Uniform Cluster Distribution with Anchor Free Node Localization for Energy Efficient Data Gathering in Wireless Sensor Networks

Bhaskar P. Deosarkar, Member, IACSIT and R.P.Yadav, FIET*E*

Abstract—Data gathering is an important research area in wireless sensor networks, being it the core operation. As an efficient topology management approach clustering is shown to have advantage, over other network architectures, due to its capability to reduce the communication burden and utilization of the processing features of the sensor nodes for local data aggregation. The algorithms proposed for this purpose may be classified as centralized or distributed based on the place of decision making. Centralized algorithms are revealed to lack the scalability in terms node density and distributed algorithms have an edge over them in such scenario. For distributed algorithms, it is needed to distribute the clusters uniformly over the network area, for them to be more energy efficient. Acquisition of the knowledge about their locations, by the sensor nodes, is expected to prove to be useful for this distribution but due to their limited energy resources and the associated cost providing location finding capabilities, like GPS or anchor nodes, is not advisable. If provided such facility, the error if occurs, due to its own limitations, may get propagated over the entire network. So, a scheme is proposed in this paper, for uniform distribution of the clusters over the network area in which the sensor nodes, utilizing anchor free and distributed localization mechanism, estimate their relative positions and based on these estimates assign to different clusters with an expectation from RSSI to provide the necessary support for required distance estimation. Sensor nodes, with knowledge of their distances to the base station, estimate their locations through collaborative efforts of small percentage of total nodes, with minimum of communication burden. Results of the simulations carried out in NS2, show that the proposed scheme has a potential to distribute the clusters uniformly over network area.

Index Terms—Clustering, cluster distribution, Node localization, RSSI

I.INTRODUCTION

Wireless sensor networks composed of large number of tiny sensor nodes, that are capable of sensing the environment, processing the data and communicating with each other or with the centralized administration, are envisioned, over recent past years, to have numerous applications (like environment monitoring, traffic monitoring, health monitoring [1-3] etc.) and the supportive advances in device technology [4] are boosting their potential. The self organizing capability, dynamic topology and the suitability for multihop routing of the sensor nodes attracted many researchers. These networks are characterized by their available limited bandwidth, limited computational capability of the sensor nodes and their limited battery power. The excessive computations and communications at the sensor nodes and the harsh deployment environment make them severely energy constrained and their size imposes restrictions on the battery replacement. Therefore it becomes necessary to make the wireless sensor network, once deployed, survive for more time by placing a minimum communication burden on these nodes, to achieve the energy efficiency and increase the network lifetime, as their communication activities are considered to be more energy consuming as compared to computational activities.

Many of the applications of wireless sensor networks require the sensor nodes to transport the field data to a distant centralized administration called base station. The process of collecting the data from multiple sensor nodes, at the base station, is termed as data gathering and local processing of data as aggregation.

Selection of the network architecture for a particular data centric application is a crucial task, being it the most important factor affecting the energy efficiency of the wireless sensor network. Out of the various architectures, clustering is shown to be scalable and energy efficient architecture for wireless sensor networks, in the presence of large number of nodes [5],[6].

In conventional clustering, by dividing the network, sensor nodes are organized into small groups called clusters and then one sensor node in each of them is selected to act as clusterhead, to coordinate the activities of the sensor nodes in its cluster. The clusterhead then allows other sensor nodes to join it, to form the cluster. During data gathering phase, clusterhead collects the data from all of its members and after aggregation transmits it to the base station. When the clusterhead energy reduces to a level at which it can not reach the base station, the cluster data may not be communicated. While utilizing conventional clustering for data gathering, the sensor nodes acting as clusterheads consume more energy compared to other cluster members, due to their long range transmissions to the base station. This ends with uneven energy consumption of the sensor nodes in the network. To address this issue, Low Energy Adaptive Clustering Hierarchy (LEACH) was proposed in

Bhaskar P Deosarkar is Lecturer in Electronics with Government Polytechnic, Nanded, Maharashtra State, India and is presently on deputation for PhD at Malaviya National Institute of Technology, Jaipur, Rajasthan, India.
(phone: +91 9887322955; e-mail: bhaskar44_nanded@yahoo.co.in).
R.P.Yadav is Professor and Head, Department of Electronics and Communication Engineering, Malaviya National Institute of Technology, Jaipur, Rajasthan, India. (e-mail:rp_yadav@yahoo.com).

[7], which for distribution of energy consumption over the network, suggested the rotation of clusterhead role among the sensor nodes and provided a major breakthrough in the field of cluster based data gathering in wireless sensor networks. Many of its limitations are reported in the literature [8][9][10][11][12][13] and most of them seem to stem from its basic clusterhead selection strategy.

In LEACH scheme, the sensor nodes select themselves as clusterheads with the evaluation of a threshold and its comparison with a random number generated. Due to this randomization used, the clusterheads are not guaranteed to be distributed uniformly over the network area and during some rounds it may happen that the selected clusterheads may lie in only a part of a network or all the eligible sensor nodes may select themselves as clusterheads or none of the sensor node may select itself as a clusterhead. When the clusterheads are concentrated in a part of a network, many of the non clusterhead nodes in the network have to transmit their data to a clusterhead located at a longer distance. Energy consumption being proportional to the squared communication distance when clusterheads are not uniformly distributed over the network area, many of the sensor nodes may consume large amount of energy and reduce the network lifetime.

Apart from the disadvantages listed in the literature [14], another factor affecting energy efficiency of the clustering algorithm is the number of nodes selecting themselves as clusterheads. All the clustering schemes use different CDMA codes in clusters to avoid the interference which needs the signal to be spread, increasing the number of bits required to be transmitted, and results in increased energy consumption. As with LEACH scheme, the number of clusterheads during different rounds is not fixed, the spreading factor needs to be set corresponding to the maximum number of clusterheads expected during some data gathering round, to avoid loss of data due to collisions.

Hence, there is need of a scheme which can select exactly equal number of nodes as clusterheads, during different rounds and from different part of the network, so that the selected clusterheads get uniformly distributed.

To distribute the clusterhead uniformly over the network area, the most obvious choice is to divide the nodes in groups based on their physical locations and the node density. But none of the distributed algorithm in the literature, to the best of author's knowledge, uses location information to form the clusters. There exist some centralized schemes which assume the nodes to be location aware and these nodes communicate their location information to the base station [15][16] but with respect to these schemes, perception of authors is that the location information available at the sensor nodes is in the form of their distances to their neighbors. The centralized schemes are not as much scalable, so the discussion in this paper is restricted to distributed schemes only.

