

# An Efficient Approximation Algorithm for Data Aggregation in Wireless Sensor Networks

Zhang ShuKui \*<sup>1,2</sup>, Cui ZhiMing<sup>1,2</sup>, Gong ShengRong<sup>2</sup>, and Fan JianXi<sup>2</sup>

<sup>1</sup> JiangSu Province Support Software Engineering R&D Center for Modern Information Technology Application in Enterprise, Suzhou, China, 215104

<sup>2</sup>Institute of Computer Science and Technology, Soochow University, Suzhou, China, 215006  
Email: {zhangsk, szzmcai, shrgong, jxfan}@suda.edu.cn

**Abstract**—Data aggregation is an efficient primitive in wireless sensor network (WSN) applications, it can reduce the communication cost, thereby extending the lifetime of sensor networks. The goal of our work is to design techniques and algorithm that lead to efficient data aggregation without explicit maintenance of a structure. As packets need to converge spatially and temporally for data aggregation, an efficient approximation algorithm is proposed to achieve the following goals: monitoring data of any portion of the region can be obtained at one time by querying the root instead of flooding those regions, thus incurring significant energy savings. Using simulations experiments, we study the performance and potential of data aggregation.

**Index Terms**— wireless sensor network, in-network process, data aggregation, efficient approximation.

## I. INTRODUCTION

In sensor networks, the communication cost is often several orders of magnitude higher than the computation cost. For optimizing the communication cost, in-network data aggregation is considered an effective technique. The inherent redundancy in raw data collected from the sensors can often be eliminated by in-network data aggregation. In addition, such operations are also useful for extracting application specific information from raw data. To conserve energy for a longer network lifetime, it is critical for the network to support high incidence of in-network data aggregation. The existing researches have analyzed the aggregation algorithm for the application of the sensor network[1],[2],[3]. Taking the memory access of aggregation algorithm and other factors into account, an optimized compression algorithm[4] was proposed, but there is no consideration of data multiple-hop transmission. Data compression algorithms based on wavelet transforming for sensor networks were proposed in Literature[5]. It can reduce the energy cost of nodes in data transferring efficiently for sensor networks, so, it can prolong the lifetime of the whole networks to a greater degree. But it had not considered the algorithm processing energy consumption and multi-hop path. Literatures [6], [7] consider the energy optimization separately from the angle of the path transmission quality and the path energy consumption to extend the lifetime, but they have not considered the data aggregation.

Analysis of data aggregation algorithm indicates that<sup>[1]</sup> seeking for the optimal aggregation tree on the condition of complete aggregation equates to solving NP-Complete problem of the minimum Steiner tree. According to this NP-Complete problem, literature [8] has considered the balance of the computation processing energy consumption and the transmission energy consumption, and the case without complete aggregation, but it does not involve the overall multi-hop energy consumption and by this constructing aggregation tree. Moreover, considering the computation of measure was done by the Sink node. Literature [9] proposed the shortest path tree algorithm, and in this algorithm each source node transmits data along the shortest path to the gathering node. If these paths overlap with each other, and carry on data aggregation alternately in the overlap section, this algorithm is less complex and with less delay of network time. But its energy saving effect was greatly affected by the network topology and cannot win great satisfaction in most cases.

Through constructing the coverage like the Voronoi, and choosing the appropriate quantity and the position to optimize the data aggregation, it may reduce the data quantity transmitting to the Sink node<sup>[10],[11],[12]</sup>. Literature [10] proposed the algorithm of the greedy aggregation tree. Its shortest path was constructed between the first source node the Sink node arrives and the nearest source node of the tree afterward. However, using this method, Sink can not learn the sensed value but through high cost flooding in the given scope, and once the gradient vector was established, it will not change in the implementation. In LEACH[12], a node set was chosen according to clusters, which clusters each node will join to rely on the node and the clusters communication cost. However, as only a very few nodes act as the role of clusters, from which the appropriate Sink node is far away, clusters will consume excessive energies as a result of the transmission data to the base station. In literature [11], a boundary node possibly belongs to more than one voronoi unit; in this case, if Sink sends out the related data inquiry in the interest region, if necessary, this boundary node must route enquiry request, which will form the bottleneck.

