saves energy.

THE SYSTEM MODEL

There are various models for wireless sensor networks. In this study, the CBDS mainly considered a wireless sensor network consisting of a number of sensors and a sink node that are randomly dispersed in the interested area. All of the sensor nodes are homogeneous and energy constrained. The locations of sink and sensors are fixed. A sensor can transmit data to any other sensor and can communicate directly with the sink node if necessary (when it is elected as cluster head node). The sensor nodes monitor their vicinity area periodically and collecting data. The Cluster Head (CH) are gathered from sensors which can be in the same the cluster at each round, where a round is defined as the process of gathering all the data from sensor nodes to the sink node, regardless of how much time it takes. The cluster head can perform data aggregation to reduce the number of redundant data if data compression algorithm is used. Finally, the CH can send the messages to the sink node for further processing. To further reduce energy consumption, in each cluster, all nodes except the head can switch to the sleeping/activating mode.

The CBDS assumed a simple model for the radio hardware energy dissipation where the transmitter dissipates energy to run the radio electronics and the power amplifier and the receiver dissipates energy to run the radio electronics (Heinzelman *et al*., 2002). For this study, we will use both the free space (d^2 power loss) and the multipath fading (d^4 power loss) channel models, that depending on the distance between the transmitter and receiver as shown in the Eq. 1 and 2. In this experiments, power control can be used to invert this loss by appropriately setting the power amplifier, the Free Space (FS) model will be used when the distance is less than a threshold d_0; otherwise, the multipath (mp) model will be used. Therefore, when nodes want to transmit a *l*-bit data packet a distance d, the radio expends:

$$E_{Tx}(l,d) = E_{Tx-elec}(l) + E_{Tx-amp}(l,d) = \begin{cases} l \cdot E_{elec} + l \cdot e_{fs} \cdot d^2, & d < d_0 \\ l \cdot E_{elec} + l \cdot e_{mp} \cdot d^4, & d \geq d_0 \end{cases} \quad (1)$$

And want to receive this data packet, the radio expends:

$$E_{rx}(l,d) = E_{rx\text{-}elec}(l) = l.E_{elec} \quad (2)$$

In the specified radio model, E_{elec} indicates the electronics energy which depends on factors such as the digital coding, modulation, filtering and spreading of the signal. $\varepsilon_{fs}.d^2$ or $\varepsilon_{mp}.d^4$ indicate the amplifier energy which depends on the distance to the receiver and the acceptable bit-error rate. For the experiments described in this paper, these energy parameters are the same as those used in LEACH, as follow:

$$E_{elec} = 50 \text{ nJ/bit}$$

$$\varepsilon_{fs} = 10 \text{ pJ/bit/m}^2$$

$$\varepsilon_{mp} = 0.0013 \text{ pJ/bit/m}^4$$

The initial energy is 2 J for every sensor node. CBDS scheme also assumed that the radio channel is symmetric, in other words, the cost of transmitting a message from A to B is the same as the cost of transmitting a message from B to A.

The CBDS scheme

The cluster construction: Clustering provides an effective method to prolonging the lifetime of a wireless sensor network. There are usually two methods used in current clustering algorithms: selecting cluster heads with more residual energy and rotating cluster heads periodically (each round) to distribute the overhead among nodes in each cluster and extend the network lifetime.

During the cluster head selection phase, nodes decide whether they should take the role of a cluster head or not. Concretely, every node is initialized not to be a cluster head and does not have an associated cluster at the beginning of a cluster head selection phase. Firstly, the head selection probability, denoted by s_i is calculated by node p_i. This probability is calculated based on two factors. The first one is called the cluster count factor which is a system wide parameter, denoted by f_c. It is a value in the range [0, 1] and defines the average fraction of nodes that will be selected as cluster heads. The factors that can affect the decision on the number of clusters and, thus, the setting of f_c, include the size and density of the network. The second factor is called the relative energy level of the node p_i, denoted by $e_i(t)$. It means the energy available at node p_i at time t. The relative energy level is calculated by comparing the energy available at node p_i with the average energy available at the nodes within its one-hop neighborhood.