With this motivation, the problem now gets converted to that of finding near accurate location of the sensor nodes, which is a widely discussed research topic in wireless sensor networks and is termed as node localization. The localization schemes proposed in the literature may be classified as RSSI based, accurate measurement based or angle and distance measure based [17]. Each of these types has its own merits and demerits, but where energy is the major constraint RSSI based schemes is a better solution but the distances estimated with RSSI measurements are not said to be accurate. But the recent study in [18] motivates the use of Received Signal Strength Indicator (RSSI) as distance estimator and the correction mechanisms proposed in [19] and [20] further biases the view in its favor. Further the localization schemes in the literature may be classified as: with and without anchor nodes. In the schemes using anchor nodes, some nodes capable of generating location information are placed inside or outside the network and by performing the triangulation of the distance estimates from the receipt of beacons from these anchor nodes, nodes estimate their locations. But, the cost of these anchor nodes raises an additional constraint and the inaccuracies associated with the system utilized by these anchor nodes cause their estimated location to drift from their actual position in the network. As these drifted measurements are used as references by other nodes in the network, their localization estimates may drift from their actual locations.

So to distribute the clusters uniformly over network area, a scheme is proposed in this paper in which sensor nodes use an anchor free node localization, utilizing RSSI, for estimating the near accurate node locations and use them for cluster distribution. Novelty of this scheme lies in that, it is fully distributed and through the collaborative efforts of only the limited number of sensor nodes makes it possible for all the sensor nodes to generate their relative location information, without any support of anchor nodes or location finding hardware like GPS. The simulations carried out in NS-2 [21], with incorporation of MIT uAMPS project sensor network framework [22], show that the proposed scheme has a great potential to compete with other existing clustering algorithms in terms of cluster distribution and energy efficiency.

Rest of the paper is organized as follows. Section II discusses the related previous research efforts in clustering as well as node localization. After explaining the preliminaries in section III, the proposed cluster distribution scheme is described in section IV. Simulation parameters are presented and the performance metric is discussed in section V. Simulation results are presented in section VI and finally, section VII concludes the paper.

II. RELATED WORK

This section discusses the background of this paper. Most of the research in energy efficient data gathering in wireless sensor networks is guided by LEACH, in which the sensor nodes organize themselves into local clusters and select one sensor node to act as clusterhead in each of them. The sensor nodes probabilistically decide to become clusterheads at any given time based on, percentage number of nodes required to be selected as clusterheads, the current data gathering round and a random number generated.

It is expected with this approach that the sensor nodes with more energy remaining select themselves as clusterheads. In order to save the battery power, the clusterhead position is rotated randomly among the sensor

nodes and to reduce the amount of data to be transmitted to the base station, local data fusion is performed in LEACH. For forming the clusters, the selected clusterheads advertise their selection and allow other sensor nodes to join the clusters. The clusterhead is responsible for the collection of data individually from all the nodes in its cluster abiding to the TDMA schedule and its transmission, after aggregation, to the base station. Such one cycle of data collection is termed as '*round*'.

Motivated from LEACH, many proposals are reported in the literature highlighting its various limitations and suggesting improvements [14]. But one of the major drawbacks has not got the proper attention. In LEACH, with increase in round number, the probability of node selecting itself as clusterhead go on increasing, if it had not selected as clusterhead during some specified number of rounds, dependent on the set expected number of clusterheads. And due to randomized approach used by the nodes to decide on their role, the number of clusterheads is not fixed during different data gathering rounds. When number of clusters is variable, the factor corresponding to amount of spreading required needs to be set equal to maximum expected number of nodes that may select as clusterheads during some data gathering round, increasing the data size and resulting in increased energy consumption. So to increase the energy efficiency, the number of clusterheads during different rounds shall be same. Another inherent drawback is that the clusterheads may get concentrated in only a part of the network and in such scenario many of the sensor nodes may find nearest clusterhead at a longer distance and their data transmission to clusterhead may consume more energy.

Thus, there is a need for development of a distributed scheme that can distribute the clusters uniformly over the network area to ensure selection of equal number of clusterheads during different rounds and to avoid the concentration of clusterheads in only a part of the network.

The base station assisted schemes are guaranteed to distribute the clusterheads uniformly over the network but the efficacy of centralized approaches is limited in large scale networks where collecting all the necessary information at the central authority is both time and energy consuming [5].

In the survey, presented recently in [14], thirty eight different clustering schemes are classified based on their clusterhead selection approach and are compared with respect to their required reclustering during each round, assistance in clusterhead selection, parameters considered, distribution of the clusterheads and creation of balanced clusters. It is highlighted that among the distributed schemes EECS [23] may distribute the clusterheads uniformly over the network area with selection of a node with highest energy remaining in a communication range as clusterhead, during a particular round.

Y. Liu, J. Gao, L. Zhu and Y. Zhang in [24] recently proposed CABCF, in which a new metric termed as communication facility (CF) is introduced, considering the parameters like distance of the node to the base station, node degree, remaining energy and size of the cluster. The neighbors communicating each other combine themselves to form a combinative node and one with the larger evaluated CF is termed as key node and this way the cluster formation process continues till the cluster of desired size is not formed. It is claimed that the proposed algorithm distributes the clusters and outperform LEACH.

Most of the distributed schemes assume the nodes to be location unaware and rely on either the probabilistic approach as in LEACH or communication among the nodes to form the clusters because of the energy efficiency and cost of the hardware required to have the provision for deriving the locations. Most economical and energy efficient solution among the localization techniques is the Received signal strength (RSS) method which determines the location of unknown node by measuring the received signal strength between the two communicating devices but the determined locations are said to be inaccurate. However, the recent work in [18], [19] and [20] motivates its use. In [18] experimental work was carried out to analyze the principle of RSSI based node localization, with two different node modules, using trilateration and Maximum Likelihood Estimation (MLE), which claims to have the accuracy of MLE estimation to be comparatively better and highlights that the multihop accuracy is greatly affected by the error in distance measurement. Localization algorithm in [19] combines difference correction method with MLE and claims the localization error to be less than 2.5 meters. However, these techniques [18] [19] use beaconing nodes to estimate the locations. The anchor free localization proposed in [25] estimates the angle to avoid the flip ambiguity in which the authors carried out the simulations in NS-2 assuming transmission range of 250 meters. Thus, the status of the work being carried out in the localization area gives a fair support to assume that, with RSSI the nodes can estimate their distances to the transmitter if they know the propagation conditions.