The distributional nucleus regression<sup>[13]</sup> share similar aspects with this paper's algorithm, but there are great differences. As for the former, every node has its approximate coefficient in its local region scope, thus it

\* Corresponding author: S.-K Zhang, Tel.:+86 0512 65241247,  
E-mail address:zhangsk@suda.edu.cn, Postal address: No 1  
Shizhi Street Suzhou China 215006.

cannot reply to the inquiry correctly which involves outside its local region. But in the algorithm of this paper, the coefficient is transmitted upward after the child node is compressed and aggregated. Therefore, the root node of splay tree obtains the final data set of approximate coefficient about its entire covered region. Now, Sink obtains monitor value of any interested region position through the direct inquiry to the root node, and each node sends information containing a vector, which is used to describe the coverage of its local area; the size of this vector increases with massive neighbor nodes which share the nuclear variable along with it. However, in this paper, in view of data aggregation algorithm of event triggering driving based on splay tree, the quantity of the data packet which transmits through each node is constant, and the demand about the node function is simple. It works only if we guarantee that the node can correspond with the constant power in a small scope, and it has certain memory function. There is no need to add the special function node in the sensor network.

## II. DATA AGGREGATIONS

The main idea of based on tree data aggregation algorithm is to use the transmission model  $M$  being able to fit more monitor data instead of the monitor data of transmission nodes, to reduce the capacity of data transmission, thus saves the energy of the sensor node. Therefore it needs to consider the relations between the cost of the return model and data quantity it may fit. The smaller the cost of transmission model, the more data it can express, the more energy-saving. Because the monitor value of node is often subject to many factors, we expect to fit the most data with the minimum cost mode, and the multiple linear regression models is exactly in line with this goal.

In splay tree each node receives and stores data reported by the recent non-tree node cyclically to it, namely the  $NT$ (Non Tree) node is responsible for the sensation, but  $AT$  node is responsible to store, here, the value saved in  $AT$  node is regarded as the function value of the  $x$ - $y$  coordinate. This process describes by three-tuple  $(f, x, y)$ , i.e.  $f$  is the attribute value transmitted by node located at  $(x, y)$ . Data tuple of node  $i$  stored in  $AT$  (Aggregation Tree) produces the approach function  $f_i(x, y)$ , and the progressive function  $f(x, y)$  by the input of the three variables  $(z, x, y)$  forms the implementation of multi-polynomial functions, data in such tree node may denotes by multiple regression polynomial function. The following is to discuss the process of carrying out the data aggregation through the polynomial regression on the splay tree.

In general, the form of multi-dimensional linear regression function is as follows<sup>[16]</sup>:

$$y = f(x_1, x_2, \dots, x_m) = a_0 + \sum_{k=1}^m a_k x_k \quad (1)$$

Where  $x_1, x_2, \dots, x_m$  is independent variable of the forecast model,  $y$  is the sample value with  $n$  dimensional vector, which denotes from specific level  $x_k$  to estimated the value of node  $a$ , and  $\vec{a}$  is  $(m+1) \times l$  dimensional vector estimated value of  $a$ . Using the least square criterion, causes the quadratic difference to be smallest.

$$F(\vec{a}) = (X\vec{a} - \vec{y})^T (X\vec{a} - \vec{y}) \quad (2)$$

$$\text{where, } a = [a_1, a_2, \dots, a_n]^T, \vec{y} = (y_n, y_n, \dots, y_n)^T \quad (3)$$

