ANOVA-based Clustering Approach For Similarity Aggregation In Underwater Wireless Sensor Networks Using An Enhanced K-means Algorithm

Namrata R Mire[1], Prof. S D Patil[2]

[1] Student of Electronics and Telecommunication Department, Rajiv Gandhi Institute of Technology Mumbai, Maharashtra, India

[2] Associate Professor of Electronics and Telecommunication Department, Rajiv Gandhi Institute of Technology Mumbai , Maharashtra, India

---***---

Abstract - *Wireless sensor network is one of the emerging field of communication world. These networks collect the information from the environment and deliver the same to the application to determine the characteristics of the environment and to detect an event. Saving energy and thus extending the wireless sensor network lifetime entails great challenges. In general, all sensor nodes directly send the information to the Base station so the energy requirement is very high. Clustering is used to decrease energy consumption and collision. In this paper, we present a new clustering method to handle the spatial similarity between node readings. In general the readings are sent periodically from sensor nodes to their appropriate cluster-heads (CHs). At the first level, each node periodically cleans its readings in order to eliminate redundancies before sending its data set to its CH. Once the CH receives all data sets, it applies an enhanced K-means algorithm based on a one-way ANOVA model to identify nodes generating identical data sets and to aggregate these sets before sending them to the sink. In our proposed work we also use distributed energy-efficient clustering algorithm for efficient communication between cluster head and base station. We perform MATLAB simulation to observe the results in network stability, no. of nodes dead ,no. of nodes alive, and data transfer between baste station to sink. Our proposed approach is validated via experiments on real sensor data and comparison with other existing clustering and data aggregation techniques.*

***Key Words*: Similarity Data Aggregation, Hierarchical K-means Clustering, Underwater Wireless Sensor Network , Distributive Energy Efficient Clustering Algorithm**

1. INTRODUCTION

Wireless Sensor Networks (WSN) are one of the innovative technologies that are widely used today. They have many advantages, namely the ease of deployment and the capacity of self-organization. However, the main challenge of these networks concerns the limited resources in terms of energy, communication and computation. Yet, focus in the present paper will be on the former challenge (i.e. energy consumption). To deal with this issue, a number of proposals have been put forward, among these are clustering algorithms that are widely used to save WSN energy .Though sensors are usually powered by batteries, it is not always practical to recharge or replace them because they are often deployed in hostile environments. Furthermore, in a WSN a large amount of energy is consumed when communications are established between the networks [1]. Hence, frequent and long distance transmissions should be minimized to extend the lifetime of the network [2]. An effective approach would be to divide the network into several clusters; each of these elects one node as its cluster head (CH) [3]. The CH collects data from the same cluster nodes, aggregates them and transmits them to the base station (BS). Many routing protocols based on clustering in which CHs are elected periodically and alternately [4, 5], have been designed; otherwise, the CH dies quickly. Moreover, the most commonly used approach for clustering is the Low-Energy Adaptive Clustering Hierarch (LEACH) algorithm [6]. The main idea of the LEACH algorithm is to randomly and alternately select the CHs. After CHs are selected, each node in the WSN will receive. warning messages from all CHs. But, two problems are brought about: (1) what is the optimal number of clusters to consider? and (2) how can we define CHs? To deal with these issues, and to improve the lifetime of a network, our purpose in this paper is to propose a novel clustering protocol based on similarity data aggregation approach called k-ways. In addition, many clustering protocols, based on the principle of this algorithm, have been developed in the two categories of Wireless Sensor Networks: homogeneous and heterogeneous WSNs. In the first category, all nodes have the same initial energy whereas in the second set, they have different energies. For the first category we give some homogeneous clustering protocols: centralized LEACH (LEACH-C)[7], Power-Efficient Gathering in Sensor Information Systems (PEGASIS)[8], and Distance-Energy Cluster Structure Algorithm (DECSA)[9]. Concerning the second category, we cite the examples of Developed Distributed Energy- Efficient Clustering (DEEC) [10], Equitable Distributed Energy-Efficient Clustering (EDEEC) [11].Here we are combining the two algorithm for the betterment of result. First is K-means algorithm for clustering purpose and DEEC algorithm to enhance the life time of the network. In section 2, we review the related works in this field. Section 3 will depict the energy model that could be used by the sensors. Section 4 contains simulation results to compare all the three under various performance measure matrices. Finally, in section 5, we present the conclusion.

