International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.2, 2015

RESEARCH ARTICLE

ISSN: 2321-7758

BIGDATA GATHERING AND MAP REDUCE FRAMEWORK FOR UNDERWATER WSN ENVIRONMENT

M.CHITRA GANESH[1], K.KALAIVANI[2]

[1]PG Computer Science and Engineering, Arasu Engineering College, Kumbakonam

[2] Assistant Professor, Department of Computer Science, Arasu Engineering College, Kumbakonam

Article Received: 11/03/2015

Article Revised on; 15 /03/2015

Article Accepted on:19/03/2015

M.CHITRA GANESH

K.KALAIVANI

ABSTRACT

To maximize network lifetime in Wireless Sensor Networks (WSNs) the paths for data transfer are selected in such a way that the total energy consumed along the path is minimized. To support high scalability and better data aggregation, sensor nodes are\often grouped into disjoint, non-overlapping subsets called clusters. Clusters create hierarchical WSNs which incorporate efficient utilization of limited resources of sensor nodes and thus extends network lifetime. As society develops quickly and people attach more importance to environmental protection, it is of great research significance to intelligently monitor the environment. Based on the sensor network, doing research on live monitor of water quality, first of all, this paper designs data collecting nodes under water ,which realize communication and organization by means of sound waves; and then it adds dada collecting nodes which automatically form networks. In the entire, sensor node spread in the critical area in an unplanned manner. Sensor network are the collection of sensor node which co-operatively send sensed data to sink node. After sensing the each sensor to deployed densely data to the base station so that a WSN can successfully operate in the presence of component failures or Densely Traffic Collusion Attack. In this environment it's very difficult to continuously surveillance. It's currently research potential in securing data aggregation in the WSN. It offers various efficient ways for collecting sensor data and managing multiple sensor nodes by launching specific map reduce applications on sensor nodes which upload data of sensor nodes to DFS and retrieve sensor data periodically from DFS

Index Terms- optimization, energy consumption, routing, clustering,

©KY Publications

I.INTRODUCTION

While the promise of Big Data is real -- for example, it is estimated that Google alone contributed 54 billion dollars to the US economy in 2009 -- there is currently a wide gap between its potential and its realization. Whereas in previous generations, a relatively small volume of analog data was produced and made available through a limited number of channels, today a massive amount of data is regularly being generated and flowing from various sources, through different channels, every minute in today's Digital Age. It is the speed and frequency with which data is emitted and transmitted on the one hand, and the rise in the number and variety of sources from which it emanates on the other hand, that jointly constitute the data deluge. The amount of available digital data at the global level grew from 150 exabytes in 2005 to 1200 exabytes in 2010.7 It is projected to increase by 40% annually in the next few years, 8 which is about 40 times the much-debated growth of the world's population. This rate of growth means that the stock of digital data is expected to increase 44 times between 2007 and 2020, doubling every 20 months [1].

Twitter data, mobile phone data, online queries, etc. These types of data can firmly be called 'Big Data', as popularly defined (massive amounts of digital data passively generated at high frequency). And, while these streams of information may not have traditionally been used in the field of development, but they could prove to be very useful indicators of human well-being Therefore, we would consider them to be relevant Big Data sources for development. Big Data for Development sources generally share some or all of these features:

(i) Digitally generated – i.e. the data are created digitally (as opposed to being digitized manually), and can be stored using a series of ones and zeros, and thus can be manipulated by computers.

(ii) Passively produced – a by product of our daily lives or interaction with digital services.

(iii) Automatically collected – i.e. there is a system in place that extracts and stores the relevant data as it is generated.

(iv) Geographically or temporally trackable – e.g. mobile phone location data or call duration time.

(v) Continuously analysed – i.e. information is relevant to human well-being and development and can be analysed in real-time.

Gathering the large volume and wide variety of the sensed data is, indeed, critical as a number of important domains of human endeavor are becoming increasingly reliant on this remotely sensed information. For example, in smart-houses with densely deployed sensors, users can access temperature, humidity, health information, electricity consumption, and so forth by using smart sensing devices. In order to gather these data, the Wireless Sensor Networks (WSNs) are constructed whereby the sensors relay their data to the "sink". However, in case of widely and densely distributed WSNs There are two problems in gathering the data sensed by millions of sensors.

