Online ISSN 1848-3380, Print ISSN 0005-1144
ATKAFF 54(4), 459–470(2013)

Xiao Wei, Qing Li, Feiyue Ye, Jun Zhang, Rongfang Bie

Building the Data Association Network of Sensors in the Internet of Things

DOI 10.7305/automatika.54-4.419
UDK 681.586.5:004.738.5; 621.396:004.738.5
IFAC 2.8.2; 4.0.5

Original scientific paper

In the Internet of Things, wireless sensor networks (WSN) is in charge of gathering and transferring environment data. It is an essential work to mine data semantic in WSN in the data derived from sensors to improve the WSN. This paper proposes the Data Association Network of sensors (DAN) to organize the mined association semantic relations among sensors into an effective form. Because DAN holds the rich data semantic of WSN, it can improve WSN in some aspects, such as detecting the abnormal sensors, simulating the data of faulty sensors, or optimizing the topology of WSN. Experimental results show that the proposed method can mine the associated relations among sensor nodes effectively, and the DAN is helpful in solving some problems of WSN.

Key words: Data Mining on IoT, Data Association of Sensors, Data Association Network of Sensors

Izgradnja podatkovne mreže senzora u Internetu stvari. Govoreći o Internetu stvari, bežična mreža senzora (WSN) ima ulogu prikupljanja i slanja podataka o okolini. Osnovni je zadatak analizirati semantiku podataka u WSN-u u podacima dobivenim sa senzora u svrhu unaprijeđenja bežične mreže senzora. U ovom radu predloženo je mrežno udruženje podataka (DAN) sa senzora u svrhu organiziranja analiziranih udruženja semantičkih relacija između senzora u djelotvorne forme. S obzirom da DAN sadrži dosta semantičkih podataka s WSN-a, može unaprijediti WSN u određenim aspektima kao npr. detekcija neispravnih senzora, simuliranje podataka sa senzora u kvaru ili optimiziranje topologije WSN-a. Eksperimentalni rezultati pokazuju da predložena metoda može efektivno analizirati udružene relacije između senzorskih čvorova te da je DAN korisno u rješavanju određenih problema WSN-a.

Ključne riječi: rudarenje podataka u internetu stvari, udruženje podataka senzora, mrežno udruženje podataka sa senzora

1 INTRODUCTION

The Internet of Things (IoT) has rapidly developed and changed people's life greatly recent years [1][2]. Radio Frequency Identification (RFID) & Wireless Sensor Network (WSN) are the most important technologies of IoT [3]. WSN consists of spatially distributed autonomous sensors to monitor physical or environmental conditions, such as temperature, sound, pressure, etc. and to cooperatively pass their data through the network to a main location [4].

Sensors in WSN belong to two types of relations: one is the adjacent relation of physical location, and the other is the topology relation of data transfer. In WSN, when sensors are fixed in location, adjacent relations among sensors are static. When sensors are mobile, adjacent relations among sensors are dynamic. Different data transfer protocols may make a sensor send its data to different target sensors, which leads to different topologies of WSN, varying from a simple star network to an advanced multi-hop wireless mesh network[5][6].

There are already many works on WSN [7], such as the organization of WSN, the data transfer protocol of WSN, and so on. These works mainly focus on the physical location relation or the data transfer protocol relation, which cannot solve some special problems in WSN. For example, when a sensor cannot work, a direct idea is to simulate its data with the data of its neighbors of physical or protocol location. However, two neighbors of physical location may have a big difference in data. In Figure 2, the sensor nodes S#8 and S#54 are very near in physical location, but the value of them are quite different. The reason is that S#8 and S#54 are deployed in two independent rooms; the environments of the two rooms are quite different which leads to the difference between the data of two sensors. So, two neighbor sensors in physical location cannot replace each other in some situations. Specially, in mobile WSN, adjacent relation between sensors is dynamic, and it is more

difficult to find neighbor sensors to simulate the sensor.

Therefore, we need to mine the association relation between sensors from sensor data. This relation is based on the statistics from the sensor data in WSN, which has no relevance with the physical location and protocol network. If two sensors have data relationships, they can replace each other in most cases.

There are some works concern analyzing data of sensors or mining the correlation from sensor data. However, these association relations are independent. If we could integrate all these associations, the more valuable semantic can be minded. We method is using Data Association Network of sensors (DAN) to organize the association semantic between sensors into an effective form. Because DAN holds the data semantic of WSN, it can improve WSN in some aspects. For example, we can find the abnormal sensor nodes according to the associated sensors; we can verify the data sent by a sensor according to its associated sensors' data; we can also reduce the redundant sensor nodes and optimize the deployment of sensors with the supporting of data association network.

So the mainly works of this paper are mining the data association relation from sensor data, building the data association network(DAN) for WSN, and exploring their help to WSN.

The data set used in this paper is the data collected from 54 sensors deployed in the Intel Berkeley Research lab. The data set consists of 2 million of records, the physical locations of all sensors, and the protocol connection of all sensors.

This paper is organized as follows. In section 2, the data association between sensors are defined and mined. In section 3, the Data Associated Network of sensors is defined and the building of DAN is discussed in detail. In section 4, three applications of DAN are shown to prove DAN is helpful to WSN. Some experimental results are presented in section 5. We introduce the related work in Section 6. Finally, we conclude in Section 7.

2 DATA ASSOCIATION BETWEEN SENSORS

In this section, we first analyze the different types of relations between a pair of sensor nodes in IoT. And then the definition of *Data Association between Sensors* is proposed. At last, how to mine the data association between sensors is discussed.

2.1 Relations between Sensors

a) Physical Location Relevancy (*PLR*)

The physical locations of sensors determine the relation between sensors. If two nodes are physical adjacent, they are called physical location relevancy.

