Polynomial Regression for Data Gathering in Environmental Monitoring Applications

Guojun Wang[1,3], Jiannong Cao[2,*], Huan Wang[2,3], Minyi Guo[1,4]

[1] School of Computer Science and Engineering, University of Aizu, Aizu-Wakamatsu City, Fukushima 965-8580, Japan
[2] Department of Computing, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
[3] School of Information Science and Engineering, Central South University, Changsha, Hunan Province, 410083, China
[4] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200030, China

Abstract - **How to prolong the lifetime of wireless sensor networks is one of the most important design issues. In order to tackle this issue, we propose an energy-efficient polynomial regression-based data gathering algorithm in environmental monitoring applications. Each sensor node in the network fits a regression function with its sensed data in most recent rounds, and sends coefficients of the regression function and some related parameters to the sink node instead of sending the sensed data. Theoretical analysis and simulation studies show that the proposed algorithm can greatly reduce data transmissions among the sensor nodes, with significant energy savings on the sensor nodes and thus extending lifetime of the entire network.**

Keywords - wireless sensor networks; data gathering; regression; energy efficiency; environmental monitoring

I. INTRODUCTION

Recent advances in wireless communications, electronics, micro-sensor and new battery technologies enable the development of small, low-cost sensors with sensing, computation and wireless communication capabilities [1]. However, in many situations such as hostile or hazardous environments, recharging battery of sensor nodes is too expensive or even impossible. Therefore, the data gathering schemes in Wireless Sensor Networks (WSNs) must be energy efficient in order to prolong lifetime of the entire network.

Taking the characteristics of WSNs into consideration, many data gathering schemes are proposed in the literature. To save energy of sensor nodes, those schemes utilize various routing techniques [5,6,7,8,10,11] to select energy-saving paths for forwarding data packets, and data aggregation techniques [2,3,4,9] for reducing spatial-temporal correlation of the sensed data. To further reduce the communications between sensor nodes and sink node, we propose an energy-efficient dual prediction-based data gathering protocol for environmental monitoring applications in [12], which avoids most data transmissions by adopting dual prediction both at the sensor node and the sink node.

LEACH [6] consists of two phases, i.e., setup phase and steady state phase. At the setup phase, the clusters are organized autonomously and the Cluster Heads (CHs) are selected randomly. At the steady state phase, the sensor nodes sense the surrounding environments and transmit the sensed data to the CHs. After receiving all the data, each CH aggregates the data before sending the data to the sink node. After *n* rounds of data gathering, the network goes back to the setup phase and enters another round of selecting new CHs.

In the distributed kernel regression [3], the sensing field of a sensor network is divided into several sub-regions according to spatial correlation, and sensor nodes of each sub-region collaborate to optimally fit a global function to each of their local measurements. The measurements of any location in each sub-region can be decided by the function, and sensor nodes in each sub-region update the base function with new measurements. The sensed data in the sensing field are assumed to be spatially correlated, and each area with spatial correlation is treated as a sub-region. Although spatial correlation areas do exist, it is difficult to differentiate them from each other. In fact, for most applications, it is impossible to pre-configure spatial correlation areas. So, it is impractical to construct the structure in most applications.

DUMMYREG [2] also divides the sensing field into several sub-regions, and each sub-region corresponds to an aggregation tree. There are two types of sensor nodes, i.e., sensing nodes and tree nodes. The sensing nodes report the sensed data to the tree nodes closest to them, and then each tree node calls the regression function and obtains the coefficients which are then passed to the higher level instead of sending the sensed data. Thus sensor nodes at each level use the coefficients of their children to improve the approximation function and this procedure stops at the root. The sink node has access to an approximation of the sensed data at any point in the region spanned by the tree. DUMMYREG aggregates data based on spatial correlation, but our scheme aggregates data based on temporal correlation.

The regression algorithms in the above two papers have the following limitations. Both algorithms have to construct and maintain an assistant structure on top of the routing protocols, resulting in large overhead. Both algorithms use multi-variant polynomial regression, and the computation is proportional to the density of sensor nodes. In our scheme, sensor nodes only transmit data after certain gathering rounds, and the

[*] Prof. Jiannong Cao is with the Department of Computing, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, Phone/Fax: (852) 27667275/27740842, Email: csjcao@comp.polyu.edu.hk.

1930-529X/07/$25.00 © 2007 IEEE
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2007 proceedings.

computation of sensor nodes is constant irrespective of the density of sensor nodes. Since we adopt single-variant polynomial regression in our scheme, the computation is less than the above algorithms.