III. PRELIMINARIES OF THE PROPOSED SCHEME

A wireless sensor network consisting of P nodes (s1, s2, s3,…,…,Sp) is considered to be deployed over a rectangular region of size M x N with coordinates in the range of [0, 0] to [M, N] such that M > 0 and N > 0. The nodes are considered to be deployed in a random fashion across the network and, once deployed, are assumed to be static. They are considered not to have any location information or location finding hardware installed with them. The sink node or the base station is assumed to be located outside the network at [Bs(x), Bs(y)] with a minimum distance d_{min} from the network such that, the line passing through the base station and the physical center of the network divides the network in two equal parts.. In data centric applications, like environment monitoring, the sensor nodes are densely deployed over a comparatively smaller geographical region and in such scenario, the nodes are considered to be within the communication range of each other and are expected to be capable of reporting to the base station. Hence, without loss of any generality, they are suitably assumed to have sufficient range to reach each other. The nodes further are assumed to estimate their accurate distances to the base station with the RSS of its transmitted beacons and the base station is considered to be constraint free. The sensor nodes

are further assumed to be equipped with CDMA facilities.

Now, with the knowledge of base station coordinates, network coordinates and their distances to the base station, the possible range of sensor node locations is the part of the
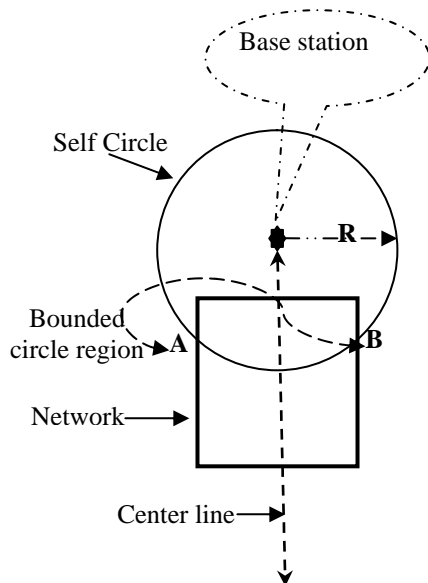

Fig.1: The self circle of a node and the bounded region giving the range of its possible locations

circle (henceforth is referred as *self circle*), centered at the base station and, of radius equal to its estimated distance to the base station and bounded by the network coordinates, as shown in Fig.1. The nodes initially are assumed to take their position at the intersection of the *self circle* and the line (henceforth referred as *center line*) passing through the physical center of the network and the base station. With this initial position, actual position of the node lies either on this line or to the right or to the left of this line on the respective *self circle*. For example, a node estimating its distance to base station as R, may have its actual position any where on the part of the *self circle* between A and B and assumes its initial position as [Bs(x), Bs(y)-R] or [Bs(x)-R, Bs(y)]. Now to find the actual location, if one of the nodes temporarily is allowed to assume and declare to have its position estimate as accurate then other nodes in the network, listening to this declaration, can easily find the amount of shift required using simple trigonometric calculations as follows:

Let us consider that the radius of the self circle of the voluntary leader node *A*, hereafter in this paper referred as Localization Leader (LL), be R2 and that of the sensor node *B*, which has to find its relative location with respect to LL is R1. With voluntary declaration by LL, B is assumed to be capable of estimating its distance to LL by using RSS of the declaration message received. With the information received in the form of message in this declaration, B can get the assumed location of LL. Using this knowledge, its own initial position and estimated distance to LL, A can calculate the amount of shift required in its initial X location and Y location and thus can find its best position as follows.

Let the perpendicular from best position of B on the central line cuts it at a distance P from the base station *O* and at a distance Q from LL, as shown in Fig. 2. As the

assumed locations of both LL and B lies on the center line:

$$R2 = P + Q$$

(1)

From trigonometric identities, it can be written that
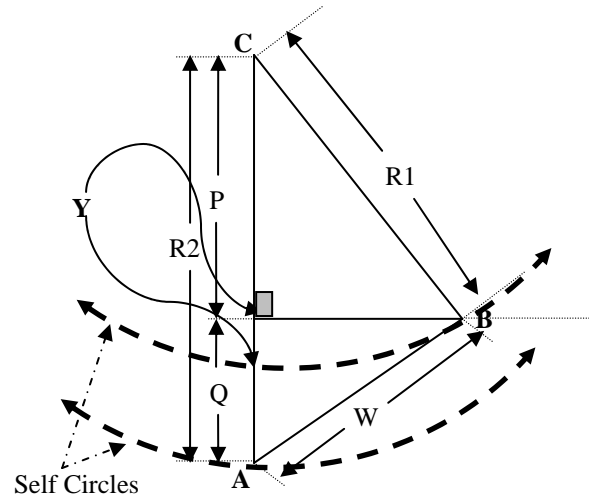
$$R1^2 - P^2 = W^2 - Q^2$$

(2)

and

Fig.2: Calculation of required X shift and Y shift to move from initial to best location

$$R1 = P + Y$$

(3)

With equations 1 to 3, shift in Y location may be written as:

$$\Delta Y = R1 - \frac{R1^2 + R2^2 - W^2}{2 \times R2}$$

(4)

Now by knowing Y, R1 and Z, X shift may be calculated as:

$$\Delta X = \pm\sqrt{R1^2 - (R1 - Z)^2}$$

(5)

Thus, with this shift from initial position there exists two solutions and when the initial position of A is itself the best position results in a single solution.

Now, if a node having the base station distance closer to the base station distance of the LL and estimated distance to LL more than the difference between two radii greedily is allowed to assume a best position, by discarding one of the two possible solutions randomly and declare itself as a *Network Director (ND)* and to send a message including its assumed best position and direction it has assumed (to the left of the center line or to the right), remaining sensor nodes then can easily assume their best positions, discarding one of the solutions. As these nodes already are assumed to be capable of estimating their distances to the transmitter by measuring RSS, they can estimate the distance to ND and with the information in the received message will now be aware of the assumed best position of ND. With the comparison of calculated and estimated distances to ND, the nodes may follow ND if the estimated distance is less than the calculated distance and choose the opposite direction otherwise.

But this concept gives rise to many questions.

Who shall take over as LL? Who shall be the ND? Which

node is more suitable for selecting itself as LL? What criterion is to be used by the nodes to take on a particular role? What, if ND chooses a wrong solution?

