A Survey on Big-Data Gathering using Mobile Collector in Densely Deployed Wireless Sensor Network

Amruta S. Pattanshetti
ME -II(Computer Engineering)
PVPIT
Pune, India

Mr. N D. Kale
Department of Computer Engineering
PVPIT
Pune, India

***Abstract* --- With tremendous growth of information and communication technology (ICT) Wireless sensor network has major contribution in big data gathering. Even if data generated by individual sensor is not significant, the overall data generated by all sensors in the network is contributed for generation of significant portion of big data. So energy efficient big data gathering is challenging task in the densely deployed wireless sensor network. Also cluster formation before data collection from sensors in the network is additional challenge. Recent research addressed these challenges with mobile sink, which in turn raise the challenge of determining the sink node's trajectory. In this paper we have proposed new solution, M-mobile collector based data gathering with network clustering based on improved Expectation maximization technique. Mobile collectors traverse a fixed path to collect data from cluster centroids and sensors in the clusters. Collected data is transferred amongst M-collectors to reach to the static sink node. Also we derive optimal no of clusters to minimize the energy consumption.**

Keywords --- Big data, Wireless Sensor Networks (WSNs), clustering, optimization, data gathering, and energy efficiency.

Challenges:

1. The energy-efficient big data gathering in the densely distributed sensor networks is, therefore, a very challenging research area.
2. Cluster formation prior to collection of data.
3. Mobile sink presents additional challenges such as determining the trajectory of the sink node.

I. INTRODUCTION

Now a day's development of Information communication and Technology has contributed to fast growth of volume of data. The concept of big data has emerged as new and widely accepted trend which is attracting much attention from the different research, academic and government industries. The current services such as social networks, cloud storage, network switches are generating significant portion of big data. It is predicted that peta bytes of data will be generated by sensors and RFID devices [1]. Collection of the large volume and wide variety of the sensed data is a critical task. As number of important domains of human endeavor are becoming increasingly dependent on these remotely sensed information [5]. The total volume of data generated by an individual sensor is not that significant; individual sensor requires a lot of energy to relay the data generated by surrounding sensor nodes. In case of dense sensor networks, the life time of sensors will be very short because each sensor node relays a lot of data generated by neighboring sensors.

There are two problems in gathering the data sensed by sensors in the network [1].

1. The network is divided to some sub-networks because of the limited wireless communication range.
2. The wireless transmission consumes lot of energy of the sensors. Even though the total volume of data generated by an individual sensor is not that significant, each sensor requires a lot of energy to relay the data generated by surrounding sensors.

Sensor nodes are resource constrained in term of energy, processor and memory and low range communication and bandwidth [2]. Sensor nodes utilize their energy during receiving, transmitting and relaying the packets. So, designing routing algorithms that maximizes the life time until first battery expires is an important consideration.

However, cluster heads will consume more energy than other sensor nodes [2]. Due to continuous resource consumption problem failing cluster head and M-collector may arise. In order to tackle this problem some resource which reach sensor must be used which will make sensor network heterogeneous.

Wang et al. "Mobile Sinks (MSs) are mobile nodes which are the destination of messages originated by sensors, i.e., they represent the endpoints of data collection in WSN-MEs. They can either autonomously consume the collected data for their own purpose or make them available to remote users by using a long range wireless Internet connection."

II. RELATED WORK AND MOTIVATION

Daisuke Takaishi et al. in [1] ”The big data is very difficult to capture, form, store, manage, share, analyse, and visualize by the conventional database tools. Furthermore, the main characteristics of big data are namely variety, volume, and velocity. The main objective of the work was to enable seamless exchange of feeds from large numbers of heterogeneous sensors.”

In sensor network minimizing data transmission is difficult in a distributed clustering algorithm. If we divide the WSN into sub networks, a node may not be able to have all the information about all the nodes, thus optimization cannot be achieved. We need to implement the minimum energy clustering. The centralized clustering algorithm, which is supervised by a super node, is much suitable for the mobile sink scheme.