First, the network is divided to some sub-networks because of the limited wireless communication range. For example, sensors deployed in a building may not be able to communicate with the sensors which are distributed in the neighboring buildings. Therefore, limited communication range may pose a challenge for data collection from all sensor nodes.

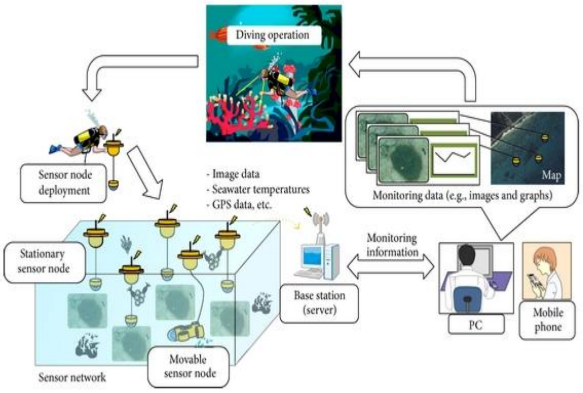
Second, the wireless transmission consumes the energy of the sensors. Even though the volume of data generated by an individual sensor is not significant, each sensor requires a lot of energy to relay the data generated by surrounding sensors. Especially in dense WSNs, the life time of sensors will be very short because each sensor node relays a lot of data generated by tremendous number of surrounding sensors. In order to solve these problems, we need an energy-efficient method to gather huge volume of data from a large number of sensors in the densely distributed WSNs.

Fig. 1, the big data comprises high volume, high velocity, and high variety information assets [3], which are difficult to gather, store, and process by using the available technologies. The variety indicates that the data is of highly varied structures (e.g. data generated by a wide range of sources such as Machine-to-Machine (M2M), Radio Frequency Identification (RFID), and sensors) while the velocity refers to the high speed processing/analysis (e.g., click-streaming, fast database transactions, and so forth). On the other hand, the volume refers to the fact that a lot of data needs to be gathered for processing and analysis. Water is the source of human life, water environment monitoring is the management and protection of water resources is an important meansin our country, the shortage of water resources,water pollution is serious, how efficient, real-timeaccess to water environment parameters, theresearch and development of new monitoring

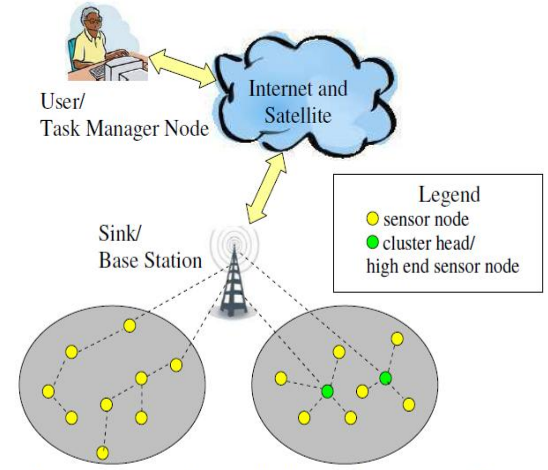
method of water environment, water environmentmanagementand protection has become an important task of water, environmental monitoringis important.

To achieve energy-efficient data collection in densely distributed WSNs, there have been many existing approaches. For example, the data compression technology [7] is capable of shrinking the volume of the transmitted data. Although it is easy to be implemented, the data compression technology requires the nodes to be equipped with a big volume of storage and high computational power. In addition, the topology control technology can evaluate the best logical to

pology and reduce redundant wireless transmissions [8], [9]. When the redundant wireless transmissions are reduced, therequired energy for having high remaining energy [10], [11] However, these technologies are not able to deal with the divided networks problem. it provides a flexible management scheme for sensor node by reconfiguring a firmware or updating configurations and data formats of sensor nodes based on mapreduce framework.



Fig(1) Big data gathering in under water WSN



(a) Flat sensor network; (b) Clustered sensor network.

