International Journal of Computer Applications Technology and Research
Volume 3– Issue 7, 409 - 414, July 2014

Review on Clustering and Data Aggregation in Wireless Sensor Network

Pooja Mann
M.Tech(Computer Science & Engineering),
Geeta Institute of Management and Technology,
Kanipla, Kurukshetra

Tarun Kumar
Deptt. of Computer Science & Engineering,
Geeta Institute of Management and Technology,
Kanipla, Kurukshetra

Abstract: Wireless Sensor Network is a collection of various sensor nodes with sensing and communication capabilities. Clustering is the process of grouping the set of objects so that the objects in the same group are similar to each other and different to objects in the other group. The main goal of Data Aggregation is to collect and aggregate the data by maintaining the energy efficiency so that the network lifetime can be increased. In this paper, I have presented a comprehensive review of various clustering routing protocols for WSN, their advantages and limitation of clustering in WSN. A brief survey of Data Aggregation Algorithm is also outlined in this paper. Finally, I summarize and conclude the paper with some future directions.

Keywords: Wireless Sensor Network, Clustering, Data Aggregation, LEACH

1. INTRODUCTION

A wireless sensor network (WSN) is an ad-hoc network composed of small sensor nodes deployed in large numbers to sense the physical world. Wireless sensor networks have very broad application prospects including both military and civilian usage. They include surveillance [1], tracking at critical facilities [2], or monitoring animal habitats [3].

In general, a WSN consists of a large number of tiny sensor nodes distributed over a large area with one or more powerful sinks or base stations (BSs) collecting information from these sensor nodes. All sensor nodes have limited power supply and have the capabilities of information sensing, data processing and wireless communication [4].

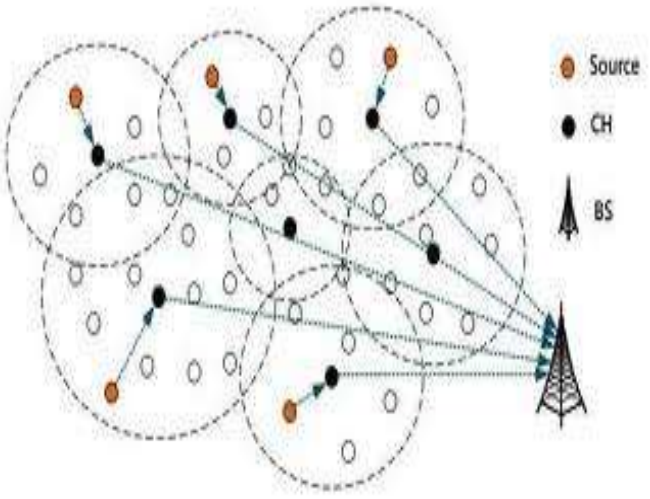

Figure 1 Sensor Network Architecture

WSN has various characteristics like Ad Hoc deployment, Dynamic network topology, Energy Constrained operation, Shared bandwidth, large scale of deployment. Despite of these characteristics routing in WSN is more challenging. Firstly, resources are greatly constrained in terms of power supply, processing capability and transmission bandwidth. Secondly, it is difficult to design a global addressing scheme as Internet Protocol (IP). Furthermore, IP cannot be applied to WSNs, since address updating in a large-scale or dynamic WSN can result in heavy overhead. Thirdly, due to the limited resources, it is hard for routing to cope with unpredictable and frequent topology changes, especially in a mobile environment. Fourthly, data collection by many sensor nodes usually results in a high probability of data redundancy, which must be considered by routing protocols. Fifthly, most applications of WSNs require the only communication scheme of many-to-one, *i.e.*, from multiple sources to one particular sink, rather than multicast or peer to peer. Finally, in time-constrained applications of WSNs, data transmissions should be accomplished within a certain period of time. Thus, bounded latency for data transmissions must be taken into consideration in this kind of applications.

