Data-Aware Clustering Hierarchy for Wireless Sensor Networks*

Xiaochen Wu, Peng Wang, Wei Wang, and Baile Shi

Fudan University, ShangHai, China
{052021120,pengwang5,weiwang1,blshi}@fudan.edu.cn

Abstract. In recent years, the wireless sensor network (WSN) is employed a wide range of applications. But existing communication protocols for WSN ignore the characteristics of collected data and set routes only according to the mutual distance and residual energy of sensors. In this paper we propose a Data-Aware Clustering Hierarchy (DACH), which organizes the sensors based on both distance information and data distribution in the network Furthermore, we also present a multi-granularity query processing method based on DACH, which can estimate the query result more efficiently. Our empirical study shows that DACH has higher energy efficiency than Low-Energy Adaptive Clustering Hierarchy (LEACH), and the multi-granularity query processing method based on DACH brings more accurate results than a random access system using same cost of energy.

Keywords: wireless sensor network, communication protocol, data distribution, multi-granularity query.

1 Introduction

In recent years, the wireless sensor network.(WSN) [1, 2] is employed a wide range of applications in military security, environmental monitoring, and many other fields. Except some high accuracy required applications (for example, applications in military), most applications of WSN are cost driven. Users want to acquire more information with less energy cost. In order to minimize the energy consumption and maximize the life span of the network, clustering techniques based on data fusion [3] such as LEACH [4], LEACH-C [5], BCDCP [6] etc. have been proposed.

All above cluster-based protocols try to find the shorter routes for data transmission and spread energy consumption around all the sensors more evenly. The sensors are organized into clusters according to the mutual distance and residual energy of them. In such scheme, the process of data collection is independent with the characteristics of collected data. But in many applications of WSN, data collected from some adjacent sensors are similar. The sensing field can be divided into regions with different characteristics. For example, in a building site, the temperature data col-

* This research is supported in part by the National High-Tech Research and Development Plan of China under Grant 2006AA01Z234 and the National Basic Research Program of China under grant 2005CB321905.

T. Washio et al. (Eds.): PAKDD 2008, LNAI 5012, pp. 795–802, 2008.

lected from indoor sensors and outdoor sensors may be similar respectively. However, during the running of the wireless sensor network, we can estimate data distribution in the network using some data mining methods. And based on this information, the clusters can be organized not only according to the mutual distance, but also the characteristics of collected data. We can build the clusters so that data collected from sensors in a same cluster are similar. This method can compress data volume more efficiently after data fusion and prolong the network's life span further. Users can acquire more information from the network with less energy cost.

In this paper, we propose a Data-Aware Clustering Hierarchy (DACH), which is not only energy-efficient, but also capable of obtaining data distribution from the network. In DACH, data distribution is estimated by a data mining process based on collected data and the sensors are distributed into a clustering hierarchy according to the discriminations between the collected data. Furthermore, we introduce a multi-granularity query processing method based on DACH to estimate the query results using a few sensors' data instead of all of them.

Our Contributions

1. We propose a data mining method to estimate the data distribution in wireless sensor network, and based on it we introduce a new clustering structure and also a new communication protocol for WSN.
2. We propose a multi-granularity query processing method based on DACH to estimate the query results using a few sensors' data instead of all of them.

2 Data-Aware Clustering Hierarchy

2.1 Data Distribution in Wireless Sensor Network

In many applications of wireless sensor network, the sensing field can be divided into a series of regions with different characteristics. In the example mentioned in section 1, the whole building site contains the indoor regions and the outdoor regions, and the space inside a building still can be divided into different building stories, rooms and areas. It is possible that the temperature is very different between some regions (for example, the indoor regions and the outdoor regions). And on the other hand, it is similar in some regions (for example, the areas in a same room). We refer to this property as the "Regional Property".

As the regional property of the sensing field, data collected from sensors deployed in the field also have the regional property. In above example, data collected from outdoor sensors may be very different from data collected from indoor sensors, as the difference of the physical conditions between outdoor and indoor regions. And in the mean time, data collected from sensors in a same room may be similar.

