http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet

Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (23): 7069-7075, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.7069.7075

Research on Handling the Sensor Data Aggregation Based on Dynamic Counting Sketches

Sheng Huang
Hunan International Economics University, Changsha, China

Abstract: Sensor networks have received considerable attention in recent years. Users do not only collect the sensor node data in many kinds of network system applications, query a group of nodes data aggregation is also often needed, such as average, sum, count and MAX. The in-network aggregation is adopted to save the limited sensor network resources. The common in-network aggregation method is the tree-based data aggregation. However, the shortcoming of the tree-based aggregation is that the packeting missing can cause the lower reliability. One of the solution is the multi-routing data aggregation. A sensor data can produce many copies which can be transmitted in the sensor network. However, multi-routing aggregation leads to the problem of the sensor recounting. In this study, we propose a sensor data aggregation method based on dynamic counting sketches for providing user to elastically set the following system parameters: (1) The want of performing data aggregation function, (2) The request of data aggregation accuracy and (3) The request of protecting privacy right. According to the system parameters set by the users, our system will effectively perform the needed sensor data aggregation query based on multi-path routing method and dynamic counting sketches. The method proposed in the study shows that the packeting space can be effectively saved in the certain accuracy by adopting the simulated data, the real data , the efficiency and the elasticity of the verification method in the wireless sensor network environment.

Key words: Sensor data, in-network aggregation, multi-path routing, dynamic counting sketches

INTRODUCTION

Users do not only collect the sensor node data in many kinds of network system applications, query a group of nodes data aggregation is also often needed, such as average, sum, count and MAX. The common applications of the data aggregation query are the returning of the average rainfall in-return, the searching of the maximum humidity node, the returning of the effective sensor nodes and others. Although the sensor data aggregation query is needed in the wireless sensor network applications, the related technologies at the present stage still have many shortcomings in many technological and application levels so that the related technologies should be improved. The following discussions should be divided into two views: the technical level and the actual application requirements. The existing technical deficiency and the application requirements are also be discussed:

Requirement of the technical level: The most common data aggregation method in the present sensor system is the tree-based data aggregation counting. The tree-based data aggregation counting firstly builds the extending tree regarding the host as the root node in the whole sensor system to connect with each network node in the wireless sensor system. Then, the sensor data aggregation starts to be performed layer by layer with the leaf node. Each network node receives the partial aggregation values transmitted by the sub-nodes. The newly partial aggregation values will be counted and transmitted to the parent node with the observed data and the received partial aggregation values. The final aggregation values will be counted by the network root nodes through the layer-by-layer method. The mainly shortcoming of the tree-based data aggregation counting is that the fault-tolerant capacity of the communication is poor. The current wireless sensor node is adopted to the wireless communication with the low cost and the low power. The communicating framework actually improves the use of the sensor energy, but the wastage rate of the packeting among the sensors is up to 30%. In the framework, the partial aggregation counting of many nodes through the tree-based data aggregation counting will be lost with the failure of the sensor node commnunication. The transmitting failure in the partial aggregation values of the signal sensor node. If the node is next to the network host, a large number of

Corresponding Author: Sheng Huang, Hunan International Economics University, China the sensor nodes will be lost and then the final aggregation values is deviated from the actual aggregation values.

The direct method of improving the communication fault-tolerance capacity in the snesor data aggregation counting is adopted to the network communicating protocol with the higher reliability. However, the use of the network communicating protocol with the higher reliability needs to par out much sensor energy consumption and caueses the bear of managing sensor energy. Therefore, the sensor data aggregation counting method based on the multi-path routing is adopted to improve the communicating fault-tolerance capacity in the premise of using the underlying communicating protocal in the current snesor network. In the recent years, the sensor data aggregation counting method based on the multi-path routing attracts the related research scholars' attention and utilization. The sensor system adopts the directed acyclic in the multi-path routing data aggregation counting to connect each sensor nodes. The sensor nodes broadcast the partial aggregation values to the previous layer nodes. Many copies in a signal partial aggregation values are transmitted and counted in the network. The partial aggregation values in the sensor node will be lost in the situation of all copies are lost. Therefore, the communicating fault-tolerance capacity in the sensor data aggregation counting is substantially improved in the communicating protocal with the low cost and power.