Based on network structure, routing protocols in WSNs can be coarsely divided into two categories: flat routing and hierarchical routing. In a flat topology, all nodes perform the same tasks and have the same functionalities in the network. Data transmission is performed hop by hop usually using the form of flooding. In small-scale networks flat routing protocols are relatively effective. However, it is relatively undesirable in large-scale networks because resources are limited, but all sensor nodes generate more data processing and bandwidth usage. On the other hand, in a hierarchical topology, nodes perform different tasks in WSNs and typically are organized into lots of clusters according to specific requirements or metrics. Generally, each cluster comprises a leader referred to as cluster head (CH) and other member nodes (MNs) or ordinary nodes (ONs), and the CHs can be organized into further hierarchical levels. In general, nodes with higher energy act as CH and perform the task of data processing and information transmission, while nodes with low energy act as MNs and perform the task of information sensing.

2. CLUSTERING

A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. A good clustering algorithm is able to identify clusters irrespective of their shapes. Other requirements of clustering algorithms are scalability, ability to deal with noisy data, insensitivity to the order of input records, etc [5]. In the hierarchical network structure each cluster has a leader, which is also called the cluster head (CH) and usually performs the special tasks referred above (fusion and aggregation), and several common sensor nodes (SN) as members. The cluster formation process eventually leads to a two-level hierarchy where the CH nodes form the higher level and the cluster-member nodes form the lower level. The sensor nodes periodically transmit their data to the corresponding CH nodes. The CH nodes aggregate the data and transmit them to the base station (BS) either directly or through the intermediate communication with other CH nodes. However, because the CH nodes send all the time data to higher distances than the common (member) nodes, they naturally spend energy at higher rates. A common solution in order to balance the energy consumption among all the network nodes is to periodically re-elect new CHs each cluster. CH nodes aggregate the data and transmit them to the base station (BS) either directly or through the intermediate communication with other CH nodes. A typical example of the implied hierarchical data communication within a clustered network is illustrated in Figure 2.

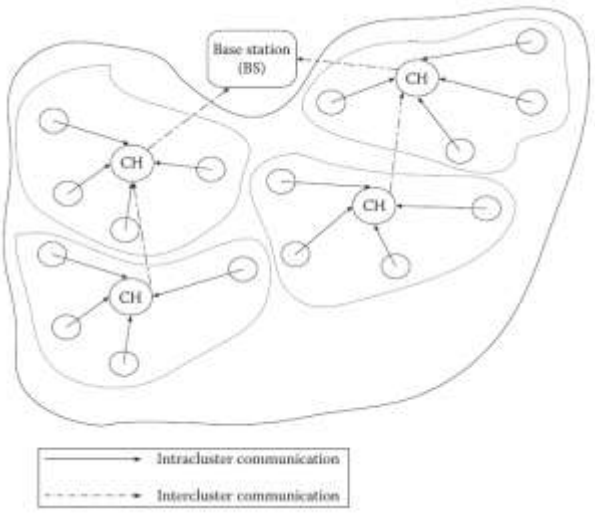

Figure 2 Data Communication in Clustered Network

2.1 Classification of Clustering

- **One Hop Model**

This is the simplest approach and represents direct communication. In these networks every node transmits to the base station directly. This communication implies not only to be too expensive in terms of energy consumption, but it is also infeasible because nodes have limited transmission range [6],[7],[8]. Most of the nodes in networks with large area coverage usually are far enough thus their transmissions cannot reach the base station. Direct communication is not a feasible model for routing in WSN.

- **Multi-hop Planar Model**

In this model, a node transmits to the base station by forwarding its data to one of its neighbors, which is closer to the base station. The latter passes on it to neighbors that is even closer to the base station. Thereby the information travels from source to destination by hop from one node to another until it reaches the destination. Regarding to the energy and transmission range node limitations, this model is a viable approach. A number of protocols employ this approach like [9][10][11][12], and some use other optimization techniques to enhance the efficiency of this model. One of these techniques is data aggregation used in all clustering-based routing protocol, for instance in [13] and [14]. Even though these optimization techniques improve the performance of this model, it is still a planar model. In a network composed by thousands of sensors, this model will exhibit high data dissemination latency due to the long time needed by the node information to arrive to the base station [15], [16].

- **Clustering-based Hierarchical Model**

A hierarchical approach for the network topology breaks the network into several areas called clusters as shown in figure 3. Nodes are grouped depending on some parameter into clusters with a cluster head, which has the responsibility of routing the data from the cluster to other cluster heads or base stations. Data travels from a lower clustered layer to a higher one. Data still hops from one node to another, but since it hops from one layer to another it covers larger distances and moves the data faster to the base station than in the multi hop model [17],[18],[19],[20].

