AN ESTIMATION ALGORITHM FOR MISSING DATA IN WIRELESS SENSOR NETWORKS

Nan Yan, Ming-zheng Zhou, Li Tong*
School of Computer and Information
Anhui Polytechnic University
24100, Wuhu, P.R.China
Emails :{ jiffyan; mzzhou; litong}@ahpu.edu.cn

***Submitted:* Sep. 19, 2012** **Accepted: May 7, 2013** **Published: June 5, 2013**

Abstract- For estimating the missing data of wireless sensor networks, an estimation algorithm called HD method, which can make use of sensoring space-time correlation of the data, was proposed based on mathematical Hermite and DESM statistical models. The algorithm not only can adaptively adjust the time and space weights, but also can accurately estimate the missing or unavailable data. The experimental results show that the algorithm has good stability and relatively high estimation accuracy.

Index terms: Wireless sensor network, missing data, mathematical model, estimation algorithm, space-time correlation

I. INTRODUCTION

In recent years, as a typical application of pervasive computing ideas, wireless sensor network (WSN) combines the logical information world with the objective physical one to address the data sources through pouring more data into the information world. The key challenge is to minimize energy consumption and handling of missing data to extend the network lifetime and improve network reliability.

With the development of wireless sensor technology and the continuous generation of large quantities of data, the uncertainty prevalently exists since the external environment, characteristics and sources of data, etc, are more diverse. So the inevitability of data uncertainty brings invisible barriers to the wide range of WSN applications. In main data collection and applications, the problems of loss of data has been much concerned and how to best deal with missing data has become a current focus of research, hence the best way is to accurately estimate the missing data.

From the current technologies, the data uncertainty can be broken down as the uncertainty of tuples and their attribute values[1], therefore, in this paper, for the uncertainty brought about by non-existing tuples, the missing data can be estimated by taking advantage of historical data stored in WSN nodes to improve the reliability of the wireless network.

The causes of the missing data are more complex, which are probably due to inaccurate raw data or the adopting coarse-grained data collection, or arise in dealing with missing values and the data integration process in order to meet specific application purposes.

(1) The original data is not accurate. This is the most direct factor of generating uncertain data. Firstly, the instrument's accuracy constraints the accuracy of the data collected by the physical instruments. Secondly, in the process (especially wireless network transmission) of the network transmission, the accuracy of the data is subject to factors such as bandwidth, transmission delay, energy. In addition, in the sensor network and the RFID applications, the surrounding environment will also affect the accuracy of the original data.

(2) Usage of coarse-grained data collection. Obviously, the conversion process from the data collection of the coarse-grained to fine-grained will introduce uncertainty. For example, supposing that a population distribution database records nationwide population with township-based units, a certain application needs an inquiry with the village-based units, the query should result in uncertainty.

(3) Meeting specific application purposes. For privacy and other special purposes, some applications can only be able to get the conversion inaccurate data other than the original precise data.

(4) Dealing with missing values. A lot of causes such as equipment failures, unable to obtain information, inconsistent with other fields, historical reasons can produce the missing values. A typical processing method is the interpolation and the data can be seen as to obey specific probability distribution after interpolation. In addition, one can also delete all records with missing values, however, this operation also vary to some extent the distribution characteristics of the original data.

(5)Data integration. Information of different sources may be inconsistent and it will introduce uncertainty in the data integration process. For example, Web contains a lot of information; however, the content of many pages is not consistent due to factors such as page updates.

With the widespread applications of WSN networks and the increasing amount of data, a variety of applications query algorithms have been also proposed [2],[3],[4],[5]. By detecting outliers of local sensor networks, Yozo Hida improved fault-tolerant ability of the data set query to error reading of nodes or node failures. Nevertheless, there are a few researches focusing on estimation of missing data in WSN by present.

M. Halatchev [6] once proposed a WARM (Window Association Rule Mining) algorithm dealing with data of associated nodes as estimates of the missing values with data mining techniques. However, this algorithm can only deal with discrete data instead of continuously changing data. Y. Li minimized energy consumption at the cost of estimation accuracy of the missing values to reach an aim, with which data evaluation model was established with the minimum data as possible [7]. In addition, Pan [8] employed methods of piecewise low-order interpolation and multiple linear regressions to estimate the missing values. From this point, we put up a new algorithm model to smoothen the curve by derivation, with which a good stability and relatively high estimation accuracy can be achieved to better meet the industrial requirements.

In this paper, section II describes the related works and presents a new estimation algorithm of missing data based on different assumptions. Section III discusses the principle of missing data estimating and gives the models and algorithms. Section IV discusses the simulation and experimental results, section V discusses the field study and the last part gives a summary of this paper.

II. RELATED WORK

In WSN, the node itself is fine small and has functions of storage, communications; when the network is unstable or the node energy is low, data loss or damage may occur resulting in incomplete data. Because the uncertainties of data tuples are dependent on two factors of time and space, the Hermite mathematical model was firstly adopted to consider the estimation algorithm of missing data associated with time [9]. Since the monitoring is usually continuous, i.e. it is time-correlated; this model can reach stationary estimation of missing data so that the smoothness of function curves is better.

In recent years, the fault-tolerant of region detection gradually become a hot topic in wireless sensor networks. Among them, Krishnamachari B [10] firstly proposed a distributed local algorithm for event region detection, Bayesian Fault Recognition Algorithms (BFRA). In this BFRA, the event is assumed to be space-related yet the error is irrelevant with space and the probability of error of each sensor is the same. The sensor reading is 0 or 1 and when the output is 1, it is determined that an event or an error has occurred.

As a local algorithm, BFRA only needs each sensor exchanging reading with their neighbors (the neighbor is defined as all the sensors located in the radius of radio communication of the specific sensor) to obtain the statistical probability of detected events of all neighbors. BFRA takes advantage of this statistical probability to describe the spatial correlation of the events, to judge conditional probability of an event together with sensor error probability. Finally, based on the sensor readings and Bayesian analysis, the final decision on whether the error or event can be deduced.

Based on the works of Krishnamachari B, Chen Q. [11] corrected some errors in theoretic analysis and Luo X. [12] improved the algorithm with consideration of the error due to sensor errors and errors caused by the event determination and considered how to choose the number of neighbors in order to achieve the purpose of fault tolerance while reducing data exchange. All the above methods are based on probability analysis need to assume that the error probabilities of each sensor are the same and need sensors to complex computing. Chen J. proposed a relatively simple non-probability detection algorithm through two rounds of voting and the subordination of the minority to the majority rule (majority-voting) to determine whether an error or event occurs or not.