Based on the regional property, we can estimate the data distribution by the discrimination between data collected from different sensors or sensor sets (discrimination of sensors or discrimination of sensor sets for short). We can consider data collected from each sensor as a time series and define the discrimination of sensors by the discrimination of the time series. In this paper, the time series, denoted by TS, with length n is: $TS=TS_1, TS_2, \ldots, TS_n$.

Def 1: Discrimination of sensors: Data collected from a sensor in a time interval compose a time series. We use Euclidean distance to define the discrimination between sensors. Furthermore, the value of the discrimination is divided by $n^{1/2}$ to eliminate the influence of the length of time series:

$$disc(i, j) = disc(TS^i, TS^j) = \sqrt{[\sum_{k=1}^{n} (TS_k^i - TS_k^j)^2]/n} \quad (1)$$

Moreover, we can estimate the discrimination between two sensor sets by the discrimination between centroids of corresponding time series of sensors in each set.

2.2 Data-Aware Clustering Hierarchy

As an example to illustrate our motivation, we use a sensor network to monitor the temperatures of a building site as shown in figure 1a. The gray region indicates outdoor regions, and the white part indicates indoor ones. Using traditional clustering methods, the clustering structure in a certain round may be organized as Figure 1b [4]. It shows that there are 17 nodes in the cluster A, nine of which are in the gray region and other eight are in the white one. Data collected from sensors in this cluster may be very different between each other.

An ideal clustering structure is shown in figure 1c. It still contains 5 clusters. The difference from figure 1b is that each cluster represents a section of outdoor or indoor regions. So this structure is more data-aware. Data collected from sensors in same clusters are similar. Based on this property, we can acquire more information from WSN using less energy. We can compress data volume more efficiently after data fusion and prolong the network's life span further. Moreover, in some applications of approximate queries, we can only query a few sensors' data instead of all of them.

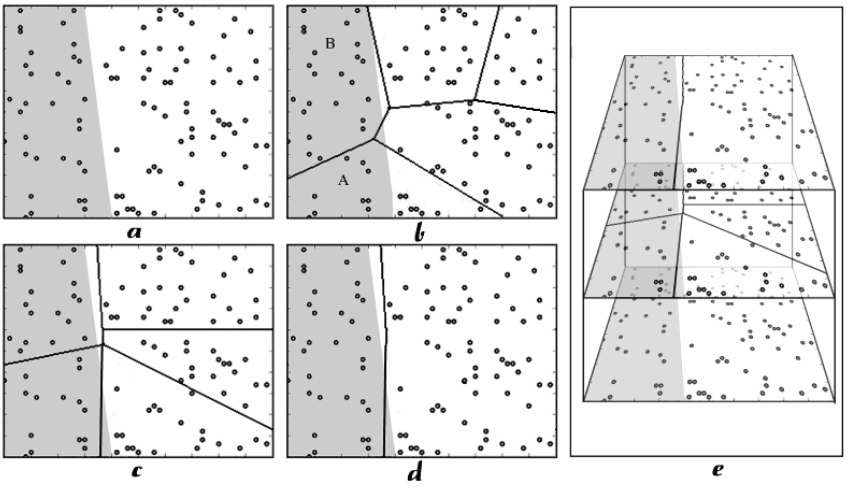

Fig. 1. (a) A region where we deployed wireless sensors to monitor temperature; **(b)** The clustering structure in a certain round using traditional cluster method; **(c)** The clustering structure based on data distribution; **(d)** The clustering structure on the higher level; **(e)** A data-aware clustering hierarchy.

In the figure 1d, similar clusters in figure 1c are merged to larger clusters respectively. The left cluster contains all the sensors in the outdoor regions, and the right one contains sensors in the indoor regions. The cluster structures in these two figures constitute a clustering hierarchy showed in figure 1e.

2.3 Algorithm for Building Data-Aware Clustering Hierarchy

In this section we propose an algorithm for building the data-aware clustering hierarchy based on topological structure of network and data distribution. To facilitate our discussion, we first give some general definitions:

Def.2 Relay: The indirectly transmission from node A to node B through node C is called Relay. The node C is called Relay node.