For example, in Figure 2 the sensor nodes S#8 and S#54; S#8 and S#9; S#8 and S#4 are PLRs. According to the distance among them in Figure 2, we know that $PLR(\#8, \#9) \approx PLR(\#8, \#54)$ and $PLR(\#8, \#54) > PLR(\#8, \#4)$.

In general, physical neighbors are more likely to be connected in WSN and to gather similar data from environment. As well, it's not always true and there are certain exceptions to this case.

b) WSN Protocol Relevancy (*WPR*)

In WSN, based on its protocol, some sensors will connect and transfer data to each other. This connection between two sensor nodes is called protocol relevancy. In general, considering to the power of sensors, the neighbors of PLR are more likely to have the WSN protocol relevancy.

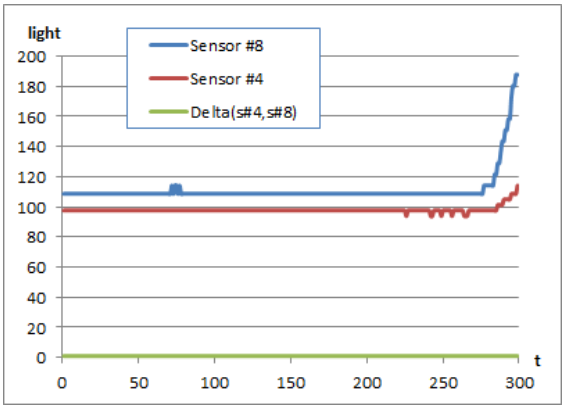
The data set has aggregated connectivity data averaged over all time. The probability of a message from a sender successfully reaching a receiver can describe the protocol relevancy. For example, in the data set, *WPR(#8,#4)=0.258*, and *WPR(#8,#54)=0.584*. *WPR(#8,#54)* is bigger than *WPR(#8,#4)* obviously.

c) Data Semantic Relevancy (*DSR*)

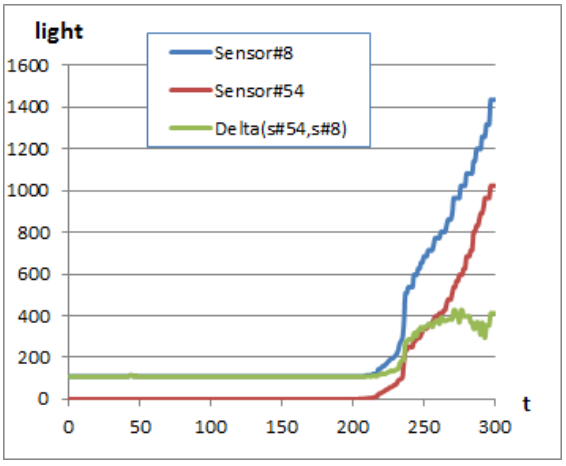
If the data provided by two sensors have relevancies, such as the same trend of changing, the similarity of values, the similar probability of changing, and so on, they are data semantic relevancy. When one sensor node goes wrong, it can be simulated by the nodes of DSR approximately.

Generally speaking, when two sensors of the same version are put into a same environment, the data they gathered will be very same. With the increasing of distance between them, the data they gathered may be more differences. But this case is not always true. It is also influenced by other facets. For example, in Figure 2, $PLR(\#8, \#54) > PLR(\#8, \#4)$ is due to that the distance between sensor node pair (#8, #54) is smaller than that of node pair (#8, #4). However, Figure 1 shows that $DSR(\#8, \#54) < DSR(\#8, \#4)$and the detail analysis is as follow.

We randomly select three hundreds of timeslots, and then the sampled data of sensors #4, #8 and #54 at the selected timeslots are gotten respectively from the data set. Figure 1-(a) shows the comparison between the data of nodes #4 and #8. The delta line shows the difference of the sample data. The changing trend of the two sensor nodes is accordant. To evaluate the similarity of changing trend, we calculate the correlation coefficient of each pair of sensors. The correlation coefficient between #4 and #8 is 0.993283. Figure 1-(b) shows the sampled data of sensor #8 and #54 and this pair of sensors also have the same changing trend. The correlation coefficient between #54 and #8 is 0.992230. That is to say both (#8,#54)



(a) Sensor #8 and Sensor#4



(b) Sensor #8 and Sensor#54

Fig. 1. The data relevancy between sensors.

and (#8,#4) are accordant. However the difference between sensors (#8, #54) is much bigger. So the conclusion $DSR(\#8, \#54) < DSR(\#8, \#4)$ can be made. After checking the room map in Figure 2, we find that there exists a wall between #8 and #54, which makes the two sensors be in different environment. Although the distance between them is small, the data they gathered are lower in Data Semantic Relevancy than (#8, #4).

The above analysis shows that the Data Semantic Relevancy is related to the Physical Location Relevancy and WSN Protocol Relevancy, but it is not entirely due to these two factors. It needs to consider many factors at same time and mining the data relevancy from sensor data.

2.2 Definition of Data Association between Sensors

Association Rules (AR) are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. For example, in text data mining, association rules are the frequent patterns in a text set which satisfy the predefined minimum support and confidence. The association of two texts can be calculated according to the mined ARs.

In WSN, the meaning of association between a pair of sensors is quite difference. Generally speaking, the association between two sensors describes how relevant they are. The precise meaning of association is determined by its application.

In the scenarios shown in the introduction section, such as finding the replaceable sensors, the data association between sensors can be defined as,

Definition 1 (Data Association between Sensors) Data Association between Sensors ***DAS*** *means two sensors are relevant in the changing of data, denoted as*$DA(X, Y)$.

$$DA(X, Y) = \alpha c_{XY} + \beta p_{XY} + \gamma d_{XY} \qquad (1)$$

where c_{XY} is the *correlation coefficient of two sensors data,* p_{XY} *is the probability of simultaneous data changing,* d_{XY} *is the difference between the values of sensors.*

The detail definitions of c_{XY}, p_{XY}, d_{XY} are represented in(2),(3), and (4) respectively .

