Abstract—In this paper, based on LEACH protocol, improvement is made and the improved protocol DCDA-LEACH is proposed. DCDA-LEACH makes improvements to LEACH with data correlation and data aggregation as the core, introducing the thinking of regionalization. It divides the nodes into regions in accordance with the relevant characteristics of the data to improve the data correlation level within the clusters. Clustering process works with the regional restrictions to enhance data aggregation efficiency and reduce energy consumption. Multi-skip routing is used to reduce the number of cluster-heads which communicate with base station directly. The protocol, therefore, can balance the node energy and prolong the network life.

Keywords—Wireless Sensor Networks; Data Aggregation; Data correlation

I. INTRODUCTION

Wireless Sensor Networks (WSN) are wireless self-organizing networks that consist of a large number of randomly scattered sensor nodes. WSN nodes are usually battery-powered, and when the energy runs out, the node will not work. Therefore, saving energy and prolonging the life cycle is a essential technology[1]. WSN nodes scattered densely, as a result, the effective detection ranges of the nodes overlap. Meanwhile, in the usual WSN applications, data collected by neighboring nodes have high temporal correlation and spatial correlation. In WSN, the communication cost is hundreds of times as much as that of computing. Under normal circumstances, within or near the measured object, a number of nodes are deployed, so the data they collected on the same event are similar or the same. Taking the above factors into consideration, data aggregation is effective, that is to use the local processing ability of sensor nodes to do advance compression, to deal with the data collected by the nodes or transmitted from other nodes, to remove redundant information, and then to transmit processed data.

Based on LEACH[2], this paper presents a protocol centered on data aggregation: LEACH based on Data Correlation and Data Aggregation (DCDA-LEACH). In the protocol, we introduced data correlation feedback strategy. The base station divides all the nodes into regions based on the data characteristics of each node, which makes the nodes within the same region have a high data dependence. In this way, it gives full play to the energy saving features of data aggregation and to extend the network life cycle, which will help the expansion of network size. This protocol applies to large-scale applications in which nodes periodically report data to the base station.

II. WORK OF THIS PAPER

The operation of DCDA-LEACH is also divided into rounds. It does completely the same operation with LEACH during the first round to enable the base station to know the data features of each node. Then the base station divides the node coverage area into regions according to the data correlation of each node, ensuring that the nodes within the same region have high data correlation. In this protocol, the nodes coverage area is divided into a fan-shaped layout of polar coordinates. Each region is defined as an atomic region. The formation of clusters is only carried out separately within the atomic region. In this way, nodes within the same cluster will have a very high correlation of data, and the efficiency of data aggregation carried out by cluster-heads will be greatly enhanced. The amount of data needed to be sent by cluster-heads is greatly reduced, which can effectively reduce the energy consumption of cluster-heads.

After the data correlation area division, the cluster-heads form a spanning tree. As shown in Fig. 1. Cluster-heads collect the data sent by nodes within their own clusters, finish data aggregation, and transmit the compressed data to the base station along the spanning tree by hops. The data correlation area division facilitates the construction of the spanning tree. In the analysis below, we can see that the division of fan-shaped layout brings power consumption advantages. It reduces the number of cluster-heads which communicate directly with the base station, makes the power load of the sensor nodes more balanced, and extends the network life cycle.

Figure 1. Data correlation area division

978-1-4244-5539-3/10/$26.00 ©2010 IEEE
A. Network Model

Suppose N sensor nodes randomly distributed in a L×L square area. And the sensor nodes have the following properties: (1) A fixed base station (BS) far away from the sensor area. (2) We have homogeneous sensor nodes with the same finite energy, the same processing and communication capabilities. (3) Network nodes are static. (4) Node can perceive its geographical location, which is expressed as coordinate. (6) Node can communicate directly with the base station or indirectly through multi-hop path. Node can control the wireless transmit power, that is, node can adjust the transmit power according to the distance. (7) Node periodically collects data and sends back to the base station.