The latency in this model is theoretically much less than in the multi-hop model. Clustering provides inherent optimization capabilities at the cluster heads, what results in a more efficient and well structured network topology. This model is more suitable than one hop or multi hop model. The multi-hop model is a more practical approach than in one hop. In this case, data is forwarded by hops from one node to another until it reaches the base station.

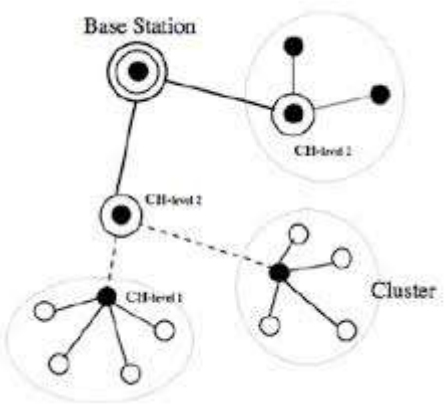

Figure 3 Hierarchical clustering-based Model

Some drawbacks of this model are the high latency in networks comprised of thousands of sensors and the serious delay that data experiences. Perhaps the most important drawback is that the closest nodes to the base station would have to act as intermediaries to all traffic being sent to the base station by the rest of the network.

3. DATA AGGREGATION

Data aggregation is a process of aggregating the sensor data using aggregation approaches. The general data aggregation algorithm works as shown in fig 4. The algorithm uses the sensor data from the sensor node and then aggregates the data by using some aggregation algorithms such as centralized approach, LEACH(low energy adaptive clustering hierarchy),TAG(Tiny Aggregation) etc. This aggregated data is transfer to the sink node by selecting the efficient path [21].

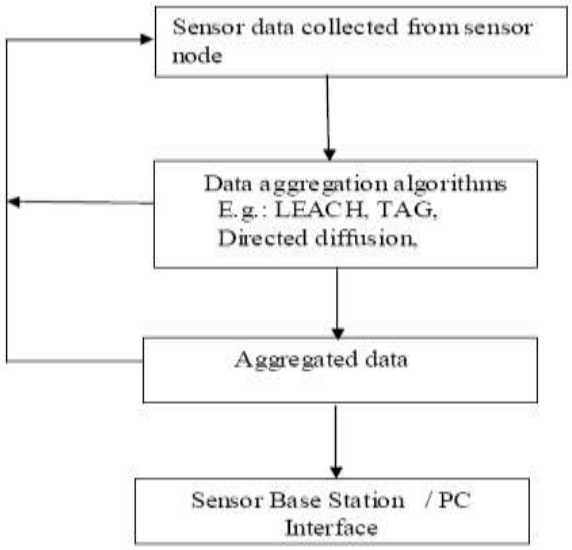

Figure 4 General architecture of the data aggregation algorithm

Data aggregation, which is the process of aggregating the data from multiple nodes to eliminate redundant transmission and provide fused data to BS, is considered as an effectual technique for WSNs to save energy. The most popular data aggregation algorithms are cluster-based data aggregation algorithms, in which the nodes are grouped into clusters and each cluster consists of a cluster head (CH) and some members, each member transmits data to its CH, then, each CH aggregates the collected data and transmits the fused data to BS. The cluster-based WSNs have an inherent problem of unbalanced energy dissipation. Some nodes drain their energy faster than others and result in earlier failure of network. Some researchers have studied this problem and proposed their algorithms which have both advantages and disadvantages. Our motivation is to propose a novel solution to this problem in the cluster-based and homogeneous WSNs, in which the CHs transmit data to BS by one-hop communication, with an objective of balancing energy consumption by an energy efficient way and prolonging network lifetime.

4. DATA AGGREGATION PROTOCOLS BASED ON NETWORK ARCHITECTURE

4.1 Flat Networks

In flat networks, each sensor node plays the same role and is equipped with approximately the same battery power. In such networks, data aggregation is accomplished by data centric routing where the sink usually transmits a query message to the sensors, e.g, via flooding and sensors which have data matching the query send response messages back to the sink. The choice of a particular communication protocol depends on the specific application at hand.

4.1.1 Push Diffusion

In the push diffusion scheme, the sources are active participants and initiate the diffusion while the sinks respond to the sources. The sources flood the data when they detect an event while the sinks subscribe to the sources through enforcements. The *sensor protocol for information via negotiation* (SPIN) [22] can be classified as a push based diffusion protocol.