The degree of DSA is influenced by the following conditions:

1. High correlation coefficient (HC). The correlation coefficient between the data of two sensors should be big enough. Satisfying this condition can ensure that two sensors have the same trend of data changing macroscopically.

2. High probability of simultaneous data changing (HP). This condition can ensure two sensors are the same trend of data changing in microcosmic.

3. Small difference between the values of sensor data (SD). This condition can ensure the data of two sensors are nearly equal in the values.

In Figure 1, conditions 1 and 2 make two lines are similar in the shape. And condition 3 makes two lines are near enough.

2.3 Mining Data Association between Sensors

According to the definition of Data Association between Sensors, the Data Association is determined by three parameters.

1) Correlation coefficient of two sensors data (c_{XY}).

Each sensor can be considered as a variable, and the sampled data on different timeslots are the values of the

variable. So the correlation coefficient can be used to evaluate the association in one aspect. The correlation coefficient of two sensors can be computed as the follow equation.

$$c_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \quad (2)$$

where X,Ydenote a sensor respectively, covdenotes the covariation, σ_X and $\sigma_Y$are standard deviations, $n$is the number of sampled times.

2) Probability of simultaneous data changing (p_{XY})

The simultaneous data changing of sensors means the sampled data of two sensors changes in a same timeslot. In all the sampled timeslot, the probability of the simultaneous data changing of sensors *X* and *Y* can be computed by the following equation.

$$p_{XY} = \frac{n}{N} \quad (3)$$

where *N* is the total number of sampled timeslots, *n* is the number of timeslots where both *X* and *Y* obviously change. The obvious changing is evaluated by the ratio of changing, denoted as$\frac{\Delta X}{X}$. Set the threshold isα, $\frac{\Delta X}{X} > \alpha$and $\frac{\Delta Y}{Y} > \alpha$should be satisfied.

3) The difference between the values of sensors(D_{XY})

The difference between the values of sensors describes the similar level of the two sensors. It can be computed by the average of the data differences of each

$$D_{XY} = \frac{1}{n}\sum_{i=1}^{n}|X_i - Y_i| \quad (4)$$

3 DATA ASSOCIATION NETWORK OF SENSORS

Section 2 defined the data association relation between two sensors and proposed the method to mine the degree of association relation. Using the proposed method, the association of any pair of sensors can be calculated. However, these association relations are independent. If we could integrate all these associations, the more valuable semantic can be minded.

In our previous work[16], association link network (ALN) is proposed to organize the association semantic relations among webpages. In this paper we employ the form of ALN to organize the data association relations among sensors as Data Association Network (DAN). Beside the two networks have the same form; they are quite different in the meaning, building, and applications.

In this paper, the WSN is discussed in different views and three kinds of network of WSN appear in this paper. For the convenience of discussion, we distinguish the three kinds of networks here.

(a) Physical Location Network (*PLN*) records the physical location of sensors in WSN. In PLN, each node denotes a sensor, and each edge records the physic distance of a pair of sensor nodes. In fixed location WSN, the coordinate of each sensor is known, and it is an easy job to compute the distances.

(b) Data Protocol Network (DPN) records the adjacent relations of sensor nodes that transfer data to each other in WSN. If two sensors transfer data in WSN, they are adjacent and there is an edge between them. In DPN, each node denotes a sensor, and each edge denotes the adjacent of a pair of sensors. The weight of each edge is the probability of successfully linking between these two nodes.

(c) Data Association Network (DAN) records the data relevancy among sensors as defined in Definition 1, which is mainly discussed in this paper.

3.1 Definition of Data Association Network of Sensors

Definition 2 (Data Association Network of Sensors) Data Association Network of Sensors *DAN is an undirected graph, in which the node represents a sensor, the edge refers to the data association relation between two sensors, and the weight of edge denotes the strength of the data association relation. It can be denoted by*

$$DAN = \{s_i | s_i \in S; (s_i - s_j) | a_{s_i - s_j} > \delta\} \quad (5)$$

where $S$is the sensors set of DAN; $s_i - s_j$ is an edge between sensor $s_i$and sensor s_j; $a_{s_i \to s_j}$denotes the association weight of edges$_i - s_j$; $\delta$is the threshold of the weight of edge. Only those edges whose weights are bigger than δcan be reserved.

Generally all the edges are stored in an adjacency matrix, denoted by

$$DAN = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mm} \end{pmatrix} \quad (6)$$

According to the application of DAN in WSN, DAN should satisfy the following rules, which are the basis of building DAN.

Rule 1 *Isolated nodes in DAN should be as few as possible.*

The isolated node means there are no data association nodes with it in WSN, and it cannot be monitored by its data associated neighbors.

Rule 2 *DAN should have an appropriate network density.*

Low network density means each node in DAN has few neighborhoods, which may lead to a bad result of being hard to find a data relevant node to replace the node out of work. Contrarily, high network density means too many neighborhoods, which may lead to the high complexity of networks and lacks of focus when deal with the neighbors.

Rule 3 *DAN should be a balance network in density distribution.*

The appropriate network density of DAN does not mean each part of the DAN has the appropriate network density. The bad case is some parts of the DAN have high density and some parts of the DAN have low density, but the average density of the entire DAN is appropriate. So the balance network in density distribution ensures each part of the network has the similar density.

Therefore, according to above three rules, the ideal DAN is a random network but not a scale free network, and each sensor has a number of data association neighbors. But the topology of DAN is influenced by the actual environment and the follow work is to build a DAN of better topology.

3.2 Building the Data Association Network of Sensors

The simply method to build the DAN is all pairs algorithm which needs calculate the association between each pair of sensors. When the quantity of sensors is too much or the time for calculating is too limited, all pairs algorithm cannot be used for its high complexity.

This section discusses how to build the DAN with low complexity of algorithm and satisfying the three rules of DAN.

