TAGPP: a tiny aggregation algorithm with preprocessing in local cluster

Yang Xiao-fei, Wu Xiao-bei, and Huang Jin-an
Automation school
nanjing university of science and technology
Nanjing,China
yxfei_0809@163.com

Abstract-hundreds or thousands of sensor nodes are deployed to WSNs (wireless sensor networks). They contribute to collect data in various environment, but with limited battery power. There are some deficiencies in WSNs. Many sensors share limited wireless channel bandwidth, which leads to some sensors lose their data sent to cluster-head. Many cheap sensors are equipped to nodes for low cost and they usually have poor capability of sensing. They are easy to be interfered by harsh environment they work, leading to measurement error and noise disturbance. For these deficiencies,a tiny data aggregation algorithm with preprocessing applied to individual local cluster, based on Grubbs criteria, fuzzy clustering, is proposed in the paper. Clustering is also adopted, as an important mechanism, to reduce energy consumption and make better network performance. Individual cluster sensors' data can be aggregated in cluster-head and Just the accurate result sent to sink node The simulation results show that It can efficiently eliminate outlier (abnormal data), make up for the impact of missing data in communications and improve data preciseness.

Keywords-wireless sensor networks; data fusion; outlier; grubbs criteria; fuzzy clustering

I. INTRODUCTION

The advance of micro-electro-mechanical system technology, wireless communication and digital electronics leads to the emergence of wireless networks of sensor devices [1], which are capable of sensing, processing and communicating. Sensor networks can be deployed in diverse environment to collect and process useful information. Sensor networks consist of many sensor nodes equipped micro-controller, RF circuits and different sensors, such as temperature, press, humidity, acoustic sensor and so on. Thus, it has a wide range of applications in the fields of military, disaster monitoring, precision agriculture and medical health care.

In all typical applications, hundreds or thousands of sensor nodes are arrayed in monitoring filed and data acquired periodically via various sensors are transmitted back by wireless communication. Via analyzing these data, observers can obtain information on physical environment. Generally, sensors equipped in networks have low accuracy and poor performance in order to cut down the cost of networks. Therefore, they are also easily affected or interfered,such as harsh environment, leading to noise or disturbance. On the other hand, Different from other ad-hoc networks, limited energy resource is a key characteristic of the sensor networks due to the fact that sensor nodes are generally powered by two batteries and cannot be recharged. So how to minimize energy consumption and maximize lifetime of networks have been a hot topic for a long time. In the existing work, solutions to energy-efficient routing algorithm have been extensively studied in both general multi-hop wireless networks [2,3,4] and the particular backdrop of sensor networks [5,6]. These algorithms can achieve the goal of either minimizing energy consumption or maximizing the networks lifetime. Data aggregation is one basic operation in sensor networks, which can extract useful information from redundancy data. Thus, few data need to be transmitted to sink node and transmission cost decreases dramatically. The LEACH algorithm presented in [7] is an elegant solution where clusters are formed to fuse data before transmitting to the base station. PEGASIS is also a near-optimal chain-based algorithm that minimizes energy [6]. However, it is worth mentioning that the problem of preprocess the raw data is not considered in these paper. In fact, several factors make wireless sensor networks acquired data prone to be outlier. For instance, these data are collected from the real world using imperfect performance sensing devices. These sensors are batteries powered and thus their performance tend to deteriorate as power exhausted. In large scale networks, sensors number is great and may cause accumulated errors. Data loss is also a problem which cannot be ignored. A large number of sensors and limited transmission bandwidth may lead data to collide in multi-nodes transmission case and cause data loss. Though, clustering is an effectual mechanism to improve quality, accuracy and fault tolerant of the sensor networks. We also can make further approach to obtain more accurate data. For example, we can eliminate outlier prior to aggregate them into final results. If not, these data will reduce the accuracy of the results. Data fusion can take from the impact of some sensors data loss in individual cluster.

