Data Aggregation and Query Processing in WSN

Ayoni Mukherjee , Sanjit Setua

Abstract - A wireless sensor network (WSN) has a wide range of important applications such as remote environment monitoring , target tracking etc. This has been enabled by the availability of sensors that are smaller, cheaper and intelligent. These sensor nodes are equipped with wireless interfaces with which they can communicate with one another to form a network. The design of a WSN depends significantly on the application, and it considers factors like the environment, the application's design objectives, cost, hardware and system constraints. The goal of our work is to present a comprehensive approach of cluster based data aggregation and processing of various aggregation queries in WSN. The aggregation operations in WSN is an attractive research topic now -a- days. We have considered variety of queries involving aggregate functions and proposed an algorithm to calculate the 'MAX' aggregate function .Our algorithm successfully eliminates the limitations of existing algorithms to perform the same operation. We have used temperature sensing application for this purpose.

Index Terms - aggregation, base station(BS), Clusters , Cluster head (CH), MAX, nodes , query ,WSN.

1 INTRODUCTION

A WSN typically has little or no infrastructure. It consists of a number of sensor nodes (few tens to thousand) working together to monitor a region to obtain data about the environment. Data gathering, processing and aggregation belong to the management and control services of WSN and aggregation process enhances the overall efficiency of WSN. Data aggregation is necessary for WSN applications which have large amount of data to send across the network. Data is collected from multiple sensor nodes and combined together to transmit to the base station. Here aggregated data is more important than individual readings. Depending on the importance of the data and criticality of the application a trade-off among different methods is necessary. Each of these methods addresses the issue of energy, robustness, scalability, accuracy and efficiency.

The general requirements for sensor network is presented in section 2 . The query processing and data aggregation technique are in section 3 and Section 4 concludes this paper.

2 GENERAL REQUIREMENTS FOR THE SENSOR NETWORKS

Low Power Consumption : Sensor nodes are usually battery powered. Manual replacement of batteries is often not possible, which makes sensor nodes dependent on their battery life.As a result, minimization of energy consumption becomes critical to achieve a robust system.

Scalability: WSNs with thousands of sensor nodes can become common. Although stationary in many cases, mobile sensors may also be used in the military or environmental application.The scalability of the system hence becomes a major concern.

Self - Organization Ability: Wireless sensor networks are large in size and work in the environment that causes the increase in failures of individual nodes. Mechanisms are needed for joining the network randomly, as well as reorganizing the network upon failures and hence, self-organization ability is essential.

Resilience: Sensor nodes may unpredictably stop operating due to environmental reasons or to the battery consumption. Routing protocols should cope with this eventuality so when a current-in-use sensor node fails, an alternative route could be discovered.

Device Heterogeneity: Although most of the civil applications of wireless sensor network rely on homogenous sensor nodes,the introduction of different kinds of sensor nodes could report significant benefits. The use of sensor nodes with different processors, transceivers, power units or sensing

- *Ayoni Mukherjee is currently pursuing M.Tech in Computer Science & Engineering at University of Calcutta, India. Email:mukherjee.ayoni@gmail.com*
- *Sanjit Setua is an Associate Professor of Computer Science & Engineering at University of Calcutta, India. Email : sscomp@caluniv.ac.in*

components may improve the characteristics of the network.The scalability of the network, the energy drainage or the bandwidth are potential candidates to benefit from the heterogeneity of sensor nodes .

Querying Ability : Due to the network size, the amount of the aggregated data may be too large for transmitting through the whole network.Because of that, the data collection in a particular region or from certain sensor nodes is needed instead. Certain sensor nodes need to be dedicated for collecting the data from regions,creating a summary and forwarding information.Querying function is used to identify collection sensor nodes and the corresponding regions.

3 DATA AGGREGATION

The basic aggregation functions[1], [2],[3],[4] are MIN , MAX , AVERAGE , SUM and COUNT. Among them only MIN and MAX are duplicate insensitive and all others are duplicate sensitive. It is considered that each sensor node has an unique identification number. For numbering the sensor nodes, a global sequential counter is used ,starting from the root node of the network tree .