3.2.1 Reducing the number of computing

Referenced to all the sensors, the phenomenon of data association is observed in part of sensors, so it is no need to compute between all pairs.

The data association of two sensors is related with the physical location and protocol relation, but it is not totally dependent on these two factors. If taking the PLN and DPN as the background knowledge for building the DAN, the computation between sensor pairs can be limited to a local scope in PLN or DPN with few nodes.

Supposing that there is a special sensor *Si*, and its PLN neighbors denoted as NPLN,

$$\mathrm{N_{PLN}} = \{v | D(s,v) < \alpha\} \tag{7}$$

Its DPN neighbors denoted as NDPN,

$$\mathrm{N_{DPN}} = \{v | T(s,v) < \beta\} \tag{8}$$

Then the candidate neighbors of *Si* in DAN is

$$N = \mathrm{N_{PLN}} \bigcup \mathrm{N_{DPN}} \tag{9}$$

The computing of data association of node *Si* is only need to do in the candidate neighbors, which is far fewer than the total number of nodes in DAN.

3.2.2 Reducing the number of isolated nodes in DAN

According to *Rule_1,* the number of isolated nodes should be as few as possible. When building *DAN*, a threshold of the weight of edge is set to determine if an edge should be reserved in the network. Normally, the threshold is fixed. A node becomes the isolated node when the weights of all its edges are smaller than the threshold and no edges are reserved. In this condition, we can decrease the threshold for this node in order to reserve a number of edges for it.

It is an available method that using dynamic threshold mechanism to reduce isolated nodes. Dynamic threshold mechanism can be realized in different methods. Here we use cycle process testing method to control the threshold. The basic idea of the method is as follow:

Supposing that the threshold to control entire network is *t,* if the degree of this node is zero, decrease the threshold *t* to $\lambda t (0 < \lambda < 1)$, repeat this process until the degree is bigger than zero.

Another cause for generating isolated nodes is when the edges are zero, which means these nodes are really isolated and have no need to be dealt with.

3.2.3 Controlling the density distribution of DAN

According to Rule_2 and Rule_3, DAN should be a network with appropriate density and degree distribution balance.

Based on the definition of graph density, the network density of DAN is defined as follow. **Definition 3 (Network Density)** Network Density ND *is the radio of the actual number to the maximum possible of edges. ND describes the density of edges and be denoted as*

$$ND = \frac{l}{n(n-1)/2} \tag{10}$$

where n is the number of nodes in DAN, l is the actual number of edges in DAN.

Network density describes the number of edges macroscopically and can't ensure every part of DAN has the appropriate network density. Therefore, the network should be evenly distributed. There are several definitions and methods of network balance degree. In this paper, with the consideration of the complexity of algorithm, we use the *mean square error* of the degrees of nodes to describe the network balance degree. **Definition 4 (Network Balance Degree)** Network Balance Degree NBD *means square error of the degrees of the nodes and be denoted as*

$$NBD = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})^2} \tag{11}$$

where n is the number of nodes in DAN, x_k*is the kth node's degree,* $\overline{x}$ *is the average degree of all nodes.*

That the NBD value of DAN is small means the network is more balanced.

3.2.4 DAN building algorithm

Considering to the above factors, the algorithm to build DAN is shown in Algorithm 1.

Algorithm 1 takes the PLN and DPN as the background knowledge, which can effectively decrease the time complexity of algorithm. And the final DAN satisfies the three rules of DAN. The effect of the algorithm will be evaluated by the experiments in Section 5.

Algorithm 1: DAN Building Algorithm

-Input: *SD, PLN, DPN*
-Output: *DAN*
-Description: To build the DAN from the sensor data with the PLN and DPN supporting

```
// build the initial DAN_t
1. for each s_i ∈ WSN(s_1, s_2, s_3, ..., s_m) do
2. begin
3. find neighbors N(s_i) by Equation (9)
4.  for each s_j ∈ N(s_i) do
5.    calculate a(s_i − s_j)
6.    set DAN_t(a_ij) = a(s_i − s_j)
7. end
8. if a_ij > t then copy a_ij from DAN_t to DAN
// Control the density distribution
9. calculate the density N of DAN
10.  if (N- γ ) → 0 goto step14
11.  if N > γ dec (t) else inc(t)
12.  if t → 0 goto 14
13.  go to step 8
14.  for each s_i ∈ DAN(s_1, s_2, s_3, ..., s_m) do
15.  begin
16.   get the subnet SD around s_i by its first order
      and second order neighbors
17.  calculate the density Ns of DAN
18.  adjust the subnet density
19.  end
20.  calculate NBD by Equation (11)
21.  if NBD > χ goto step 14
// remove the isolated node.
22.  For each isolated node s_x in DAN
23.  select new edges from the method of section 3.2.2
24.  end
```

4 APPLICATIONS

DAN records the data association relation of sensors, which is helpful to WSN and there are many applications based on DAN. To prove the conclusion, we will discuss three examples of the DAN applications.

The first application is to detect the abnormal sensors and the second is to simulate the faulty sensor. Both of them use the part of knowledge of DAN.

The third application is to optimize the distribution of sensors, which uses the macroscopic knowledge of DAN.

4.1 Detect the abnormal sensor

In *WSN*, it is easy to find that a sensor does not work when it does not send any data. It is also easy to find an obviously abnormal sensor when it sends an abnormal value beyond the reasonable scope. For example, when the value of a temperature sensor is 100°C, it will attract the attention of manager. However, it is not easy to find the small error of sensor. Based on the knowledge proposed by DAN, it is possible to find the abnormal sensor with small error which can't be noticed by human.

Supposing that the threshold to control entire network is *t,* if the degree of this node is zero, decrease the threshold to $\lambda t(0 < \lambda < 1)$, repeat this process until the degree is bigger than zero.

