DATA AGGREGATION USING PROBABILITY THEORY FOR WIRELESS SENSOR NETWORKS

Anuradha M P[1] Gopinath Ganapathy[2]
School of Computer Science, Engineering and Applications
Bharathidasan University, Tiruchirapalli, TamilNadu 620 023, India.
[1]anushivam@yahoo.com
[2]gganapathy@gmail.com

ABSTRACT: **Wireless Sensor Networks (WSN), generally consists of nodes capable of sensing, processing and transmitting of data, detect and tracks the required information. Recent advances in digital electronics and wireless communication have led to the emergence of WSN's. In WSN for monitoring / obtaining some information, the geographically distributed sensor nodes can cooperate with each other to improve the performance of data processing. The position of the various nodes in a network can be distributed or hierarchical. The nodes may possess same or different characteristics. The main objective of this paper is to develop aggregation techniques suited for all varied conditions as well as overcoming the real time issues. This paper develops a mathematical approach using probability functions and distances to restore the data. This paper also proposes a highly secure network as it involves mathematical operations and can be carried out only by the authorized users. This methodology provides efficient, secure, economical, reliable and more productive than existing techniques. This work, in all provides the importance of data aggregation in WSN, the possible efficient methodology and its real time applications.**

Keyword - WSN, Aggregation, Probability Theory

1. INTRODUCTION

Each sensor node in the WSN [1] composed of sensing unit, data processing unit, data communication unit and power unit. Sensor nodes prove the physical phenomenon and transmit the processed data via wireless transceiver. The information processing in sensor nodes deals with the change of information from one representation to another representation. It consumes lot of energy. WSN can be classified into two types called homogenous WSN and heterogeneous WSN. Typically some nodes will have more resources, such processing power and energy level in the network but in homogeneous, all the nodes are the same in terms of resources. But in heterogeneous all the nodes are different in terms of all resources. The hardware representation of sensor node is shown in Fig 1.

The WSN is divided into several fields in which each field forms cluster [2] with multiple sensor nodes, the cluster head namely as the leader of the cluster. Depends upon the network, the direct communication and hop by hop communication is taking place between the cluster head and base station. Among sensor nodes, multi_hop communication is used.

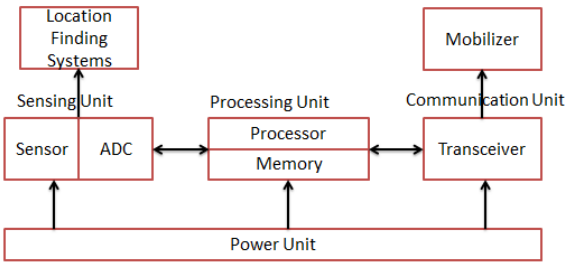

Fig. 1. Sensor Node: Hardware

Sensor nodes in each cluster send a message to the cluster head which combines data and transform the processed data to the sink or base station. The cluster head [3] is responsible for coordinating all activities among the sensor nodes in the WSN within the cluster. The cluster head also responsible for routing packets, perform duplicate packet suppression and also performs data aggregation.

A cluster consists of different types of sensor nodes for data processing specifically called special purpose nodes, generic sensor nodes, high bandwidth sensor nodes and gateway nodes. It supports tiny OS development platform, higher layer Zigbee standard, multi hop routing also supports embedded Linux and windows. Energy optimization is the most important extent in WSN. The activity of the sensing unit, processing unit and communication unit in terms of sensing, processing, data transmitting and receiving will consumes the battery power. From the analysis, data transmission, data processing and data receiving consumes more energy than data sensing.

To reduce the overall energy consumption [4] of entire WSN is achieved through a perfect data processing system, optimized routing techniques, efficient resource management techniques, data management and data aggregation. The nodes performing data aggregation functions require high computational rate than processing and data forwarder nodes. Table 1 shows the comparison of different sensor nodes with respect to processor, power, memory and radio. Instead of hardware configuration and requirements, in this paper energy efficient data aggregation technique using probability theory is discussed.

Table. 1 Sensor Node Comparison

Sensor Nodes	Processor	Power	Memory	Radio
Special Purpose Sensor nodes	8 bit 4-8Mhz	3mw - 3µw	3K RAM	50-100 Kbps
Generic Sensor Nodes	ATMEGA 128	0.036mw –32mw	4K RAM 128K Flash	75Kbps-250Kbps
High Bandwidth Sensor Nodes	ARM 7TDMI 12-48Mhz	50mw – 120mw	64KB SRAM 512KB Flash	Bluetooth
Gateway Sensor Nodes	X86	-	64KB SRAM 32KB Flash	Serial Connection

The data processing scenario of WSN is discussed in Fig 2.