To answer these questions, with the knowledge of the distance to the base station and the network coordinates, the sensor nodes can easily decide on the part of the network in which they do lay. The nodes estimating there position around the network center, if are allowed to compete for the LL and ND roles, above questions are most likely to be answered.

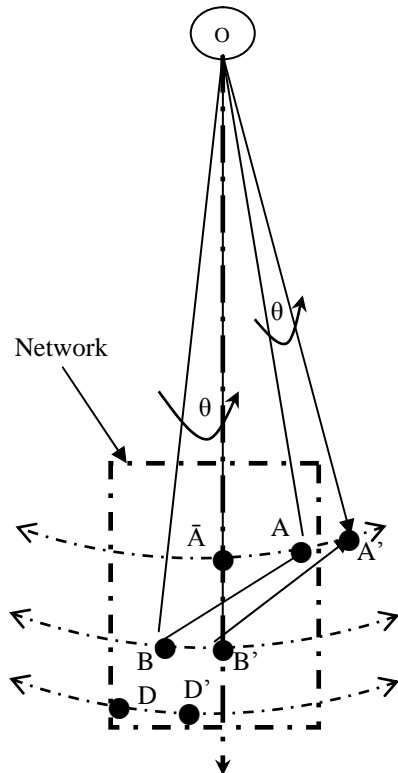By exchanging some control messages, these competing

Fig.3: The angular shift of the LL location from its initial position and the relative displacement of sensor nodes

nodes may get distances to each other using RSSI with which, each node can calculate a distance based fitness value for claiming LL role and exchange it with the competing nodes. Comparison of the fitness values received from competitors with their own fitness value may be used by the nodes, to decide their role as LL or ND or general node.

Sum of the distances to all the competing nodes seems to be a good function for evaluation of fitness value as this value will always be less for node near the center line and maximum for the node located near the network edge. With this separation, if the node with minimum fitness value is allowed to be LL and with maximum value to be as ND then, with the messages received from these nodes, other nodes can easily locate themselves with respect to LL. As LL is more near to the center line, though not guaranteed to be on center line, the drift in its actual location is expected to be very small. Let us now discuss the effect of the voluntary assumption of LL, about its location. Let in Fig. 3, A shows the actual position of node A and Ā is its assumed initial position. The location marked B' shows the assumed location of LL whereas B is its actual position. With the

advertisement received from LL, A estimates its distance to LL and shifts its location to same distance from B' (as B' is known to A, as the location of LL) and locates itself at A'. Let O be the base station location, then from trigonometric fundamentals it can be easily proved that ΔOAB ≡ ΔOA'B'. Due to similarity of triangles and one of the vertexes being common, if triangle ΔOAB is rotated around the common vertex O, then with a rotation by some angle θ both the triangles will overlap. This shows that the angular displacement of θ in the location of LL causes an equal angular displacement in the actual location of A. From figure it is clear that the angular displacement in the
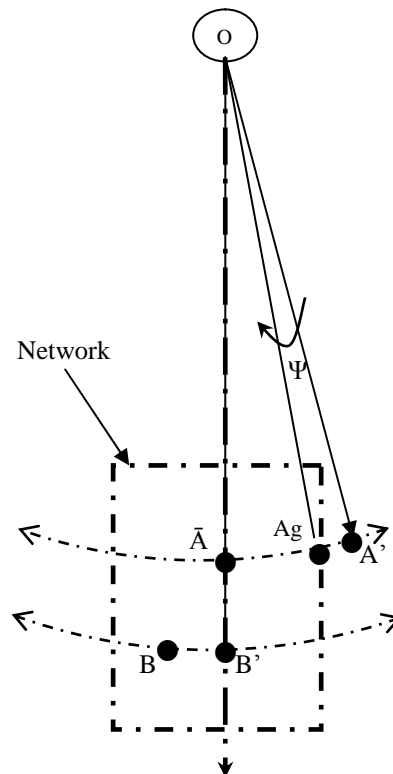

Fig.4: Calculation of the offset value by the nodes found outside the specified network boundaries.

location is anticlockwise. If somehow, a correction equal to similar distance but in clockwise direction is applied to the calculated best location then node A may better approximate its location. But with this, new questions arise. How do nodes in the network will come to know about the correction in terms of displacement? How much correction shall they apply?

The nodes in the network may compare their locations with network coordinates and can easily get the status about their calculated best location to be inside or outside the network. The nodes found them outside can easily find the angular displacement required in their calculated best location to reach the nearest network edge. As LL selected is closer to the center line, very few nodes are expected to find themselves outside the network. The node very much near to the edge will expectedly find the required minimum angular displacement for it to be within the network boundaries and if the correction corresponding to this displacement is applied to their calculated best locations by all the sensor nodes then with a maximum probability, all the sensor nodes will find themselves within the specified

network boundaries. Looking at Fig. 4 it can be observed that, for node A to be within the network a correction corresponding to angular displacement Ψ needs to be applied in clockwise direction. Thus all sensor nodes can locate themselves within the network boundaries.

At this stage, following question needs to be answered. Are the adjusted locations, the true approximations of their actual locations? Answer is, not guaranteed. Then, how to correct the locations, if they are wrong?

This happens due to arbitrary selection of the shift in the location by ND. In Fig. 3, it can be observed that LL is aligned to the left of the center line, and it is assumed that ND takes correct solution direction. Keen observation shows that the nodes in other part find themselves outside the network, following ND, and the nodes from remaining half shift their location inside the network. Node D, actually located at D, near the network edge in left half, moves closer to the center line and settles at D' and node A, actually located at A, finds itself outside the network. It can be concluded from this scenario that nodes in the half part of the network in which LL lies, have more probability of finding their locations inside the network and nodes near the network boundary in the half part of the network in which ND lays, always tend to find their best locations outside.

With the proposed formulation it is expected that the network will lean towards the direction in which ND assumes its best position and forces all the nodes from its half to follow it. Thus, if ND takes a wrong direction then there expected is the complete flip in locations of all the sensor nodes. However, this avoids the individual location flipping of the nodes and also the associated ambiguity. Whenever there is a flip, it is network wide. As the purpose of this paper is to distribute the clusters uniformly over the network area, flip of locations around the center line will not affect the performance of clustering algorithm, as the relative neighborhood remains intact.