DAN holds the knowledge of the association relation among sensors. In DAN, the neighbors of a sensor are the data association sensors. In normal condition, the changes of data of these association sensors are relevant according to the association weight. This phenomenon reflected in the DAN is the topologic of these sensors relatively stable with the time changing. If a sensor is abnormal, the associations with its neighbors are changing which leads to the changing of the topology of DAN. It is to say simply that its neighbors will change or the association weights with

its neighbors are obviously changing.

Therefore, monitoring the topological or the weight of the edge of DAN is an available way to detect the abnormal sensors. But the threshold of the standard to judge a sensor is abnormal is related to the special type of *WSN*.

The effect of the application is proved in the Experiment 2 in section 5.3.

4.2 Simulate the faulty sensor

Another application of *DAN* is to simulate the data of the sensor when it is out of work and there is no redundant sensor for it.

In *DAN* the neighbors have the data association relation, and the data of the faulty sensor can be simulated by the data of associated neighbors approximately.

Simple; the similar of data of two sensor is increase with the increasing of the weight of association between two sensors. But when a sensor has several neighbors and each neighbor has different weights of association, how to simulate the data according to the neighbors is more complex.

Here, we can use the weighted average values of neighbors as the simulation value.

$$V_{Sk} = \frac{1}{n} \sum_{i=1}^{n} V_{Si} * W_{(Si-Sk)} \tag{12}$$

where $Sk$is the error sensor, Si is theith neighbor of Sk. $V$is the value of sensor, $n$is the number of neighbors, $W_{(Si-Sk)}$is the association weight between sensor $Si$and Sk.

The effect of the application is proved in the Experiment 3 in section 5.4.

4.3 Optimize the distribution of sensors

In *WSN*, the number of sensors and the physical location of each sensor are determined by the scenario of the application. In the implementation of *WSN*, in order to improve the reliability of *WSN*, certain quantities of redundant sensors are added to *WSN*.

As the application in section 4.2 shows, a sensor can be simulated by its associated sensors. So the redundant sensors could be reduced.

The optimization of *WSN* is a complex task, because the optimization is concerned in the three networks of WSN, *PLN*, *DPN* and *DAN*. In this paper, we only give some advices for optimization by hand according to the analysis of *DAN*, while the automatic optimization algorithm is our ongoing work and will be discussed in the next paper.

1) Redundant Degree

A sensor can be simulated by its associated neighbors does not mean it can be removed directly, because when the sensor is removed, the topology of the DAN will be changed. Only when the removing of a sensor does not influence the topologic of DAN, it can be removed.

The redundant degree describes how many sensors can be removed, which is defined as the ratio of the number of removable sensors and the total number of sensors.

2) Candidate redundant sensors

Checking all the sensors, the candidate redundant sensors are composed of all the removable sensors. It should be noted that when a sensor is combo sensors, it need consider of several DAN together.

For example, in the data sets this paper used, each sensor can gather humidity, temperature, light and voltage values; there is a DAN for each value. By the influence of environment each DAN have its own candidate redundant sensors. A sensor can be removed only when it is removable in each DAN.

5 EXPERIMENTS

5.1 Data Sets

The data sets used in this paper is "Intel Research Berkeley Sensor Network Data" which can be downloaded from the database group at MIT [17]. Here we first thank the people who contribute to the data set. They are Peter Bodik, Carlos Guestrin, Wei Hong, Sam Madden, Mark Paskin, and Romain Thibaux .

This dataset contains the information collected from 54 sensors deployed in the Intel Berkeley Research lab between February 28th and April 5th, 2004.

Mica2Dot sensors with weather boards collected time stamped topology information, along with humidity, temperature, light and voltage values once every 31 seconds.

The sensors were arranged in the lab according to the diagram shown in Figure 2.

5.2 Experiment 1: Building the DAN

Algorithm 1 is used to build the DAN, and it takes the PLN and DPN as the background knowledge, which can effectively decrease the time complexity of algorithm.

There are *54* nodes in the data set. All pairs algorithm needs *1548* times of association degree computing. Using the background knowledge, the average size of candidate set is 4.6, which means a sensor has 4.6 neighbors in PLN and DPN. The proposed algorithm only computes 248 times and the increasing is 82.9%.

Figure 3 is the final DAN of temperature sensors and Figure 4 is the final DAN of light sensors. The two figures show that the topologies of the two DANs satisfy the rules. Both of the networks are density balance. Figure 4 has more isolate nodes than Figure 3, which is determined by the characteristic of different sensors.