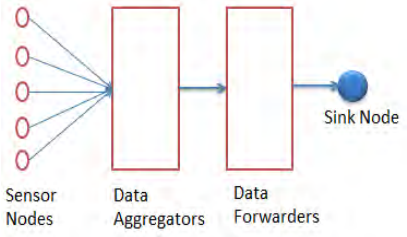

Fig. 2. Data Processing in WSN

Network aggregation and query processing generally involve query propagation and data aggregation. To load a query into a sensor node in WSN, an efficient network structure and data aggregation algorithm has to be established. Once the query is loaded and distributed to all the nodes, it satisfies the query conditions, data are collected, aggregated and communicated to the sink node. The sensor nodes can be trained either individually or they can be trained together. The fusion is the technique used to combine the different classifications result from various sensor nodes. Aggregations of raw data's from various sensor nodes are more useful than individual sensor node reading. In aggregation, the set of sensor readings is summarized into a single data report. Whenever executing the queries, to extract the data from randomly distributed sensor nodes.

Related with data aggregation, He et al. [5] proposed the scheme for data aggregation called "Slice Mix AggRegaTe" (SMART). In this scheme each sensor node slices its reading into n pieces. n-1 pieces are securely distributed to n-1 neighbors near end sensor node. This scheme has a high computational cost, high communication overhead and it has high message collision rate. In this paper, the energy efficient and highly secured data aggregation techniques for WSN are introduced based on probability theory and aggregated query. Hence the computational cost and communication overhead are reduced. General simulations are conducted and the comparison is made for the following parameters like message transmitted, the accuracy level and level of energy left are analyzed and compared with the result SMART system.

The remainder of this paper as follows: Design objectives and network aggregation architecture are presented in section 2. Section 3 provides simulation results are detailed and analyzed. Finally we summarize our work and conclusion is made in section 4.

2. DESIGN OBJECTIVES – NETWORK AGGREGATION

Data aggregation in WSN [6] is a highly efficient technique to save energy and power. There are many proposed methods of data aggregation. The previously proposed techniques do have some drawbacks which cannot be avoided. This paper suggests an efficient, secure and economic data aggregation technique [7], [8] using probability theory and also suggests nested aggregation queries. Since it employs mathematical functionality for data aggregation process in WSN. The proposed architecture is shown in Fig 3.

The proposed technique makes use of probability function. The probability function described in this paper is a ratio of distances. It can also be named as "unique identifier function" since it is a combination of another function (described later) results in a unique value of the function.

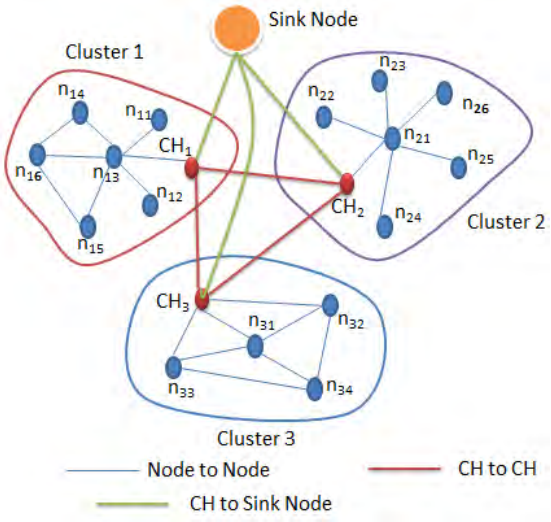

Fig. 3. Proposed WSN Architecture

The probability function also mentioned as Unique identifier function provides a unique value for nodes located at a particular distance say d from the sink. This is highly effective as it does not involve tedious mathematical calculations but performs a simple division operation. The question arises is that why do we need to use a probability function? For any system there is a particular case say theoretical, where we achieve 100% output. So here a comparison is made between the actual expected value with the 100%value. Also the value of this function always lies between 0 and 1. So the user may expect a particular range of values. This procedure helps greatly in time reduction, increased efficiency and increased rate of data retrieval. In the proposed paper, the probability function is defined as the ratio of the optimum distance of the node from the actual distance of the node from the sink node/cluster head as the case may be. In this paper the probability function is defined as the ratio of the optimum distance of the node from the aggregate node to the actual distance of the node. If these two distances are the same then the value of the probability function becomes unity. If 'x' is the optimum distance and 'd' is the actual distance then the probability function becomes

u_{ij}= P (n) = x/d, where n is the node number

The problem that may seem to appear here is the calculation of the distance which is used in the unique identifier function. It may appear that calculation is not practically possible. The following sequence of analysis using mathematics gives the value of distance. The next problem that may arise that the value of this function may have a same value for two different nodes as it is a mere ratio. The solution to those provided after the mathematical analysis of distance.