IV. Proposed Cluster Distribution Scheme

In the proposed cluster distribution scheme, sensor nodes in the network after listening to the localization beacon from the base station, estimate their distance to the base station, and based on this estimate, assume some initial position and decide to compete for the ***Localization Leader role.*** The eligible nodes for competition are referred as contenders. The contenders put their claim by broadcasting a LEADER_CLAIM using CSMA/CA and listen to the similar claims from other contenders. With the receipt of claim from other contenders, the contending sensor nodes compute a distance based fitness for LL role and broadcast it to the competing contenders. With the comparison of fitness values of other contenders, the sensor node with minimum fitness value decides to be the *Localization Leader (LL)* for the network and the sensor node with maximum fitness value decides to be the *Network Director (ND)*. LL then broadcasts its status and assumed initial position to the network and other sensor nodes in the network listen to it. After receiving *LL message* from *LL, ND* takes its position and broadcasts a *ND_Message.* Other nodes, except LL, listen to this message and decide to locate themselves

towards left or right of the LL, considering their estimated distances to the base station, LL and ND to assume their best position. To take into account the drift into location due to drift in the location of LL from its actual location, the nodes which found themselves out of the network coordinates calculate the amount of correction required and broadcast it to the network. All sensor nodes, except the transmitting node, listen to this broadcast and then all sensor nodes adjust their locations by applying a shift corresponding to the maximum correction received in these messages.

Thus the proposed scheme consists of following phases:

A. BS Hello

The sink node or the base station is considered not to have any resource constraint and transmits a start message for the sensor nodes in the network with higher power.

B. BS Distance Estimation and Contender Decision

If the channel state information or the type of propagation medium is known, a priori, to the sensor nodes then the distances may be estimated with the received signal strength. In this study, it is assumed that the nodes can estimate their accurate distance to the base station. Based on the minimum and maximum distances to the base station the network is divided in concentric strips and the nodes from one of the strips in which network center is expected to fall are allowed to compete as contender nodes.

Let *D* be the estimated distance of the sensor node to the base station and *Max* be the maximum of the two base station coordinates. Then the nodes satisfying following criteria select themselves as contender nodes.

$$0.40 \times (Max - d_{min}) < D < 0.60 \times (Max - d_{min})$$
(6)

With this equation, approximately 20% of the sensor nodes in the network are expected to select themselves as contenders to ensure the decision of sufficient number of nodes from the middle portion of the network (as seen by the base station) for this role. The reason for targeting the middle portion is that, if the nodes from the nearest part or farthest part are allowed to form this expected zone then there remains a possibility of very few nodes selecting themselves as contenders. The number of nodes allowed to compete may be varied with adjustment in the strip width.

Based on the location of the base station the nodes assume their position as:

If { BS(x) > network(x) and BS(y) < network(y)}
{
Node(x) = BS(x) – Node to BS_distance
Node(y) = 0.5 * network (y)
} elseif { BS(y) > network(y) and BS(x) < network (x)}
{
Node(y) = BS(y) – Node to BS_distance
Node(x) = 0.5 * network(x)
}

C. Competition for LL Role

The nodes decided to be contenders broadcast a short LEADER_CLAIM message to put their claim for LL role and listens to the similar messages from competing

contenders. Based on the received signal strength of these messages the contenders estimate their distance to every other contender.

D. Fitness calculation

Contenders calculate their fitness for the LL role with the evaluation of fitness value (F) as:

$$F = \frac{\sum_{i=1}^{C} D_C(i)}{C}$$

(7)

where, C is the number of contender nodes and $D_c(i)$ is the distance to the ith contender.

E. Exchange of fitness values

Again using CSMA/CA, the contenders broadcast their fitness value and listen to the fitness values of other contenders.

F. Decide LL and ND

The nodes around the center of the selected region will always have the smaller distances to all other nodes in the selected region and the nodes near the edge will find themselves at a longer distance from some of the nodes, particularly the nodes located near the opposite boundary of the network.

Thus the node with minimum sum of the fitness value is expected to be at the center of the selected region and the node with maximum fitness value can expect it to be near to one of the two edges, as can be observed from Fig.3.

So, with comparison of the own fitness value with the fitness values of competing contenders, the contender decide to be LL or ND as follows:

if (fitness_value < fitness values of all other contenders}
{Decide to be LL}
elseif {fitness_value > fitness values of all other contenders}
{Decide to be ND} else {
decide as a general sensor node}

G. Advertise LL

The sensor node decided as LL advertises its selection, including its assumed initial location, over the network and all other nodes, listening to it, estimate their distance to LL based on RSSI measurement and the assumed propagation model.

H. Advertise ND

The node selected as ND makes another assumption about its location as it lies towards left of the LL (Perceived with respect to LL looking towards base station) and estimates its best location using "*Find* Best *Location*" discussed in the next subsection-I and then after listening the LL advertise, ND advertises its status over the entire network. Nodes receiving it make a decision to locate them to the left or right of their initial position, looking towards the base station.

I. Find Best Location

With the knowledge of two location references, in the form of base station and LL along with their estimated respective distances, the estimation of position by the sensor nodes will not get a trivial solution as their always exists two solutions to such problem. To choose one of them, the ND message is used.

Though, it seems that when third node location in the form of ND is made available with triangulation other nodes can easily estimate their location but it may force the sensor node to find its location away from its self circle.

J. Find Offset

It is expected that the LL lies around the center line but due to random distribution of the sensor nodes there is a very little possibility of sensor node lying exactly on the center line and the division of the network in concentric strip reduces this possibility by a factor equal to number of strips the network is divided in. So there is large probability of drifting of the assumed location of LL from its actual location. All other sensor nodes find best location with their estimated distance from this drifted assumption and hence there may be drift in their locations from actual locations. This causes some of the nodes to locate themselves out of the network boundaries. With the comparison of estimated best positions with network coordinates, the nodes finding themselves outside calculate the minimum amount of angular shift required, in their estimated location, for them to be on the network edge and broadcast it over the network, using CSMA/CA.

K. Offset adjustment

The sensor nodes listening to the offset values from nodes found them outside the network, apply a shift to their location corresponding to maximum of the offset reported.

L. Assign to clusters

Now, with the estimated locations, clusters of any size and in any part of the network may be formed. In this study the network is divided in four parts with assumed network to be a square area and the number of nodes to be 100. When more number of clusters is required to be formed, then it may become difficult to divide the network in parts with equal geographical area but then also, the knowledge of location information is useful for uniform distribution of clusters over the network area.

V. SIMULATION PARAMETERS AND PERFORMANCE METRIC

The simulation model used in this study and the metric used for evaluating the performance of the proposed cluster distribution scheme are explained in this section.