3.1 Query processing

We have followed cluster based approach for data aggregation which indicates efficient data collection ,management and aggregation in WSN. Each cluster will consist of a number of small sensor nodes(SN), one cluster head (CH), one beacon node(BN) . There is a base-station (BS) which acts as an user interface for the WSN. Here BS is analogous to the root node. In reality we require different BSs for distributed processing and communication for any WSN. The sensor nodes deployed in the wide area, form many cluster groups for efficient network organization. The processing of queries is independent within a cluster group. Different clusters take part in parallel computation and finally different CHs communicate with each other to produce the result required at the BS. The queries are provided at the base station(BS) through user interface or generated automatically at the BS. The BS broadcasts the query to reach each CH (within the transmission range of BS).Here the processing of the queries is not same for all sensor nodes in the network. The responsibility is hierarchically distributed among CH and other regular sensor nodes present in a cluster. Finally the aggregated result is obtained at the BS. We have proposed algorithms for communication and distribution of responsibilities for computation, result transmission , among the CHs to provide a single valid result of an aggregation query to the BS . The main goal is efficient query processing and reduction of response time. Time required for result calculation by a cluster group is predicted beforehand by the depth of each cluster .We have considered different types of queries:

```
i.   SELECT  MAX( temperature )
          FROM sensors ;


ii.  SELECT  AVERAGE( temperature )
              FROM sensors ;

iii. SELECT  MAX( temperature)
     FROM sensor
     WHERE  room in ( SELECT  room
           WHERE   floor = '3' ) ;

iv.  SELECT  MAX( temperature)
      FROM sensor
     EPOCH  DURATION  10 s;
```

3.2 MAX function implementation in WSN

It is known that monotonic and exemplary digest functions [5]can be computed efficiently by localized information exchanges between one hop neighbors. MAX and MIN functions are both monotonic and exemplary digests.

A digest function is denoted by $f(v_1, v_2,\_ \_ \_ ,v_n)$, where v_i is the value contributed by i^{th} sensor node. Additionally, f is decomposable by a function g , such that :

$f(v_1, v_2,\_ \_ \_ ,v_n), = g(f(v_1, v_2,\_ \_ \_ ,v_k), f(v_{k+1}, \_ \_ \_ ,v_n))$

A decomposable digest function is one in which the final result can be calculated from partial results. The values v may either be scalars or vectors. The problem of digest computation is that each node i provides a value v_i as its contribution to the digest function f, where v_i may change over time. The goal of the digest computation mechanism is for each sensor node in the network to contain a continuous estimate for the current value of f.

A digest function is monotonic if only if, when two partial results r_1 and r_2 are combined by a function r =

$g(r_1, r_2)$, such that the result r satisfies the relation for all $i = 1, 2.. , r >= r_i$ for an ordering relationship > =. It is exemplary if the final result can be determined from one single contribution value.

Here in-network aggregation [4] approach has been considered instead of naive centralized approach. Each sensor node computes a partial result of the digest function and passes that result to other neighboring sensor nodes. In-network aggregation has better energy-efficiency characteristics, communication overhead is less and the computation is evenly distributed.

3.2.1 Aggeregation Method

Initially, each sensor node i sets its perceived maximum value $m_i = v_i$, source of maximum $s_i = i$, hop distance $h_i = 0$ and periodically sends a tuple $M = (m_i; s_i; h_i)$ to its neighbors. Upon receiving a message $(m_j ; s_j ; h_j)$ from neighboring sensor node j with $m_j > m_i$, a sensor node i sets $m_i = m_j$, $s_i = s_j$, $h_i = h_j + 1$, parent $p_i = j$. If $m_j = m_i$, it further checks if $s_j > s_i$, which guarantees strict monotonicity. Node i may switch its parent node from j to node k, when k provides the same maximum value but a shorter hop distance $h_k < h_j$.

This method gives an idea of topology of the entire network and an implicit tree (digest tree) is built as each sensor node keeps a track of its parent during message exchange.

Here a lot of power is exhausted during message exchange and nodes continually computes the stated digest function. So , nothing can be concluded about the finiteness property of this algorithm. To ensure finiteness property of the above algorithm , we must do the following things. The basic assumption is that all nodes share a global clock.

Let a query for calculating the MAX value of data provided by the sensor nodes , is either generated at BS or accepted from an user through the user interface at time instant "t" . So BS expects to obtain the result of the state of the network at that time ie : "t". The computation proceeds by considering the value supplied by each sensor node at time "t". There is also a provision to store new values (at time " t+1" and next according to the application) without modifying the value of instant "t". The network topology, being dynamic in nature , can change in the mean time before supplying the result of aggregation query to the BS. That change in topology is not considered till the result is given to the BS. So the result of the query is obtained by considering the state of network at time "t" only. Here the network state is virtually freezed for any computation as soon as the computation cycle is started. The change in topology is taken into account only after supplying the result of all previously started operations.