The main contribution of the paper is two-fold. First, the grubbs criteria is introduced in wireless sensor networks,

978-0-7695-3610-1/09 $25.00 © 2009 IEEE
DOI 10.1109/NSWCTC.2009.122

which can efficiently detect outlier in the raw data before aggregation. Second, a tiny data fusion algorithm based grubbs criteria and fuzzy clustering algorithm is proposed.

The rest of the paper is organized as follows. Section Ⅱ briefly presents relevant previous work.. Section Ⅲ describes relative knowledge including local clustering architecture, grubbs criteria and fuzzy clustering algorithm. Section Ⅳ describes the algorithm. Section Ⅴ simulates and verifies TAGPP algorithm. Section Ⅵ concludes the paper and directs for future work.

II. RELATED WORK

Outlier detection, an essential step preceding any data analysis, is used to suppress outlier. Outlier detection improves robustness of the data analysis. Barnett and Lewis make a survey on outlier detection methodologies in the statistics community[8]. J. Branch et al. develop an algorithm for the computation of outlier based on parametric, unsupervised methods [9].

Data fusion strengthens accuracy and robustness of the acquired data. Considerable attention has attracted from the research community [6, 7, 10]. In sensor networks, data aggregation inside the networks can drastically cut down the communication cost and ensure the desired bounds on the quality of data. Madden et al. propose an effective aggregation tree (TAG-tree), which works on well-defined epochs, and reduces the communication by using an optimum tree structure[10]. Sharaf et al. propose more efficient means exploiting group by queries, which works on top of TAG and COUGAR aggregation [11]. In addition to the tree aggregation, cluster-based aggregation approaches have also been investigated in LEACH and PEGASIS. In LEACH, randomly selected cluster heads perform aggregation, and communicate directly with the base station to reduce energy consumption. PEGASIS selects the cluster head by organizing nodes in a chain. Using only one cluster head at a time conserves energy at the expense of introducing latency issues. Compression is used recently as an aggregation method for sensor networks. Lazaridis and Mehrotra propose using a piecewise constant approximation algorithm to compress the time series with quality guarantees [12]. Since the method proposed is lossy compression for single dimension time series only, applying it would not keep the valuable correlation information in multidimensional data. Deligiannakis et al. propose extracting a base signal from data and further using piecewise linear regression to compress original signal, using the base signal [13].

III. RELATIVE KNOWLEDGE ON TAGPP ALGORITHM

We consider such sensor networks that each sensor periodically produces and transmit data to sink for further processing. Harsh environment, low performance sensors and unreliable communication generally cause data error or data loss. Since sensor nodes possess computation abilities, part of the computation can be off-loaded to in-networks

A. Hierarchical Clustering Architecture

LEACH is a perfect mechanism to build network protocol that minimizes energy dissipation in sensor networks. The key features of LEACH are:

- Localized coordination and control for cluster set-up and operation
- Local compression to reduce global communication

The same hierarchical clustering architecture and means to classify the nodes as LEACH are used in the paper. Thus, all nodes are divided into several local clusters. Once all nodes are organized, each cluster-head creates a schedule to non-cluster-head nodes in local cluster. After that, the non-cluster-head nodes can transmit their data to cluster-head as their orders. Once the cluster-head has data from all the nodes in local cluster, the cluster-head aggregates the data based on data fusion algorithm and then transmits the aggregated data to the sink node.

B. Aggregation Architecture in Local Cluster

Sensor nodes are with the capability of acquiring, computing and transmitting. They acquire monitoring targets' information and store them in their flash or memory periodically. These sensor nodes can be viewed as many tiny distributed Databases. They can transmit data to high level nodes. The cluster-head node pre-processes and aggregates them. Data aggregated will be sent to sink node. Observer can get data from it and make further analysis to master useful information on the monitored region.TAGPP algorithm works in individual Local cluster illustrated in fig.1. Data preprocessing is added to TAGPP algorithm.