Figure. 2. Major trends of big data gathering

II. RELATED WORKS AND OUR MOTIVATION

Big data and its analysis are at the core of modern science and business. Identified a number of sources of big data such as online transactions, emails, audios, videos, images, click-streams, logs, posts, search queries, health records, social networking interactions, mobile phones and applications, scientific equipment, and sensors. Also, it was pointed out, in their work, that the big data are difficult to capture, form, store, manage, share, analyze, and visualize via conventional database tools. Furthermore, the three main characteristics of big data, namely variety, volume, and velocity are discussed in those works that were briefly described in Section I.

The concept of big data is stimulating a wide range of industry sectors. Specific examples of big data generated by sensors were provided in the report. For instance, manufacturing companies usually embed sensors in their machinery for monitoring usage patterns, predicting maintenance problems, and enhancing the product quality. By studying the data streams generated by the sensors embedded in the machinery allow the manufacturers to improve their products. The numerous sensors deployed in the supply lines of utility providers generate a huge volume of data, which are consistently monitored for production quality, safety, maintenance, and so forth. Other examples of sensors generating a bulk of the big data consist in electronic sensors monitoring mechanical and atmospheric conditions. In addition, sensors used for healthcare services (to monitor bio-metrics of the human body, patient’s conditions, healthcare diagnoses, treatment phases, and so forth) are identified to be a rich source of big data in the report presented in [12]. However, how to gather the sensed data from these numerous sensors in an energy efficient manner remained beyond the scope of the report.

The work in [13] presented a cloud-based federated frame-work for sensor services. The main objective of the work was to enable seamless exchange of feeds from large numbers of heterogeneous sensors. Various applications using big data generated by densely distributed WSNs have also emerged in literature. In addition, in [14] , big data in terms of the healthcare information (e.g., blood pressure and heart rate) sensed by numerous sensors are used to realize remote medical care services. Furthermore, patient’s location information are used to arrange prompt dispatch of ambulances. Large volume of data gathered from location-sensors attached to animals enabled researchers to observe various animal habitats. Because widely and densely distributed WSNs collect various types of data, the overall data which are gathered is, indeed, overwhelming. To efficiently gather the big data generated by the densely distributed WSNs is, however, not an easy task since the WSNs may be divided into sub-networks because of the limited wireless communication range of the sensors. Apache Hadoop is an open-source implementation of mapreduce framework that supports the running of appli-cations on large clusters of commodity hardware.

In conventional research works, data gathering using the mobile sink in WSNs has been widely studied in literature. Data Mobile Ubiquitous LAN Extensions (MULEs) is the one of the most prominent and earliest studies on the mobile sink scheme. Data MULEs follow the basic steps of all the mobile sink schemes. First, it divides sensor nodes into clusters. Second, it decides the route for patrolling each cluster. The work in 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 is not based on the nodes location, might result in inefficient data gathering. If there is no sensor node remaining in the cluster, patrolling the empty cluster results in waste of time and degraded efficiency. Also, patrolling based on randomness might result in unbalanced visits to clusters with different numbers of sensor nodes. Thus, the mobile sink might fail to collect information.

Low-Energy Adaptive Clustering Hierarchy (LEACH) is one of the most famous clustering algorithms in WSNs using the static sink node. In LEACH, the clustering algorithm is executed by the each sensor node. Sensor nodes exchange information on their residual energies, and the nodes with higher residual energy are given a higher probability of becoming a cluster head. By doing periodical re-clustering, energy consumption of each node becomes eventually equal. However, LEACH still has several shortcomings. For example, because LEACH is based on the assumption that each node can communicate with all other nodes, the WSNs deployed in wide areas are not able to use the algorithm. Most of the distributed algorithms like LEACH naturally consider the limitation of the nodes communication range. *K*-hop Overlapping Clustering Algorithm (KOCA) and *k*-hop connectivity ID (*k*-CONID) are examples of the distributed clustering algorithms. Authors of KOCA focused on multiple overlap-ping clusters, and designed the KOCA algorithm based on a probabilistic cluster head selection and nodes location. The *k*-CONID algorithm is also a probabilistic algorithm. The nodes exchange their random IDs with each other, and the node that has the minimum ID within k-hop is selected as a cluster head.