B. Wireless Communication Model

Reference [3] proposed two kinds of communication model: Free space model and multi-path fading model. Free space model is applicable when the distance between sending node and receiving node is less than some value d0, otherwise we use multi-path fading model. When the sender sends l bit data to the receiver and the distance is d, the energy consumption of sender and receiver can be calculated with the following two formulas:

\[
E_{tx}(l, d) = \begin{cases} 
1 \times E_{elec} + l \times \eta_{fs} \times d^2, & d < d_0 \\
1 \times E_{elec} + 1 \times \eta_{mp} \times d^4, & d \geq d_0 
\end{cases}
\]

(1)

\[
E_{rx}(l, d) = 1 \times E_{elec}
\]

(2)

where \(E_{tx}(l, d)\) = the energy consumed when the sender sends l bit data to the receiver d distance away, \(E_{rx}(l)\) = the energy consumed when the receiver receives l bit data, \(E_{elec}\) = energy consumption of radio transceiver circuit, \(\eta_{fs}\), \(\eta_{mp}\) = channel model parameters.

In DCDA-LEACH we assume that within the cluster, free space model is applicable and we use multi-path fading model in the communication between BS and cluster-heads.

C. DCDA-LEACH algorithm

1) Data-related zoning

In the first round, DCDA-LEACH does the same things as LEACH. The cluster-heads send data back to BS. BS explores the data relevance of each node. Accordingly, the node coverage area is divided into a number of data-related areas. Starting from the second round, clustering only happens separately within the data-related areas. Division process is as follows.

(a) Based on the distance to the base station, the sensing area is divided into several belt areas. BS sets its own hops to 0. Suppose that the distance of the nearest node to BS is \(d_0\). The base station sets the width of each belt area basing on the scale of the whole detection area, supposing it to be w. BS broadcasts an area dividing message, which contains a value of hops to BS, with the radius of \(d_0 + w\). Nodes which receive the message set their own hops to be 1 plus the hops value in the message. Then BS broadcasts again, and the radius rises to \(d_0 + 2w\) and the value of hops in the message plus 1, and so on. Each node uses the hops value as the belt area number, so the belt area number will be 1, 2, 3… from the central belt to peripheral ones.

(b) Each belt area is divided into a number of fan-shaped regions (called atomic region). The layout is carried out by BS, and then the results will be broadcasted to each node.

We choose a borderline of the whole area (e.g. the bold edge at the left in Fig. 1) as reference. Each node can calculate its relative curvature using its coordinates and the coordinates of BS. This fan-shaped layout carried out by BS can be expressed by a linked list group recording the relative curvature. The basic idea of fan-shaped regions division algorithm is as follows: There is a threshold of data discrepancies and a threshold of angle range. When either threshold is reached, an atomic region is determined. In order to make the atomic regions of different belt areas having the same size, the threshold of angle range decreases from central belt to peripheral ones.

We name the two borderlines of No. i belt area as \(R_i-1\) and \(R_i\), and the threshold of angle range of No. i belt area as \(P_i\). Then we know \(P_i\) has some initial assignment. The determination of \(P_i\) is based on the principle that the size of atomic regions should be more or less the same. We derived the following result from the equation:

\[
P_{i+1} = P_i \times \frac{R_{i+1}^2 - R_i^2}{R_i^2 - R_{i-1}^2} (i>1)
\]

After zoning, BS counts the total node number of every region, and then broadcasts the zoning layout and the node number to all nodes. Each node that has received the broadcasting can calculate the relative angle of its connecting line to BS, and ensure which atomic region it belongs to, while recording the node number in its memory.

2) Routing Establishment

Starting from the second round, clustering only happens separately within the data-related areas. Cluster-head selection algorithm is the same as LEACH protocol.

The nodes chosen as cluster-head broadcast the cluster-head advertising message (ADV), which contains the atomic region ID of the sender. When non-cluster-head nodes receive the ADV message, they have to see if the atomic region ID in the message is the same as theirs. If not, they ignore the message. If some nodes receive several ADV messages during some time, and the atomic region ID in the message is the same as their own, they choose a cluster to join according to the intensity of the ADV message.

After cluster-head selection, inter-cluster routing is to be established. At this stage, the non-cluster-head nodes turn into sleep. The cluster-heads broadcast a Tree-Advertising message (T-ADV) with a radius of 2w, which contains the belt area number of the sender. Cluster-heads which receive the T-ADV messages keep the message in which belt area number minus 1 equals their own belt area number, and ignore other ones. After a certain time, cluster-heads select the highest intensity sender among the messages they have received as their parent node, and then send to it a T-Join message. The tree routing structure is shown in Fig. 2.

III. REALIZATION
In each frame, the cluster-heads collect the data of all the non-cluster-head nodes and perform data aggregation. Then the aggregated data travel along the tree route to BS by a multi-hop way.