The rest of the paper is organized as follows. In Section II, an energy-efficient regression-based data gathering algorithm is presented. We make theoretical analysis and simulation studies compared with existing algorithms in Section III. Finally, we conclude the paper in Section IV.

II. REGRESSION-BASED DATA GATHERING ALGORITHM

The basic idea of Energy-efficient polynomial Regression-based Data gathering algorithm (ERD) works as follows: each sensor node senses its surrounding environment and stores the sensed data in its buffer. After certain rounds of sensing, it fits a regression function with the data in its buffer, and sends coefficients of the regression function and some related parameters to the sink node instead of sending all the sensed data. It clears its buffer after above operations and begins a new regression loop. The sink node regenerates the sensed data using the coefficients of the regression function and some related parameters.

A. Basic Assumptions

We assume that each sensor node is assigned a unique identification and all the sensor nodes are randomly deployed in the network. And we assume that the sink node has unlimited computation and communication capabilities.

B. Polynomial Regression Model

We assume to use an m-degree polynomial

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m \qquad (1)$$

to approximate a given set of data, (x_1, y_1), (x_2, y_2), … , (x_n, y_n), where $n > m$. The best fitting curve *f(x)* of the set of data has the least square error, that is,

$$\Pi = \sum_{i=1}^{n} [y_i - f(x_i)]^2 = \sum_{i=1}^{n} [y_i - (a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_m x_i^m)]^2 = \min \quad (2)$$

Note that a_0, a_1, a_2, …, and a_m are unknown coefficients while all x_i and y_i are given by the sensed data. To obtain the least square error, the unknown coefficients a_0, a_1, a_2, …, and a_m must yield zero first derivatives:

$$\frac{\partial \Pi}{\partial a_0} = 2\sum_{i=1}^{n} [y_i - (a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_m x^m)] = 0$$
$$\frac{\partial \Pi}{\partial a_1} = 2\sum_{i=1}^{n} x_i [y_i - (a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_m x^m)] = 0$$
$$\frac{\partial \Pi}{\partial a_2} = 2\sum_{i=1}^{n} x_i^2 [y_i - (a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_m x^m)] = 0 \qquad (3)$$
$$\vdots$$
$$\frac{\partial \Pi}{\partial a_m} = 2\sum_{i=1}^{n} x_i^m [y_i - (a_0 + a_1 x_i + a_2 x_i^2 + \cdots + a_m x^m)] = 0$$

The unknown coefficients a_0, a_1, a_2, … , and a_m can thus be obtained by solving the above linear equations.

C. Algorithm Descriptions

In the proposed algorithm, suppose sensor nodes can buffer n rounds of the sensed data, where n can be adjusted according to the requirements of applications and the capability of sensor nodes. We define the duration of n rounds of data gathering as a *Regression Round*. Each sensor node senses its surrounding environment and stores its sensed data in its buffer. At the end of each regression round, it fits a regression function with the data in the buffer, and generates an m-degree polynomial $f(x)$ according to the methods mentioned in the above section. It then sends the coefficients of the regression function and some related parameters to the sink node. It then clears its buffer and prepares for the next regression round. The proposed algorithm is shown in Fig. 1.

```
ERD_Algorithm()
{
  initialize node;
  $i$ = 0;
  While (node is active)
  {
    While ($i < n$)
    {
      node senses environment;
      node stores sensed data ($x_i$, $y_i$) into buffer;
      delay (interval);
    }
  run polynomial regression on the data set of buffer;
  send coefficients of polynomial and related parameters to sink node;
  clear buffer;
  $i$ = 0;
  }
}
```

Figure 1. The ERD algorithm

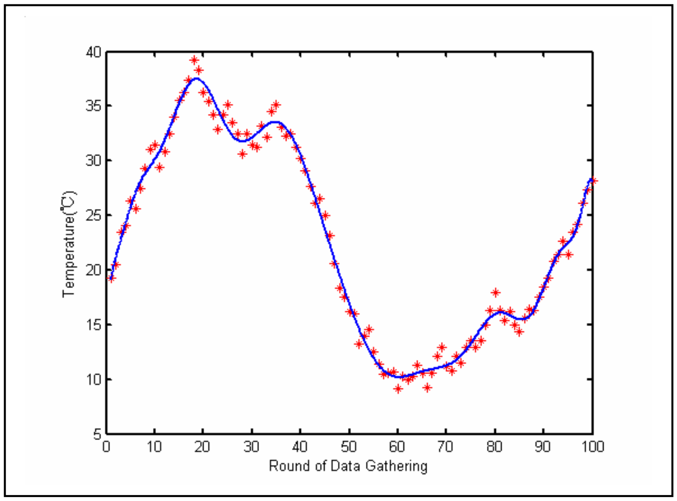

Figure 2. The curve of regression on the sink node

The sink node regenerates the m-degree polynomial according to the coefficients of the regression function and other related parameters, and then obtains approximation of the sensed data. There is an example in Fig. 2 in which the sink node regenerates a curve according to the coefficients and related parameters sent by a sensor node. The asterisks