4.1.2 Two Phase Pull Diffusion

Directed diffusion is a representative approach of two phase pull diffusion. It is a data centric routing scheme which is based on the data acquired at the sensors. The attributes of the data are utilized message in the network. Figure 5 illustrates the interest propagation in directed diffusion. If the attributes of the data generated by the source match the interest, a gradient is set up to identify the data generated by the sensor nodes. The sink initially broadcasts an interest message in the network. The gradient specifies the data rate and the direction in which to send the data. Intermediate nodes are capable of caching and transforming the data. Each node maintains a data cache which keeps track of recently seen data items. After receiving low data rate events, the sink reinforces one particular neighbor in order to attract higher quality data. Thus, directed diffusion is achieved by using data driven local rules.

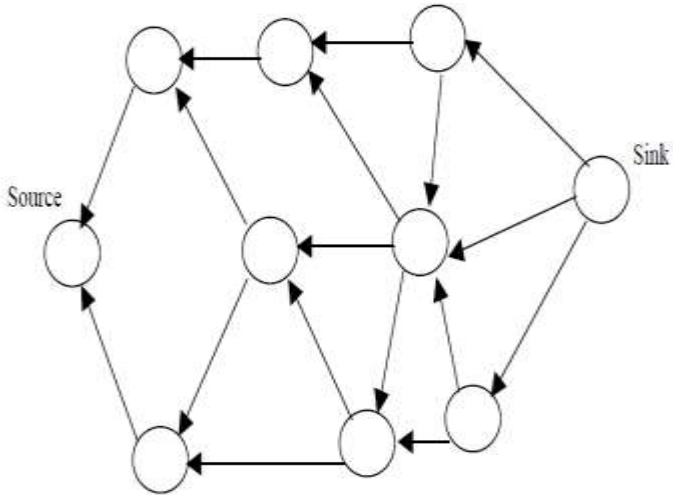

Figure 5 Interest propagation in directed diffusion

4.1.3 One Phase Pull Diffusion

Two phase pull diffusion results in large overhead if there are many sources and sinks. Krishnamachari et al. [23] have proposed a one phase pull diffusion scheme which skips the flooding process of directed diffusion. In one phase pull diffusion, sinks send interest messages that propagate through the network establishing gradients. However, the sources do not transmit exploratory data. The sources transmit data only to the lowest latency gradient pertinent to each sink. Hence, the reverse route (from the source to the sink) has the least latency. Removal of exploratory data transmission results in a decrease in control overhead conserving the energy of the sensors.

4.2 Hierarchical Networks

A flat network can result in excessive communication and computation burden at the sink node resulting in a faster depletion of its battery power. The death of the sink node breaks down the functionality of the network. Hence, in view of scalability and energy efficiency, several hierarchical data aggregation approaches have been proposed. Hierarchical data aggregation involves data fusion at special nodes, which reduces the number of messages transmitted to the sink. This improves the energy efficiency of the network.

4.2.1 Data Aggregation in Cluster based networks

In energy constrained sensor networks of large size, it is inefficient for sensors to transmit the data directly to the sink. In such scenarios, sensors can transmit data to a local aggregator or cluster head which aggregates data from all the sensors in its cluster and transmits the concise digest to the sink. This results in significant energy savings for the energy constrained sensors. Figure 6 shows a cluster based sensor network organization. The cluster heads can communicate with the sink directly via long range transmissions or multi hopping through other cluster heads.