So the head selection probability can be calculated by multiplying the cluster count factor with the relative energy level:

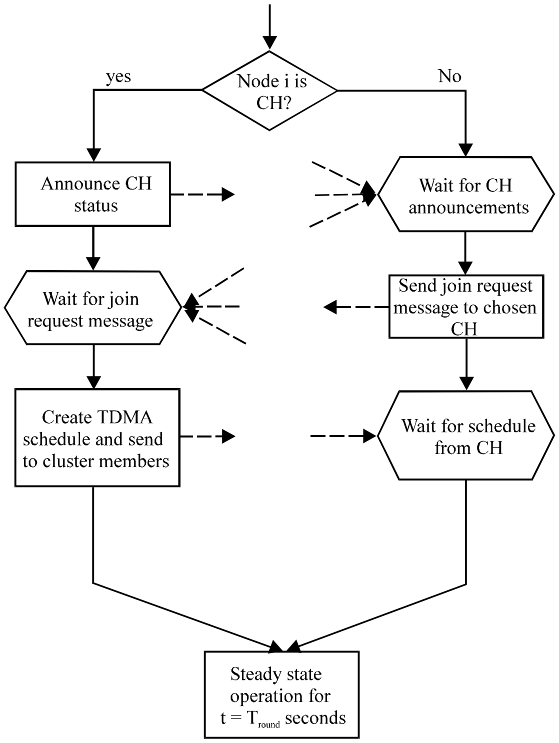

Node i is CH?
yes
No
Announce CH status
Wait for CH announcements
Wait for join request message
Send join request message to chosen CH
Create TDMA schedule and send to cluster members
Wait for schedule from CH
Steady state operation for t = T_round seconds

Fig. 2: Flowchart of the clustering algorithm of CBDS

$$s_i = f_c * \frac{e_i(t) * (|nbr(p_i)| + 1)}{e_i(t) + \sum_{p_j \in nbr(p_i)} e_i(t)} \quad (3)$$

The Eq. 3 enables CBDS to select nodes with higher energy levels for cluster head. Once s_i is calculated, node p_i is chosen as a cluster head, with probability s_i. If selected as a cluster head, p_i send the cluster formation messages to neighbor nodes within its transmission range. And the clustering round is start. If p_i is not selected as a cluster head, it waits for some time to receive cluster formation messages from other nodes. If no such message is received, it repeats the whole process, starting from the s_i calculation.

After electing cluster heads, a node decide to join a cluster according to the strength of the signal. The CBDS used a similar clustering algorithm with LEACH to form cluster, the flowchart of the cluster construction algorithm is shown in Fig. 2. Within a sensor node, the dominant energy consumer is the radio unit. When the network is partitioned into clusters, data transmission can be classified into two stages: intra-cluster communication and inter-cluster communication. In order to save energy, we assume every node has two interfaces to adapt to different transmission range: a short range interface and a long range interface. The short range interface enables nodes to communicate with other nodes within the cluster and is what is used most of the time. The long range interface enables cluster head nodes to directly route messages to other cluster heads or the sink, as Fig. 3 shown.