Def.3 Relay Region: Given a node s and a relay node r, the relay region of s with respect to r is defined as follows:

$$R_{\alpha,c}(s,r) = \{x \mid \|sx\|^{\alpha} > \|sr\|^{\alpha} + \|rx\|^{\alpha} + c\} \ . \tag{2}$$

where ||xy|| denotes the distance between node x and node y, α and c are two constant parameters which equal to 4 and E_{TX}/ε_{amp} respectively according to the above radio model. Obviously, the nodes in the relay region of s with respect to r can be reached with least energy by relaying r.

Def.4 Neighbor: The node not in any relay region of s is called neighbor of s. Formally, we define it as follows:

$$N_{\alpha,c}(s) = \{u \mid \forall r, u \notin R_{\alpha,c}(s,r)\} \ . \tag{3}$$

Furthermore, two sensor sets A and B are called neighborhood sets if there are two neighborhood sensors a and b, where sensor a is in the set A and b is in set B.

Before building a k-level clustering hierarchy, we define a series of thresholds, δ_0, δ_1, …, δ_{k-2} satisfying $\delta_{k-2}>\delta_{k-1}>\ldots>\delta_1>\delta_0=0$. The thresholds can be specified according to the sensing scenarios. For example, in a temperature monitoring system, the thresholds can be specified as 0, 0.5, 1, 2, 4 (℃).

Assume that the base station save data collected from every sensors as a time series. In the lowest level (level 0), we initialize a set for each sensor in the network and compute the discrimination between each pair of sets using following equation:

$$d(S_A, S_B) = \begin{cases} \infty & \text{A and B are not neighbors} \\ disc(S_A, S_B) & \text{A and B are neighbors} \end{cases} \ . \tag{4}$$

Then, we build clusters on higher level in a bottom-up way by following steps:

1. Find the pair of sets with minimum mutual discrimination d_{min};
2. If d_{min} is larger than the threshold δ_i, output current sets as the clustering structure of level-i;
3. Combine these two sets into a new set, compute the centroid of all the time series in the new set and update its discrimination with other sets
4. Repeat steps 1-3 until the minimum discrimination is larger than the maximum threshold δ_{k-2}

5. Combine all the remaining sets into an only set as the cluster on the highest level, which contains all the sensors in the network.

After that, we obtained a clustering hierarchy. Every set on every level contains a series of subset on the lower level which is called as "descendent sets". And in the mean time, it is also a part of a set on the higher level which is called as "ancestor set".

3 A Communication Protocol Based on DACH

Based on the DACH proposed in last section, we introduce a novel communication protocol for WSN. We also call this communication protocol as DACH for short in the context of not leading to any ambiguity. DACH operates in three phases: initialization, setup and data transmission.

Initialization Phase: When the sensors are deployed on the field or the topological structure of the network is changed (e.g. when the energy of some sensors is exhausted or the properties of the circumstance are changed), the network enters the initialization phase.

During the initialization phase, the base station receives data from all sensors in a given period and generates a time series for every sensor. Based on these time series, the base station computes the discriminations between each pair of sensors and builds the clustering hierarchy using the method proposed in section 2.3.

Setup Phase: When the system is initialized, the network enters the setup phase. The main task in this phase is to generate routing path and schedule for each node. The base station receives information of the current energy status from all the nodes in the network. Based on the feedback and the clustering hierarchy, base station generates the routing path in a bottom-up way.

For each cluster, the algorithm selects one node as the cluster-head. For level 0, since each cluster only contains one node, each node is a cluster-head. For each cluster C on level-i (i>0), the cluster-head must satisfy following two conditions:

1) It is a cluster-head of one of its children clusters;
2) Its residual energy is the highest among cluster-heads of all its children clusters.

Now the routing path of each sensor node can be obtained easily. It first transmits data to the corresponding cluster-head which subsequently transmits data to the cluster-head of its parent cluster. This process continues until the data is transmitted to the only cluster-head on the highest level. And the data is sent to the base station finally.