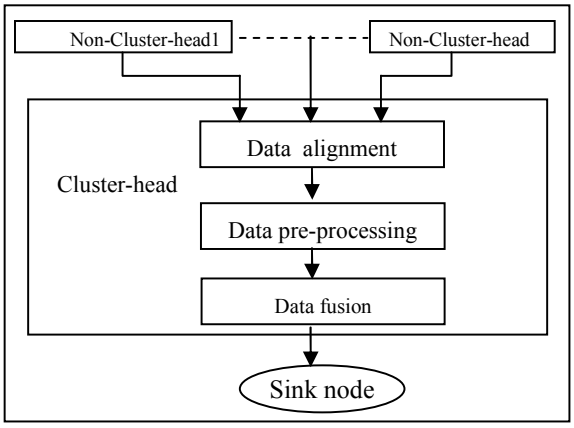

Fig.1 Architecture of local cluster

C. Data Alignment

Data alignment generally includes coordinates transform and time adjustment. The aim is to set common reference point. Actual means adopted in project should be rested with special application. If applied to high acquiring frequency case, estimation and plug algorithm can be employed. In low frequency case, data acquired from sensor node can be alternative, which next to reference time of data aggregation center. The second method is adopted by the paper's MTALAB simulation, because indoor temperature can't

greatly change in short time. That's to say, using equation $Z_{kT} = x_i$, KT is the reference time and x_i is datum next to KT time mostly

D. Grubbs Criteria to Detect Outlier

In sensor networks, so many causes would lead outlier into acquired data. If not kicked out and aggregated them into result, they would make the accuracy of the result lower. Grubbs criteria contributes to eliminate outlier in our TAGPP algorithm.

Suppose we have got a set of raw data from non-cluster-head sensor nodes $x_1, x_2 \cdots x_{n-1}, x_n$, obeying the law of normal distribution curve in general. N is the number of sensor nodes sensing in local cluster. Firstly, they are sorted in order from small to large as (1).

$$x_1 \le x_2 \le x_3 \cdots \le x_{n-1} \le x_n \qquad (1)$$

Then, Set

$$\begin{cases} \bar{x} = \dfrac{1}{n}\sum x_i \\ v_i = x_i - \bar{x} \\ S = \sqrt{\dfrac{1}{n}\sum v_i^2} \end{cases} \qquad (2)$$

$\bar{x}, S$ are the notations of mean and standard deviation respectively. According to principle of sequential statistics, find out the exact distribution of statistics (3).

$$g_i = \frac{x_i - \bar{x}}{S} \quad (\text{i=1 或 n}) \qquad (3)$$

If significance level α and n have been given, $p[g_i \ge g_0(n,\alpha)] = \alpha$ is a small probability event and will not appear when $X_i(i = 1, 2, \cdots n)$ conform to normal distribution. The critical value $g_0(n,\alpha)$ can be looked up via G table. Compareing with critical value can percolate outliers. Outliers generally exist at the two sides of (1), that's to say, the maximum or the minimum is the most suspicious value within data set. If $g_i < g_0(n,\alpha)$, it indicates that there are no outlier and all data can be used to aggregate. Otherwise, if $g_i \ge g_0(n,\alpha)$ the data set includes outlier which should be deleted from data set. Such program can be iterated until none outlier exists.

E. Fuzzy Clustering with Weighted Coefficients

It's a good method to weight various values' accuracy using weighted coefficient, which reflects values' relative importance or contribution to the fact. Many sensors are used to measure the same object in monitoring region and send us many various data in accordance with sensing accuracy of sensors and distance with object under test. Poor performance of the sensors and long distance with object generally lead measurement error into acquired data. Every one can reflect the fact of the object to some extent. Higher accurate data are assigned bigger coefficients and lower accurate data are assigned smaller. The bigger coefficients of the data are, the greater contributions to the result the data make. It's no doubt that the aggregated result can be more closer to the fact by weighted aggregation. Fuzzy clustering algorithm is used to produce coefficients and aggregation result.