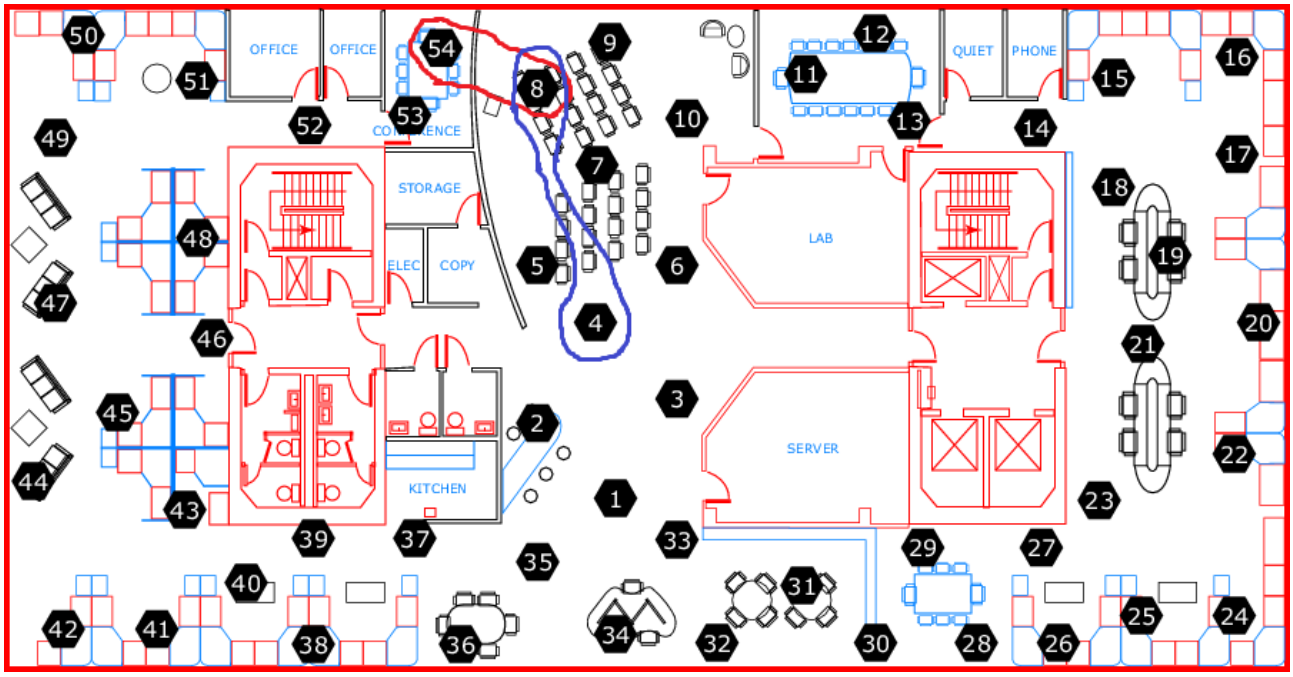


Fig. 2. The Sensor Deploy Map in the Intel Berkeley Research Lab

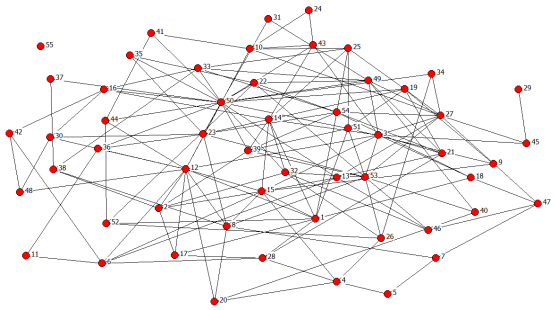


Fig. 3. The result DAN on temperature sensors

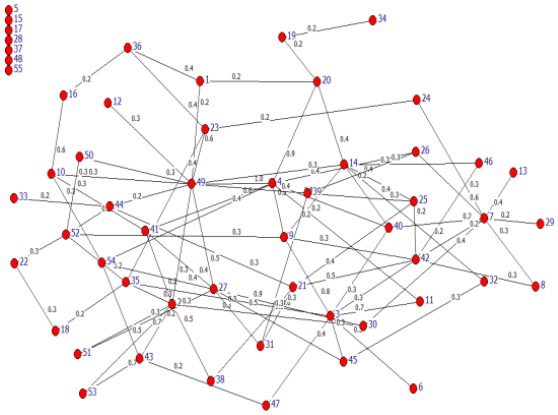


Fig. 4. The result DAN on light Sensors

5.3 Experiment 2: Detecting the abnormal sensor

In this experiment, fifteen sensors are randomly selected from *DAN* of light sensors, and the results are the average of all the selected sensors. The experimental result of the relations between the abnormal sensors and the changing of DAN are shown in Figure 5.

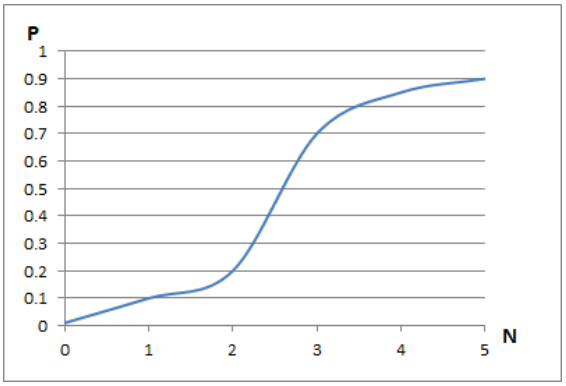
Figure 5-(a) shows the relation between the changing of neighbors and the probability of the sensor being abnormal. The abscissa *N* is the number of the changed neighbors, and the ordinates P is the probability of a sensor becoming abnormal.

Figure 5-(b) shows the relation between the changing of the association with its neighbors and the probability of the sensor being abnormal. The abscissa *A* is the changing of association value with its neighbors and the ordinates P is probability of the sensor is abnormal.

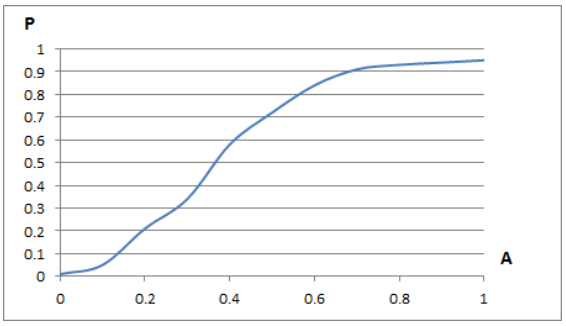
From Figure 5 (a) (b), it is obviously that with the changing of topology of DAN the sensor is more likely to be abnormal.

5.4 Experiment 3: Simulate the faulty sensor

In this experiment, ten sensors are selected from *WSN*, and each sensor is simulated both according to *DAN* and



(a)



(b)

Fig. 5. The relation between the abnormality of sensors and the changing of DAN topology

PLN.

In *DAN*, the sensor is simulated by Equation (12) . In *PLN*, the sensor is simulated by the average of the value of near sensors. The evaluation result is described by the similarity between the actual value and the simulated value. The similarity is defined as,

$$s = 1 - \frac{|Va - Vs|}{Va} \tag{13}$$

where $Va$is the actual value of the simulated sensor, $Vs$is the simulated value of the sensor, $|Va - Vs|$is the absolute value of the difference between simulated value and actual value.