In WSNs, minimizing data transmission is difficult for a distributed clustering algorithm. If a WSN is physically divided into sub-networks, a node cannot possess information about all the nodes in the WSNs. Thus, the algorithm cannot achieve optimization. To realize minimum energy clustering, we need to use the centralized clustering algorithm. Moreover, the centralized clustering algorithm, which is conducted by a super node, is suitable for the mobile sink scheme. Power-Efficient Gathering in Sensor Information Systems (PEGASIS) and KAT mobility (*K*-means And TSP mobility) are one of the centralized clustering algorithms PEGASIS algorithm constructs chain clusters of nodes based on location, and repeats cluster head selection. PEGASIS algorithm considers the limitation of the communication range, and achieves uniform energy consumption. However, the algorithm still does not achieve minimization of energy consumption because the clustering algorithm uses greedy algorithm. KAT mobility divides the nodes into clusters by using k-means algorithm. Because *k*-means algorithm is the centralized clustering algorithm based on the nodes location, the clustering result is closer to the total optimization. While the result is the optimal cluster that reduces energy consumption, the KAT mobility algorithm is designed without considering the communication range limitation. Therefore, the mobile sink might fail to collect information from all nodes.

Contemporary research on the sensor node clustering algorithm can be classified into three types, namely centralized algorithms without considering nodes information (i.e., location or communication range), distributed algorithms without considering nodes information, and distributed algorithms that consider the nodes location and communication range. However, to achieve both minimization of data transmission and data collection from all the nodes, we need to use a centralized algorithm, which considers the nodes location and communication range. Unlike existing algorithms, our proposed clustering algorithm achieves both minimization of data transmission and data collection.Earlier research works on sensor node clustering algorithms demonstrates that the increasing number of clusters reduces energy consumption for data transmission. Certainly, the idea holds since increasing the number of clusters decreases the cluster-sizes and shortens the transmission length. Some re-searchers consider that certain limitations on the number of cluster can be decided by other factors. For example, in the limitation is the maximum acceptable latency of data collection. While these limitations are realistic assumptions, they do not consider the energy consumption for data requests. In our paper, we first focus on the effect of data request messages by increasing the number of clusters. Based on a simple and common data gathering model of the densely distributed WSNs, we demonstrate that the number of data request messages has a noticeable impact on the energy consumption of the sensor nodes. When the connectivity of the nodes becomes bigger, the impact becomes larger also. In this paper, we present how to evaluate the optimal number of clusters to minimize the energy consumption of the sensor nodes. monitoring indicators improve, monitoring environment is increasingly complex background, the Sensor network technology into the solution of water environment automatic monitoring of an ideal scheme. Sensor network is a kind of important network monitoring tools, according to environment independent completion of various monitoring task "smart" system. Wireless sensor network based on the water environment monitoring system has the advantages of: 1, low cost; 2, monitoring points distribution range; 3, network flexible structure; 4, little influence on the surrounding ecological environment etc..

III. CLUSTERING-BASED BIG DATA GATHERING IN DENSELY DISTRIBUTED WSN

In this section, we first outline the clustering problem in WSN using mobile sink and the challenges in solving this problem. After that, we introduce the considered network model and the overview of EM algorithm for clustering. Based on EM algorithm, we proposed our clustering method and the procedure to gather data using the proposed method.

A. Clustering problem

When considering the scheme of data gathering in WSN using mobile sink, the biggest challenge in reducing energy Consumption is how to decide the location where data gathering is conducted. In other words, this problem has same meaning as answering the following two questions. 1) What is the best algorithm for dividing nodes into clusters? 2) How many clusters is optimal in terms of reducing energy consumption? As we assume that required energy for data transmission of node is proportional to the square of trans-mission distance, the best consumption for data transmission must minimize the sum of square of data transmission distance in a network. EM algorithm is powerful and well-known tool to solve the clustering problem by repeatedly calculate the simple math formula. Since the EM algorithm can minimize the sum of square of distance between every node and cluster centroid, we adopt EM algorithm over the 2-dimensional Gaussian mixture distribution. However, there is a limitation of the maximum communication range in the realistic situation. Not all nodes can connect to each other and also to the cluster centroid. Nodes that cannot directly communicate with the cluster centroid need to communicate in a multi-hop manner. In multi-hop