Many copies in the signal partial aggregation values are transmitted and counted, the same data may be received and counted many times so that the double counting problem will be produced. The query result will not be affected by the double counting in several aggregation queries, such as gaining the sensor number with the maximum temperature value in the system. Many same data counted many times may cause the false data aggregation query result in some aggregation queries, such as returning the sensor node numbers in the sensor network activities. The method of directly avoiding the double counting problem is to transmit the sensor number together with the sensor observing values. In this way, the sensor node can know whether a signal sensor data is a copy. Furthermore, the double counting will be avoided. However, the method of distinguishing the copies by transmitting the sensor number can cause the extra energy transmitting and energy consumption for each transmitting needs to transmit the sensor number. The tree-based data aggregation and the multi-path data aggregation have the advantages and disadvantages, respectively. The tree-based data aggregation possesses the accurate query result and the multi-path data aggregation just provides the approximating or error results without causing the wrong network. The tree-based data aggregation losses a large number of the sensor observing values and the muti-path data aggregation remains the optimal fault-tolerance capacity and can provide a more accurate aggregation query result when the network is poor.

Requirement of the application level: The present sensor data aggregation query also has the shortcoming in the actual application layer, except for the shortcoming in the technical procession. When the future sensor system largely enters human's environment, the right of privacy protection read in the individual sensors can not be neglected. However, the existing sensor data aggregation technology aims at the outside data collection and large-scale scientific application. In the application, the right of reading privacy protection read in the individual sensors is not the key point in the mainly sensor system design. However, when the wireless sensor network is applied in the home application or community management, the data received in the sensor node may involve in the sensitive individual privacy or sicurity consideration. The system manager may also needs some data aggregated in the sensor data and the data may contribute to the energy planning or the energy-saving application. Therefore, how to efficiently and safely conduct the sensor data aggregation query in the premise of considering the right of privacy protection read in the individual sensors is an important consideration.

RELATED RESEARCHES AND METHOD FEATURES

Several studies (Considine *et al*., 2004; Fan and Chen, 2008) propose the Duplicate-Insensitive Sketches to solve the double counting problem produced in the multi-path routing data aggregation counting. The mainly spirit in the study (Considine *et al*., 2004) is to adopt the Duplicate-Insensitive Sketches to express the broadcasting data in the sensor when the data is transmitted through the sensor. The same signal data finally just be counted once for the copy of the sketches structure has no influencing characteristics, Therefore, the double counting problem is solved. The study has three mainly disadvantages: (1) The study adopts the FM sketches as the Duplicate-Insensitive Sketches so that the accuracy of the gained estimating values is very low and the degree of the variation in the obtained data aggregation estimating values is very high. The hush function in the FM-Sketches simulates a series of binomial tests, the opportunity of the values in the i counted from the left of the FM-Sketches dyadic array doubles compared to the values in the i+1. Therefore, during the estimation, the corresponding RN from the left to the right in the dyadic array actually has half opprtunity respectively of corresponding to the (RN+1) and (RN-1) so that FM-Sketches has the low accuracy and high variation degree, (2) The utilization of FM-Sketches needs to maintain a group of randomly hush functions in each sensor nodes so that the extra memory expense and the sensor energy consumption will be caused, (3) The utilization of FM-Sketches can not guarantee the accuracy in the data aggregation estimating values. Paper (Fan and Chen, 2008) adopts LC Sketches as the Duplicate-Insensitive Sketches to solve the double counting problem caused in the muti-path routing data aggregation counting. The data aggregation values provided in the LC Sketches is more accurate than the FM sketche's through the verification of simulating the experiments and the analysis of the theory. The LC Sketches and the FM sketches both face the problem of the data structure length is decided by the upper limit of the actual data aggregation values (such as the sensor numbers in the network). The data structure length is very long and the most of values transmitted in the node far away from the network host are zero can cause a large number of the sensor energy consumption.