R. C. Shah et. al. [3] “Data MULEs follow the basic steps for all the mobile sink schemes. Firstly, the sensor nodes are divided into many clusters. Secondly, a route for patrolling the cluster is decided. The work in [3] assumes a simple data collection scheme whereby the mobile sink node divides sensor nodes into grids regardless of the sensor nodes’ location, and patrols the grids by using random walk between the neighboring grids. However, this type of clustering, which does not take into account the nodes’ location, may result in inefficient data gathering.”

Heinzelman et. al. [4] “Low-Energy Adaptive Clustering Hierarchy (LEACH) [4] is one of the most famous clustering algorithms in WSNs which uses the static sink node. In this type, the clustering algorithm is executed by each of the sensor nodes. Sensor nodes exchange information about their residual energies, and the nodes that have higher residual energies are given a higher probability to become the cluster head.”

Earlier research works on sensor node clustering algorithms demonstrates that the increasing number of clusters reduces energy consumption for data transmission. Definitely, the idea holds since increasing the number of clusters decreases the cluster-sizes and shortens the transmission length. Some researchers consider that certain limitations on the number of cluster can be decided by other factors.

III. PROPOSED MODEL

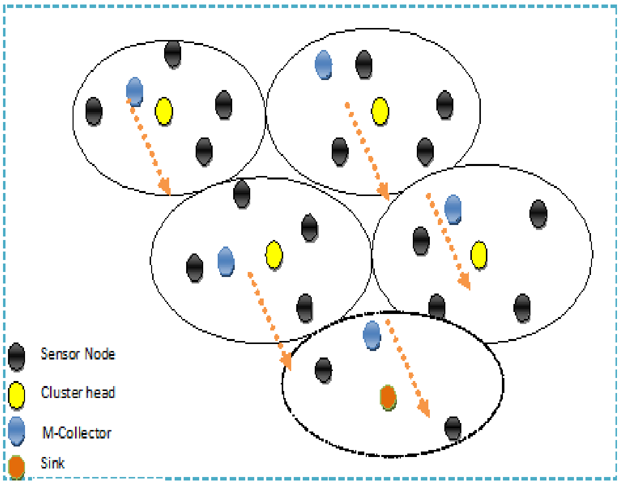

Fig 1: Proposed Model –M-Collector Based Big-Data gathering.

Fig. 1 shows the proposed model for big data gathering in wireless sensor network. Once the sensors are deployed, firstly the clustering formation is done using modified expectation maximization technique. After formation of cluster, one sensor is elected as cluster head. Cluster head will be responsible for data collection from all nodes in the cluster. In order to save the battery power of cluster heads, cluster heads are chosen rotationally from all nodes in the same cluster. In order to collect data from all cluster heads a resource rich Mobile collector is added to each cluster. M-Collector traverse a fixed path from one cluster to another cluster to send collected data to static sink via intermediate M- Collector.

A. Overview of the modified EM-algorithm

EM algorithm assumes that nodes are distributed according to Gaussian mixture distribution [1],

$$p(x) = \sum_{k=1}^{k} \prod_{k} N(x|\boldsymbol{\mu}k, \boldsymbol{\Sigma}k) \qquad (1)$$

Where,
K : Total number of clusters and
Πk: indicate the total number of clusters and the mixing coefficient of the *k*th cluster.

N(x|μ,∑) =

$$\frac{1}{(2\pi)|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathrm{x}-\mu)^{\mathrm{T}}\ \Sigma^{-1}\ (\mathrm{x}-\text{⍰})\right\} \ (2)$$

Where,
X: The position vectors of all nodes.
μk: the position vector of centroid of cluster *k*.
Σ*k*: 2×2 covariance matrix of the *k*th cluster.

At first EM algorithm [1] calculates each node's value of degree of dependence that is referred to as responsibility. The value of responsibility shows how much a node depends on a cluster. Following equation gives *n*th node's value of degree of dependence on *k*th cluster:

$$\gamma nk = \frac{\prod_k N(x_n \mid \mu_k, \Sigma_k)}{\sum_{j=1}^{k} \pi_j N(x_n \mid \mu_j, \Sigma_j)} \qquad (3)$$

The responsibility takes values between 0 and 1.