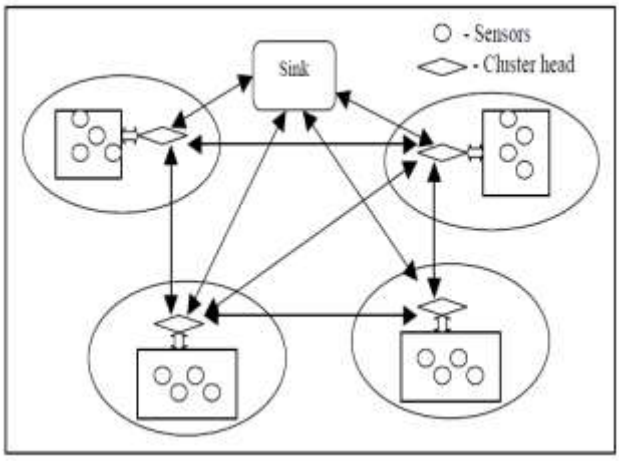

Figure 6 Cluster based Network

4.2.2 Chain based Data Aggregation

In cluster-based sensor networks, sensors transmit data to the cluster head where data aggregation is performed. However, if the cluster head is far away from the sensors, they might expend excessive energy in communication. Further improvements in energy efficiency can be obtained if sensors transmit only to close neighbors. The key idea behind chain based data aggregation is that each sensor transmits only to its closest neighbor. Lindsey et al. [24] presented a chain based data aggregation protocol called power efficient data gathering protocol for sensor information systems (PEGASIS). In PEGASIS, nodes are organized into a linear chain for data aggregation. The nodes can form a chain by employing a greedy algorithm or the sink can determine the chain in a centralized manner. Greedy chain formation assumes that all nodes have global knowledge of the network. The farthest node from the sink initiates chain formation and at each step, the closest neighbor of a node is selected as its successor in the chain. In each data gathering round, a node receives data from one of its neighbors, fuses the data with its own and transmits the fused data to its other neighbor along the chain. Eventually the leader node which is similar to cluster head transmits the aggregated data to the sink. Figure 7 shows the chain based data aggregation procedure in PEGASIS.

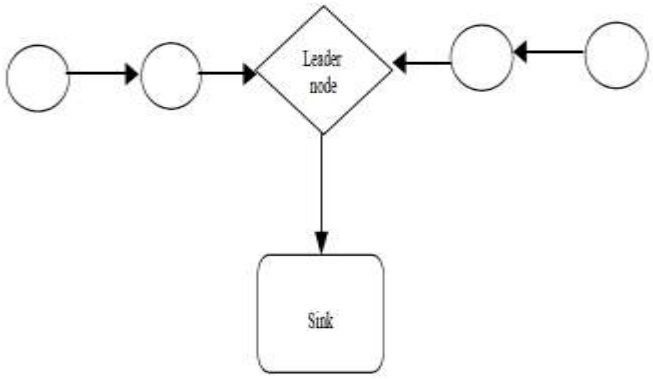

Figure 7 Chain based organization in a sensor network

The PEGASIS protocol has considerable energy savings compared to LEACH. The distances that most of the nodes transmit are much less compared to LEACH in which each node

transmits to its cluster head. The leader node receives at most two data packets from its two neighbors. In contrast, a cluster head in LEACH has to perform data fusion of several data packets received from its cluster members. The main disadvantage of PEGASIS is the necessity of global knowledge of all node positions to pick suitable neighbors and minimize the maximum neighbor distance.

4.2.3 Tree based Data Aggregation

In a tree based network, sensor nodes are organized into a tree where data aggregation is performed at intermediate nodes along the tree and a concise representation of the data is transmitted to the root node. Tree based data aggregation is suitable for applications which involve in-network data aggregation. An example application is radiation level monitoring in a nuclear plant where the maximum value provides the most useful information for the safety of the plant. One of the main aspects of tree-based networks is the construction of an energy efficient data aggregation tree.

4.2.4 Grid based Data Aggregation

In grid-based data aggregation, a set of sensors is assigned as data aggregators in fixed regions of the sensor network. The sensors in a particular grid transmit the data directly to the data aggregator of that grid. Hence, the sensors within a grid do not communicate with each other. In this aggregation, the data aggregator is fixed in each grid and it aggregates the data from all the sensors within the grid. This is similar to cluster based data aggregation in which the cluster heads are fixed. Grid-based data aggregation is suitable for mobile environments such as military surveillance and weather forecasting and adapts to dynamic changes in the network and event mobility. Figure 8 shows that in grid based data aggregation, all sensors directly transmit data to a predetermined grid aggregator.