1930-529X/07/$25.00 © 2007 IEEE

represent the data sensed by the sensor node, and the curve is regenerated by the sink node.

III. Performance Analysis & Simulation Studies

A. Performance Metrics

We use three metrics to analyze and compare the performance of the proposed algorithm as follows:

Average Traffic: This metric shows the average number of data transmissions per node during a regression round. According to First Order Radio Model in [6], energy consumption is proportional to the number of communications. The energy consumption by computation and sensing can be negligible compared with energy consumption by communications, so we don't consider the energy consumption by computation and sensing in order to simplify the model.

Mean Square Error (MSE): To evaluate precision of the regression function, we define MSE as follows:

$$MSE = \frac{1}{n}\sqrt{\sum_{i=1}^{n}(f(x_i)-y_i)^2} \qquad (5)$$

$f(x_i)$ and y_i represent the value of regression function and the sensed data at the time x_i respectively, and n is the number of data gathering rounds in each regression round. When the approximation of the sensed data generated by the polynomial approaches the sensed data, the MSE will become small, and vice versa.

Max Error (ME): To obtain the bound of regression error, we define the maximum difference between the sensed data and the data generated by the regression polynomial within a regression round as Max Error.

$$ME = \max\{|f(x_i)-y_i|\} \quad (0 < i \le n) \qquad (6)$$

Here $f(x_i)$, y_i and n are the same with those in (5). The ME indicates how much the error has happened during a regression round. The ME is meaningful for those applications which are sensitive to maximum error, because such applications can adopt appropriate parameters of regression polynomial according to the ME which they can tolerate.

B. Performance Analysis

In this section, we compare average traffic of ERD, LEACH, and DUMMYREG during a regression round. Sensor nodes send data to the sink node once for each regression round in ERD, but sensor nodes have to send data to the sink node every data gathering round in LEACH. Suppose the number of coefficients of the regression polynomial and related parameters is m, and each coefficient is s bytes. Each sensor node sends at least two parameters (sensed data and sensing time) to the sink node every round in LEACH, so the total transmissions over n rounds in LEACH are as follows:

$$T_{LEACH} = 2\cdot s\cdot n \qquad (7)$$

The total transmissions in ERD are as follows:

$$T_{ERD} = m\cdot s \qquad (8)$$

Hence, the ratio of average traffic between ERD and LEACH is as follows:

$$R_{EL} = \frac{m\cdot s}{2\cdot s\cdot n} = \frac{m}{2\cdot n} \qquad (9)$$

Table I presents the ratio of average traffic between ERD and LEACH (R_{EL}) under different regression round and different degree of the regression polynomial. The R_{EL} is proportional to the degree of the regression polynomial, and is inverse proportional to the regression round. When the degree of regression polynomial and the regression round are set to 10 and 120 respectively, and R_{EL} is only 4.17%. R_{EL} is 37.50%, even though the degree of regression polynomial rises to 30 and regression round decreases to 40.

TABLE I. THE RATIO OF AVERAGE TRAFFIC BETWEEN ERD AND LEACH

R_{EL}	m=10	m=20	m=30
n=40	12.50%	25.00%	37.50%
n=80	6.25%	12.50%	18.75%
n=120	4.17%	8.33%	12.50%

In order to compare ERD with DUMMYREG, we assume ERD adopts the same topology as DUMMYREG. The monitoring field consists of several sub-regions, where each sub-region corresponds to a complete binary aggregation tree. We only compare ERD with DUMMYREG under complete binary aggregation tree, because the performance of DUMMYREG with complete binary tree is the best among topologies proposed in [2]. Assuming that p is the depth of complete binary aggregation tree, and t is the number of sensor nodes in the aggregation tree. n_s is the average number of the sensing nodes reporting to each tree node, and k is the number of coefficients and related parameters of the regression polynomial. The average traffic of DUMMYREG in a sub-region during n rounds is as follows:

$$\begin{aligned} T_{DUMMYREG} &= (t\cdot n_s\cdot 2s + t\cdot k\cdot s)\cdot n \\ &= (2\cdot n_s + k)\cdot t\cdot s\cdot n \\ &= (2n_s + k)(2^{p+1}-1)\cdot s\cdot n \end{aligned} \qquad (10)$$