Practically speaking, it is not easy to find the distance between the node and the aggregate node geographically. Hence we use the following analysis technique for enhanced calculation of distances.

$$u_{ij} = P(n) = \begin{Bmatrix} \frac{x}{d}, x < d \\ \frac{d}{x}, d < x \end{Bmatrix}$$

We know, Speed = distance/time; Distance = speed*time

The speed of the signal sent by the user will be in a particular range and it will be known to the user sending the signal in which data is encrypted. Let this signal in which the data is encrypted be sent which has a speed 'v'. The time at which the signal is sent is noted as 't_1' and the time of reception of signal is 't_2'. The difference, i.e. (t_2-t_1) is the time of travel of the signal from node to aggregate node. The product of the above terms, i.e. v and t_2-t_1 gives the required distance between the two nodes.

Usage of probability function alone results in ambiguity of the function. This is because there is a possibility that the probability function value may be of the same value in some cases as it is just a ratio. So the introduction of distance function along with the probability function becomes a necessity to erase the drawback of ambiguity. In data aggregation technique there will be retrieval of data. The message to be sent is encrypted along with the distance function to enable efficient retrieval of data after the process of data aggregation is completed. Hence this distance function is combined with the probability function as it results in a unique value for each node.

In data aggregation technique [9], [10], during the process similar data will not be aggregated more than once as it corresponds to the same message signal. To ensure that this property of data aggregation is maintained, we use properties of set theory before we proceed to data aggregation. If a union operation is performed on a set of elements only distinct elements will be included in the output. The same property is to be used here.

In the proposed work, the probability function for each node, cluster head and base station is predefined. Based on the preloaded, "unique identifier function" performs the aggregation process securely. If a data is to be retrieved at a later date, it will not be necessary to find the node required, instead it can be retrieved from the cluster head itself. Whichever is nearer to the base station. Hence it's efficient.

Consider a network of nodes with the aggregate node AN, divided into i clusters with the cluster heads CHi and the j nodes of each cluster by n_{ij}.

Let $n_{11}(m)$, $n_{12}(m)$, $n_{13}(m)$..... $n_{ij}(m)$ be the messages sent by the aggregate node to the nodes.

Hence the probability function is defined as

$u_{11} = x/d_{11}$, $u_{22}=x/d_{22}$ ... $u_{nn} = x/d_{nn}$

Say if the message sent by the aggregate node be $n_{11}(m)$, $n_{12}(m)$, $n_{13}(m)$..... $n_{ij}(m)$ which includes same as well as distinct messages. Now if the union function is applied on the set

A = { $n_{11}(m)$, $n_{12}(m)$, $n_{13}(m)$..... $n_{ij}(m)$ }

Then the output becomes

B = $n_{11}(m)$ U $n_{12}(m)$ U $n_{13}(m)$..... U $n_{ij}(m)$

= {M (n_1), M (n_2)....., M (n_n)}

The messages of set B are totally distinct. Hence it ensures accurate data aggregation process.

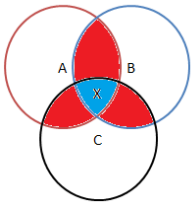

Fig. 4. Representation of Union Function

For 3 clusters,

A ≡ Message in Cluster 1

B ≡ Message in Cluster 2

C ≡ Message in Cluster 3

X≡ Message is only transmitted.

Now consider the following scenario. Suppose the message sent by the node n_{12} and the node n_{14} are the same. Then the above function eliminates one of the data depending on the user's choice and aggregates with the remaining data. Let us consider $n_{15}(M)$ $n_{12}(M)$ $n_{14}(M)$ then upon application of the above function and say the user decides to get from n_{14}, the aggregation result will be like $n_{15} \| \{ n_{12} (M) U n_{14} (M) \} = n_{15} \| n_{14}$

MATHEMATICAL FORMULATION FOR n-NODES PLACED IN ROWS AND COLUMNS:

Let an aggregate node be placed in a system which is surrounded by n-nodes which are present in i^{th} position and j^{th} level. Let D (M) be the distance function along with encrypted message, u denote the probability

function. D_{ij} represents the distance of the node in i^{th} position and j^{th}level. Let CH_1, CH_2, CH_3 be the cluster heads with a Sink node. Let the cluster nodes be n_{11}, n_{12}, to n_{34} under the node CH_1 and similarly for CH_2 and CH_3will also have the node. The CH consist the value of distance function as well as probability functions. This also applies for the base station and sink node. To which the cluster heads process the data. For CH_1 the data aggregation will be n_{11} (m_1) || n_{12} (m_2) || n_{13} (m_3).....