As communication cost is much larger than computational cost, the contribution of computations to the energy consumption is considered to be negligible in this analysis. The assumed energy required for running the transmitter and receiver electronic circuitry E(elec) is 50nJ/bit and for acceptable SNR required energy for transmitter amplifier E(amp) is 100pJ/bit/m^2.

To evaluate the performance of the proposed cluster distribution algorithm different node topologies were generated in the network simulator NS-2, with incorporation of MIT uAMPS project (NS-2 Extension) sensor network framework. The proposed algorithm was implemented using

the node structure as in MIT framework which works at 914 MHz. The channel is assumed to have only system losses and not the propagation loss. The header packet size to 25 bytes and data packet size to 500 bytes were set. The nodes were assumed to have the capability to reach each other in a single hop. For the proposed cluster formation scheme trace files were generated to record the actual and estimated node locations, number of nodes in each cluster and their IDs. With the help of this data various performance metrics are evaluated.

To the best of author's knowledge, this is the first effort of its kind to utilize distributed node localization for cluster formation and hence the results obtained only are presented and are not compared with anyone else.

Simulation parameters used are listed, below, in Table 1.

Table 1: Simulation Parameters

Parameters	Values
Network area	100m x 100m
Clusters to be formed	4
Number of nodes	100
Base station location	(50,175)
Distance of the base station from network center	125m
Bandwidth	1Mb/s
Packet Size	500 bytes
Init energy	2J
Propagation model	Two ray ground

Several runs of the scenario are simulated to collect the desired data at steady state and to obtain statistically confident averages.

In order to evaluate the performance of the proposed cluster formation scheme, following performance metrics were focused.

Localization error: It is the Euclidian distance between the actual location of the sensor node and its estimated location. This metric gives the efficiency of the localization process. Less the localization error, more uniform will be the cluster distribution.

Number of contenders, their locations and location of LL: This plot shows the locations of the nodes that select to compete for LL and ND roles. It gives the number of nodes selecting to compete for LL and ND role and their distribution over the specified region. Location of LL is very important and distances of the sensor nodes from it may indicate the suitability of LL as a network center.

Cluster distribution: This is a scatter plot showing the X and Y locations of the nodes in each cluster. It gives the distribution of the clusters over the network and the number of nodes in each cluster. Non-overlapping and distributed clusters are expected to increase the energy efficiency of data gathering algorithm.

Distance to LL (Center Distance): This is a plot of node number versus their distance from the center. This plot gives the closeness of the LL to the network center. Average value of these distances is the indicative measure of this closeness.
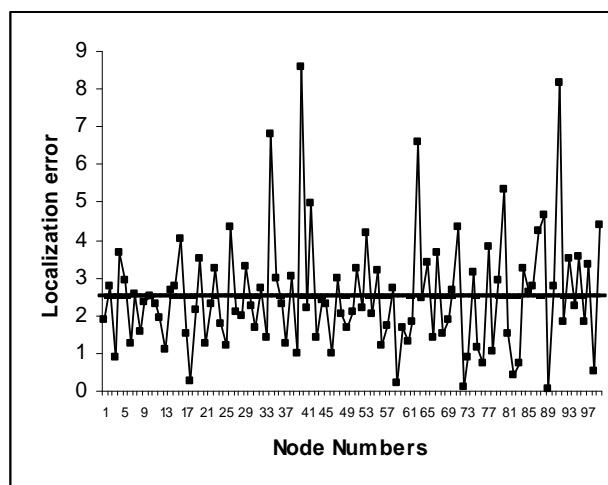

Figure 5: Node Localization Error

Offset values: This is the amount of angular displacement, in radians, calculated with respect to the base station, by the nodes found them out of the network boundaries, with the calculation of the shift required to move to the best position from their initial position. This is also a measure of closeness of the LL to the center line. More is the distance of LL from the center line more may be the nodes that find themselves out of the network boundary and more is the angular displacement required in the assumed best positions of the sensor nodes. With the increase in the distance of LL from the center line the number of nodes finding out of the network boundaries increase and these nodes are expected to broadcast their calculated required angular displacement. As its consequence, increased number of transmissions may result in increased energy consumption.

VI. SIMULATION RESULTS

The results of the simulation carried out in NS2 are presented in this section:

A. Localization Error

Fig. 5, given below, shows the plot of node number versus their Euclidian distance between the actual and estimated locations. The Euclidian distance is used as a measure of error due to localization process used. From above graph it is clear that the error varies from 0.044 meter to 8.58 meter with an average of 2.51 meter error per node location, as shown by the trend line. But careful observation of the generated trace file shows that one of the reference point used by other sensor nodes, the LL itself, is in error of 2.17 meters. This may be one of the reasons of this error. Relative to the location of LL, the average location error then may be seen as 0.34 meter. This may be because of the offset adjustments.

When none of the node is exactly on the network boundary and some nodes close to, but inside the network boundary, locates them out of the network and calculate the offset to locate on nearest network boundary. Accordingly, corresponding to the maximum offset, nodes adjust their final best positions. So, it is possible that the calculated offset may be less than the actual offset. Secondly, due to propagation model distances estimated may not be exact.

B.Distribution of Sensor Nodes:

Following Fig. 6 shows the scatter plot of the nodes according to their actual positions and estimated positions. It can be observed from the figure that, the estimated locations lie well inside the network boundary and node positions are well spread and nearer to their actual positions with a maximum Euclidian distance between the actual and estimated position of 8.58 meters. It can be concluded that the proposed scheme gives better localized estimates of the node positions and may be used as basis for physical location based evaluations. In the case of flip i.e. when nodes find their position near exactly mirror location of their actual position, due to use of ND, there occurs a complete flip of locations for all the nodes and error though seem to be very large but as discussed earlier, is expected not to matter a much as far as data centric applications
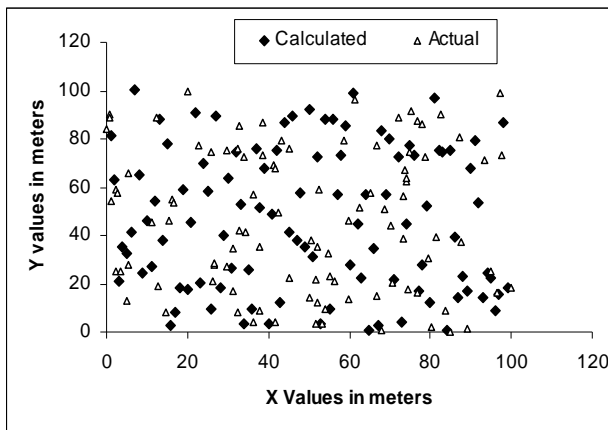

Figure 7: Sensor nodes selected as contenders and location of LL

Figure 6: Actual and estimated locations of the sensor nodes

Figure 8: Cluster distribution

is concerned, where data is more important than the actual location of nodes.