The average traffic of ERD during n rounds is shown as follows:

$$\begin{aligned} T_{ERD} &= [t\cdot n_s\cdot m\cdot s + \sum_{i=0}^{p}(n_s+1)\cdot(i+1)\cdot 2^i\cdot m\cdot s] \\ &= m\cdot s\cdot[t\cdot n_s + \sum_{i=0}^{p}(n_s+1)\cdot(i+1)\cdot 2^i] \\ &= m\cdot s\cdot[(2^{p+1}-1)\cdot(2n_s+1) + (n_s+1)\sum_{i=0}^{p} i\cdot 2^i] \end{aligned} \qquad (11)$$

So, the ratio of average traffic between ERD and DUMMYREG is as follows:

$$R_{ED} = \frac{T_{ERD}}{T_{DUMMYREG}} = \frac{m\cdot[(2^{p+1}-1)\cdot(2n_s+1)+(n_s+1)\sum_{i=0}^{p} i\cdot 2^i]}{(2n_s+k)(2^{p+1}-1)\cdot n} \qquad (12)$$

Assuming that n_s=12 and p=4, Table II presents the ratio of average traffic between ERD and DUMMYREG (R_{ED}) under different regression round and different degree of regression polynomial. Table II indicates that R_{ED} is proportional to the degree of the regression polynomial and inverse proportional to the regression round. In most cases, traffic of ERD is less than DUMMYREG. When the degree of the regression polynomial and the regression round are set to be 10 and 120 respectively, R_{ED} is only 16.20%. When ERD adopts high degree polynomial and short regression round, the transmissions in ERD will be more than DUMMYREG. From simulation results in the next sub-section, we can also deduce that it is unnecessary to adopt high degree polynomial and short regression round.

TABLE II. THE RATIO OF AVERAGE TRAFFIC BETWEEN ERD AND DUMMYREG

R_{ED}	m=10	m=20	m=30
n=40	48.60%	75.11%	91.80%
n=80	24.30%	37.55%	45.90%
n=120	16.20%	25.04%	30.60%

C. Simulation Results

We evaluate the factors that affect the performance of ERD in this sub-section, which include the degree of the regression polynomial and the regression round. Assume that trends of the sensed data are continuous in most cases. We can find many natural phenomena which conform to this assumption, for example, the temperature and humidity change continuously in natural environments.

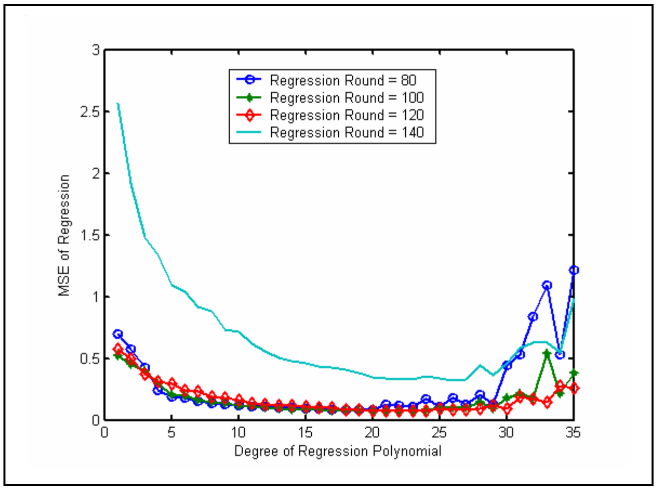

Figure 3. The trends of MSE of regression over the degree of regression polynomial