The essential condition of existence minimum is the  $F(\vec{a})$  partial derivative is zero, then

$$\nabla_{\vec{a}} F(\vec{a}) = \nabla_{\vec{a}} (X\vec{a} - \vec{y})^T (X\vec{a} - \vec{y}) = \vec{0} \quad (4)$$

Again

$$\begin{aligned} & \nabla_{\vec{a}} (X\vec{a} - \vec{y})^T (X\vec{a} - \vec{y}) \\ &= (\nabla_{\vec{a}} (X\vec{a} - \vec{y}))^T (X\vec{a} - \vec{y}) + (\nabla_{\vec{a}} (X\vec{a} - \vec{y}))^T (X\vec{a} - \vec{y}) \\ &= 2X^T (X\vec{a} - \vec{y}) = 2X^T X\vec{a} - 2X^T \vec{y} = 0 \end{aligned} \quad (5)$$

$$\text{Then: } X^T X\vec{a} = X^T \vec{y} \quad (6)$$

If  $X^T X$  is irreversible, there is a solution. In the equation (11) on both sides is multiplied by  $(X^T X)^{-1}$ , has  $a = (X^T X)^{-1} X^T \vec{y}$ . (7)

Using the polynomial regression, we can obtain the following equation.

$$X = \begin{pmatrix} 1 & y_1 & y_1^2 & x_1 & x_1 y_1 & x_1 y_1^2 & x_1^2 & x_1^2 y_1 & x_1^2 y_1^2 \\ 1 & y_2 & y_2^2 & x_2 & x_2 y_2 & x_2 y_2^2 & x_2^2 & x_2^2 y_2 & x_2^2 y_2^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & y_n & y_n^2 & x_n & x_n y_n & x_n y_n^2 & x_n^2 & x_n^2 y_n & x_n^2 y_n^2 \end{pmatrix}$$

$$Z = [z_1, z_2, \dots, z_n]^T, \beta = [\beta_1, \beta_2, \dots, \beta_n]^T \quad (8)$$

$$\text{where, } \vec{b} = (X^T X)^{-1} X^T \vec{Y} \quad (9)$$

$$\begin{aligned} p(x, y) &= \beta_0 + \beta_1 y + \beta_2 y^2 + \beta_3 x + \beta_4 xy \\ &+ \beta_5 xy^2 + \beta_6 x^2 + \beta_7 x^2 y + \beta_8 x^2 y^2 \end{aligned} \quad (10)$$

From the equation (10), we can compute  $\vec{\beta}$  with a given location  $(x, y)$ , and obtain the value of  $z = p(x, y)$  is property value of  $(x, y)$  nodes. Set  $\vec{\beta}$  is  $(m+1) \times 1$  - vector, then  $X^T X$  certainly is the  $m+1$ [14] step non-singular. In other words,  $n \gg m+1$  and  $X$  cannot denote for weighted linear combination of any other row set. In this paper, the data aggregation algorithm according to the input of the width priority, each tree node has a coefficient from the formula (10) and sends the coefficient set to its parent node. Nodes of each level use the coefficient which obtains from its child to renew sensor attribute value, and these data combine the detection value of node itself to calculate the new coefficient set, and then transfer to a higher level. In the process, to identify the even attribute value in the region is the crux of the matter, because they have a direct bearing on the accuracy of the aggregation, and through the upper and lower bounds the of coordinates of the  $\{x_{min}, y_{min}, x_{max}, y_{max}\}$  to identify of the region, where the minimum and maximum value from the Son of the father of the current node in the tree under all the sensor nodes.

As in Fig. 1, set  $a$  as the current aggregation node, and data value in the region updates through node  $a$ . The scope of the region defines by the subtree of node  $a$  to which passed through the smallest and largest coordinates

of the sensor nodes. Thus node  $a$  gets the border coordinates of the region from its child. Through based on construction of the splay tree and the above

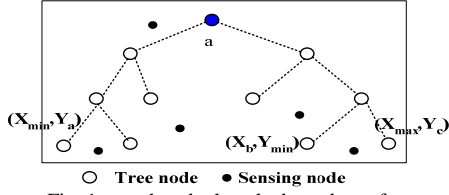


Fig. 1 a node calculate the boundary of the region for data regeneration

description about the process of regression, answer queries every specified time.