2. REVIEW OF CLUSTERING ALGORITHM FOR WIRELESS SENSOR NETWORKS

John Heidemann, et al.[1] from Information Sciences Institute, University of Southern California Research did research on Challenges and Applications for Underwater Sensor Networking here they explores applications and challenges for underwater sensor networks. They highlight potential applications to off-shore oilfields for seismic monitoring, equipment monitoring, and underwater robotics. They identify research directions in short range acoustic communications, MAC, time synchronization, and localization protocols for high-latency acoustic networks, long duration network sleeping, and application-level data scheduling. but after implementing this they find a range problem for data transmission

Moumita Pramanick, et al.[3] Department of Computer Science and Engineering Jadavpur University, Kolkata, India Northumbria University United Kingdom here they implement energy aware sleeping clustering based routing scheme (EASSCER)for wire less sensor node is produced. In this system they overcome the disadvantage of previous paper but after implementing they got problem regarding the security .so introduces new application

S. Siva Ranjani Kalasalingam University Krishnankoil,et al.[5] develop new algorithm. Because of the open deployment, sensors are vulnerable for security threats. In this paper they address the data aggregation and security issues together. They modify Energy efficient Cluster Based Data Aggregation (ECBDA)[1] scheme to provide secure data transmission. for the security aspect Bayesian fusion algorithm to enable security

Hassan Harb, et al.[7], did research and created paper on A Suffix-Based Enhanced Technique for Data Aggregation in Periodic Sensor Networks, 10th IEEE Int. Wireless Communications and Mobile Computing Conference(IWCMC 2014), 2014. The research on acoustic underwater sensor networks has significantly advanced in the last decades, energy consumption still remains the major challenge to overcome. A two-tier data aggregation based transmission-efficient technique for periodic UASN which applies at each cluster separately in a clustering network. but increasing network life time is still a problem.

3. Problem formulation and Methodology Used

The wireless sensor network nodes sense the environment and transmit data to the BS. The latter analyzes data and gives some conclusions about the activities in the supervised area. In our work, we use the energy mode and analysis that are presented in [6, 7].The radio energy dissipation model is illustrated in Figure 1 (extracted from [6]

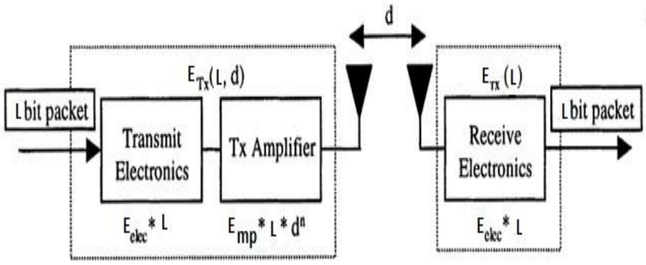

Figure 1: Radio Energy Dissipation Model

In this figure, E_{Tx} (L, d) presents the energy spent to transmit L-bits over a distance d, and E_{rx} is the energy spent to process L-bit message. In this model, both the free space (d2 power loss) and the multipath fading (d4 power loss) channel models were used, depending on the distance between the transmitter and receiver [2,9]. Power control can be used to invert this loss by appropriately The parameter E_{elc} denotes the energy per bit dissipated to run both the transmitter and the receiver circuits. This parameter depends on many factors such as digital coding, modulation, filtering, and the spreading of the signal [7].