Paper (Manjhi *et al*., 2005) proposes the wireless sensor data aggregation with the high reliability in the wireless sensor network by connecting the muti-path routing data aggregation with the tree-based data aggregation. The multi-path routing data aggregation will be adopted when the network communicating failure rate is high. The tree-based data aggregation will be adopted when the network communicating failure rate is low. The method adeptly combines the advantages of the muti-path routing data aggregation and the tree-based data aggregation. However, the mainly price of the method is that the data aggregation model maintained in each region should be initiatively managed and dynamically switched. When the network is unstable, the frequently switching of the remained data aggregation model in each region can cause extra energy consumption and the descension of the energy utilization ratio.

Paper (Chen *et al*., 2006) adopts Data Exchanging Process to count the data aggregation values in the wireless sensor network. The sensor node randomly becomes the dominant node in a Data Exchanging Process. The sensor node which becomes the dominant node invites the neighboring nodes to form the sensor group through the broadcasting. The node in the sensor group will transmit the stored partial data aggregation values to the dominant node.

After the dominant node receives the partial data aggregation values transmitted by the sensor group, counting the newly partial data aggregation values and broadcasting the nodes in the sensor group. The partial data aggregation values stored in each sensor node will become the correct data aggregation value with the use of the Data exchanging process. The shortcoming of the method is that the multiple Data Exchanging Process will cause the sensor energy consumption and the delaying of the query result returning. Paper (He *et al*., 2007) proposes the requirement of the right of reading privacy protection read in the signal sensor node. The study has three shortcomings: (1) The safe encryption between the two nodes communication should be used. The manner is to prevent from the attack of bugging, but the extra price for the safe mechanism should be payed, (2) The signal sensor node reading can easily be exposed when a certain number of the sensor nodes are invaded, (3) The study does not consider the sensor communicating failure factor.

The sensor aggregation data query handling system with the efficiency and the elasticity is proposed. The proposed system provided the users elasticity to set the following system parameters: (1) The want of performing data aggregation function, (2) The request of data aggregation accuracy and (3) The request of protecting privacy right. The innovation is a follows.

The dynamic sketches structure whose length can be altered

The previous data sketching methods should configure the pre-using data space in advance before the sensor aggregation values counting. The overdue configurations are usually not avoided so that the meaningless sensor energy consumption is caused and the application lifetime in the network is shortened. The proposed concept of the dynamic configuring space makes the sensor network aggregation values counting more elastic and more sensor energy-consumption saving.

The counting between the sensor reading and the sensor aggregation values is represented by the use of the dynamic counting sketches structure.

The sensor reading of the aggregation counting is showed by the dynamic counting sketches structure. The counting method is different from signally transmitting the aggregation counting sensor reading and the final counted sensor data aggregation can not be influenced. Namely, The propose sensor aggregation values counting will largely improve the aggregation values accuracy based on the multi-path routing data transmitting method in the premise of not increasing the extra sensor energy consumption.

The technology of protecting the right of the reading privacy in the individual sensor node.

The proposed technology of protecting the right of the sensor reading privacy is unnecessary to use any safe encryption technologies, not afraid of any sensor nodes are invaded and not considering the node bugging technologies. In the actual applications, the technology of protecting the right of the sensor reading privacy is feasible and efficient. In addition, the sensor node with the privacy protection and the sensor node without the privacy protection in the sensor data aggregation can be together counted without being divided. The method improves the energy handling efficiency in the sensor dara aggregation counting when partial sensor nodes just need to be adopted in the privacy protection application.

SYSTEM METHOD

The Fig. 1 shows the method flowchart. After the users transmits the related parameters, Base-station transmits the order to the sensor end. After the snesor end receives the order, the dynamic counting sketches data structure aggregation algorithm is conducted and then the values will be returned to the users.