C.Number of contenders, their locations and LL location:

Locations of the nodes selected themselves to compete for the role of LL and ND is shown in the following Fig. 7. It is observed that within the selected region 17 nodes were there which contended for LL and ND roles. The node shown with dark circle was observed to select itself as LL which can be observed approximately at the center of the region selected. Thus it can be concluded that, with the proposed scheme it is possible to find a node located at the center of the region selected. The value of the fitness evaluated were observed as 465 meter (minimum) for selection of LL and 758(maximum) meter for the node selected as ND.

D. Cluster Distribution

Following Fig.8 shows the classification performed by the nodes to assign themselves to the clusters using the estimated locations. It can be observed that the clusters are uniformly distributed over the network area. Thus it can be concluded that the proposed scheme is capable of distributing the clusters uniformly over the network area. The network was assumed to be divided in four clusters. Division of an area in parts of equal circular areas or square areas is a challenging task. Based on the node density, the network area and number of clusters to be formed or
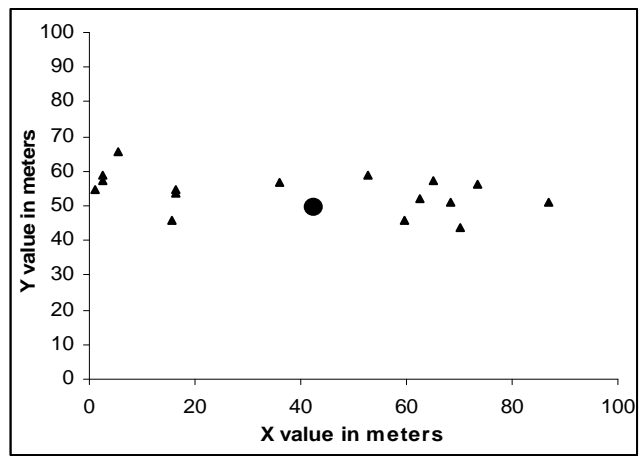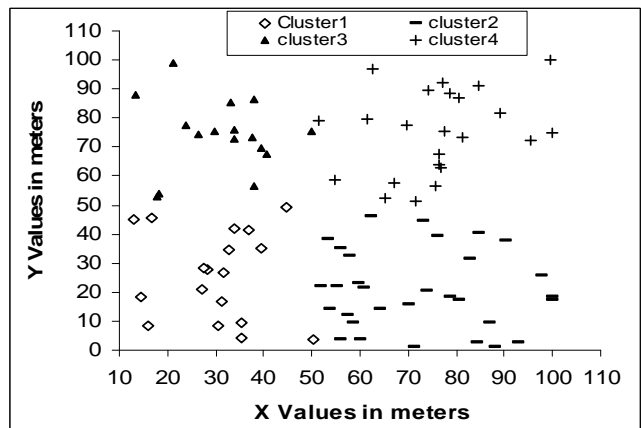
approximate number of expected nodes in each cluster, a rule shall be framed and supplied to sensor nodes.

E.Distance to LL (Center Distance)

Fig. 9, given below, shows the distance of the nodes in the network from LL. It is observed that this distance varies from minimum of 8 meter (Distance of LL to LL is not considered) to maximum of 75 meter with an average distance of 39.19 meter, as shown by the trend line. It can be inferred from this plot that the network will not drift a much and the correction required in terms of angular displacement will be very small and the number of nodes that have to broadcast their offset values is also small. Node 41 with zero distance is LL.

F. Offset Values

Locations of the nodes that found themselves out of the network after calculating the shift, from their initial positions and the amount of offset correction required, for them to be on the nearest network boundary from the calculated best position are presented in the following Table-2. With the setup used, for a particular simulation run, seven nodes were reported to found themselves outside and their relative offset calculations are as shown in this table. It is important to note that the actual position of the nodes may not be exactly on the network boundary but for to be within the specified boundary the corresponding node assumes its final best location to may be on the boundary.
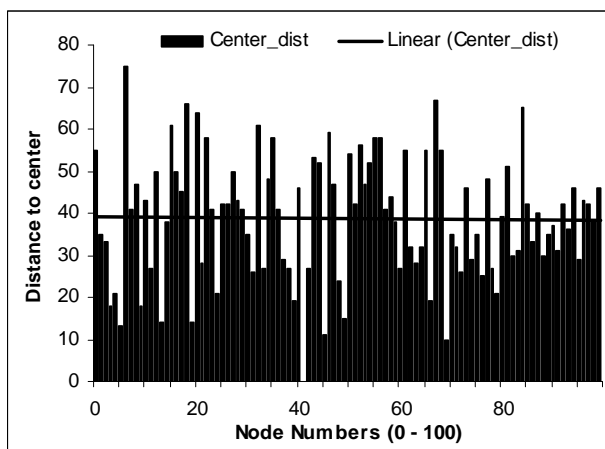
Figure 9: Distance of the sensor nodes from LL

Table 2: Nodes found out of network boundaries and their offset values

Node ID	Offset in radians
52	0.0068979047340724797
6	0.04408092222567761
32	0.050590256504748465
67	0.0082744082198960012
18	0.072689169241994422
20	0.052619336433805008
35	0.035523992718828667

But the nodes have to correct their location according to the maximum of the offset correction observed by any node. So, at least one of the nodes assumes its final location on the boundary and correction corresponding to this error could not be accommodated with the proposed scheme. The node (Node number 18) corresponding the maximum displacement assumed its position at (100, 18.0898882) when its actual position was (99.7459075, 18.0091471).

VII. CONCLUSIONS AND FUTURE SCOPE

A novel cluster formation scheme is proposed in this paper, to benefit from node localization. The scheme is observed to have following features:

1) It is fully distributed.
2) It requires less number of transmissions to form the clusters and does not require any extra hardware or bandwidth.
3) With the knowledge of only one location in the form of base station, sensor node localize relatively through their collaborative efforts.