Wireless sensor networks are self-organized with coordination skills and provide the end-users with intelligent and good understanding of the environment [6]. Literature [7] further elaborated intelligent wireless sensor network system and devoted to collaborative

information processing methods in multi-agent systems and synergistic approaches of wireless sensor networks based on based on a multi-agent theory.

Similarly, the uncertain data appear widely in many application fields such as sensor networks. Chen et al. also believed that the uncertainty exists in the changing of sensoring data [11] and gave a frame to describe the uncertainty of the sensing data, with which different levels of uncertainty qualitatively and quantitatively represented data query were assessed and the balance among the data uncertainty, query accuracy and computational cost was discussed.

Through multiple-redundant sensors and fusion algorithm of different data, Lap lante [4] alleviated the sensing uncertainty. Based on the intrinsic association between WSN and MAS as well as the treatment of RST uncertainty, Dai Zhi-feng et al. explored new ways of the integration of WSN, MAS and RST to a certain extent and built a distributed reconcile framework of the adaptive sensing uncertainty treatment.

In fact, the uncertainty and certainty can transform into each other to some extent, a certain level of uncertainty may be higher levels of certainty and these uncertainties may hide certain regularity [1].

On the basis of the related researches, for the data management of Wireless sensor networks especially the validity strategies of uncertainty treatment, people should not only focus on the uncertainty of many dispersed, incomplete, inconsistent or unreliable sensor data with noise but the different forms of uncertainty from the perspective of a higher level.

By exploring the inherent relationship, the combined effects and layering and hierarchical coordination of different types of uncertainty implied in an uncertain situation, it is needed to expand the potential and space of uncertainty intelligent processing studies of WSN, MAS and RST by cross-integration and seek the certainty from uncertain sensoring data and the coordination and unity from conflicts and discords so that new models and methods can be established by acquiring effective decision-making information achieved by data fusion from the local to the global.

This article focuses on the fault-tolerant algorithm of how to estimate the missing data in the area, That is, by considering the sensors how to consume as small as possible resources (computing and communications), we can achieve the detection of the area when the portions of the sensor error occurs. As described above, the existing related works are based on the assumption that the event is the spatial correlation however the error is space-irrelevant and the fault tolerance is carried out with the information redundancy in the space of a sensor network, but it requires frequent exchange between the sensor readings.

However, in reality, the event is not only a spatial correlation but correlation with time. In other words, the event will continue for some time and its characteristics have some statistical characteristics variation over time. Therefore, the property of data characteristic varying over time can be described with a random process the exchange of data between the sensors can be reduced by using the information redundancy in the time of the data collected by the sensors for fault tolerance [13].

In addition, all the previous works assumed that the probability of error in each sensor was the same. But in reality, there are many factors causing the sensor error probability not the same:

(1) The error probability of different batches or different factory-made sensors is often different;

(2) In the sensor network deployment process, the sensors may be affected by the environment related damage, leading to different sensor error probability. For example, different sensors in the sensor network dispensed by aircraft, different ground conditions will produce different degrees of damage.

(3) The error probability of sensors may be affected by the environmental changes. Among them, some factors will lead to the probability of sensor error occurs before the actual deployment, which cannot be informed [14].

Except for strong time correlation, the collected data also has spatial correlation, thus DESM (Data Estimation using Statistical Model) model [3] against spatial correlation was secondly taken to estimate the real-time data together with the historical data from nodes having missing data between neighboring nodes.

Finally, an HD (Hermite and DESM) algorithm sensoring data space-time correlation was proposed based on Hermite and DESM model, this model can adaptively adjust the parameters of weights in estimating equation according to characteristics of the collected changing data to obtain better estimate results.

III. MODEL DEFINITION & ALGORITHM DESIGNING

This article assumes that a large-scale wireless sensor network evenly covers an interested area for detecting the missing data within this area and make estimates. The network has n sensors, each sensor Si can perceive their logical location through topology relation xi, $1 \leq i \leq n$. Among the sensors, synchronous data sampling and news communication are ensured through the bottom of mechanism.

The sensor is sampling once every ΔT time. When the readings of sensor Si exceed the threshold Rth, it indicates that error occurs at Xi or Si. The Rth value is usually based on the formula (1):

$$R_{th}(t) = \frac{Exp_n(t) + Exp_e(t)}{2} \tag{1}$$

Where Expn (t) is the expectation function of the readings of the correct sensor in the normal area. In a relatively stable environment (such as temperature in the natural environment), Expn (t) is a constant. Expe(t) is the expectation function of the readings of the correct sensor in the event area and t is the sampling time of the sensor. Because the characteristics of the event can only be perceived after the event, and the event characteristics variation is dependent on the time interval from the moment of the event other than the specific time when the error occurred, therefore, if an event occurs at time t0 and the event (features) can last at least Tth time, the expectation and variance functions meet the Expe(t) = Expe(t-t0) = Expe(τ), Vare(t)=Vare(t-t0)=Vare(τ), 0≤τ=t-t0≤Tth. Expe(0) is defined as the characteristics of expectations when the event just occurs and Rth (0) as a threshold value to determine whether an event occurs. Under normal circumstances, the greater the Tth is, the more accurate of the event judgment.

When the periodic readings of the sensors seen as a sequence {r(t)} arranged according to the temporal order, the sequence essentially is sample values of the event process. Therefore, the theoretical study of the sequence is based on the corresponding random process. Although the finite dimensional distribution function family can completely characterize the statistical properties of the random process, it is often very difficult to determine the family of finite dimensional distribution of the random process in practical problems. Therefore, the researchers often turn to study some important digital features of the random process, such as expectation and variance functions.

Taking into account that the sensors only have limited computing and storage capacity, therefore only the square root of the expectation and variance function values numbered by m=Tth/ΔT can be saved within the Tth time on the sensors. For simple calculation, it is assumed that ΔT can be divisible by Tth and the error process is a second-order moment, that is, expectation and variance functions exist in the event procedure. Sensor networks are commonly used to detect specific errors, so we can assume that the expectation function Expe(t) and the square root of variance function VarSRe(t) can be stored in the sensor memory before deployment or the Expe(t) and VarSRe(t) can be distributed to various sensors through the meeting points (sink node) by message after deployment[17-23].