The designed system mainly contains the following two critical technologies:

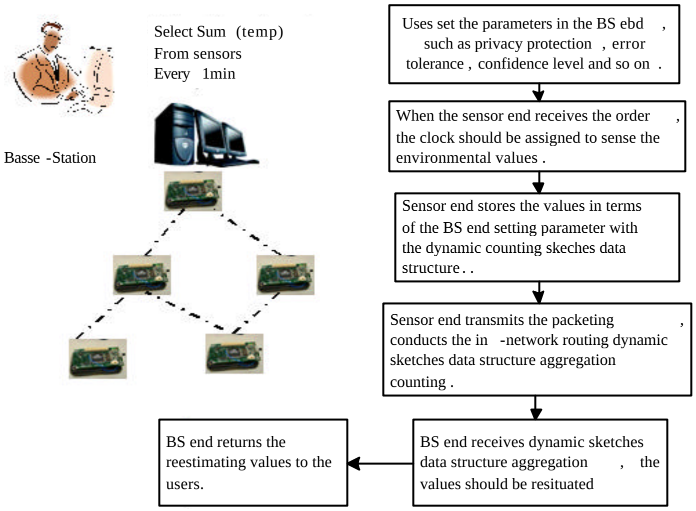
- **Dynamic counting method:** Figure 2 is the concept figure of Dynamic counting sketches. According to the reading, a group of one-dimensional array space will be dynamically configured, the binary bit represents the storage of the reading and the bit-wise OR operation replaces the aggregation additive operation for the same data can be avoided recounting. The proposed dynamic configurating space concept can make the sensor netwwork aggregation counting and the use of the needed space more elastic and more sensor energy-consumption saving

The steps are as follows:

- **Step 1:** Using the dynamic data skeches structure to represent the sensor data:

$$m_i = \left[\frac{v_i.L(k,\varepsilon,\delta)}{5(e^t - t - 1)} \right] .5(e^t - t - 1) \qquad (1)$$

Firstly, According to the sensor data values vi, all participated data aggregation node ui is configured as the following dynamic data sketches structure length:



Uses set the parameters in the BS ebd , such as privacy protection , error tolerance , confidence level and so on .

When the sensor end receives the order , the clock should be assigned to sense the environmental values .

Sensor end stores the values in terms of the BS end setting parameter with the dynamic counting skeches data structure . .

Sensor end transmits the packeting , conducts the in -network routing dynamic sketches data structure aggregation counting .

BS end receives dynamic sketches data structure aggregation , the values should be resituated

BS end returns the reestimating values to the users.

Fig. 1: Sensor method and system flowchart

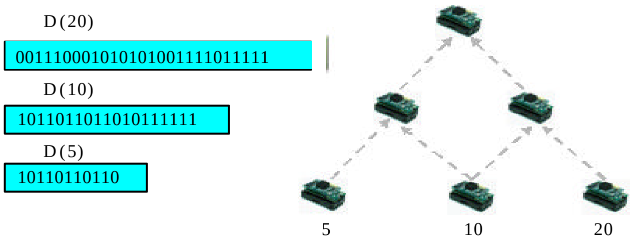


Fig. 2: Concept figure of the dynamic data sketches structure

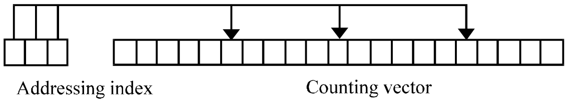


Fig. 3: Dynamic sketches data structure

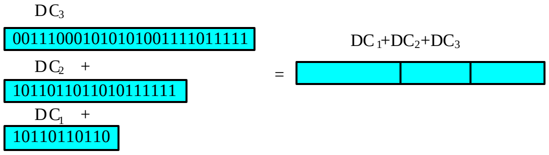


Fig. 4: Dynamic sketches data structure aggregation counting

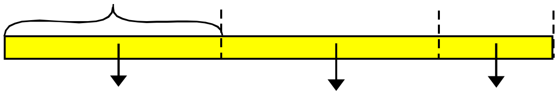


Fig. 5: Opproximate data aggregation result counting

$$m_i = \left[\frac{v_i.L(k,\varepsilon,\delta)}{5(e^t - t - 1)}\right].5(e^t - t - 1) \qquad (2)$$