$$E_{Tx}(L, d) = \begin{cases} \mathrm{L.\,E_{elec}} + \mathrm{L.\,E_{fs}.\,d^2} & \text{if d} < do \\ \mathrm{L.\,E_{elec}} + \mathrm{L.\,E_{amp}\,.\,d^4} & \text{if d} > do \end{cases}$$

(1)

where E_{fs} and E_{mp} present the amplifier energy respectively in a free space (with d^2 power loss) and in a multipath fading (with d^4 power loss) channel models. They depend on the distance between the transmitter and the receiver. If this distance is less than a threshold d_0, then the free space model is used; otherwise, the multipath model is used. The value of the threshold d_0 has been given by Heinzelman et al. in [7]. It is defined as follows:

$$do = \frac{E_{fs}}{E_{amp}} \qquad (2)$$

Furthermore, considering a WSN which consists of N nodes uniformly distributed the total dissipated energy during a round is determined by:

$$E_{Round} = \sum_{k=1}^{K} E_{CH_k} + \sum_{j=1}^{N-K} E_{NCH_j} \quad (3)$$

where E_{CHk} is the consumed energy when the CH of the cluster labelled k, receives, aggregates, and transmits data to the base station. Whereas E_{NCHj} is the consumed energy by a non CH labelled j, and K is the total number of cluster heads

3.1 Two-level heterogeneous networks

We have used two types of nodes in the network, normal and advanced nodes. Eo is the initial energy of the normal nodes, and m is the fraction of the advanced nodes, which own a times more energy than the normal ones. Thus there are m.N advanced nodes equipped with initial energy of Eo.(1 + a), and (1 - m).N normal nodes that are equipped with initial energy of Eo. The total initial energy of the two-level heterogeneous networks is given by

$$\mathrm{E_{total}} = \mathrm{N.(1-m).Eo + m.N.(1+a).Eo}$$
$$= \mathrm{N.Eo.(1+a.m)} \qquad (4)$$

Therefore, the two-level heterogeneous networks have a. m times more energy and virtually a .m more nodes [7]

3.2 Multilevel heterogeneous networks

In multi-level heterogeneous networks, initial energy of sensor nodes is randomly distributed over the close set [Eo, Eo (1 + amax)], where Eo is the lower bound and amax determine the value of the maximal energy. Initially, the node si is equipped with initial energy of Eo. (1 + ai), which is ai times more energy than the lower bound Eo. The total initial energy of the multi-level heterogeneous networks is given by [7]

$$E_{total} = \sum_{i=1}^{N} Eo(1 + a_i) = Eo(N + \sum_{i=1}^{N} a_i) \qquad (5)$$

4. PROPOSED ALGORITHM

Here from the various clustering approach we are using k-means algorithm for clustering the data by selecting appropriate cluster head. K-means algorithm consists of dividing data into K disjoint classes based on the K eigenvectors related to K largest Eigen values of a Laplacian matrix. In our study, we consider a network with N nodes, uniformly distributed within a M×M square region. Moreover we assume that the network topology remains unchanged over time.

The three steps of our proposal are

4.1 Pre-processing step

So as to avoid that each node needs to know the global knowledge of the network (which is quite unrealistic for WSN), in the proposed algorithm, the base station collects the different node positions and applies the clustering process. We note that each node knows its own location, which can be obtained at a low cost by a Global Positioning System or by using other localization systems [18, 19]. Then, the WSN nodes transmit their location in a short message to the Base Station. The degree matrix $D \in R^{N\times N}$ of G is a diagonal matrix defined by $D=[d_{ij}]$ and the N×N Laplacian matrix of the graph is defined by

$$L = D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \qquad (6)$$

The objectives of the current step are to define the optimal number of clusters and to form them. Based on the Laplacian matrix L defined above, we form a new matrix U composed of the K eigenvectors related to K largest Eigen values of L. In order to determine the K clusters of the WSN, we apply the classification algorithm k-means to the matrix U.