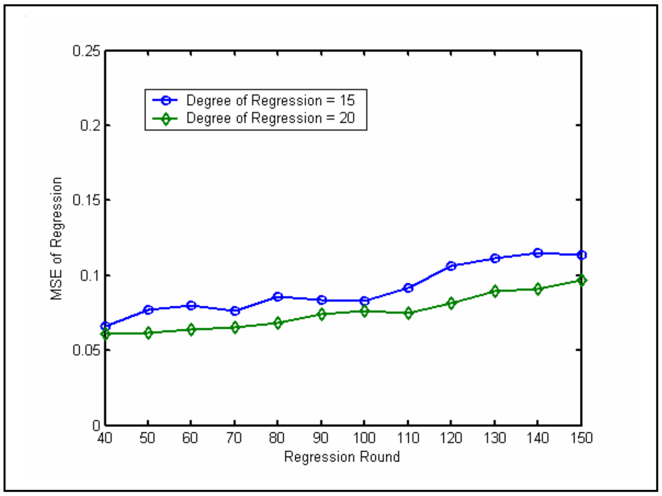

Figure 4. The trends of MSE of regression over the regression round

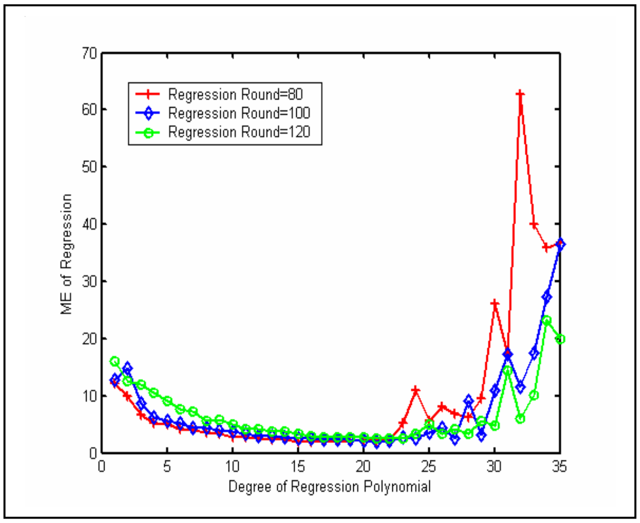

Figure 5. The ME of regression over the degree of regression polynomial

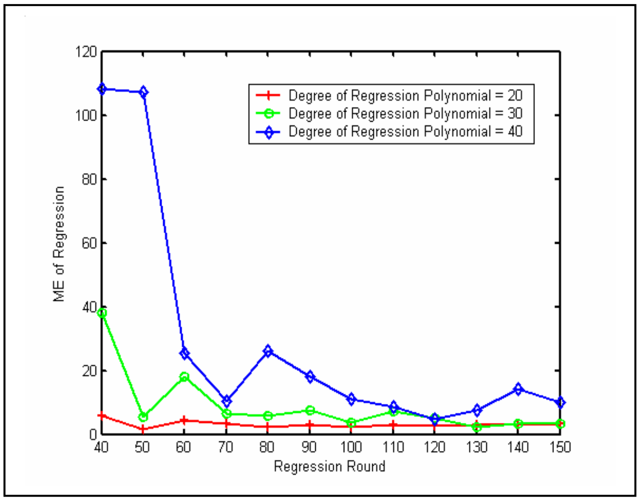

Figure 6. The ME of regression over the degree of regression polynomial

1930-529X/07/$25.00 © 2007 IEEE

This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE GLOBECOM 2007 proceedings.

Based on extensive simulation studies, we observe that the degree of the regression polynomial has significant impact on the performance of ERD. When the degree of the regression polynomial is small, the sensor node can reduce data transmissions. However, if the degree of regression polynomial is small, it is difficult to describe the trends of the sensed data since the MSE will be high. The MSE will decrease when the degree of the regression polynomial increases, but the MSE will not decrease all the time. Interestingly, when the degree of the regression polynomial increases to a certain threshold, the MSE will then increase. The relationship between the degree of the regression polynomial and the MSE with regression round n=80, n=100, n=120, n=140 are shown in Fig. 3. When the degree of the regression polynomial is between 15 and 30, the MSE decreases to minimum. The MSE will increase rapidly when the degree of the regression polynomial exceeds 35, so the MSE is not proportional to the degree of regression polynomial. The choice of the degree of the regression polynomial for a specific application is dependent on the regression and its tolerance of errors.

When the degree of the regression polynomial is determined, the regression round also has impact on the precision of approximation of the sensed data. Since each sensor node has to send coefficients and related parameters to the sink node in each regression round, the regression round is proportional to computation, and is inverse proportional to communication cost. Fig. 4 presents the MSE of the regression polynomial changing with regression round when degrees of the regression polynomial are set to 15 and 20 respectively. The MSE of the regression polynomial is non-sensitive to regression round, because the MSE keeps stable when regression round varies from 40 to 150 in Fig. 4.