The detection principle of missing data based on statistical hypothesis testing is as follows:

How to identify events: sensor detection of missing data is based on statistical hypothesis testing. If the sensor reading sequence {r(tj)} is over the threshold Rth(0) of the event occurring and the numbers of {R(tj)} satisfying the Equation (2) exceeds C (the threshold of statistical hypothesis testing method) in the after Tth time, then {r(tj)} satisfy the hypothesis testing conditions. Equation (2) gives the degree of correlation between the statistical characteristics of the event process and the sequence of sensor readings, which is called hypothesis testing conditions. After the Tth time, if the majority of all the neighboring sensors also detects that the event occurs (because the event is spatial correlation), then it is considered that event occurs. It can be seen from the following theoretical analysis that the selection of δ and C can have effects on the algorithm.

$$\frac{|r(t_j) - Exp_e(t_j)|}{VarSR_e(t_j)} < \delta \tag{2}$$

However, the above detection mechanism is not applicable for the sensors located in the boundary area of the event because such neighboring sensors always inconsistently detect the events. For the uniform distribution, since the number of sensors located in the boundaries of the event area much is smaller than that of sensors in the inner of the event area, the impact of event detection of the sensors located in the boundary area can be ignored under normal circumstances.

How to identify misjudgment of missing data: If the sensor reading sequence {r(tj)} exceeds the threshold value Rth (0) of the event but it does not meet hypothesis testing conditions, or {r(tj)} meets the conditions but the majority of neighboring sensors do not detected the event occurring after Tth time, then the missing data of the sensor occurs.

Hermite model as known to all, although many scientific computing problems can be expressed with a function or functions, most of which cannot be directly calculated by computers and need to be replaced by approximate functions implemented on the computers to fulfill numerical calculations. Methods of interpolation and the least squares based on data-fitting are fundamental and effective solutions for approximation replacements and calculations. The following gives the problem definition, marks and some constraints.

Assuming the data sets collected from nodes of the wireless sensor network as time-series are expressed as $H = < x_0, y_0 >, < x_1, y_1 >, ... < x_n, y_n >$, among which y_n is the collected data at the time of x_n . Then this model presents the function $f(x_i)$ and its first derivative

$f'(x_i) = y_i'$ from the difference values between two points with derivatives by interpolation, which requires interpolation polynomial H (x) satisfy:

$$\begin{cases} H(x_i) = f(x_i) & (i = 0,1,.....,n) \\ H'(x_i) = f'(x_i) = y_i' & (i = 0,1.....,r)(r \le n) \end{cases} \tag{3}$$

As polynomial interpolation, cubic interpolation has been high enough and the Runge phenomenon, poor approximation results in minor intervals generated by high order interpolation within equidistant nodes may occur if the number is higher. Therefore, when the number of nodes is large and the convergence and stability of high-order interpolation polynomials cannot be guaranteed, it is needed to employ the piecewise interpolation method.

Lemma: the piecewise cubic functions $H_h(x)$ satisfy:

(1) $H_h(x)$ is the cubic polynomial in each sub-interval $[x_i, x_{i+1}]$;

(2) $H_h(x)$ is first-order continuously differentiable in $[x_0, x_n]$;

(3) $H_h(x_i) = y_i, H'_h(x_i) = y_i'$.

Then the piecewise cubic function Hh(x) is expressed as:

$$H_h(x) = \sum_{k=0}^{n} (y_k \alpha_k(x) + y_k' \beta_k(x)) \tag{4}$$

Where interpolation basic functions $\alpha_k(x)$、$\beta_k(x)$、(0 1, …,n) do not zero in the sub-interval $[x_{k-1}, x_{k+1}]$ and zero in the other intervals.

$$\alpha_k(x) = \begin{cases} \left(1 - 2\frac{x - x_{k-1}}{x_k - x_{k-1}}\right)\left(\frac{x - x_{k-1}}{x_k - x_{k-1}}\right)^2, x \in [x_{k-1}, x_k], k \ne 0, \\ \left(1 - 2\frac{x - x_{k+1}}{x_k - x_{k+1}}\right)\left(\frac{x - x_{k+1}}{x_k - x_{k+1}}\right)^2, x \in [x_k, x_{k+1}], k \ne n \end{cases} \tag{5}$$

$$\beta_k(x) = \begin{cases} x - x_k \left(\frac{x - x_{k-1}}{x_k - x_{k+1}}\right)^2, x \in [x_{k-1}, x_k], k \ne 0, \\ (x - x_k)\left(\frac{x - x_{k+1}}{x_k - x_{k+1}}\right)^2, x \in [x_k, x_{k+1}], k \ne n \end{cases}$$

a. DESM model

DESM model estimates the indetermination by examining the collected data of adjacent sensoring nodes. We assume that the data of node xi is lost and this node has n neighboring

nodes so that the spatial correlation of the node m and its perception nodes is the strongest for the layout of different positions.

The estimation value of this model is:

$$\hat{y}_{ik} = (1-\alpha)\hat{y}_{i(k-1)} + \alpha\hat{z} \tag{6}$$

$$\text{Where} \quad \hat{z} = y_{i(k-1)}(1 + \frac{y_{mk} - y_{m(k-1)}}{y_{m(k-1)}}) \tag{7}$$

$\hat{y}_{ik}$ is the estimate value of the perceived data of node i at time k, $\hat{y}_{i(k-1)}$ is the mathematical expectation of the collected data of node i before time k-1, $\hat{z}$ is the estimate value of the perceived data from node m to node i at time k, α is correlation coefficient between the two nodes, also known as weight.

$$\alpha = \varphi(y_m, y_i) = \frac{Cov(y_m, y_i)}{\sigma_{y_m}\sigma_{y_i}} \tag{8}$$

b. HD algorithm

Analysis of the above models shows that when the collected data is missing or unavailable, the data can be estimated through space-time correlation. For the intermittent loss of data, it can be estimated using Hermite model. When the lost data is more, we can use DESM model to estimate to meet the practical applications. However, the changes of data within a certain period of time cannot be predicted in practice, therefore it cannot be determined which model is more accurate.