Such as "SELECT temperature FROM sensors WHERE location = (x, y)", or "the highest temperature in target scope "of issues. In the latter case, generate set of (x, y) coordinates in the designated area, Sink firstly informed of the attribute value of each point location to calculate the maximum value. When Sink needs to know the data of (x, y), it will send this inquiry to the root, the inquiry by the AT spreading down until the leaf nodes of the last layer. A data aggregation algorithm based on splay tree  $SPAT(p, n_s)$ . where  $n_s$  is the average number of sensor nodes reporting to the tree node.

- 
- 1: Begin
  - 2: For each of leaf node  $i$  of the tree  
file node " $i$ " dat is read  
multivariate polynomial regression is performed on each data file and the coefficients are stored in the each of the arrays  $\beta_0, \beta_1, \dots, \beta_8$  each of size  $N$   
End For
  - 3: Initialize level to  $2^p$   
While  $p$  is greater than 0  
sum = level +  $2^{p-1}$ ;  $k$  = level  
While  $k$  < sum
  - 4: for each of the non leaf nodes  $k$  of the tree computes random x-y points for each of its 2 children  $i$  and  $(i+1)$  where  $(x_{min}, y_{min})$  and  $(x_{max}, y_{max})$  are the coordinates of the leftmost and down most node and rightmost and top most node respectively reporting to the node  $i$  and  $i+1$ .  
End For
  - 5: Using  $(\beta_{i0}, \beta_{i1}, \dots, \beta_{i8})$  and  $(\beta_{(i+1)0}, \beta_{(i+1)1}, \dots, \beta_{(i+1)8})$  new attribute values are calculated and appended to node " $k$ ". node  $k$  then calls the regression function to calculate  $(\beta_{k0}, \dots, \beta_{k8})$  and passes it to its parent.
  - 6: End While
  - 7: Level = sum
  - 8: End While
  - 9: End
- 

### III. SIMULATION RESULTS AND DISCUSSION

In this paper, discrete event simulation platform  $NS2$  was employed to conduct simulation tests; the simulation parameters are shown in table 1 and the focus is to make performance evaluation of the data aggregation algorithm in the following aspects: (1)compression ratio (2) comparison between the sizes of the aggregated packets and non-aggregated packets in the root.

The definition of the variables was shown in table 1. Supposing the total number of node in region  $A$  is  $D$ , then the density of nodes  $\rho=D/A$ ,  $A_s$  is the sub-region including the single aggregation tree  $T_c$ , so the average number  $U$

of node is determined by  $A_s$  in the sub-region. For complete binary tree, the total number of nodes is  $t$ , has

$$t = 2^{(p+1)} - 1 \quad (11)$$

Table 1 Values of simulation parameters used

| Parameter | Variation        |
|-----------|------------------|
| $A$       | $800 \times 800$ |
| $R$       | 40m              |
| $D$       | 1630             |
| $A/D$     | 0.0025           |
| $A_s$     | $400 \times 400$ |
| $P$       | 0.33             |
| $p$       | 4                |
| $n_s$     | 12               |

In addition,  $n_s$  set by the front, that is, the average number of sensor nodes reporting to the tree node, in the sub-region the upper bound of node number  $U$ [15] is:

$U = n_s \times t + t$  or  $t = U / (n_s + 1)$ , replacement of  $t$  using formula (15), results in an optimal solution with the depth of AT:  $p = \ln(\frac{U}{n_s + 1} + 1) - 1$ , Set  $D=1630$ ,  $A=800 \times 800$ .