In the equation, ε represents relative error proportion, δ represents confidence level and the 1's complement, k is the sum of the sensors, t is (k/L(k,ε,δ))

Figure 3 shows that the dynamic sketches data structure DC (BI, CV) in a sensor node contains a Border index and a counting vector. Border Index provides the counting vector length in different nodes to index the position and Counting Vector is an array whose initial value is zero. Then, all nodes ui randomly set the vi element in the configured dynamic data sketches struture counting vector as 1 and the Border Index points to the mi position

- **Step 2:** In-network dynamic sketches data structure aggregation counting

Next, the in-network dynamic sketches data structure aggregation counting will be conducted. The counting starts from the bottom sensor node the sensor network topology and then broadcasting the dynamic sketches data structure to the above layer node. When the node in the dynamic sketches data structure is received, the dynamic sketches data structure itself and the received dynamic sketches data structure will be conducted in-network aggregation. When the two dynamic sketches data structure are aggregated, the definition of the aggregation motion is as follows: the two dynamic sketches data structure is defined as DC1 (BI1,CV1) and DC2 (BI2,CV2). If |CV2| is larger than or equivalent to |CV1|, the sum DC3 (BI3,CV3) of the DC1 and DC2 is:

- BI3 = BI1 ∪ BI2 and
- CV3[i] = CV1[i] ∨ CV2[i],for i = 0, …, |CV1|-1 and CV3[i] = CV2[i], for i = |CV1|, …, |CV2|-1

Figure 4 shows that the node will make the aggregated dynamic sketches data structure to the above layer nodes for being transmitted and handled, until all dynamic sketches data structure focuses on the network host node

- **Step 3:** The approximate data aggregation results counting:

$$DC_{final} = \sum_{i=1}^{k} DC_i(v_i, m_i), \quad \text{where} m_1 < \ldots < m_i < \ldots < m_{max} \qquad (3)$$

After all dynamic sketches data structure concentrates on the network host node, dynamic sketches data structure make the network host node produce a final dynamic sketches data structure DC_{final}:

$$DC_{final} = \sum_{i=1}^{k} DC_i(v_i, m_i), \text{where} m_1 < \ldots < m_i < \ldots < m_{max} \qquad (4)$$

Figure 5 shows that the final approximate data aggregation result counting will be conducted in terms of the Border Index and a Counting Vector in the dynamic sketches data structure and then returning the final result to the users who transmits the query:

$$\begin{aligned}\hat{n} &= -B_3 \mathrm{In}(V_3) - B_2 \mathrm{in}(V_2 / V_3) - B_1 \mathrm{In}(V_1 / V_2) \\ &= -120\mathrm{In}(0.6) - 100\mathrm{In}(0.38/0.6) - 50\mathrm{In}(0.16/0.38) = 150\end{aligned} \qquad (5)$$

$$\hat{n} = \sum_{i=1}^{max} \hat{v}i = -m_{max} \mathrm{In}(V_{m_{max}}) - \sum_{i=1}^{ax-1} m_i \mathrm{In}(V_{mi} / V_{mi+1}) \qquad (6)$$

Dynamic random protection rules: According to the requirement of the data aggregation accuracy given by users, the utilized data aggregation accuracy in the sensor dynamic sketches data structure can be dynamically arranged and all sensor data aggregation can be guaranteed. All data is in the requirement of the data aggregation accuracy defined by the uses.

The steps are as follows:

- **Step 1:** Sensor node receives order and then timing transmits the sensor environmental values

Sensor node receives order and then the sensor node uses the sensor module. The sensor reading v_i will be obtained in terms of the sensor clock rate given by users

- **Step 2: Random numbers are produced:** After the sensor module obtains readings, the sensor node randomly produces a v-random in terms of the reading v_i in the sensor module. The range of v-random is in $[0, v_i]$ and then the V_i value is produced by summing v_i and v-random
- **Step 3:** Using the dynamic counting method to represent the v-random reading