Expanding on this concept and applying the formula we get, D (d_{11}, M_{11}) * u_{11} + D (d_{12}, M_{12}) * u_{12} + …… This is for the cluster nodes. Now if the aggregate node is considered, expanding the formula it will be D (CH_{11}, M) * u_{11} + D (CH_{22}, M)*u_{22} +.....For example:D_{11} represents the function of node in first position of the first cluster level. The mathematical formulation becomes$\sum_{j=0}^{j=n} \sum_{i=o}^{i=n} (Dij(M) \times uij)$. The message can then be decrypted from the distance function using sensing elements.

3. SIMULATION STUDY AND PERFORMANCE ANALYSIS

In this section, the performance evaluation of SMART [5] system and proposed system are made through theoretical analysis and simulation. This simulation is done using the NS2 simulator. In this simulation the network consists of 500 nodes and its randomly deployed across the 250m X 250m area. The data rate 2.5Mbps and transmission range varied from 20m to 30m. The communication cost and overhead, energy consumption and aggregation accuracy are evaluated. As the time interval is increased, the distance increases, hence the probability drops as it is x/d. For a particular time interval as probability drops, the number of messages sent or received will be inversed. This will have a higher value. Hence number of messages can be sent over a time interval, thus increasing the efficiency of the transmission.

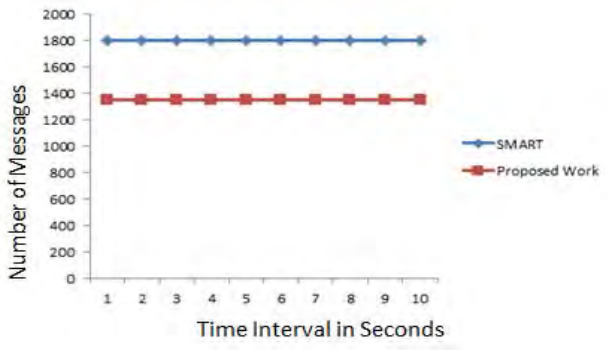

Fig. 5.Communication Overhead with respect to unique identifier function

Data aggregation reduces the number of messages transmitted among the nodes and leads to a substantial reduction in energy. The communication overhead with respect to the unique identifier function of the proposed work with SMART scheme is described in Fig 5. The accuracy with respect to the unique identifier function of the proposed work with SMART scheme is described in Fig 6.

In any network system energy saving is the most important criteria to be considered. The signal is usually sent in the form of waves, the speed of which is almost a constant. The total energy is split into potential and kinetic energy. The potential energy is defined with respect to the sink which is taken as the reference. As the distance increases from the sink, the potential energy also increases. This can be taken as energy supplied to the system.

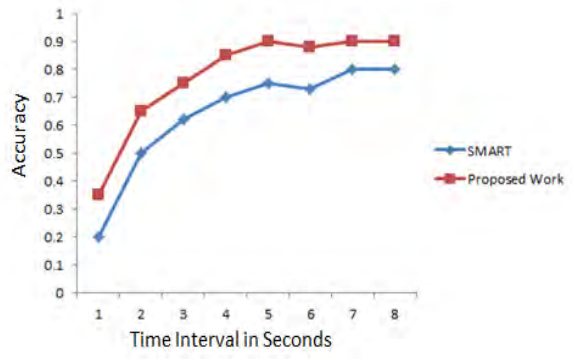

Fig. 6. Accuracy with respect to unique identifier function

As the value of probability function lies between 0 and 1, definite range of value exist for the aggregation function. For example t1 to t2. Hence the users need not check for -∞ to t1 and t2 to ∞.