$\rho = 1630 / 800^2 = 0.0025$ ,  $A_s = 400 \times 400$ . Then the average number of nodes in the region is  $U = 0.002469 \times 400 + 400 = 408$ . set the depth of the tree  $p=4$ , nodes number  $T_c = t = 2^{(4+1)} - 1 = 31$ ,  $n_s = 12$ .  $\therefore$  The total number of sensor nodes in this region  $= 31 \times 12 = 372$ . In fact, the total number of nodes in the region  $U = 31 + 372 = 403 < 408$ . Therefore, the definition of parameters is effective.

If we place about 400 sensor nodes in the square of  $400 \times 400$  units, changes in the temperature  $39^\circ\text{C} - 49^\circ\text{C}$  will be considered as abnormal, which might indicates that a fire occurs in the region. The following is the results of an aggregation, the premise is the assumption that inquiry has been extended to all the leaf nodes of the splay tree.

#### (1). Compression Ratio

Fig. 2 shows the compression ratio changes with the depth of the tree, as expected, almost constant for 0.02. The decline of the curve shows that with the depth increases, the compression ratio reduces. The deeper the depth of the tree is, the better the degree of compression turns, and the less the output becomes. The high compression ratio reduced the whole information content and thus has saved the correspondence band width and the total energy.

#### (2). Size of Root Output Packet

Fig. 3 shows the data traffic after data aggregation though  $SPAT$  algorithm in the root, as well as the root receives data non-aggregation and the normal one-to-many communication. Though the depth of the splay tree is different, when all the nodes of the splay tree implement data compression, the size of packet sent from root to the Sink is a constant, that is, fixed  $(w_x + w_y + w_c)$  bytes. The packet size sent from one tree to another tree node by node is almost constant, and it is nothing to do with the size of the network, which

makes the total energy of data kept within reasonable bounds. This confirms the above assumptions, that is, each tree node sends the packet which only contains coefficients and  $x$ - $y$  coordinates to its parent node, and the size of the packet is independent of the number of nodes in the tree. In the traditional many-to-one communications, all of the leaves must send data to the root node, so that when the network node increases, the size of data packet transmission growth through the root node with no limits. Through the implementation of this algorithm for data compression, the largest Data communications reduce by the amount of 85% in comparison with [16].

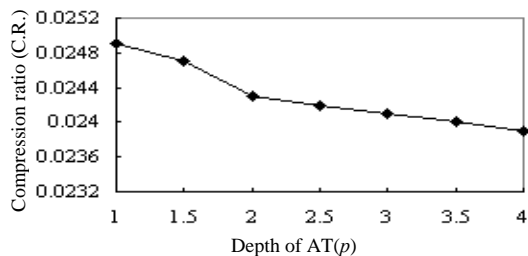


Fig. 2. Dependence of compression ratio

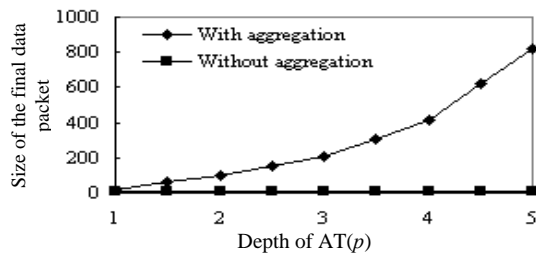


Fig. 3 Dependence of size of data packet (at root node)

#### IV. CONCLUSION

In this paper, we proposed a novel data aggregation algorithm through the construction of the splay tree, and this algorithm will also be able to detect the event attribute value in the positions where the sensor nodes are lacked. In the construction phase of the tree, the root choice is distributional. It eliminated the request for the overall situation root positional information by Sink. By limiting the number of communications, fixed-size information and without taking the depth of the aggregation tree into account, its percentage of error can be controlled within an acceptable range when data compression ratio remains constant. Simulation results show that the algorithm can effectively improve the perception capacity of the overall network and reduce the energy consumption.