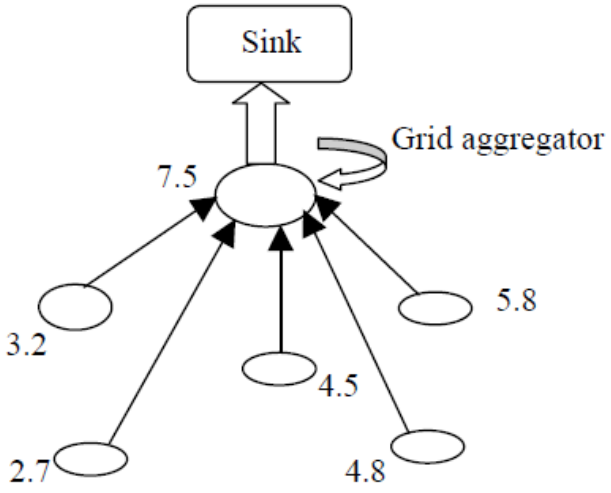

Figure 8 Grid based Data Aggregation

5. LEACH PROTOCOL

LEACH performs local data fusion to "compress" the amount of data being sent from the clusters to the base station, further reducing energy dissipation and enhancing system lifetime. Sensors elect themselves to be local cluster-heads at any given time with a certain probability. These cluster head nodes broadcast their status to the other sensors in the network. Each sensor node determines to which cluster it wants to belong by choosing the cluster-head that requires the minimum communication energy. Once all the nodes are organized into clusters, each cluster-head creates a schedule for the nodes in its cluster. This allows the radio components of each non-cluster-head node to be turned off at all times except during its transmit time, thus minimizing the energy dissipated in the individual sensors. Once the cluster-head has all the data from the nodes in its cluster, the cluster-head node aggregates the data and then transmits the compressed data to the base station.[25]
LEACH is self adaptive and self-organized. This protocol uses round as unit, each round is made up of cluster set-up stage and steady-state stage, for the purpose of reducing unnecessary energy costs, the steady state stage must be much longer than the set-up stage. The process of it is shown in Figure 9.

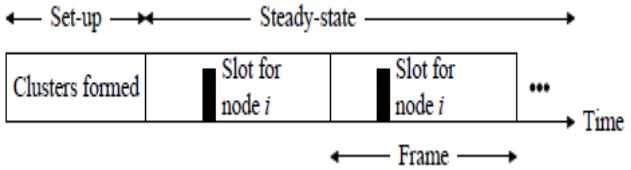

Fig.9 LEACH Protocol process

At the stage of cluster forming, a node randomly picks a number between 0 to 1, compared this number to the threshold values $t(n)$, if the number is less than $t(n)$, then it become cluster head in this round, else it become common node. Threshold $t(n)$ is determined by the following:

$$t(n) = \begin{cases} \dfrac{p}{1 - p * (r \bmod \frac{1}{p})} & if\ n \in G \\ 0 & if\ n \notin G \end{cases}$$

Where p is the percentage of the cluster head nodes in all nodes, r is the number of the rounds, G is the collections of the nodes that have not yet been head nodes in the first 1/P rounds. Using this threshold, all nodes will be able to be head nodes after 1/P rounds.[26]

6. ACKNOWLEDGEMENT

I would like to thanks the faculty member of the Computer Science & Engineering Department of Geeta Institute of Engineering and Techology, kanipla (District Kurukshetra).

7. REFERENCES

[1] D. Culler, D. Estrin, and M. Srivastava, "Overview of Sensor Networks," *IEEE Computer*, August 2004.