Nonetheless, the most important question raised by the proposed strategy concerns the optimal number of clusters (K) to be used. With the aim to respond to this question, we consider the total consumed energy in each round (equation (9)).We note that by considering K cluster, this energy depends on the distances between the CH and the non cluster heads of each cluster. i.e. ERound . The objective function that allows to decide whether to reconsider the partitioning process or not of the WSN, is defined by the distance matrix M_{dis}^{k} ($M_{dis}^{k} = [dis_{ij}^{k}]$; with dis_{ij}^{k} is the distance between the node i and the node j of the cluster labeled k) of each cluster. The allowed threshold to this function is d_0. Hence, if at least one element of any M_{dis}^{k} is greater than d_0, the considered number of clusters will be incremented (K+1) and the k-mean algorithm will be reused. Otherwise, the optimal number of clusters is K. In the proposed algorithm we first determine the clusters before specifying the CHs. Besides, the optimal number of cluster partitions is as well defined automatically. Therefore, our algorithm is completely different from the others (such as LEACH, LEACH-C)

4.2 Cluster head election step

Once the clusters are determined, the next step consists in defining the CHs. Note that numbered node id (identification) will be in some random position on the cluster. Thus, the cluster head in each round of communication will be at a random position on the cluster. It is so important that nodes die at random locations of the network. In the round r of the simulation, we use the number c_k=(r mod $|S_k|$) to select the suitable cluster head for the appropriate cluster; where $|S_k|$ represents the total number of nodes in a defined cluster k. Besides, if the residual energy Eng_{ck} of the node, with id=ck, is greater than a threshold Θ_{Eng}, this node will be the CH of the cluster k in the round r. We define Θ_{Eng} as the minimum residual energy required for a given node to be a CH. This Θ_{Eng} is given below

$$\theta_{Eng} = L.((\, |S_k| + 1).E_{elec} + \, |S_k| \, .E_{DA} \;\; + E_{mp}.d^4{}_i$$
(6)

where d_i is the distance between the node i and the BS. Consequently, each cluster head will be able to collect data from the cluster nodes and will transmit the aggregate information to the BS. Thus, the number of the direct transmissions is efficiently reduced and the whole network lifetime is extended. In addition, energy consumption will be distributed with more equitability between all nodes

4.3 Data transmission

Once clusters and cluster heads are created, each CH knows which nodes it is supervising. Based on the node's id in the appropriate cluster, a Time Division Multiple-Access MAC protocol schedule assignment will be generated automatically. If we suppose that the node id=i is elected as CH, the node id = (i + 1 + $|S_k|$)mod $|S_k|$ will take the first time slot to transmit; where $|S_k|$ is the total number of nodes in cluster k. Here, we avoid the techniques applied by traditional algorithms which consume more energy and ask for more synchronization when the CHs are elected. Moreover, this technique guarantees that there are no collisions among data message and also allows the radio components of each non-cluster head node to be turned off at all times except during its transmit time, thus reducing the energy consumed by the individual sensors [7]. Assuming that all nodes can transmit, with enough power, to reach the BS, if the distance between any node and the BS is less than the distance between this node and its corresponding CH, the node will transmit data directly to the BS. Now, each non cluster head sends its data during its allocated transmission time to its respective CH. The latter must keep its receiver on to receive all data. When all data is received, the CH performs signal processing functions to compress the data into a single signal. Once this phase is completed, each CH sends the aggregated data to the BS. In this sub-phase, each non CH can turn off to the sleep mode in order to reduce the consumed energy

5. SIMULATION AND OUTPUT

In this section, we evaluate the performance of DEEC protocol using MATLAB. We consider a wireless sensor network with N = 100 nodes randomly distributed in a 100m X 100m field. Without losing generalization, we assume the base station is in the center of the sensing region. To compare the performance of DEEC with other protocols, we ignore the effect caused by signal collision and interference in the wireless channel. The parameters used in our simulations are shown in Table 1. The protocols compared with DEEC and LEACH is also shown in results. We will consider following scenarios and examine several performance measures. The table 1 shows the simulation parameters used