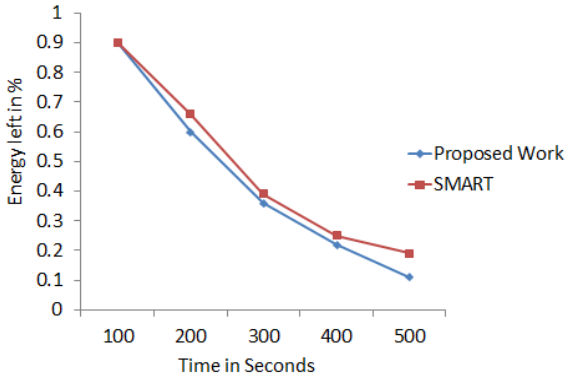

Fig. 7.Percentage of Left Energy with respect to unique identifier function

The kinetic energy is almost a constant. The usage of probability function reveals the nature of the graph. The combination of the proposed relations shows the fact that the system is energy efficient. Fig 7 exemplifies the energy left of proposed work and SMART with respect to time. In a network, speed of signals sent by the aggregation node is constant. Potential energy is defined with respect to aggregation node. As distance of node increases, potential energy increases. For stability, the addition of energy consumed and energy exhausted is equal to zero. As potential energy increases, energy consumed drops. The usage of probability theory reveals the nature of the graph (exponentially decreases). The combination of the above relations shows that the system is energy efficient.

4. CONCLUSION

The proposed technique is efficient, reliable, secure, and economical as it employs a pure mathematical approach for data aggregation. This also ensures enhanced security as the speed of the signal cannot be known. Also the time taken cannot be estimated by an outsider as it requires the exact time at which the signal is sent. It requires simple equipment's and not too much of mathematical calculations. Also this process makes use of probability function, as the range reduces as mentioned earlier. So the range of the final result also tends to be in a particular range. So the range for the checking of the result of data aggregation values also will be finite, hence reducing sufficient amount of time, energy and power. From analysis, it is proved that the proposed work is providing better results in terms of communication overhead, accuracy, and level of energy left from sensor nodes during the aggregation process in WSN.

REFERENCES

[1] F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, ''A Survey on Sensor Networks,'' IEEE Communications, Aug. 2002, pp. 102–114.
[2] D. Wei and A. H. Chan, "Clustering algorithm to balance and to reduce powerconsumptions for homogeneous sensor networks," in Wireless Communications,Networking and Mobile Computing, 2007. WiCom 2007. International Conferenceon, 2007, pp. 2723–2726.
[3] R. Krishnan and D. Starobinski, "Efficient clustering algorithms for self-organizingwireless sensor networks," Ad Hoc Networks, vol. 4, no. 1, pp. 36–59, January 2006.
[4] S. Chatterjea, L. Hoesel, P. Havinga, A framework for a distributed and adaptive query processing engine for wireless sensor networks, Trans. Soc. Instrum. Control Eng. 1 (2006) 58–67.
[5] W. He, X. Liu, H. Nguyen, K. Nahrstedt, T. Abdelzaher. PDA: privacy-preservingdata aggregation in wireless sensor networks, in: IEEE INFOCOM, 2007.
[6] X. Tang, J. Xu, Extending network lifetime for precision constrained dataaggregation in wireless sensor networks, INFOCOM, 2006.
[7] B. Krishnamachari, D. Estrin, S. Wicker, The impact of data aggregation inwireless sensor networks, in: Proceedings of the ICDCS Workshops, 2002, pp.575–578.
[8] C. Castelluccia, E. Mykletun, G. Tsudik, Efficient aggregation of encrypted datain wireless sensor networks, Mobiquitous, 2005.
[9] O. Younis and S. Fahmy, "An experimental study of routing and data aggregationin sensor networks," Mobile Adhoc and Sensor Systems Conference, 2005.IEEEInternational Conference on, pp. 8 pp.–, Nov. 2005.
[10] M. Younis, M. Youssef, and K. Arisha, "Energy-aware routing in cluster-based sensornetworks," in MASCOTS '02: Proceedings of the 10th IEEE International Symposiumon Modeling, Analysis, and Simulation of Computer and Telecommunications Systems(MASCOTS'02). Washington, DC, USA: IEEE Computer Society, 2002, p. 129.
[11] A. Mahimkar and T. Rappaport, "Securedav: a secure data aggregation andverification protocol for sensor networks," in Global Telecommunications Conference,2004. GLOBECOM '04.IEEE, vol. 4, 2004, pp. 2175–2179 Vol.4.
[12] L. Hu and D. Evans, "Secure aggregation for wireless networks," in SAINT-W '03:Proceedings of the 2003 Symposium on Applications and the Internet Workshops(SAINT'03 Workshops). Washington, DC, USA: IEEE Computer Society, 2003,p. 384.