The proposed scheme is shown to distribute the clusters uniformly over the network area. It can be concluded from the results obtained that, if perfect channel conditions are made known to the sensor nodes, with the help of RSSI, the sensor nodes can localize themselves in a distributive manner. The accuracy of the node localization scheme depends on the accuracy of the RSS measurements and hence the scheme is more suitable where channel do not vary a much. Cluster formation greatly depends on the node density. The assumption of square shaped or rectangular shaped network imposes restriction on the number of clusters to be formed. In practical situations, how well the RSSI measurement will give the near accurate results needs

to be tested.

The proposed cluster formation scheme will further be extended to implement the complete clustering algorithm to observe its scalability and energy efficiency. With this scheme, selection of one node from each cluster to act as clusterhead during each round of data gathering is expected to distribute the clusterheads uniformly and provide the desired energy efficiency. Most importantly, the computational burden on the sensor nodes is not considered so far but this work needs to be extended for its study and to solve the problem of flipping nodes.

REFERENCES

[1] I. Akyildiz , E. Stuntebeck, "Wireless underground sensor networks: Research challenges" Elsevier Science Journal of Ad Hoc Networks, Vol.4, 2006, pp. 669–686

[2] I. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirc, "Wireless sensor networks: a survey", Elsevier Science Journal of Computer Networks, Vol.38, 2002, pp. 393–422

[3] M.Nekovee, "Sensor networks on the road: the promises and challenges of vehicular ad hoc networks and grids", In Proc. *Workshop on Ubiquit. Computing and e-Research*, '05, 2005.

[4] M. Augusto M. Vieira, D. C. da Silva Junior, "Survey on Wireless Sensor Network Devices", In Proc. *IEEE Conf. Emerging Technologies and Factory Automation (ETFA03)*, Sept. 2003, pp 16-19.

[5] O. Younis, M. Krunz and A.Ramasubramanian, "Node Clustering in Wireless Sensor Networks: Recent Developments and Deployment Challenges", IEEE Network, May/June-2006, pp. 20-25.

[6] J. Yu and P. Chong, "A survey of clustering schemes for Mobile Ad Hoc Networks", IEEE Communications Surveys and Tutorials, First Quarter 2005, Vol 7, No. 1, pp. 32-47.

[7] W. .Heinzelman, A.Chandrakasan and H.Balakrishnan, "Energy-Efficient Communication Protocol for Wireless Microsensor Networks", In Proc.*33rd HICS*, 2000.

[8] A. Abbasi, M. Younis, "A survey on clustering algorithms for wireless sensor networks", Elsevier Science Journal of Computer Communications, Vol.30, 2007, pp. 2826-2841.

[9] C. Liu and C. Lee, "Power Efficient Communication Protocols for Data Gathering on Mobile Sensor Networks", In Proc. *IEEE Int. Conf.* Vehicular Technology (VTC'04), 2004, pp. 4635-4639.

[10] Y. Qian, J. Zhou, L.Qian and K. Chen, "Highly Scalable Multihop Clustering Algorithm for Wireless Sensor Networks", in Proc. *IEEE ICCCAS 2006*, Vol.3, June, 2006, pp.1527-1531.

[11] H.Chen, C. Wu, Y. Chu, C. Cheng and L. Tsai, "Energy Residue Aware (ERA) Clustering Algorithm for Leach-based Wireless Sensor Networks", In Proc. *2nd Int. Conf. Systems and Networks* Communications (ICSNC 2007),25-31 Aug, 2007, pp.40.

[12] J. Zhao, A. Erdogan and T. Arslam, "A Novel Application Specific Network Protocol for Wireless Sensor Networks",. In Proc. *ISCAS (6), 2005*, pp. 5894-5897

[13] [13] H. Lim, Sung Kim, H. Yeo, Seung Kim, and K. Ahn "Maximum energy routing protocol based on strong head in Wireless Sensor Networks", In Proc. *Int. Conf. ALPIT.*, 2006, pp. 414-419

[14] B. Deosarkar, N. Yadav and R.P. Yadav, "Clusterhead Selection in Clustering Algorithms for Wireless Sensor Networks: A Survey" , In *Proc. Int. Conf. Computing, Communication and Networking (ICCCN 2008),* Dec. 18-20, 2008, Karur, Tamilnadu, India.

[15] W. Heinzelman, *Application-Specific Protocol Architectures for Wireless Networks*, Ph.D Thesis, Massachusetts Institute of Technology, June 2000.

[16] S.Muruganathan, D. Ma, R.Bhasin and A. Fapojuwo, "A Centralized Energy-Efficient Routing Protocol for Wireless Sensor Networks", IEEE Radio Communication, March, 2005, pp. S8-S13.

[17] D. Tran, and T.Nguyen, "Localization In Wireless Sensor Networks based on Support Vector Machines", IEEE Transaction Parallel and Distributed computing, Vol. 19(7), July, 2008, pp.981-994

[18] J. Rencheng, W. Hongbin, P. Bo and G. Ning, "Research on RSSI-Based Localization in Wireless Sensor Networks", In Proc *Int. Conf. Wireless communication networking and mobile computing 2008* (Wicom 08), 12-14 Oct.,2008 pp 1-4

[19] Z. Deng, W. Ren and L. Xu, "Localization algorithm based on difference estimation for wireless sensor networks in Pervasive

Computing Environment" , In Proc. *Third Int. Conf. Pervasive Computing and Applications, ICPCA 2008, 6-8 Oct, 2008*

[20] W. Wang and Q. Xu, " RSS-based Monte Carlo localization for mobile sensor networks" , J. IET Communications, Vol..2, No.5, 2008, pp. 673-681.

[21] K. Fall and K. Vardhan, The Network Simulator (ns-2). Available online: http://www.internetworkflow.com/downloads/ ns2leach/ns-allinone-2.27.tar.gz

[22] MIT µAMPS project ns2 code extensions. Available online :http://www-mtl.mit.edu/research/icsystems/uamps /research /leach/leach_code.shtml

[23] L. Lijun, W. Hongtao and C.Peng, "Discuss in Round Rotation Policy of Hierarchical Route in Wireless Sensor Networks", In Proc. *IEEE Int. Conf. WiCOM'06, 2006*, pp. 1-5.

[24] Yuhua Liu, Jingju Gao, Longquan Zhu, Yugang Zhang, "A Clustering Algorithm Based on Communication Facility in WSN", In Proc.*WRI Int. Conf. Communications and Mobile Computing, CMC'09*, 6-8 Jan. 2009

[25] B. Tang, X.Zhu, A.Subramanian and J.Gao, "Distributed, Anchor-Free Localization in Sparse Sensor Networks by Local Angle Measurement", Available online: www.cs.wichita.edu/~bintang /papers/local_angle.pdf