Suppose n data participated, result y can be described as $y = f(x_1, x_2 \cdots x_n)$. Fuzzy math and first derivative T-S modeling are utilized to achieve data fusion. Result can be described as (4), w_i is weighted coefficient of data [14].

$$\begin{aligned} y &= \sum_{i=1}^{n} w_i x_i \\ \sum_{i=1}^{n} w_i &= 1 \end{aligned} \qquad (4)$$

The weighted coefficient can be worked out via membership function. Normal distribution membership function described in (5) is employed in the paper [15,16].

$$u(x) = e^{-(\frac{x-a}{b})^2}, \text{ a>0, b>0} \qquad (5)$$

Notation a is the center of universe of discourse, b is standard deviation of the random data. They can be easily worked out. When the membership function is known, x_i can be substituted to (6) to get their probability u_i, just as described in (6).

$$u_i = e^{-(\frac{x_i - a}{b})^2} \qquad (6)$$

Then all the probability should be normalized to produce weighted coefficient. Now the aggregated result y can be worked out by (7).

$$y = \sum_{i=1}^{n} \frac{u_i}{u_1 + u_2 + \cdots + u_n} x_i \qquad (7)$$

IV. TAGPP ALGORITHM

In this section, our tiny or lightweight aggregation algorithm with preprocessing is proposed. Based on the number of nodes in local cluster, sensor head creates a TDMA schedule broadcasted to every node when to transmit data in turn and then receives data from them. Algorithm function process describes as follows:

Step1: receive data from the non-cluster-head nodes during their transmission time .

Step2: select data needed to data alignment according to aggregation reference time

Step3: eliminate outlier using grubbs criteria described in D of Ⅲ. Such program may be iterated untill none outlier exists.

Step4: assign weighted coefficients to data set without outlier based on algorithm described in E of Ⅲ.

Step5: work out the aggregation value.

Step6: transmit the aggregation value to the sink node over transmission time.

V. SIMULATION IN MATLAB

The performance of our algorithm was evaluated via MTALAB simulation in this section. For experimentation, real-world sensor data streams available from Intel Berkeley research Lab was used[17], in which distributed data points share spatial and temporal properties. The data were comprised of sensor readings (*e.g.* heat, light, temperature) from 54 sensors, which were periodically transmitted to a base station. There exists some missing data points in the data.txt For reality, we hadn't fixed up them. The data points include the following features: (i) ID of the sensor that produced the point, (ii) epoch (sequential number denoting the data points position in the entire stream), (iii) data value, (iv) location coordinates of the sensor. But in the paper, only temperature was used to simulate our algorithm.

For the real-world sensor data streams, the sensors were arranged in the lab as shown in [17]. LEACH algorithm was employed to achieve clustering architecture. The 54 sensor nodes were divided into five local clusters. Data from two of them were used to simulate our algorithm. One cluster consisted of sensors number from 2 to 10, the other one is from 14 to 20. we selected one hour's data from time 0:59 to 1:59 and another from time 2:59 to 3:59. These data were imported our algorithm and results were plotted by MATLAB. The simulation results are illustrated in fig.2. Solid lines are the curves of sensor nodes in local cluster according to their real acquired data and short dotted line with little circle points are the aggregation data. As seen from fig.2, the fact that the temperatures of the environment changes from high to low can be inspected through all the solid lines and such change tendencies are reflected well by the curve of the aggregation data. They synthesize all the information of the sensor nodes, reducing measurement errors and more closer to the fact.

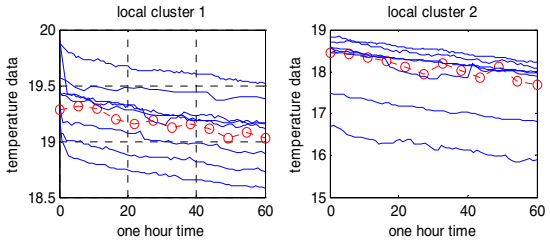

Fig.2 sensor nodes data and aggregated data

VI. CONCLUSION AND FUTURE WORK

A tiny aggregation algorithm with preprocessing has been proposed in the paper. It's effective to eliminate outlier and aggregate data via simulation using real-world data streams. Its simplicity may make it run well on hardware of sensor nodes. The algorithm will be transplanted to our real sensor nodes system in future work

REFERENCES

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam and E. Cyirci, "Wireless Sensor Networks: A survey," Computer networks, vol. 38, 2002, pp.393-422. doi:10.1.1.5.2442.