  The sensor node utilizes the above mentioned dynamic counting method to produce a length in terms of the V_i value produced in the step 2. The length is as follows:

$$mi-\left[\frac{\max(k*(v_i+r)/t, G(k*(v_i+r), \varepsilon_{naw}, \delta))}{5*e^t/t}\right]*5*e^t/t \tag{7}$$

  In the equation, r is v-random whose range is [0, vi]
- **Step 4:** In-network dynamic sketches data structure aggregation counting

  Next, the in-network dynamic sketches data structure aggregation counting is conducted. The counting starts from the bottom sensor node the sensor network topology and then broadcasting the dynamic sketches data structure to the above layer node. When the node in the dynamic sketches data structure is received, the dynamic sketches data structure itself and the received dynamic sketches data structure will be conducted in-network aggregation
- **Step 5:** Approximate data aggregation results counting.

  After all dynamic sketches data structure concentrates on the network host node, dynamic sketches data structure make the network host node produce a final dynamic sketches data structure DC_{final}:

$$DC_{final} = \sum_{i=1}^{k} DCi(v_i, m_i), \text{where } m_1 < \ldots < m_i < \ldots < m_{max} \tag{8}$$

$$\hat{n} = \left(\sum_{i=1}^{max} \hat{v}i\right)/2 = \left(-m_{max}\ln(V_{m_{max}}) - \sum_{i=1}^{max-1} m_i \ln(V_{m_i}/V_{m_{i+1}})\right)/2 \tag{9}$$

According to the information provided by Border Index and a Counting Vector in the dynamic sketches data structure and then the following dynamic counting estimation formula is utilized:

$$= \left(\sum_{i=1}^{max} \hat{v}i\right)/2 = \left(-m_{max}\ln(V_{m_{max}}) - \sum_{i=1}^{max-1} m_i \ln(V_{m_i}/V_{m_{i+1}})\right)/2 \tag{10}$$

EXPERIMENTS AND ANALYSIS

A sensor network is simulated and all methods are adopted in the simulated environment. The superiority of the method (RIA-DC) will be proved from two points, that is, the aggregation accuracy counted in the sensor aggregation and the needed space counted in the sensor aggregation. The method is compared with the following existing methods:

- **FMS:** Using FM-Sketch as the main data structure to avoid the double counting
- **RIA-LC:** Using LC-Sketch as the main data structure to avoid the double counting\

Aggregation value accuracy: Firstly, the accuracy is defined as follows:

$$\text{Accuracy} = \frac{\text{Actual valuemin} - \text{Estimation value}}{\text{Actual value}} \tag{11}$$

In the experiment proof chart, each sensor node adopts 1149 bits in the RIA-LC and FMS1, each sensor node adopts 6640 bits in the FMS2, each sensor node adopts 554 bits in the RIA-LC. Figure 6 is the accuracy analysis chart. The figure adjusts the communication failure rate in the simulated sensor network node to observe the optimal accuracy provided by each method. In the experiment proof chart, there are the two following experimental observations and conclusions. Firstly, when Lose rate is less than 0.2, the error rate in the RIA-LC, RIA-DC and FMS2 is lower than FMS1. Secondly, RIA-LC, RIA-DC and FMS2 have the approximate error rate in the fixed Lose rate. The needed space in RIA-LC and FMS2 is higher than RIA-DC's. Therefore, the proposed method has high accuracy and engrgy efficiency.

Sensor aggregation counts the needed space: In the experiment proof chart, the returning accuracy provided in each method should be fixed and the size of the simulated sensor network should be adjusted, that is. the number of the nodes in the network should be adjusted. The needed counting space in each method at the stage should be obtained through observing the experiment comparisons. Figure 7 is the experimental results. The two conclusions can be obtained from the experimental charts. Firstly, when the sensor network is smaller under the