By considering the time and space correlation, we propose HD algorithm, the main idea of which is that the default value of node xi at time k can be estimated by weighted calling the values estimated with Hermite and DESM models. So the HD algorithm can be defined as:

$$\hat{y}_{ik} = (1-\mu)\hat{y}_H + (\mu)\hat{y}_D$$

$$0 \le \mu \le 1 \quad (\text{Where } \mu \text{ is the weight}) \tag{9}$$

In the above formula, $\hat{y}_H$ and $\hat{y}_D$ are the two estimation values of the same node at the same time by Hermite and DESM models respectively. Taking into account the different physical environment, HD algorithm combines the results of the two models and can change the proportion of the estimation values of two models by adjusting the weights to improve the accuracy of the data estimation.

c. Choices of improving nearest neighbor nodes

As DESM model takes the spatial correlation into account, the key is how to choose the most relevant spatial neighbor nodes, that is to calculate $\min|\hat{y}_{ik} - y_{ik}|$. Because y_{ik} is a missing data, it is impossible to calculate the minimum targeted value. However, since the spatial change in WSN changes little in a short time, it can be considered to be relatively stable so the targeted value you can be approximated to the comparison between recent historical data, $\min|\hat{y}_{i(k-1)} - y_{i(k-1)}|$. Actually, the physical nodes are not too many laid out, therefore their nearest neighbor nodes can be determined by enumerating.

d. Weight adjustment

The adjustment algorithm of μ is selected as the basic algorithm of artificial intelligence - BP (Back propagation) algorithm, which composes of two processes of forward propagation and error back-propagation of the signals. The main purpose is to modify the weights of each model by sharing error between $\hat{y}_H$ and $\hat{y}_D$ via error back-propagation that is a weight adjustment process [15]. Generally speaking, the more samples, the results are more precise; so 10 samples are selected with the rules of thumb in this algorithm. Since the actual data cannot be perceived, sensoring the time correlation of data shows that the data will change not much within a short period so that the 10 samples have been obtained from a recent historical data.

e. Algorithm processing

The algorithm is divided into the following steps:

(1) To define the closest node with the enumeration method to determine the historical data involving calculation.
(2) To estimate the missing values with DESM and Hermite models respectively.
(3) To determine the value of μ by calling the BP algorithm to learn the estimated results.
(4) To get the final estimation results according to formula (9).

Parts of the codes are as follows:

```
y_1=interp1(x,y,xx,'cubic');
        //Calculating the estimation value with Hermite model
n=length(xx);
    for k=1:n   //Controlling the number of cycles with k
y_2(k)=DESM(x,y,yy,xx(k));
```

```
        //Calculating the estimation value with DESM model
    end
u=BP_weights(y_1,y_2,y);
        //Obtaining weights of the models
y_3=y_1*u+y_2*u;   //Achieving the final results
```

IV. SIMULATIONS AND RESULTS ANALYSIS

a. Data sets and experimental platform

The simulation data set for simulation was introduced from real monitoring data in Intel-Berkeley Laboratory [16] with MATLAB7.0 as the experimental platform.

To test the experimental results, we have estimated the non-missing data in the perceived data set and compared the estimates with its corresponding true value. Since the original data set may contain missing values, so a section of data with less missing values was firstly selected from the original data set in experiments. Then the missing values were replaced by the approximate values of the sensory data at the approximate time to form a full testing data set without missing values [4].

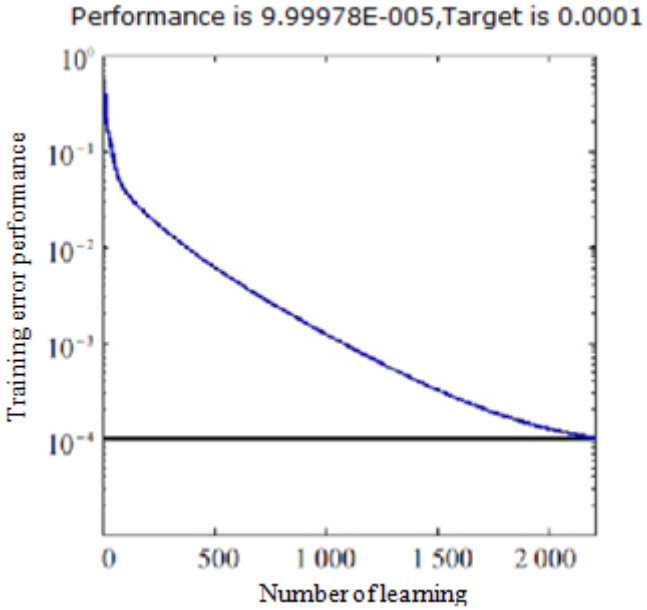

Figure 1. Training error curve

To acquire better convergence effect, the data used has been normalized by a processing principle of P= (P-Min Value)/ (Max Value – Min Value). This can effectively prevent the phenomenon from occurring that when the learning gradient reaches a minimum value to the

end yet the target error is not reached. As shown in figure 1, after learning and training for 2205 times, the error converges faster and the convergence effect is better.

Figure 2 is a snapshot of the simulation experiment. In this figure, each data point (x, y) ($1 \leq x, y \leq 32$) marked with different symbols represent each sensor and the symbol of black squares indicates the error ones. In this experiment, there are 103 error sensors accounting for about 10% of the total number of sensors. When defining the sensor error probability as the ratio of the error ones to the total sensors and defining the sensor error recognition rate as the ratio of the number of identified error sensors to all error ones, if the final state of the error sensor is Fault then the sensor identification is Error.

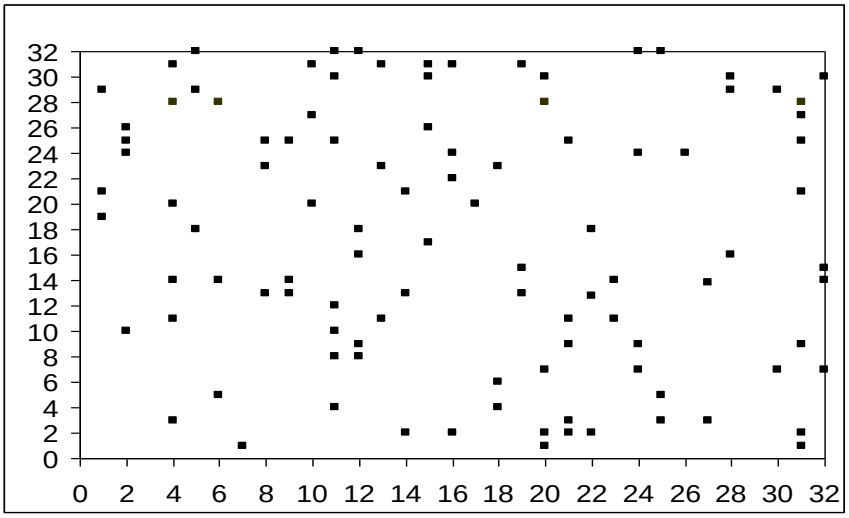

Figure 2. A snapshot of the simulation experiments

Assuming the events occur within the region in the (17, 5) to (27, 15) 11a × 11a. In simulation tests, if the final state of the sensor located in the event area is "Event", it is considered that the sensor detects the event. To define the detection probability as the ratio between the number of sensors detected and the total number of the sensors located in the event area, with which it can be estimated whether an algorithm is optimal or not.