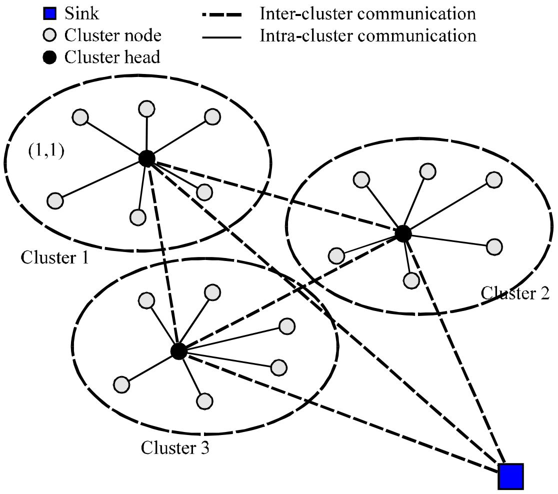


Fig. 3: The hierarchical of network

To further reduce energy consumption, in each cluster, all nodes except the cluster head are in sleep mode. Nodes in sleep mode still can sense data, but rely on the cluster head for other functions. The nodes rotate the responsibility of acting as the cluster head in a round-robin manner in the order of their identifiers to balance the workload and energy consumption among nodes. Each node keeps a countdown timer. If a node becomes a cluster head, it sets countdown timer to T_c which is the length of duty period. And the node triggers its timer to start the countdown. With the knowledge of all the nodes in its cluster, the cluster head knows the next head among clients in the round-robin manner. If its timer expires, the head transfers its duty to the next head. It also sends the information of links of cluster members and neighboring heads to its successor.

This robin based head election is suitable for the case when nodes have homogeneous computing resource, energy and mobility capacity. In the case when nodes have heterogeneous computing and energy capability, nodes with more resources have higher priority to be selected as cluster heads. The node with lower mobility has higher priority to be selected as cluster head in a dynamic environment. The head's duty period can be adjusted to balance the head change overhead and the effect of load balance.

VIRTUAL COORDINATE DETERMINATION

The entire sensor network is divided into M clusters through cluster construction algorithm; each cluster is assigned a unique number from 0 to M-1, by centralized approach. It is named Cluster ID (CID). The cluster head maintains the links to other cluster heads. It periodically exchanges "hello" messages with its clients and neighboring cluster heads. A head communicates with its neighboring heads by specifying the IDs of the neighboring zones in the messages. Each cluster member(including the cluster head) has a intra-cluster number, It is unique within a cluster. The identifiers of nodes in a cluster which has N nodes are normalized from 0 to N-1, called them Intra-Cluster Node ID (NID). So each node has 2-tuple identifiers (CID, NID) called virtual coordinate. For example, a node's identifier in the cluster is 2 and it is in the No. 5 cluster, so its virtual coordinate is (5, 2). According to the above definition, each node within the network has a unique virtual coordinate.

Data insertion method: The CBDS scheme assumed that the cluster structure can be maintained long enough stable after clustering establish. This means that the number of clusters of the network can remain unchanged.

When a sensor node (source node) detects an event (sensor reading), it forwards the event to the head node in its cluster. The scheme uses a hash function similar to the DCS Scheme. Each sensor reading (event) is mapped to a geographical location by a hashing function based on the values of the event's attribute. So the head node get a 2-dimensional coordinates represented by (X, Y). Then it needs to find the home node of this data item, so it has to find a coordinate match (X, Y). CBDS had adopted a virtual coordinate method. Assume the home node's virtual coordinates for the event type is (X_i, Y_i). Firstly, X_i is calculated as follows:

$$X_i = X\%m \tag{4}$$

where, the m denotes the number of clusters within the network, Symbol % denotes the modulus calculated. So the coordinates of the X_i is known, the data packet is send to the cluster head of CID = X_i. Secondly, Y_i can be calculated as follows:

$$Y_i = T\%n_{CID=xi} \tag{5}$$

where, n_{CID} = X_i is the number of nodes within the cluster of CID = X_i. So the Data is finally sent to the node of the NID = Y_i in the cluster of CID = X_i.

For example, there are 4 clusters in the network as shown in Fig. 4. The node A's NID = 6, it is in the cluster of CID = 0, so its virtual coordinates is (0, 6). When A sensed an event, the data packet is sent to the head of its cluster. We assumed according to the hash function, the packet is mapped to (15, 61). So the virtual coordinate X_i of the home node is 3 according to the Eq. 4, the data is sent to the cluster head whose CID = 3. This cluster has 8 nodes, take n = 8 into the Eq. 5, we get Y_i = 5. The data is sent to the node B from the head node. So node B with virtual coordinate (3, 5) is the home node of the data item. If sink node need query data, it can get the geographic coordinates of the data according to the hash function and then the virtual coordinates can also be obtained through the Eq. 4 and 5.