In [1] the second step, the EM algorithm evaluates K weighted center of gravity of a 2-dimensional location vector of nodes. This assessment uses the responsibility value as weight of the nodes. At the third step, the locations of the cluster centroids are changed to the weighted centers of gravity evaluated in the second step.
EM algorithm evaluates the value of the log likelihood as shown below.

$$P = \ln p(X|\mu, \sum, \prod)$$

$$= \sum_{n=1}^{N} \ln\{\sum_{k=1}^{k} \pi_k N\,(x_n|\mu_k, \Sigma_k)\} \qquad (4)$$

B.Proposed clustering method

We have used the EM algorithm for clustering in widely densely deployed wireless sensor network. WSNs, which have high variety and high volume of data, we need to consider "groups" [1], which refer to sets of nodes that can communicate with each other. Hence, nodes that cannot communicate with each other belongs to different groups. To gather the data from all the nodes, the number of clusters must be set to more than the number of groups [1].

As mentioned in [1], at first, the mobile sink sets the cluster centroids, μ, to random locations. Using a random position vector of cluster centroids, the communication distances of each of the node in the cluster to the respective cluster centroids, Dnk is calculated. The mixing coefficient π and the covariance matrix Σ is calculated.
After completion of the cluster initialization phase, [1] our proposed method selects a group g that has the largest value of proportion of number of nodes to the number of clusters in a group g, shown as:

$$v_g = \frac{K_g}{N_g} \qquad (5)$$

Group that has the highest value of vg, our method picks up all nodes that belong to group g and updates this node's responsibility value γnk. The responsibility value reflects how much node n belongs to cluster k. With the help of values of updated responsibility γnk, cluster centroids μ, the covariance matrix Σ are re-calculated and the number of nodes which belongs to kth cluster is calculated as shown in the following equation,

$$N_k = \sum_{x_n \in x} \gamma_{nk} \qquad (6)$$

These calculations are executed again and again until the difference between the newly calculated P and previously calculated P becomes smaller than small number ε.

Algorithm1. Proposed clustering algorithm

```
Initialize cluster centroids μ to random locations.
  Calculate cluster's parameters ∏ and ∑.
  Calculate Dnk and P.
   While |P – P^new| < ϵ do
Select a group g which has the biggest value v_g.
              for k ϵ K_g do
               for n ϵ N_g do
Calculate nth node's responsibility value γ_nk.
end for
Calculate number of nodes belong to cluster, N_k.
Update the cluster's parameters ∏, μ and ∑ by using N_k.
end for
Evaluate the log likelihood P^new.
End while
Return cluster centroids, μ, covariance matrix, ∑ and the
number of nodes that belongs to each cluster.
```

IV. DATA GATHERING PROCEDURE USING THE PROPOSED CLUSTERING TECHNIQUE

A. Data gathering procedure using the proposed clustering technique:

After clustering, the mobile collectors patrol the cluster and collect data from sensors and centroids of the cluster. Reducing the delay generated by the mobile sink is the main objective behind introducing the Mobile collector in wireless network. mobile collectors collect data from centroids and sensor nodes in the network by traversing a fixed length path. As mobile collectors traverse fixed length path which reduce the complexity generated by mobility of collectors. mobile collectors travel from one cluster to another cluster to transfer data from one mobile collector to another to reach to the static sink.

Steps To Gather Data:

1. Sensor nodes will forward data to the cluster head(or M-Collector depending upon availability of M-Collector in the coverage).
2. Mobile Collector belonging to each cluster will travel its fixed path , when there is connectivity between cluster head and M-collector, cluster head transfer data to the M-Collector.
3. M-Collector also receives data from other sensors in the cluster.
4. When M-Collector gets connectivity with another M-Collector in the direction of sink, it transmits collected data to that M-Collector.
5. Likewise all M-Collector transfers' data towards sink node.
6. Sink receives data from nearby M-collectors.