The experimental results are shown in Figure 6, from which we can see that the effect of *DAN* method is better than *PLN* method. After checking the physical location of the sensors, we find that when the sensors are influenced by the surrounding environment, DAN method is obviously better than PLN method, as shown in Figure 6.

6 RELATED WORK

Recently more and more researchers focus on the mining on sensor network data. The related work to this paper includes the organization of the correlated data of sensors, the method to mine the correlated relation from sensor data, and the applications of the correlated data of sensors.

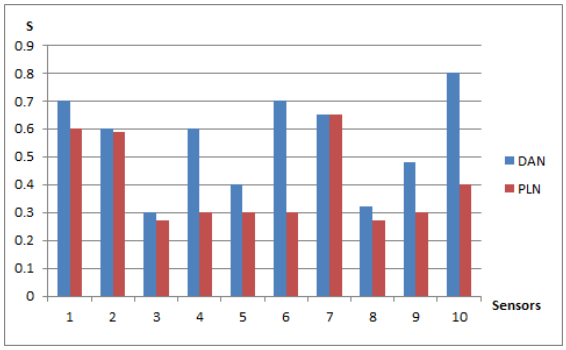


Fig. 6. The relation between the abnormality of sensors and the changing of DAN topology

6.1 Mining the correlation from sensor data

Cao et al. [8] proposed a method to mine the data correlation from multi-faceted sensor data. They transfer high dimensional multi-faceted data into lower dimensional data and detect the correlation among multi-faceted data. Their method to mine the data correlation is different with ours. Their work is just to mine the correlation among sensor data. In our work the correlations is mined to build the semantic network of all the sensors in higher level.

Gupta et al. [9] proposed the algorithm to exploit data correlations in sensor data for minimizing communication costs incurred during data. They select a small subset of sensor nodes which can be used to reconstruct data for the entire sensor network. In our method, the correlations are mined to build the Data Association Network (DAN). The same subset of sensor nodes can be also gotten from DAN and DAN can be used in more applications.

Safarinejadian et al. [10] proposed a distributed variational Bayesian algorithm for density estimation and clustering in sensor networks. It can be seen as a clustering method based on the density of sensor data. Their work is different with ours. We also use the density in our method, but it is a parameter to control the topology of Data Association Network.

There are still many other methods for mining the correlation from sensors data [11][12][13]. Besides the mining methods are different, the purposes of these methods are seldom to build the semantic network of the sensor.

6.2 Organizing the correlated data of sensors

Jindal et al. [14] proposed a model to organize the spatially correlated sensor network data. Their work is similar

with ours. Both of us concern the organization of correlated sensor data. But the models are different, their mode is based on Markovian and our model is based on semantic link network.

Bhattacharya et al. [15] proposed the method to model the high-level semantic events from low-level sensor signals. The model of sensor data is also different with ours.

Luo et al. proposed the association link network (ALN) to organize the semantic relation of webpages [16],[18],[19],[20]. Liu et al. proposed the community discovery method which can be used to find the related nodes in ALN [21][22]. The Data Association Network (DAN) in our work has the similar structure with ALN. But there are different in the building methods and the application domains.

There are still many other methods for organizing the correlated data of sensors. But few of them model the sensors in semantic network and used it to improve the WSN.

7 CONCLUSION

This paper analyzes WSN from the view of data relations, mines the data semantic relation between two sensors from sensor data, builds the Data Association Network of sensors (DAN). DAN has the semantic of the WSN which can improve WSN in some aspect. The experimental results show that the proposed method can mine the association relations among sensor nodes effectively, and the DAN is helpful in improving WSN.

Our contributions of this paper are as follows

1. Proposes a method to mine the data association relation between sensors from sensor data, which can mind the association relations among sensor nodes effectively.

2. Proposes Data Association Network of sensors (DAN) to organize the independent association semantic into an effective form.

3. Proposes an algorithm to build DAN, which has low complexity and can ensure the final DAN is a balance network in density distribution.

4. Uses DAN to solve some problems of WSN, which can improve WSN in some aspect.

The automatic optimization algorithm of WSN is our ongoing work, and the future work is to analyze DAN to get the useful knowledge to improve WSN.

ACKNOWLEDGMENT

Research work reported in this paper was partly supported by the National Science Foundation of China under grant nos. 91024012, 61071110, 61171014, U1135005, by the National High-Tech Research & Development Program of China under grant no.2009AA012201, by the Shanghai Leading Academic Discipline Project under grant no. J50103, and by the Fundamental Research Funds for the Central Universities.

REFERENCES

[1] N. Gershenfeld, R. Krikorian, D. Cohen, “The Internet of Things”, Scientific American, 2004.

[2] J. P. Conti, “The Internet of things,” Commun. Engineer. vol. 4, pp. 20-25, 2006.

[3] INFSO D.4 Networked Enterprise & RFID INFSO G.2 Micro &Nanosystems Groups in co-operation with the RFID working Group of the EPoSS, “Internet of Things in 2020”, 2008.

[4] http://en.wikipedia.org/wiki/Wireless_sensor_network.

[5] J.F. Martínez, P. Castillejo, M. Zuazua, “Wireless sensor networks in knowledge management”, Procedia Computer Science, 1(1), pp. 2291-2300, 2010.

[6] Dargie, W. and Poellabauer, C., “Fundamentals of wireless sensor networks: theory and practice”, John Wiley and Sons, 2010.

[7] Sohraby, K., Minoli, D., Znati, T. "Wireless sensor networks: technology, protocols, and applications, John Wiley and Sons", 2007.

[8] CAO Dong,QIAO Xiu-Quan,Judith Gelernter et al. “Mining Data Correlation from Multi-Faceted Sensor Data in Internet of Things”. China Communications, 8(1), pp. 132-138, 2011.