#### ACKNOWLEDGMENT

Supported by the the National Natural Science oundation of China (60673092,60873116,60873047); Natural Science Foundation of Jiangsu Province of China(BK2008161, BK2008154); The Key Programs of Ministry of Education of China (207040);Jiangsu Provincial Major Program of Science Technique support and independent innovation Foundation (BE2008044); Funded by Preresearch Project of Soochow University; The Opening Project of JiangSu Province Support

Software Engineering R&D Center for Modern Information Technology Application in Enterprise (SX200903);The Higher Education Graduate Research Innovation Program of Jiangsu Province in 2009.

#### REFERENCES

- [1] Li JZ, Li JB, Shi SF. Concepts, issues and advance of sensor networks and data management of sensor networks. *Journal of Software*,2003,14(10):1717-1727.
- [2] Ren FY, Huang HN, Lin C. Wireless sensor networks. *Journal of Software*,2003,14(7):1282-1291.
- [3] Naoto Kimura,Shahram Latifi.A survey on data compression in wireless sensor networks. In proc of the Int'l conf on Information Technology: Coding and Computing. Los CA:2005,2:8-13
- [4] Kenneth B,krste A.Energy aware lossless data compression, *ACM TransonCputer Systems*,2006,24(3):250-291
- [5] XIE Zhi-Jun , WANG Lei , LIN Ya-Ping ,et al.An Algorithm of Data Aggregation Based on Data Compression for Sensor Networks, *Journal of Software*,2006,17(4):860-867.
- [6] Olga Saukh,Pedro Jose,Andreas Lachenmann,et al.Generic routing metric and policies for WSNs, In Proc of 3th European Workshop on wireless sensor networks. Berlin,2006,99-114
- [7] Noseong Park, Daeyoung Kim, Yoonmee Doh, et al. an optimal and lightweight routing for minimum energy consumption in wireless sensor networks, In Proc of the 11th IEEE Int'l Conf on Embedded and Real\_time Computing Systems and applications. New York,2005,1533-2306
- [8] I Kadayif, M Kandemir. Tuning in-sensor data filtering to reduce energy consumption in wireless sensor networks, In Proc of Design, Automation And Test In Europe Conference And Exhibition. Los Alamitos, CA, 2004, 1530-1591
- [9] Dantu, R. Abbas K O'Neill, et al. Data Centric Modeling of Environmental Sensor Networks, *Global Telecommunications Conference Workshops, GlobeCom Workshops 2004*. 447-452
- [10] Intanagoniwiwat, Estrin, Govindan. Impact of network density on data aggregation in wireless sensor networks. Proc Of the 22th International Conference on Distributed Computing Systems, July 2002, 575-578.
- [11] Henry Dubois Ferriere, Deborah Estrin. Efficient and Practical Query Scoping in Sensor Networks, *IEEE International Conference on Mobile Ad-hoc and Sensor Systems*, Los Angeles, USA, April 2004. 564-566
- [12] W. Heinzelman, A. Chandrakasan, H Balakrishnan. Energy-Efficient Communication Protocol for Wireless Microsensor Networks, *Proc of the 33th International Conference on System Sciences*, Hawaii, January 2000
- [13] Guestrin C, Bodik P, Thibaux R, et al. Distributed Regression: an Efficient Framework for Modeling Sensor Network Data, .3th International Symposium on Information Processing in Sensor Networks (IPSN' 04). New York, 2004. 1-10
- [14] Ignacio Solis, Katia Obraczka, In-Network Aggregation Trade-offs for Data Collection in Wireless Sensor Networks, *INRG Technical Report 102*, 2003
- [15] Vuran MC, Akan OB, Akyildiz IF. Spatio-Temporal correlation: Theory and application for wireless sensor networks. *Computer Networks*, 2004, 45: 245-259.
- [16] Tilman Wolf, Sumi Y. Choi. Aggregated Hierarchical Multicast for Active Networks, *IEEE Military Communications Conference*, 2001, 2: 899-904