As can be seen from figure 3, the detection probability of the event area approximatively decreases linearly with the growth of error probability of the sensors. When the sensor error probability is less than 10%, the best result can be achieved that the sensor network can detect 93% of the events in the event area.

As can be seen from figure 4, when the error probability of the sensors increases, the recognition error rate maintains at around 90% because the algorithm can recognize the errors by taking advantage of the time-correlation of the sensor readings. Therefore, even if the

number of error sensors in the sensor networks increases, the algorithm can still ensure that of the recognition error rate to be about 90%.

b. Simulation results and analysis

Simulation results of Hermite, DESM and HD Model are shown in figure 5, figure 6 and figure 7, respectively.

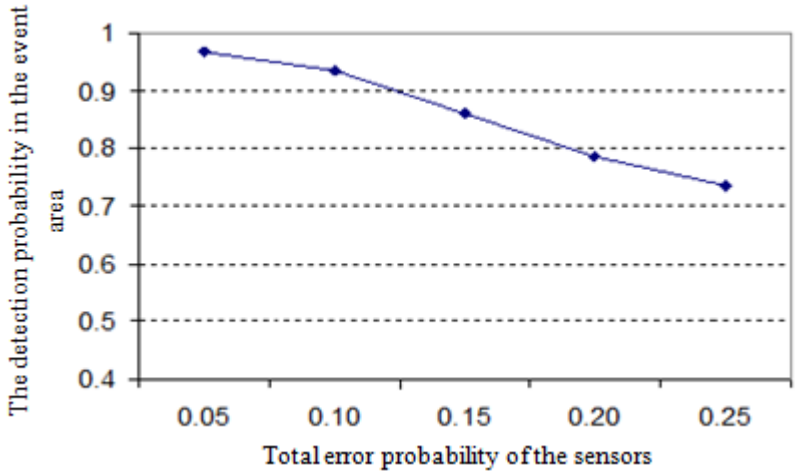

Figure 3.The variation of detection probability with the error probability of the sensors in the event area

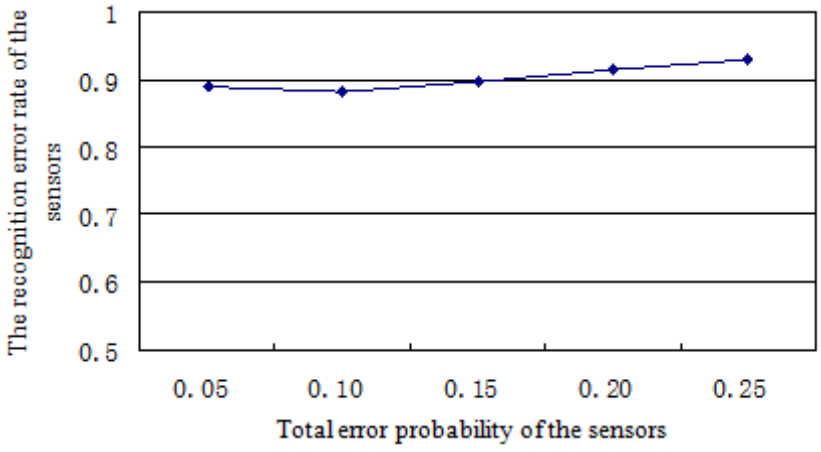

Figure 4.The variation of the recognition error rate with the error probability of the sensors

In order to reduce energy consumption in the WSN, the time interval is usually not too short. Figure 5 gives the actual data curve of a temperature sensor Sensor1 in a period of time and the curve of calculations resulting from Hermite model.

Hermite model estimates numbers of uncertainty depending on the node's own historical data, with the temporal correlations. From figure 5, it can be seen that the data curve estimated with this model is smooth, which reveals better estimate results.

Since more sensor nodes are laid out in the actual work environment, the uncertainty can be also estimated with the spatial correlation, namely the sensory data from neighbor nodes by making full use of the surrounding resources. When executing the corresponding numerical estimating.

DESM model chiefly takes advantage of the spatial correlations.

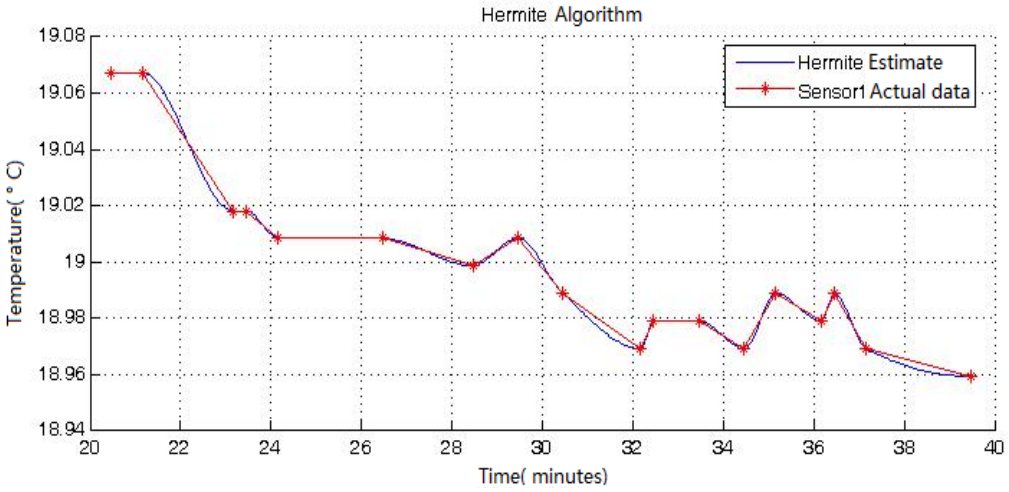

Figure 5. Simulation results of Hermite Model

Figure 6 shows the data curves of two temperature sensors Sensor1 and Sensor2 in the same period of time and the estimation results of Sensor1 with DESM model. Although it is not as smooth as that of Hermite model, the peripheral resources are fully made use of to estimate.