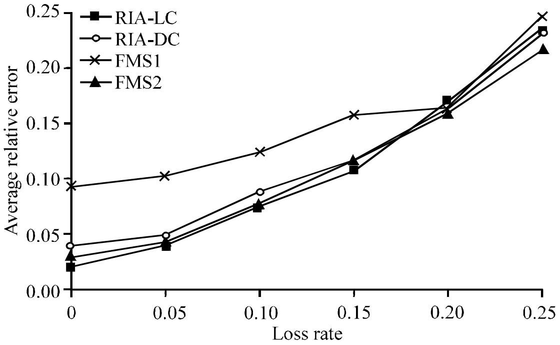


Fig. 6: Accuracy analysis chart

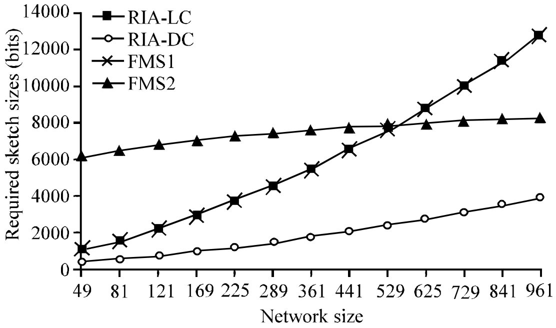


Fig. 7: Needed space analysis chart in the sensor

situation of gaining the same returning accuracy, the needed space in the RIA-LC is superior to FMS2's. When the sensor network is larger under the situation of gaining the same returning accuracy, the needed space in the FMS2 is superior to RIA-LC's. Secondly, the proposed RIA-DC is superior RIA-LC and FMS2 in the any conditions. The experiment proves the superiority in the proposed technology again.

CONCLUSION

Aiming at a group of nodes data aggregation query, such as average, sum, count and MAX, the aggregation query is the widely used critical query in the sensor network application based on the monitoring. According to the advantages and the disadvantages of the existing sensor data aggregation counting method, the strength in the sensor data aggregation counting and the aggregation accuracy are largely improved without the extra energy consumption. Different from the improvement and strenfthening in the sensor hardware bottom structure, the proposed technology is the improvement of the data aggregation algorithm upper the application layer, has nothing to do with the bottom layer structure used in any sensor communications. The proposed technology can be applied in any advanced and reliable sensor communication protocol and hardware. Furthermore, the accuracy and the strength of the sensor aggregation ccounting in the coummunication protocal or hardwareI can be improved. In addition, the application requirement of the signal privacy protection in the sensor data aggregation counting should be considered. The technological requirement in the home application, community management or other sensor applications near to human beings is common.

Compared with the existing technology the proposed sensor reading protection technology is unnecessary to adopt ant safe encryption technology, unnecessary to afraid of any sensor nodes being invaded, unnecessary to caring the node bugging technology. Additionally, the sensor node with the privacy protection and the sensor node without the privacy protection in the sensor data aggregation can be together counted without being divided.

REFERENCES

Chen, J. Y., G. Pandurangan and D. Xu, 2006. Robust computation of aggregates in wireless sensor networks: Distributed randomized algorithms and analysis. IEEE Trans. Parallel Distribut. Sys., 17: 987-1000.

Considine, J., F. Li, G. Kollios and J. Byers, 2004. Approximate aggregation techniques for sensor databases. Procedings of the 20th International Conference on Data Engineering, March 30 -2 April, 2004, Boston, MA, USA, pp: 449-460.

Fan, Y.C. and A.L.P., Chen, 2008. Efficient and robust sensor data aggregation using linear counting sketches. Proceedings of the IEEE International Symposium on Parallel and Distributed Processing, April 14-18 ,2008, Miami, Florida, USA., pp:1-12.

He, W., X. Liu, H. Nguyen, K. Nahrstedt and T. Abdelzaher, 2007. PDA: Privacy-preserving data aggregation in wireless sensor networks. Proceedings of the 26th IEEE International Conference on Computer Communications, May 6-12, Anchorage, pp: 2045-2053.

Manjhi, A., S. Nath and P.B. Gibbons, 2005. Tributaries and Deltas: efficient and robust aggregation in sensor network streams. Proceedings of the ACM SIGMOD International Conference on Management of Data, June 14-16, 2005, Baltimore, Maryland, USA., pp: 287-298.