The cluster-head of a level-i cluster and all the cluster-heads in its children cluster compose a sub-network. To improve the energy efficiency, in this sub-network the sensors transmit data using a multi-hop method.

Data Transmission Phase: The data transmission phase consists of three major activities: data gathering, data aggregation and data routing. Using the scheme described above, each sensor node transmits sensed data to its corresponding cluster-head. For each cluster-head, once receiving data from all contained nodes, it aggregates the collected data into a data of smaller volume and sends it to cluster-head on higher level. The cluster-head on highest level transmits the aggregated data to the base station.

For spreading energy consumption between sensors more evenly, after a period of data transmission phase, the network will enter the setup phase again, reselect the cluster-head and regenerate the routing path for every sensor.

4 Multi-granularity Query Processing Method Based on DACH

In most performance-driven application, WSN may have less stringent performance requirements and can be implemented at much lower cost. J. Frolik proposed random access techniques to help facilitate such requirements [9]. In [9] the quality of service (QoS) measures application reliability with a goal of energy efficiency. According to the user-defined quality of service, the random access system selects a proportion of sensors for data gathering. These sensors are called "active sensors". But as the active sensors are selected randomly, the data collected by a random access system may not be able to simulate the whole data set appropriately.

In this section, we discuss our multi-granularity query processing method based on DACH for cost-driven applications. Since in each cluster, the data of all nodes are similar, we can execute the query on the cluster-heads on certain level instead of all sensor nodes.

The multi-granularity queries have the following basic structure:

```
SELECT expr1, expr2…
FROM network
WHERE pred1 [and|or] pred2
LEVEL ON levelNum
```

The SELECT, FROM and WHERE clauses are defined as the standard SQL. The LEVEL ON clause specifies the level of the query. From the definition of clustering hierarchy, it can be seen that the data of all nodes are similar in a certain cluster, and the lower the level, the number of the cluster-heads is larger and the data are more similar. In other words, the user can specify the number of active sensors and the approximate estimating error using the LEVEL ON clause.

Assume the levelNum be k, the cluster-heads on level k estimate the data of the sensors in the same cluster based on its own data. That is, the query is processed on the cluster-heads on level-k.

Because of the similarity of data collected from sensors in the same cluster, our method can estimates the query result accurately. Data gathered by the random access system may omit data of sensors in some small and special region because the probability of sampling data in it is relatively low. On the contrary, our method will not ignore these regions.

5 Performance Evaluations

To test the performance of DACH and the multi-granularity query processing method, we simulate an environment temperature monitoring system. Using this simulated system, we compare the energy efficiency of DACH with LEACH and the estimating accuracy of the multi-granularity query processing method with the random access system presented in [9]. All the algorithms are implemented in JAVA. The test

environment is PC with AMD Athlon processors 3000+, 1GB of RAM, and running Windows XP Professional.

We use the network model and radio model discussed in [4, 5, 6] and simulate the temperature information using a bitmap. The system generates the coordinates of the sensors randomly and set the parameters of temperature based on RGB colors of the corresponding pixels. The value of R/10 represents the average temperature in one day. The value of G/10 represents the maximum temperature in one day. And the value of B/10 represents the time of the maximum temperature. We simulate the intraday temperature by sinusoid. The temperature on point (x, y) at time t is denoted by:

$$T_{(x,y,t)} = R_{(x,y)}/10 + (G_{(x,y)} - R_{(x,y)})/10 \cdot \sin((x - B_{(x,y)}/10 + 6) \cdot \pi/12) \ . \quad (5)$$

We assume that the base station locates at point (-20, -20). Each node is assigned an initial energy of 2J. The number of data frames transmitted for each round is set at 10; the data message size for all simulations is fixed at 500 bytes, of which 25 bytes is the length of the packet header.