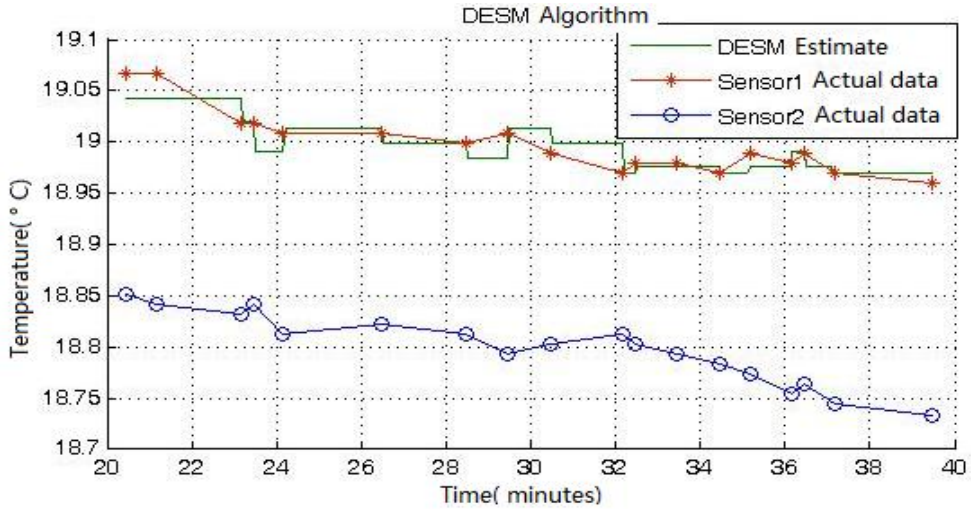

Figure 6. Simulation results of DESM Model

HD algorithm combines the advantages of above two models to make the most of time and space correlation and adjusts the values' proportion of each model through the weights. When the μ value is small, the estimate is close to that of the Hermite model and similar to the DESM model on the contrary. Therefore, the estimates of HD algorithm are always approximately optimal.

Figure 7 shows the data curves at the same time period of two temperature sensors marked as Sensor1 and Sensor2 together with the estimation results of HD algorithm for Sensor1. This algorithm is comparatively accurate to estimate the missing data over the nodes by using the time and space correlation. Simulation results also show that the algorithm is more effective for high-accuracy estimates and the curve is quite smooth to meet higher needs in some projects.

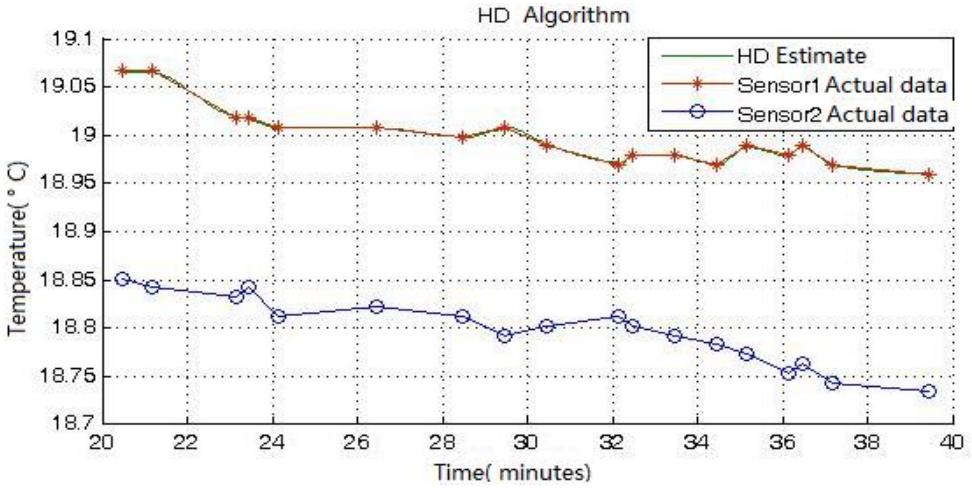

Figure 7. Simulation results of HD Model

V. FIELD STUDY

In this section, we deploy a real outdoor system to verify that method can be reliably applied to surveillance networks. We deploy 80 sensor nodes in a 75m ×20m forest. As illustrated in Figure 8, we put the nodes in a 5 ×16 manner, which enables us to easily locate the nodes and compute the links' length. Nodes' transmission power is set as 15 which guarantee each node has 10 neighbors on average [24-26].

a. Basic Observations

First we observe the distribution of faulty links. Due to hidden terminal and receiving queue overflow, a node with more neighbors is more likely to lose packets, thus degrades its links' performance. In this field study, the nodes are divided into 4 groups according to the number

of their neighbors, and we record each probe's sender ID to identify the sources of received probes for each node. Figure 9(a) shows the ratio for each group. For those nodes which have less than 9 neighbors, each of them loses 1.5 probes on average, while the nodes with more than 12 neighbors averagely lose 4 probes.

In addition, we compare every pair of nodes to find out the asymmetric pairs, i.e., only one of two nodes has received the other's probe. Similarly, we divide the nodes into 4 categories according to the number of their neighbors. Figure 9(b) compares the ground truth and our inference results.

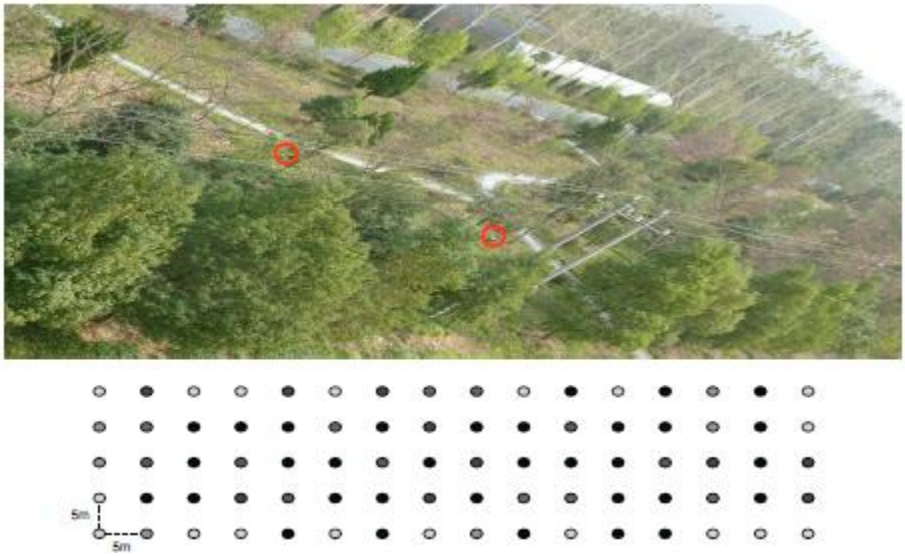

Figure 8. Field Topology

For those nodes which have less than 9 neighbors, each of them has 1.2 asymmetric links on average. For those nodes which have more than 12 neighbors, however, they averagely have 3.2 asymmetric links. In fault report, we set the prior knowledge about asymmetric link ratio is 20%, which means the inference program expects that around 20% links in the network are asymmetric. As we can see, our program has a better accuracy for the nodes with less neighbors, and it deduces that there exist average 1.3 and 3.7 asymmetric links respectively for the nodes with less than 9 neighbors and those with more than 12 neighbors.