Some applications not only require low average error between the sensed data and the regression data, but also need to restrict the max error to a certain range. So, we should evaluate the ME which is defined as the difference between the sensed data and the regression data. Fig. 5 shows the ME changing with degree of regression polynomial when the regression round is 80, 100, 120 respectively. We take sensing temperature as an example in Fig. 5 to show the relationship between the ME and the degree of regression polynomial. The ME decreases to minimum when the degree of regression polynomial varies from 15 to 25. The ME increases dramatically when the degree of regression polynomial increases. The curve of regression polynomial jitters dramatically at both ends of the curve with very high ME.

We investigate how the regression round and the degree of regression polynomial influence the ME in Fig. 6. When the regression polynomial adopts short regression round, the ME will reach high value that cannot be tolerated by most applications. That is because, when the regression round is small compared with the degree of regression polynomial, the curve of regression will jitter dramatically at both ends, which results in large difference between sensed data and regression data, i.e., large ME shown in this figure. The ME will decrease to acceptable value when the regression round becomes 100 or more. Although long regression round can achieve low ME and save more energy, the regression round can not be too large because of resource constrains of sensor nodes, such as buffer, computation, etc. As shown in Fig. 6, regression round can be adopted as long as possible if the sensor node can deal with all the data in time. When the regression round becomes 100 or more, the ME keeps stable in Fig. 6.

IV. Conclusion

We proposed a novel energy-efficient data gathering algorithm based on polynomial regression model. The proposed protocol can greatly reduce the number of data communications in the network by transmitting parameters of the regression polynomial instead of the sensed data, and thus prolong the lifetime of the entire network.

Acknowledgment

This work is supported by the Hong Kong Polytechnic University under the ICRG Grant G-YE57 and A-PF77, the National Natural Science Foundation of China under Grant Nos. 60740440032 and 60533040, and the Program for New Century Excellent Talents in University under Grant Number NCET-06-0686.

References

[1] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, “A survey on sensor networks,” IEEE Communications Magazine, Vol. 40, Issue 8, pp. 102 - 114, 2002.

[2] T. Banerjee, K. Chowdhury, and D.P. Agrawal, “Distributed data aggregation in sensor networks by regression based compression,” Proceedings of MASS 2005, pp. 283 - 290, 2005.

[3] C. Guestrin, P. Bodik, R. Thibaux, M. Paskin, and S. Madden, “Distributed regression: an efficient framework for modeling sensor network data,” Proceedings of IPSN 2004, pp. 1 - 10, 2004.

[4] T. He, B.M. Blum, J.A. Stankovic, and T. Abdelzaher, “AIDA: Adaptive Application Independent Data Aggregation in wireless sensor networks,” ACM Transactions on Embedded Computing Systems, Vol. 3, Issue 2, pp. 426 - 457, 2004.

[5] W.R. Heinzelman, J. Kulik, and H. Balakrishnan, “Adaptive protocols for information dissemination in wireless sensor networks,” Proceedings of MobiCom 1999, pp. 174 - 185, 1999.

[6] W.R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, “Energy-efficient communication protocol for wireless microsensor networks,” Proceedings of HICSS 2000, pp. 1 - 10, 2000.

[7] C. Intanagonwiwat, R. Govindan, and D. Estrin, “Directed diffusion: a scalable and robust communication paradigm for sensor networks,” Proceedings of MobiCom 2000, pp. 56 - 67, 2000.

[8] J. Kulik, W. Heinzelman, and H. Balakrishnan, “Negotiation based protocols for disseminating information in wireless sensor networks,” ACM Wireless Networks, Vol. 8, pp. 169 - 185, 2002.

[9] B. Krishnamachari, D. Estrin, and S. Wicker, “Modelling data-centric routing in wireless sensor networks,” Proceedings of INFOCOM 2002, pp. 42 - 49, 2002.

[10] A. Manjeshwar, and D.P. Agrawal, “TEEN: a routing protocol for enhanced efficiency in wireless sensor networks,” Proceedings of IPDPS Workshops 2001, pp. 35 - 43, 2001.

[11] G. Wang, T. Wang, W. Jia, M. Guo, H.-H. Chen, and M. Guizani, “Local update-based routing protocol in wireless sensor networks with mobile sinks,” Proceedings of ICC 2007, June 2007, Glasgow, Scotland, UK

[12] G. Wang, H. Wang, J. Cao, and M. Guo, “Energy-efficient dual prediction-based data gathering for environmental monitoring applications,” Proceedings of WCNC 2007, pp. 3516 - 3521, 2007.

1930-529X/07/$25.00 © 2007 IEEE