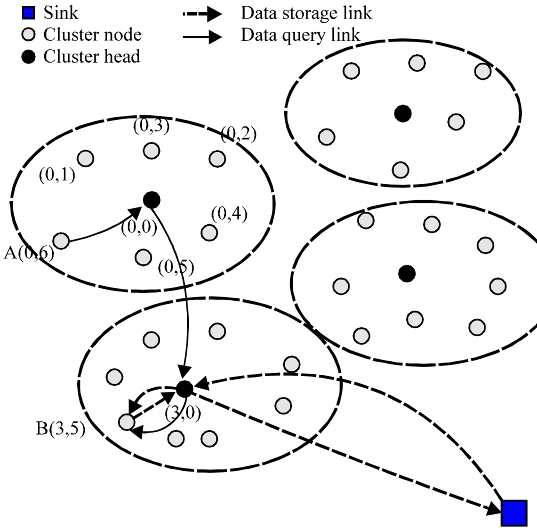


Fig. 4: A CBDS storage and query example

Performance evaluation: Here, we conducted simulations to evaluate the CBDS scheme. In order to evaluate the performance of this scheme, two different storage algorithms (GHT and proposed CBDS) simulated by NS-2 simulator.

Seriously simulations were performed to evaluate the performance of the proposed CBDS. 100 nodes are randomly deployed in a square area 100×100 m^2 as shown in Fig. 5. Every node generates 50 events that are normal distribution of values and a percentage of 70% of the events falls into a percentage of 20% of the reading range. In the simulations, each node has only 2 J of energy at the beginning of experiments. The sink node is located far away from the region, at point (50, 175).

We measured the energy level of each sensor node and compared the results with that of GHT scheme which is one of the most well-known data centric storage.

Table 1 shows the average residual energy of each sensor node in the entire area and hot-spot area. As shown in the table, entire sensor nodes of the GHT consume more energy than those of the proposed CBDS

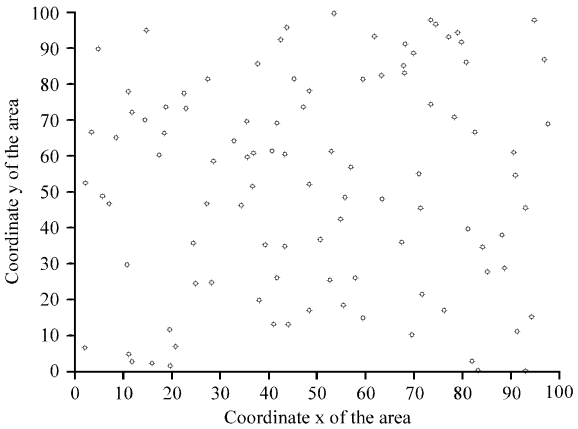


Fig. 5: 100 nodes random deployed in a square area

Table 1: Average energy of each sensor node LC

Energy	GHT (J)	CBDS (J)
Average energy of a sensor node	1.02	1.28
Average energy of a sensor node in hot-spot	0.63	0.95

by about 36%. In the hot-spot area, sensor nodes of GHT consume much more energy than those of the proposed CBDS by about 30%. The reasons why the proposed CBDS outperforms GHT are as follows. The first reason is that the GHT uses GPSR routing protocol. It needs to use perimeter model because of empty destination locations where there are no nodes residing. This situation often occurs, so it consumes more energy. Also, CBDS does not use GPS to get location information, it will save more energy. Moreover, CBDS do a better job in handling hot spot problem, because CBDS storage scheme uses the virtual coordinates instead of geographic location.

CONCLUSION AND DISCUSSION

In this study, we proposed a cluster based data storage scheme to support scalable handling of monitoring data in wireless sensor networks. The CBDS scheme can provide reliably responses to queries while minimizing the use of limited energy and computational resources. This data storage scheme ensures scalability and load balancing of communication as well as adaptivity in presence of dynamic changes. Analysis and simulation results show that the CBDS scheme outperforms the GTH-based DCS scheme in overall performance.