3) Timing of data aggregation

In WSN, besides determining the nodes which to do data aggregation along the data return path, the data aggregation energy-saving effect is closely related to the waiting time of different nodes before data aggregation. We need to know how long to wait for, and data from which nodes to be aggregated; namely the timing of data aggregation.

Data-aggregation nodes needs to aggregate as much data as possible, but if the waiting time is too short, it may result in no or only partial data aggregation. In the contrary, if waiting time is too long, it may result in intolerable long delay of transmission of data to the sink node. Therefore, we need to find the balanced point between delay and effect of data aggregation.

Considering the characteristics of this design, we use cascading timeout method [4] to determine the timing of data aggregation. Cascading timeout method determines the timing of data aggregation basing on the hops of nodes to the sink in the data-aggregation-tree routing. Nodes with fewer hops from the sink node needs to wait for a longer time, while ones with more hops wait for a shorter one; and the waiting time of all nodes form a cascade, which can effectively save energy while ensuring the accuracy of the data and short time delay at the same time.

Each node maintains a data aggregation timer, and sets its first timeout interval as $e$ in the first cycle. $e$ is calculated as follows:

$$e = t \cdot shd \cdot h$$

Where all nodes are requested to send data to BS every $t$ time, $shd$ = the time interval of the upstream and downstream nodes to send data, usually based on communication bandwidth, network load and channel quality, etc, $h$= the number of hops from the sink. After the first timeout of the timer, the data aggregation timer is always set to be $t$ in the subsequent cycles.

IV. PERFORMANCE ANALYSIS

Data aggregation rate [5], which we suppose to be $\alpha$, refers to the length ratio of the data after and before data aggregation. The range of $\alpha$ is $[c/Q, 1]$. Where $c$ = the number of cluster-heads, $Q$ = the total number of nodes. So if it takes $c/Q$, it means the data collected by non-cluster-head nodes has the maximal correlation. If it takes 1 it means there is no correlation in the data. If it takes any value between these two, the data have a certain degree of correlation.

Basing on temporal and spatial correlation, the data collected by the nodes will only have a relatively small change in certain rounds, therefore the layout of data-related areas can remain unchanged in several rounds.

Performance analysis of DCDA-LEACH is as follows.

The analysis is based on the following assumptions:

(a) Total number of nodes = $N$, uniformly distributed in a $L \times L$ region, the number of CH (cluster-head) = $k$. So, the average number of CM (non-cluster-head node) of each cluster = $(N/k) - 1$. The message length of each CM = $l$ bits.

(b) The energy consumption of aggregating 1 bit data = $E_{DA}$, data aggregation rate = $\alpha$. Data transmission within cluster bases on free space model. $E_{CH}$ = energy consumption of cluster-heads, $E_{CM}$ = energy consumption of non-cluster-head node, $E_{LEACH}$ = total energy consumption of LEACH in a frame, $E_{DCDA-LEACH}$ = total energy consumption of DCDA-LEACH in a frame.

Analysis of energy consumption between LEACH and DCDA-LEACH is as follows:

In LEACH the transmission from cluster-heads to BS uses multi-path fading model.

$$E_{LEACH} = kE_{CH} + (N-k)E_{CM}$$

$$E_{CH} = lE_{elec} \left( \frac{N}{k} - 1 \right) + lE_{DA} \frac{N}{k} + lN \frac{N}{k} \alpha E_{elec} + lN \frac{N}{k} \alpha_{elec}$$

$$E_{CM} = lE_{elec} + lE_{elec}$$

In the first round, DCDA-LEACH and LEACH protocol perform the same operation, and energy consumption is the same.

In the second round, DCDA-LEACH needs to do the following additional work to finish data-related area division and establish routing structure:

1) All nodes receive the layout message sent by BS. $E_l = N \cdot E_{elec} \cdot l$, where $l_1$ = the length of the message.
2) Cluster-head broadcasts to establish the routing tree, and each cluster-head sends the $T$-join message. 

$$E_2 = kE_{dec}l_2 + l_2k\omega_2(2w)^2 + kE_{dec}l_3 + l_3k\omega_2(2w),$$

where $l_2$ and $l_3$ are the length of the message.