B. Computing the optimal number of clusters

To obtain the optimal number of clusters, we need to define objective function, W(*K*), which can be defined as the sum of energy consumption in one cycle of M-Collector patrol as follows.

$$W(K) = D_{Req} E_{Req}(K) + D_{Dat} E_{Dat}(K) \qquad (7)$$

Where,
E_{Req} (*K*): The sums of the square of transmission distance of data requests.
E_{Dat} (*K*): The sums of the square of transmission distance of data messages.
D_{Req}: The data size of data request messages.
D_{Dat}: The data size of data message.

*E*Dat (*K*) is evaluated according to the following equation:

$$\text{EDat} = \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{h=1}^{H_{nk}} \gamma_{nk} \cdot l_h^2 \qquad (8)$$

Hnk is the hop count from the nth node to the kth cluster centroid and lh is communication distance of each hop.

The optimal number of clusters Kopt is defined by the following equation.

$$K_{opt} = \max(G, \arg_K \min(W(K)) \qquad (9)$$

G: Group.

We consider group of node that has Ng nodes and Kg cluster centroids. Data request message is sent from every cluster and every node re-broadcasts it one time. The total required energy to transmit data request massage is formulated as follows:

$$E_{Req} = \sum_{g=1}^{G} K_g N_g R^2 \qquad (10)$$

R is the maximum transmission range of the sensor nodes.

If there is no imbalance of location of cluster centroids, the number of nodes that belongs to each cluster is the same.

$$\frac{K_g}{N_g} = \frac{K}{N} \qquad (11)$$

Here, if the number of nodes is larger than 1, the connectivity, *C*, can be approximated as follows:

$$C = \frac{\sum_{g=1}^{G} N_g(N_g - 1)}{N(N-1)} \div \frac{\sum_{g=1}^{G} N_g^2}{N^2} \qquad (12)$$

From all above equation we can calculate the energy required.

$$E_{Req} = K N R^2 C \qquad (13)$$

Thus, it can be seen that the number of clusters has a significant effect on connectivity. By calculating the required energy for data transmission as in (14), and data request transmission as in (9), the optimal number of clusters, (10), can be calculated by using (8) and (10).

CONCLUSION

In this paper, we carried out the survey to investigate the challenging and demanding issues pertaining to collection of the Big Data generated by densely deployed wireless sensor networks. We also investigated the issue related to the mobile sink node's trajectory used in the mobile sink scheme and the cluster formation. To address these challenges we proposed a new scheme where we used Mobile collector based data gathering with network clustering based on improved Expectation Maximization technique. Mobile collectors traverse a fixed path to collect data from cluster centroids and sensors in the clusters. Collected data is transferred amongst M-collectors to reach to the static sink node. Also we derive optimal no of clusters to minimize the energy consumption.

ACKNOWLEDGEMENT

I wish to thank N D.Kale Department of Computer Engineering for his guidance and timely help given to me in completion of my survey paper.

REFERENCES

[1] Daisuke Takaishi , Hiroki Nishiyama "Towards Energy Efficient Big Data Gathering in Densely Distributed Sensor Networks" DOI 10.1109/TETC.2014.2318177, IEEE Transactions on Emerging Topics in Computing.

[2] R. C. Shah, S. Roy, S. Jain, and W. Brunette, "Data MULEs: MULEs: modeling and analysis of a three-tier architecture for sensor networks," *Ad Hoc Networks*, vol. 1, no. 2-3, pp. 215 - 233.2003

[3] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, Balakrishnan, "Energyefficient communication protocol for wireless microsensor networks," in *Annual Hawaii International Conference on sytem Sciences*, vol. 2, Jan. 2000

[4] Bisio and M. Marchese, "Efficient satellite-based sensor networks for information retrieval," *IEEE Systems Journal*, vol. 2, no. 4, pp. 464–475, Dec