Here the user gets the result for query of instant "t" , at time instant "t+k" and this additional time is actually the computation time and result propagation time.

Example:

Let us consider the following query:

SELECT MAX(temperature)
FROM sensor ;

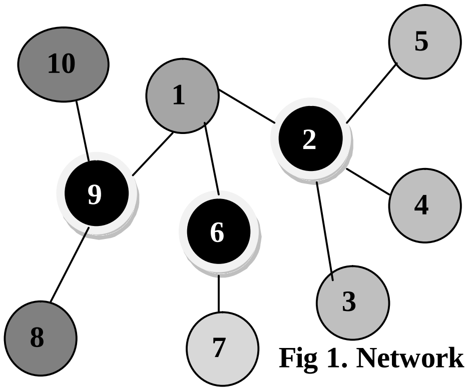

Fig 1. Network Tree

S = {(Node number , Temperature Reading)}

S={(1,25),(2,27.5),(3,26),(4,26.5),(5,26),(6,26),(7,28),(8,29),(9,27),(10,30) }

S be the set of ordered pairs (node number, Temperature in centigrade scale).The connected nodes are in transmission range of one another.Node10is the source of MAX value. TABLE 1

Time	Node	Exchange data with nodes & their readings	Max value (m_i)	Hop count (h_i)	Source of max value (s_i)	Parent (p_i)
t+4	1	9 [temp = 30] , 2 [temp = 30] , 6 [temp = 30]	30	2	10	9

	2	1[temp = 30] , 3 [temp = 27.5] , 4 [temp = 27.5], 5 [temp = 27.5]	30	3	10	1
	3	2 [temp = 30]	30	4	10	2
	4	2 [temp = 30]	30	4	10	2
	5	2 [temp = 30]	30	4	10	2
	6	1[temp = 30] , 7 [temp = 28] ,	30	4	10	1
	7	6 [temp = 30]	30	4	10	6
	8	9 [temp = 30]	30	2	10	9
	9	1[temp = 30], 10[temp = 30], 8[temp = 30]	30	1	10	10
	10	9 [temp = 30]	30	0	10	-

The TABLE 1 shows the final stage of result calculation for this network tree. At time instant " t+4" , all nodes eventually know that node 10 is the source of MAX value .Every node now knows its distance from node 10 and the necessary route to visit node 10. Actually we have shown the calculation for a single cluster consisting of 10 nodes. The WSN will consist of a large number of nodes. The same algorithm will be followed in each cluster parallely.BS will receive the MAX value for each zone (a zone is a particular geographic region which is assumed to consist of a number of clusters). The Cluster heads of each zone will communicate among themselves to find the MAX value for a zone.

4 CONCLUSION

In this paper, the main focus is on the cluster based query processing and data aggregation in WSN. All algorithms are designed after considering the general requirements of any WSN. Our proposed method is reliable as responsibility of aggregation is distributed hierarchically. We have also preserved finiteness property for MAX algorithm which is a limitation of standard distributed algorithms for MAX calculation.

5 REFERENCES

[1] Yick Jennifer, Mukherjee Biswanath, Ghosal Dipak :"Wireless sensor network survey ",Computer Networks 52 (2008) 2292–2330(science direct) ,2008.

[2] Watfa Mohamed, Daher William Al Azar Hisham : "A Sensor Network Data Aggregation Technique", International Journal of Computer Theory Engineering, Vol. 1, No. 1, April 2009.

[3] Meliou Alexandra, Guestrin Carlos, and Hellerstein Joseph, "Approximating Sensor Network Queries Using In-Network Summaries", Information Processing in Sensor Networks (IPSN,2009).

[4] Fasoloy Elena , Rossiy Michele, Widmer J¨org Zorziy Michele:" In-network Aggregation Techniques for Wireless Sensor Networks: A Survey " , IEEE Wireless Communications, Vol. 14, No. 2, April 2007, pp. 70-87.

[5] Prabh Shashi :" Data aggregation and data dissemination in Wireless Sensor networks. " www.cs.virginia.edu/~ksp2q/publications