Since the data-related area division can be used for several rounds, operation 1 is executed only once in certain rounds. And the energy consumption of receiving messages is much smaller than that of sending messages. If we don’t take the energy consumption of 1 into consideration: 

$$E = kE_{CH} + (N-k)E_{CM} + E_2,$$

Where the formula of $E_{CM}$ is (5). Cluster-heads send data to BS basing on free-space model.

$$E_{CH} = lE_{dc}(N/k - 1) + lE_{da}N/k + l\frac{N}{k}\alpha E_{dc} + l\frac{N}{k}\alpha \omega d_{BS}d_{BS}^2;$$

$$E_{CM} = 100, \text{ nodes distribute in a rectangular area of } 500 \times 500, \text{ the fixed base stations is outside the scope of the network.}$$

BS divides the node coverage area into regions according to the data correlation of each node, and the formation of clusters is only carried out separately within the atomic region, which makes sure that the nodes within the same cluster have a high data correlation. Thus data aggregation rate of DCDA-LEACH $\alpha$, which is closer to $c/Q$, is much smaller than LEACH protocol. Associated with $\alpha$ in (4), the energy consumption of $l(N/k)\alpha E_{dc} + l(N/k)\alpha E_{da}d_{BS}^4$ is far smaller in DCDA-LEACH than that in LEACH. This part is the power consumption of cluster-heads sending data to BS, which accounts for a large potion in the total energy consumption. And in each round, the energy consumption of cluster-heads is much larger than the common nodes. Reducing this part of the energy consumption can effectively extend the network life cycle. In $E_2$, the message length $l_2$ and $l_3$ are very short, and only the cluster-heads are needed to do this job, so the energy consumption $E_2$ can be ignored. To sum up, DCDA-LEACH is much better than LEACH in reducing energy consumption.

Then, we will discuss the performance improvement brought by the tree routing. In LEACH, the cluster-heads far away from BS communicate directly with BS, consuming enormous amounts of energy, which leads to the rapid depletion of the energy of these nodes and is not conducive to extend the life of the network as a whole. Using tree routing, the cluster-heads send data to BS along multi-hop path. In this way, multi-path fading can be avoided. Although the energy consumption of the nodes close to BS will increase because of forwarding additional data, the life of the network will be effectively extended. Furthermore along the course of transmission by skip, data of each cluster-head can be further aggregated, thereby reducing the amount of data returned, further reducing energy consumption.

Considering energy consumption, so long as the threshold of angle range of atomic region is not too large, we can always ensure the interior angle of the adjacent edges along the path to be large obtuse angle. According to Cosine Theorem, DCDA-LEACH tree routing algorithm is also superior than LEACH in energy consumption.

V. SIMULATION AND RESULTS ANALYSIS

We use OMNeT++ for simulation. Experiment parameters are set as follows: total number of network nodes $= 100$, nodes distribute in a rectangular area of $500 \times 500$, the fixed base stations is outside the scope of the network. Consumption parameters of the circuit $E_{dc}=50nJ/bit$, Amplifier Parameters $\varepsilon=100pJ/bit/m^2$, initialization of the node energy=0.5J, Cluster radius=80m, data length of each cluster node sending to cluster-heads =1600bit, the data aggregation rate of cluster-heads=0.3 in DCDA-LEACH and 0.6 in LEACH.

Fig. 4 compares the total energy consumption of DCDA-LEACH and LEACH in the simulation. As can be seen, in the 1s~2s stage, the power consumption of DCDA-LEACH is slightly higher than the LEACH protocol due to the establishment of routing. In the normal work stage, energy consumption of DCDA-LEACH is significantly lower than the LEACH protocol.

Fig. 5 shows the energy consumption of one cluster-head which is closer to BS in LEACH and DCDA-LEACH. DCDA-LEACH consumes more than LEACH due to the need to forward the data from other cluster-heads.

Then, we will discuss the performance improvement brought by the tree routing. In LEACH, the cluster-heads far away from BS communicate directly with BS, consuming enormous amounts of energy, which leads to the rapid depletion of the energy of these nodes and is not conducive to extend the life of the network as a whole. Using tree routing, the cluster-heads send data to BS along multi-hop path. In this way, multi-path fading can be avoided. Although the energy consumption of the nodes close to BS will increase because of forwarding additional data, the life of the network will be effectively extended. Furthermore along the course of transmission by skip, data of each cluster-head can be further aggregated, thereby reducing the amount of data returned, further reducing energy consumption.

Figure 4. Total energy consumption comparison of DCDA-LEACH and LEACH
REFERENCES