b. Performance Evaluation

We also evaluate this method in terms of false negative rate and false positive rate. We try to add the basic observations into our inference model, to help the program specify different ratios of asymmetric links for different groups of nodes. Figure 10(a) compares the false negative rate before and after the learning. As we can see, for those nodes with less than 9

neighbors, the false negative rates are both 8.9%. In contrast, this method improves its accuracy for those nodes with more than 12 neighbors, i.e., detect the links of those nodes more accurately. Before learning, this method has a false negative rate as 17%, while it achieves to 15.4% after we modifying the program according to figure 9(b), i.e., adjust the prior knowledge about the number of asymmetric links for different group of nodes.

We observe similar improvement of false positive rate in figure 10(b). The evaluations before and after learning stay the same for the nodes with less than 9 neighbors, i.e., both are 7.4%. In contrast, the links of nodes which have more than 12 neighbors are more accurately detected.

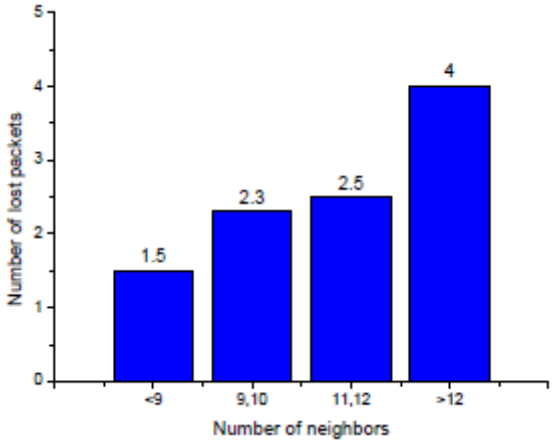

Figure 9(a). Correlation between the number of neighbors and that of lost packets.

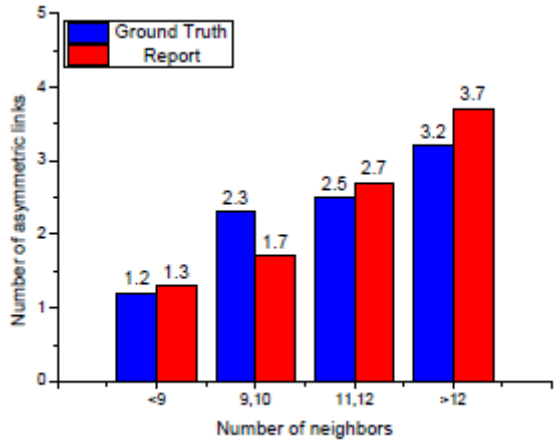

Figure 9(b). Correlation between the number of neighbors and that of asymmetric links.

Learning decreases the false positive rate from 15.5% to 13.2%. To explain the benefit generated by learning, we compare the faulty links in the fault reports of two inference

models. We find that the parts related to the links of nodes with less than 9 nodes are totally the same.

Figure 9(b) shows that the average number of asymmetric links of nodes which have less than 9 neighbors in ground truth is 1.2, while 1.3 in report, which means that tiny difference of knowledge about the number of asymmetric links couldn't change the detection significantly. In contrast, for those nodes with more than 12 neighbors, the numbers are respectively 3.2 and 3.7, thus the learning process improves the evaluation by adjusting the corresponding knowledge in the inference model.

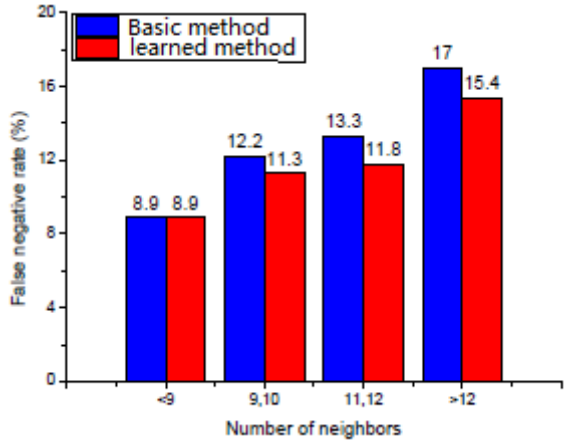

Figure 10 (a). Evaluation in terms of false negative rate

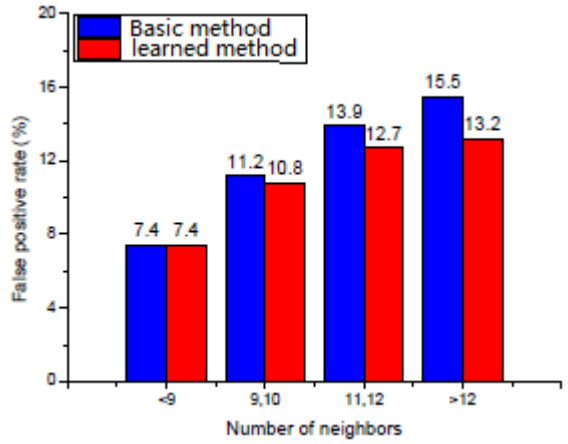

Figure 10(b). Evaluation in terms of false positive rate

VI. CONCLUSIONS

A wireless network often contains a large number of links which virtually exist in the air, but we can never directly observe whether they perform well or not. In WSN, the inevitability of

data uncertainty making data loss and its accurate estimation is of more special concern and interests.

In summary, most of the existing researches are based on spatial correlation of events to achieve fault tolerance with data exchange between adjacent sensors. However, this method will increase the amount of sensor network communication. In this paper, in order to extend the life of the sensor networks, it is considered that the sensor readings are not only space-correlated but time-correlated and the sensors may achieve fault-tolerance by making use of the time-correlation between local readings.

In this paper, an HD algorithm rising from time and space correlation of the data was put up and developed based on analyzing the mathematical Hermite and DESM models. This algorithm accurately estimated the missing data from the historical data of nodes with the most adjacent data in physical environment and the simulation results showed that a high estimating accuracy could be achieved for given testing data sets. Nevertheless, the HD algorithm achieves a high accuracy at the cost of computing time, how to enhance the efficiency needs to be gradually improved in future.