.Table -1: Simulation Parameters

Parameters	Value
Network field	(100,100)
No. of nodes	100
Eo (Initial energy of normal node)	0.5 J
Message size	4000 Bits
E_{elec}	50 n J/bit
E_{fs}	10nJ/bit/m^2
E_{amp}	0.0013pJ/bit/m^4
E_{DA}	5nJ/bit/signal
do (threshold distance)	70m
P_{opt}	0.1

Case 1 :

Eo=0.5 , a=1

Here in this case the initial energy Eo=0.5 and there are 20 advanced nodes 1 times more energy than normal nodes

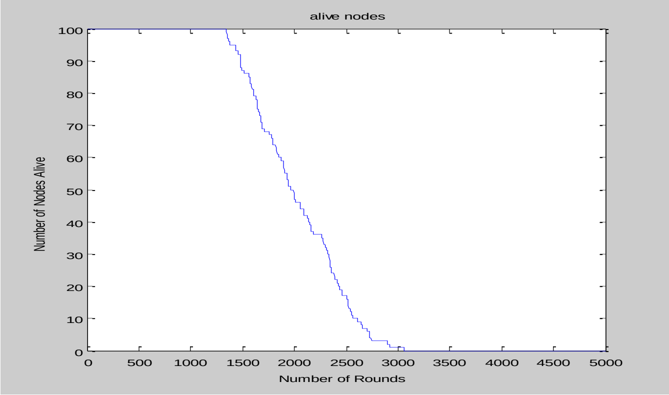

Figure 1: Number of alive nodes over rounds under two-level heterogeneity of DEEC

Figure 1 represent the number of nodes alive during the lifetime of the network. It clearly shows that by introducing super nodes lifetime increases. Stability period and lifetime of DEEC is longer as compared to other .from the figure we can see that the no of alive nodes are increasing round by round .As the last node dead at 3064 the no of alive nodes at that time will be 1936.

Figure 2 shows the conclusion in terms of number of data packets received at the base station. The results show that fort the given protocols it goes linearly for around 3000 rounds and after that the difference can be seen. It is clear DEEC has more numbers of data packets received at base station in comparison to other. From the figure we can conclude that the data transmit and received from packet to base station is varying up till 3000 round. After that the graph goes constant as the last node dies at n=3054