[2] J. Chang and L. Tassiulas, "Energy Conserving Routing in Wireless Ad-hoc Networks," IEEE Conf. On Computer Communications (INFOCOM2000), Mar. 2000, pp..22-31. doi:10.1.1.107.7720.

[3] K. Dasgupta, K. Kalpakis and P. Namjoshi, "Efficient Algorithms for Maximum Lifetime Data Gathering and Aggregation in Wireless Sensor Networks," Computer networks, vol.42, 2003, pp. 697-716.doi:10.1.1.19.4235.

[4] A. Sankar and L. zhen, "Maximum Lifetime Routing in Wireless Ad-hoc Networks," IEEE Conf. On Computer Communications(INFOCOM 2004), Mar. 2004, pp. 1089-1097. doi:10.1.1.1.4437.

[5] M. Bhardwaj and A. P. Chandrakasan, "Bounding the Lifetime of Sensor Networks via Optimal Role Assignments," IEEE Conf. On Computer Communications(INFOCOM2002), June, 2002. pp.1587-1596. doi:10.1.1.18.2661.

[6] S. Lindsey, C. Raghavendra, and K. M. Sivalingam, "Data Gathering Algorithms in Sensor Networks Using Energy Metrics," IEEE Trans. On parallel and distributed systems, Sept. 2002. Vol. 13, pp.924-935. doi:10.1.1.117.8734.

[7] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient Communication Protocol for Wireless Microsensor Networks," Proc.of Hawaii Conf. system sciences(HICSS00), Jan. 2000, pp.3005-3014. doi:10.1.1.112.3914.

[8] V. Barnett and T. Lewis, Outliers in Statistical Data, John Wiley & Sons, 1994.

[9] J. Branch, B. Szymanski , C. Giannella, R. Wolff, and H. Kargupta, "In-Network Outlier Detection in Wireless Sensor Networks," International Conf. On Distributed Computing Systems(ICDCS 2006), July,2006. pp51-58.doi:10.1.1.61.1223.

[10] S. Madden, M.J. Franklin, J.M. Hellerstein, W. Hong, "TAG: a Tiny Aggregation Service for Ad hoc Sensor Networks," Proc. of USENIX, Jun. 2002, pp. 131–146. doi:10.1.1.7.4301.

[11] M. A. Sharaf, J. Beaver, A. Labrinidis, P. K. Chrysanthis, "Balancing Energy Efficiency and Quality of Aggregate Data in Sensor Networks," Journal on Very Large Data Base. Vol.13, 2004, pp.384–403. doi:10.1.1.93.9451.

[12] I. Lazaridis, S. Mehrotra, "Capturing Sensor-generated Time Series with Quality Guarantees," International Conf. On Data Engineering(ICDE2003), March, 2003, pp.429-440. doi:10.1.1.11.5901.

[13] A. Deligiannakis, Y. Kotidis, N. Roussopoulos, "Compressing Historical Information in Sensor Networks," Proc. of Special Interest Group on Management Of Data(SIGMOD2004), June, 2004. pp.527–538. doi:10.1.1.10.1657

[14] T. Takagi and M. Sugeno, "Fuzzy Identification of Systems and Its Applications to Modeling and Control," IEEE Trans. Syst.,Man,Cybern. vol.15,1985, pp.116-132.

[15] L. Zehua, C. Chuanbo, and L. Wenhai, "Normal Distribution Fuzzy Sets,"International Conf. Fuzzy Information and Engineering(ICFIE2007),May, 2007, pp280-289.

[16] Kezhong Huang, random method and fuzzy mathmatics aplication, shanghai:tongji press, 1985.

[17] Intel Berkeley Research Lab. Wireless Sensor Data. http://db.lcs.mit.edu/labdata/labdata.html