[9] H. Gupta, V. Navda, S. Das, “Efficient gathering of correlated data in sensor networks”, ACM Transactions on Sensor Networks, 4(1), pp. 1-31, 2008.

[10] B. Safarinejadian, M.B. Menhaja, M. Karrari, “Distributed variational Bayesian algorithms for Gaussian mixtures in sensor networks”, Signal Processing, 90(4), pp. 1197-1208, 2010

[11] S. Yoon, C. Shahabi, “The Clustered AGgregation (CAG) technique leveraging spatial and temporal correlations in wireless sensor networks”, ACM Transactions on Sensor Networks, 3(1), pp. 1-39, 2007.

[12] V. Rajamani, S. Kabadayi, Julien, C., “An interrelational grouping abstraction for heterogeneous sensors”, ACM Transactions on Sensor Networks, 5(3), pp.1-31, 2009.

[13] J. Wang, Y. Liu, S.K. Das, “Energy-efficient data gathering in wireless sensor networks with asynchronous sampling”, ACM Transactions on Sensor Networks, 6(3), pp. 1-37, 2010.

[14] A. Jindal, K. Psounis, “Modeling spatially correlated data in sensor networks”, ACM Transactions on Sensor Networks, 2(4), pp.466-499, 2006.

[15] A. Bhattacharya, A. Meka, A. K. Singh, “MIST:distributed indexing and querying in sensor networks using statistical models”, Proc. of the 33rd international conference on Very large data bases (VLDB), pp. 854-865, 2007.

[16] Xiangfeng Luo, Zheng Xu, Jie Yu, Xue Chen, “Building Association Link Network for Semantic Link on Web Resources”, IEEE T. Automation Science and Engineering 8(3): 482-494 2011

[17] Intel Lab Data. http://db.csail.mit.edu/labdata/labdata.html

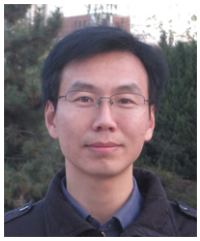
[18] Xiangfeng Luo, Jie Yu , Qing Li, Fangfang Liu, Zheng Xu, "Building Web Knowledge Flows based on Interactive Computing with Semantics", New Generation Computing, 28(2), pp. 113-120, 2010

[19] Xiangfeng Luo, Zheng Xu, Qing Li, Qingliang Hu, Jie Yu, Xinhuai Tang, "Generation of similarity knowledge flow for intelligent browsing based on semantic link networks", Concurrency Computation Practice and Experience, 21(16), pp. 2018-2032, 2009.

[20] Xiangfeng Luo, Ning Fang, Bo Hu, K Yan, HZ Xiao, "Semantic representation of scientific documents for the e-science Knowledge Grid", Concurrency Computation Practice and Experience, 20(7), pp. 839-862,2008.

[21] Jin Liu, Bo Li, Wensheng Zhang, “Feature extraction using maximum variance sparse mapping”, Neural Computing and Applications,21(8), pp. 1827-1833, 2012.

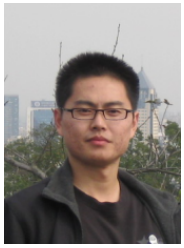
[22] Jin Liu, Jing Zhou, Junfeng Wang, Feng Zhang, Fei Liu, “Irregular community discovery for cloud service improvement”, The Journal of Supercomputing, 61(2), pp.317-336, 2012.



Xiao Wei is an associate professor at Shanghai Institute of Technology (China). Concurrently, he is a visiting scholar at City University of Hong Kong. His main research interests include Web Content Analysis, Semantic Search, and E-learning. He has published over 20 research papers in these areas.

Qing Li is a Professor at the Department of Computer Science, City University of Hong Kong. Concurrently, he is an Adjunct Professor of the University of Science and Technology of China (USTC), Zhong Shan (Sun Yat-Sen) University and Huazhong University of Science and Technology (Wuhan), and a Guest Professor (Software Technology) of the Zhejiang University (Hangzhou, China). His research interests include Object Modeling, Multimedia Databases, Web Services and e-learning. Prof. Li has published over 280 technical papers and book chapters in these areas.

Feiyue Ye received his master degree in 1995 from Shangdong University and his PH.D. degree in 2000 from China University of Petroleum. His research interests include information retrieval, database and mobile computing etc. Prof. Ye has won more than 10 scientific awards including two awards for the advanced science and technology. Moreover, he has published more than 50 papers indexed by EI or SCI, including one treatise and three teaching materials.



Jun Zhang received the bachelor’s degree in 2008 from Shanghai University, where he is currently working toward the PHD degree in the School of Computers. His main research interests include online word relation discovery, knowledge representation, topic detection and tracking.

Rongfang Bie received her Ph.D. degree in 1996 from Beijing Normal University, where she is now a professor. She visited the Computer Laboratory at the University of Cambridge in 2003. Her current research interests includeknowledge representation and acquisition for the Internet of Things,computational intelligence and model theory.

AUTHORS' ADDRESSES

Xiao Wei
School of Computer Engineering and Science,
Shanghai University, Shanghai, China
email: xwei@shu.edu.cn

Prof. Qing Li
Department of Computer Science,
City University of Hong Kong, Hong Kong
e-mail: qing.li@cityu.edu.hk

Prof. Feiyue Ye
School of Computer Engineering and Science,
Shanghai University, China, 200072
e-mail: yefy@shu.edu.cn

Jun Zhang
School of Computer Engineering and Science,
Shanghai University, China, 200072
e-mail: zhangjun_shu@shu.edu.cn

Prof. Rongfang Bie
Information Science and Technology,
Beijing Normal University, Beijing, China
email: rfbie@bnu.edu.cn

Received: 2012-11-10
Accepted: 2013-01-28