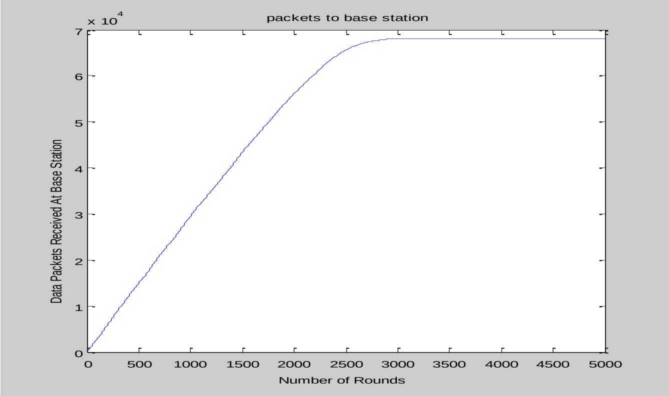

Figure 2 : Data packets over rounds under two-level heterogeneity of DEEC

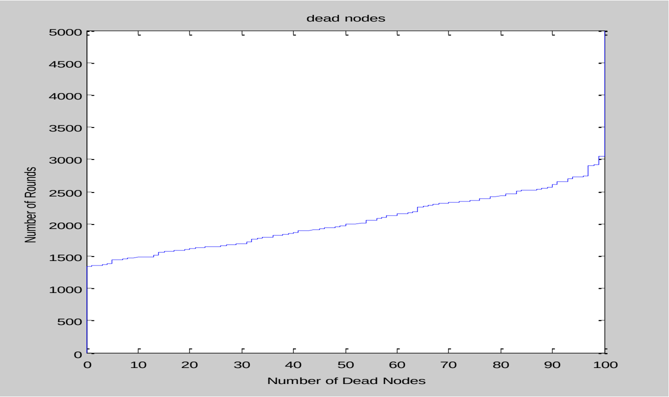

Figure 3: Number of dead nodes over rounds under two-level heterogeneity of DEEC

Figure 3 shows the result in terms of number of dead nodes over 5000 rounds under two level heterogeneity. Here we can see that over the five thousand rounds the First node dies at 1339. at this particular point all the other no of nodes shows the alive node. Similarly the tenth node dies at 1477 and lastly all nodes dies at 3054 and hence at this point as all the nodes are died the communication will get stop. and the total energy will be get calculated.

We notice that for these simulations, the energy of a node decreases each time it sends, receives or aggregates the data.

We propose to compare our proposed method detailed above to:

1. The LEACH protocol; which is the centralized version of the LEACH algorithm. In particular, we chose this protocol because our method is considered as a centralized one.
2. Method1: Here, we define a modified version of our proposed approach. It consists in considering the k-ways algorithm with K equals to 5% of the total number of nodes in the WSN; 5% is the optimal number of CHs proposed in [6]. Our purpose is to highlight the importance of the number of the considered clusters

6. Comparison Between LEACH and DEEC protocol

Here the simulation parameters are same for LEACH as used for DEEC

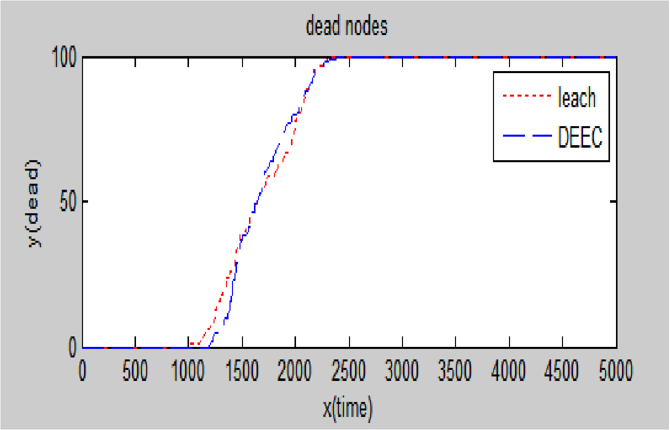

Figure 4: Impact of The Node Density N On The Performances Of The Compared Algorithms

The Figure 4 shows the effects of the node density on the compared clustering techniques as well as on the network's stable regions (First Node Dead "FND"). As shown in this figure, the percentage of dead nodes in LEACH is more than DEEC. we can see that the FND in LEACH at 1000 round where in DEEC FND at 1350.It means the DEEC provides more energy to node for the communication. It follows that even if the node density increases the new proposed approach still gives best results compared to the others protocols. The robustness of the new proposed algorithm is certainly due to the fact that the clustering process is firstly used before the process of the cluster head election. And because of the considered number of clusters that is based on the minimization of the consumed energy in the WSN. Also, it is due to the fact that the residual energy of nodes is considered in the cluster head election step