ACKNOWLEDGEMENT

This work was supported by the Natural Science Foundation of Anhui Province China under Grant (1308085MF88) and the Program of Educational Commission of Anhui Province (KJ2011B024) and the Program of Educational Commission of Anhui Province (KJ2012B012), which are gratefully acknowledged.

REFERENCES

[1] C.-Q. JIN. “urvey on the Management of Uncertain Data”, Communications of CCF, vol.5, no.4, pp.37-42, April 2009.

[2] A. F. Salami, H. Bello-Salau, F. Anwar1, A. M. Aibinu , “A Novel Biased Energy Distribution (BED) Technique for Cluster-Based Routing in Wireless Sensor Networks” , International Journal On Smart Sensing and Intelligent Systems, vol.4, no.2,pp.161-173,June 2011.

[3] J. Pei, M. Hua, Y. F. Tao, and X. M. Lin, “answering techniques on uncertain and probabilistic data: tutorial summary”, Proceedings of 2008 ACM SIGMOD International Conference on Management of Data, Vancouver, pp. 1357-1364, 10-12 June 2008.

[4] Kimelfeld B, Kosharovsky Y, Sagiv Y. Query, "efficiency in probabilistic XML models" Proceedings of the 2008 ACM SIGMOD international conference on Management of data. Vancouver, pp.701-714, 10-12 June 2008.

[5] Cormode G, Garofalakis M. "Sketching probabilistic data streams", Proceedings of the 2007 ACM SIGMOD international conference on Management of data. Beijing, pp.281-292, 11-14 June 2007.

[6] Ré C, Letchner J, Balazinska M, Suciu D. "Event queries on correlated probabilistic streams", Proceedings of the 27th ACM SIGMOD international conference on Management of data. Vancouver, pp.715-728, 10-12 June 2008.

[7] Y. Li, C. Ai, W. P. Deshmukh, and Y. Wu. "Data estimation in sensor networks using physical and statistical methodologies", Proceedings of the 28th IEEE International Conference on Distributed Computing Systems. Beijing, China, pp.538-545, 17-20 June 2008.

[8] L.-Q. Pan, J.-Z. Li, J.-Z. Luo, "A Temporal and Spatial Correlation Based Missing Values Imputation Algorithm in Wireless Sensor Networks", Chinese Journal of Computers, vol.33, no.1, pp.1-11, January 2010.

[9] F. Pei and X.-L. Han, "Thinning and Optimization of Cubic Hermit Curves", Computing Technology and Automation, vol.28, no.4, pp.68-71, April 2009.

[10]Krishnamachari B., Iyengar S. "Distributed Bayesian Algorithms for Fault-Tolerant Event Region Detection in Wireless Sensor Networks". IEEE Transactions on Computers, vol.53, no.3, pp.241-250, March 2004.

[11]Chen Q., Lam K. Y., Fan P. "Distributed bayesian algorithms for fault-tolerant event region detection in wireless sensor networks". IEEE Transactions on Computers, vol.54, no.9, pp.260-264, September 2005.

[12]Luo X., Dong M., Huang Y. "On Distributed Fault-Tolerant Detection in Wireless Sensor Networks". IEEE Transactions on Computers, vol.55, no.1, pp.58-69, January 2006.

[13]Ding M., Chen D. et al. "Localized Fault-Tolerant Event Boundary Detection in Sensor Networks". Annual IEEE Conference on Computer Communications (INFOCOM), Miami, pp.840-844, 13-15 March 2005.

[14]Ian F. A., Weilian S. et al. "A Survey on Sensor Networks". IEEE Communications Magazine, pp.102-114, August 2002.

[15] J.-P. Yang, Q. Li, Z. Liu, X.-L. Yuan, "Research of improved BP algorithm based on self-adaptive learning rate", Computer Engineering and Applications, vol.45, no.11, pp.56-58, November 2009.

[16] S. Madden, Intel Berkeley research lab data, http://berkeley.intelresearch.net/labdata, 2003.

[17]S.S. Ahuja, S. Ramasubramanian, and M.M. Krunz. "Single-link failure detection in all-optical networks using monitoring cycles and paths", IEEE/ACM Transactions on Networking, vol.17, no.4, pp.1080–1093, April 2009.

[18]R. Tan, G. Xing, Z. Yuan, X. Liu, and J. Yao. "System-level calibration for fusion based wireless sensor networks", In Proceedings of IEEE RTSS, San Diego, CA, USA, pp.143-231, 30 November- 3 December 2010.

[19] X. Wang, L. Fu, and C. Hu. "Multicast performance with hierarchical cooperation". IEEE/ACM Transactions on Networking, vol.20, no.3, pp.1-4, March 2011.

[20] Z. Li, M. Li, J. Wang, and Z. Cao. "Ubiquitous data collection for mobile users in wireless sensor networks", In Proceedings of IEEE INFOCOM, Shanghai, China. pp.230-234, 10-15 April 2011.

[21] K. Liu, Q. Ma, X. Zhao, and Y. Liu. "Self-diagnosis for large scale wireless sensor networks". In Proceedings of IEEE INFOCOM, Shanghai, China, pp.350-354, 10-15 April 2011.

[22] S. Liu, G. Xing, H. Zhang, J. Wang, J. Huang, M. Sha, and L. Huang. "Passive interference measurement in wireless sensor networks". In Proceedings of IEEE ICNP, Kyoto, Japan, pp.430-433, 5-8 October 2010.

[23] Y. Liu, K. Liu, and M. Li. "Passive diagnosis for wireless sensor networks". IEEE/ACM Transactions on Networking, vol.18, no.4, pp.1132–1144, April 2010.

[24] E. Magistretti, O. Gurewitz, and E. Knightly. "Inferring and mitigating a link's hindering transmissions in managed 802.11 wireless networks". In Proceedings of ACM MobiCom, Chicago, Illinois, USA. pp.1221-1224, 20-24 September 2010.

[25]A.-Y. Zhou, C.-Q. Jin, G.-R. Wang, and J.-Z. LI, "A Survey on the Management of Uncertain Data", Chinese Journal of Computers, vol. 32, no.1, pp.1-16, January 2009.

[26]Q. Ma, K. Liu, X. Miao, and Y. Liu. "Sherlock is around: Detecting network failures with local evidence fusion", In Proceedings of IEEE INFOCOM, Orlando, FL, USA, pp.320-329, 25-30 March 2012.