[2] N. Xu, S. Rangwala, K. Chintalapudi, D. Ganesan, A. Broad, R. Govindan, and D. Estrin, "A Wireless Sensor Network for Structural Monitoring," *Proceedings of the ACM Conference on Embedded Networked Sensor Systems, Baltimore, MD*, November 2004.
[3] A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler, and J. Anderson, "Wireless Sensor Networks for Habitat Monitoring," *WSNA'02, Atlanta,Georgia*, September 2002.
[4] Xuxun Liu, "A Survey on Clustering Routing Protocols in Wireless Sensor Networks,"*Sensors* 2012, *12*, 11113-11153; doi:10.3390/s120811113, 9 August 2012.
[5] Amandeep Kaur Mann, Navneet Kaur, "Survey Paper on Clustering Techniques," International Journal of Science, Engineering and Technology Research (IJSETR) Vol 2, Issue 4, April 2013
[6] S. Meguerdichian, S.Slijepcevic, V.Karayan and M.Potkonjak, "Localized Algorithms in Wireless Ad-Hoc Networks: Location Discovery and sensor Exposure," in Proc. Of Mob adhoc, Long Beach, CA, USA, pp. 106-116, 2001.
[7] Wang Wei, "Study on Low Energy Grade Routing Protocols of Wireless Sensor Network," Dissertation, Hang Zhou, Zhe Jiang University, 2006.
[8] Wei Bo, Hu Han-ying, Fu Wen, "A pseudo LEACH algorithm for Wireless Sensor Network" in Proc. IAENG, March 2007.
[9] Fan Xiangning, "Improvement on LEACH Protocol on Wireless Sensor Network," mt. Conference on Sensor Technologies and Applications, 7 July, 2007.
[10] Haiming Yang, Biplab Sikdar, "Optimal Cluster Head Selection in the LEACH Architecture," Performance, Computing and communications Conference, Int. 2, 2007.
[11] Haosong Gou, "An Energy Balancing LEACH Algorithm for Wireless Sensor Networks," 7th Conference on Information Technology, 3 October, 2010.
[12] Hu Junping, "A Time-based Cluster-Head Selection Algorithm for LEACH", IEEE, 1 August, 2008.
[13] A. Manjeshwar and D.Agrawal, "TEEN: A Routing Protocol for Enhanced Efficiency in Wireless Sensor Networks," in Proc. 1st Workshop on Parallel and Distributed Computing Issues in Wireless Networks and Mobile Computing, San Francisco, CA, pp. 2009-2015, April 2001.
[14] Bilal Abu Bakr, "Extending Wireless Sensor Network Lifetime in the LEACH-SM Protocol by Spare Selection," 5th Int. Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, July,2011.
[15] W.R. Heizelman, A.Chandrakasan and H.Balakrishnan, "Energy-Efficient Communication Protocol for Wireless Microsensor Networks", Proc. 33rd Hawaii Int. Conference on System Science, Vol. 2, 4-7 Jan,2000.
[16] Ye W, Heidenman J Esrtrin D, "An Energy-Efficient MAC Protocol for Wireless Sensor Network," in proc. IEEE INFOCOM, http://www.isi.edu/div7/publication files/yeO2a.pdf , 2002.
[17] M. Dong, K. Yung and W. Kaiser, "Low Power Signal Processing Architectures for Network Microsensors," in Proc. Int. Symposium on Low Power Electronics and Design, pp 173-177, Aug, 1997.
[18] Mo Xiaoyan, "Study and Design on Cluster Routing Protocols of Wireless Sensor Networks," Dissertation, hang Zhou, Zhe Jiang University, 2006.
[19] Mu Tong, "LEACH-B: An Improved LEACH Protocol for Wireless Sensor Network," IEEE, 2010.
[20] O. Younis and S. Fahmy, "HEED: A Hybrid,Energy-Efficient, Distributed Clustering Approach for Ad Hoc Sensor Networks," Trans. Mobile Computing, Vol. 3, No. 4, pp 336-379, Oct-Dec, 2004
[21] Nandini. S. Patil, Prof. P. R. Patil, "Data Aggregation in Wireless Sensor Network," IEEE International Conference on Computational Intelligence and Computing Research,2010
[22] J. Kulik, W.R. Heinzelman and H. Balakrishnan, "Negotiation-based protocols for disseminating information in wireless sensor networks," *Wireless Networks*, vol. 8, March 2002, pp. 169-185.
[23] B. Krishnamachari and J. Heidemann, "Application specific modeling of information routing in wireless sensor networks", *Proc. IEEE international performance, computing and communications conference*, vol. 23, pp. 717-722, 2004.
[24] S. Lindsey, C. Raghavendra, and K.M. Sivalingam, "Data gathering algorithms in sensor networks using energy metrics," *IEEE Trans. Parallel and Distributed Systems*, vol. 13, no. 9, September 2002, pp. 924-935.
[25] Wendi Rabiner Heinzelman, Anantha Chandrakasan, and Hari Balakrishnan, "Energy efficient Communication protocol for Wireless Microsensor Networks," Published in the Proceedings of the Hawaii International Conference on System Sciences, January 4-7, 2000, Maui, Hawaii.
[26] Chunyao FU, Zhifang JIANG, Wei WEI2and Ang WEI, "An Energy Balanced Algorithm of LEACH Protocol in WSN," IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 1, No 1, January 2013.