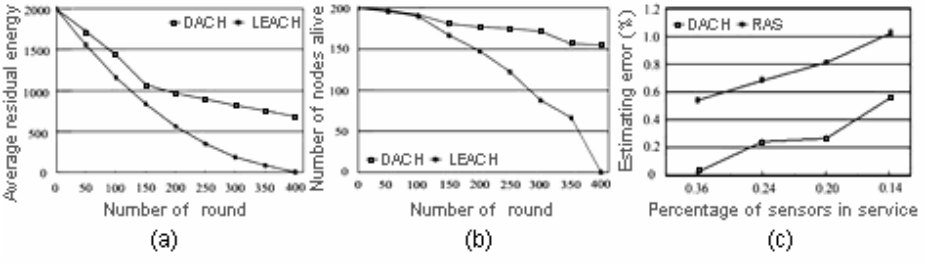

Fig. 2. (a) The average residual energy of sensors of DACH and LEACH at different number of operation rounds. **(b)** The number of nodes that remain alive at different number of rounds. **(c)** The estimating errors of the multi-granularity query processing method and that of the random access system corresponding to percentages of sensors in service.

In the first experiment we compare the energy efficiency of DACH and LEACH. We simulate a 50m×50m network with 200 sensors. Figure 2a shows the average residual energy of sensors of DACH and LEACH at different number of operation rounds. It can be seen that DACH has more desirable energy efficiency than LEACH.

Figure 2b shows the number of nodes that remain alive at different number of rounds. After 100 rounds, the sensors in the LEACH die more quickly than DACH. And when all sensors are dead in the LEACH, more than 150 sensors in the DACH remain alive. It indicates that DACH distributes the energy load among the sensors in the network more efficiently.

In the second experiment, we compare the accuracy of the query results between the multi-granularity query processing method based on DACH and the random access system. We simulate a 50m×50m network with 100 sensors and compute the average temperature in the network. The figure 2c shows that our method always

processes the query more accurately than the random access system. On average, our method reduces the estimating error by 0.5 degrees.

6 Conclusions

In this paper we proposed a data-aware clustering hierarchy for wireless sensors network (DACH) and a multi-granularity query processing method based on DACH. DACH divides the sensors into clusters according to the data distribution as well as mutual distance between sensors. Using the similarity of data collected from sensors in same clusters, the multi-granularity query processing method estimates the query result only by the data from cluster-heads on certain level. The simulation results show that DACH has much higher energy efficiency than LEACH. And the estimating errors of the multi-granularity query processing method are less than random access approach.

References

[1] Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless Sensor Network: A Survey. In: IEEE Communications Magazine (August 2002)
[2] Dong, M., Yung, K., Kaiser, W.: Low Power Signal Processing Architectures for Network Microsensors. In: Proceedings 1997 International Symposium on Low Power Electronics and Design, August 1997, pp. 173–177 (1997)
[3] Hall, D.: Mathematical Techniques in Multisensor Data Fusion. Artech House, Boston (1992)
[4] Heinzelman, W.R., Chandrakasan, A.P., Balakrishnan, H.: Energy Efficient Communication Protocol for Wireless Microsensor Networks. In: 33rd Hawaii International Conference on System Sciences (January 2000)
[5] Heinzelman, W.B., Chandrakasan, A.P., Balakrishnan, H.: An Application-Specific Protocol Architecture for Wireless Microsensor Networks. IEEE Transactions on wireless communications 1(4) (2002)
[6] Murganathan, S.D., Ma, D.C.F., Bhasin, R.I., Fapojuwo, A.A.O.: A Centralized Energy-Efficient Routing Protocol for Wireless Sensor Networks. In: IEEE Radio Communications (March 2005)
[7] Lindsey, S., Raghavendra, C., Sivalingam, K.M.: Data Gathering Algorithms in Sensor Networks using Energy Metrics. IEEE Transactions on Parallel and Distributed Systems 13(9) (September 2002)
[8] Ding, P., Holliday, J., Celik, A.: Distributed Energy-Efficient Hierarchical Clustering for Wireless Sensor Networks. In: Prasanna, V.K., Iyengar, S.S., Spirakis, P.G., Welsh, M. (eds.) DCOSS 2005. LNCS, vol. 3560, Springer, Heidelberg (2005)
[9] Frolik, J.: QoS Control for Random Access Wireless Sensor Networks. In: IEEE Wireless Communications and Networking Conference (2004)