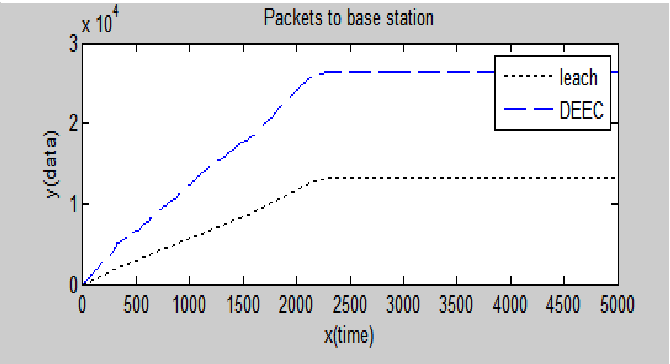

Figure 5:Data Packets over rounds under two-level heterogeneity DEEC and LEACH

Also, figure 5 illustrate that data received at the base station per round for DEEC is more as compared to LEACH. The results show that for both the protocols it goes linearly for around 2000 rounds and after that the difference can be seen. The more number of data is transmitted in DEEC protocol as compared to LEACH protocol.

REFERENCES

[1] John Heidemann, Govindan, R. and Estrin, D. Information Sciences Institute, University of Southern California Research Proceedings of the 6th Annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom00), Boston, MA, USA, pp. 56-67, August, 2000.

[2] Al-Karaki, J., N. and Kamal, A., E. "Routing techniques in wireless sensor networks: a survey", IEEE Wireless Communications, issue 6, pp. 6-28, December, 2004

[3] Moumita Pramanick ., J., Haase, M. and Timmermann, D. " energy aware sleeping clustering based with deterministic cluster-head selection", Proceedings of the IEEE International Conference on Mobile and Wireless Communications Networks (MWCN 2002), Stockholm, Sweden, pp. 368-372, September, 2002.

[4] McDonald, B. and Znati, T. "Design and performance of a distributed dynamic clustering algorithm for Ad-Hoc networks", Proceedings of the Annual Simulation Symposium, pp. 27-35, Seattle, WA, USA, 22-26 April 2001.

[5] P. S. Siva Ranjani Kalasalingam, N.H. Vaidya, M. Chatterjee, D. Pradhan, A, "Cluster-based approach for routing in dynamic networks", ACM, SIGCOMM Computer Communication Review, pp 49-65, 27 (2), 1997.

[6] Heinzelman, W., R., Chandrakasan, A. and Balakrishnan, H. "Energy-efficient communication protocol for wireless microsensor networks", Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS-33), pp. 3005-3014, January 2000.

[7] H. Harb, A. Makhoul, R. Tawil and A. Jaber, A Suffix-Based Enhanced Technique for Data Aggregation in Periodic Sensor Networks, 10th IEEE Int. Wireless Communications and Mobile Computing Conference IWCMC 2014), p. 494-499, 2014.

[8] Lindsey, S., and Raghavenda, C.S. "PEGASIS: power efficient gathering in sensor information systems", Proceeding of the IEEE Aerospace Conference}, Big Sky, Montana, March 2002.

[9] Yonga, Z. and Peia, Q.,"A Energy- Efficient Clustering Routing Algorithm Based on distance and Residual Energy for Wireless Sensor Networks", International Workshop on Information and Electronics Engineering (IWIEE), Harbin, China, March 10-11, 2012

[10] ALI JORIO, 2SANAA EL FKIHI A New Clustering Algorithm For Wireless Sensor Networks, Journal of Theoretical and Applied Information Technology 31st March 2013

BIOGRAPHIES

Namrata R Mire received her B.E. degree in Electronics and Telecommunication Engineering from Rashtrasant Tukadoji Maharaj Nagpur university, Nagpur in 2013. Presently she is pursuing M.E. degree in Electronics and Telecommunication Engineering from University of Mumbai. Her area of interest includes Wireless Communication and Communication networking. email id :- namratam545@gmail.com

Shrikrishna D Patil is an Associate Professor from the Department of Electronics and Telecommunication Engineering, Rajiv Gandhi Institute of Technology, University of Mumbai, India with 28 years of teaching experience. He's having Area of Specialization are Computer Communication Networks, Wired and Wireless Communication, Wireless Sensor Network, Digital Communication, Principles of Communication Engineering, Control System. He is fellow member of IETE. He has acquired publications in many reputable journals. email id:- sdpatil65@rediffmail.com