Our future work is to investigate an efficient data compression and data fusion algorithm deploy in cluster head to reduce communication overhead and save storage space. Another direction of future work is how to maintain the CBDS data dissemination in presence of network dynamic changes.

ACKNOWLEDGMENTS

This study is supported in part by a grant from the National Science Foundation of China (No. 6080202, No. 60773190) and the Fundamental Research Founds for the Central University of China (No. 2011TS140).

REFERENCES

Akyildiz, I.F., W. Su, Y. Sankarasubramamiam and E. Cayirci, 2002. A survey on sensor networks. IEEE Commun. Magazine, 40: 102-114.

Heinzelman, W.R., A.P. Chandrakasan and H. Balakrishnan, 2002. An application specific protocol architecture for wireless microsensor networks. IEEE Trans. Wireless Commun., 1: 660-670.

Idris, M.Y.I., E.M. Tamil, N.M. Noor, Z. Razak and K.W. Fong, 2009. Parking guidance system utilizing wireless sensor network and ultrasonic sensor. Inform. Technol. J., 8: 138-146.

Intanagonwiwat, C., R. Govindan and D. Estrin, J. Heidemann and F. Silva, 2003. Directed diffusion for wireless sensor networking. IEEE/ACM Trans. Network, 11: 2-16.

Lindsey, S. and C.S. Raghavendra, 2002. PEGASIS: Power efficient gathering in sensor information systems. IEEE Aerospace Conf. Proc., 3: 1125-1130.

Poon, C.C.Y., Z. Yuan-Ting and B. Shu-Di, 2006. A novel biometrics method to secure wireless body area sensor networks for telemedicine and m-health. Communi. Magazine, IEEE, 44: 73-81.

Pottie, G.J. and W.J. Kaiser, 2000. Wireless integrated network sensors. Commun. ACM, 43: 51-58.

Ratnasamy, S., B. Karp, S. Shenker, D. Estrin, R. Govindan, L. Yin and F. Yu, 2003. Data-centric storage in sensornets with GHT, a geographic hash table. Mobile Networks Applicat., 8: 427-442.

Ruan, Z., X. Sun, W. Liang, D. Sun and Z. Xia, 2010. CADS: Co-operative anti-fraud data storage scheme for unattended wireless sensor networks. Inform. Technol. J., 9: 1361-1368.

Seada, K., A. Helmy and R. Govindan, 2004. On the effect of localization errors on geographic face routing in sensor networks. Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks, April 26-27, 2004, Berkeley, CA., USA., pp: 71-80.

Shenker, S., S. Ratnasamy, B. Karp, R. Govindan and D. Estrin, 2003. Data-centric storage in sensornets. Comput. Commun. Rev., 33: 137-142.

Tovar, A., T. Friesen, K. Ferens and B. McLeod, 2010. A DTN wireless sensor network for wildlife habitat monitoring. Proceeding of the Electrical and Computer Engineering, May 2-5, 2010, Calgary, AB, pp: 1-5.

Wang, W., B. Wang, Z. Liu, L. Guo and W. Xiong, 2011. A cluster-based and tree-based power efficient data collection and aggregation protocol for wireless sensor networks. Inform. Technol. J., 10: 557-564.

Yao, Y.X., X.Y. Tang and E.P. Lim, 2006. In-network processing of nearest neighbor queries for wireless sensor networks. Proceedings of the 11th International Conference on Database Systems for Advanced Applications, April 12-15, 2006, Singapore, pp: 35-49.

Ye, F., G. Zhong, S. Lu and L. Zhang, 2005. GRAdient broadcast: A robust data delivery protocol for large scale sensor networks. Wireless Networks, 11: 285-298.

Yicka, J., B. Mukherjeea and D. Ghosal, 2008. Wireless sensor network survey. Comput. Networks, 52: 2292-2330.

Younis, O. and S. Fahmy, 2004. HEED: A hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks. IEEE Trans. Mobile Comput., 3: 366-379.