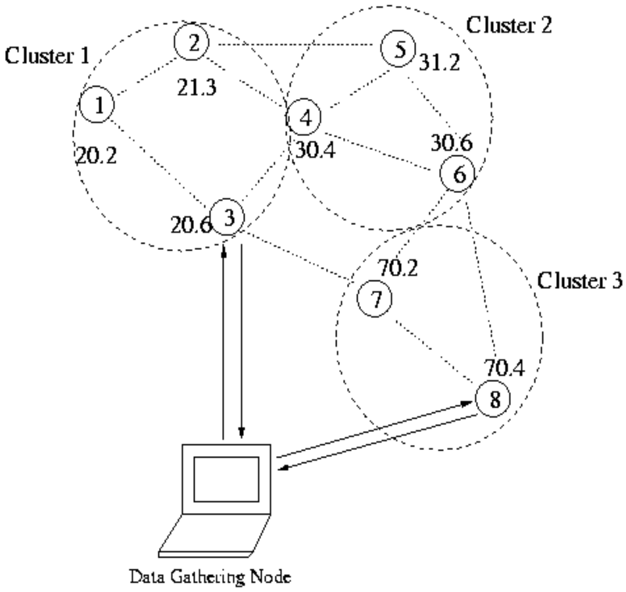


Figure. 3. An example of the considered network

communication, communication distance is a sum of distance between nodes in multi-hop path. Therefore, as shown in Fig. 2, communication distance is different from direct distance. However, the EM algorithm minimizes the sum of square of direct distance, not communication distance.

B. Considered network model

In this paper, we consider a network which consists of a mobile sink and many sensor nodes spread within a limited field. Every sensor node knows its location by using localization technology, and the mobile sink knows all nodes locations. Regardless of being a sink or the sensor, a node has a limited communication range *R* and communication is always successful if it is within *R*. The mobile sink node patrols the cluster centroid that is calculated to minimize energy consumption for data transmission, and collects data from sensor nodes. Sensor nodes are equipped with a buffer memory and store sensed information until mobile sink approaches the cluster centroid. The information is transferred to the sink node by multi-hop fashion. In this paper, we assume a densely distributed WSN in a large area such as schools, urban areas, mountains, and so forth and thus WSNs are divided into sub-networks. Fig. 2 shows a simple example of the assumed network. *N* sensor nodes illustrated by circles are distributed in the target *L* × *L* area. *K* centers of clusters illustrated by filed circle are to be visited by mobile sink. A solid area and a dotted circle means group of nodes and cluster, respectively. In this paper, group means a set of nodes that can communicate with each other. The nodes that belong to different groups cannot communicate with each other due to being far away. There are *G* groups in the field, and N_g and K_g refers to the number of nodes and number of clusters in the *g*th group, respectively. The number of groups is calculated by the nodes location and communication range *R*. In case of Fig. 2, N_1 = 7 and K_1 = 2 because there are 7 nodes and 2 clusters in group 1.

C. Overview of EM algorithm for clustering

The EM algorithm is a classical clustering algorithm, which assumes that nodes are distributed according to Gaussian mixture distribution,

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right),$$

where *K* and π_k indicate the total number of clusters and the mixing coefficient of the *k*th cluster, respectively. *N* (x/*μ*, Σ) is defined as follows,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)\,|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\},$$

where x is the position vectors of all nodes. Cluster parameters, μ_k and Σ_k, are the position vector of centroid of cluster *k* and 2×2 covariance matrix of the *k* th cluster, respectively. At the first step, EM algorithm calculates each nodes value of degree of dependence that is referred to as responsibility. The responsibility shows how much a node depends on a cluster. The *n*th nodes value of degree of dependence on *k*th cluster is given by following equation.

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right)}.$$

Because of its definition, the responsibility takes values between 0 and 1. At the second step, the EM algorithm evaluates *K* weighted center of gravity of a 2-dimensional location vector of nodes. This evaluation uses the responsibility value as weight of nodes. At the third step, the locations of the cluster centroid are changed to the weighted centers of gravity evaluated in the second step.

D. Proposed clustering method

Our objective is to propose a clustering method based on the EM algorithm. In supposed widely and densely deployed WSNs, which have high variety and high volume of data, we need to consider groups, which refer to sets of nodes that can communicate with each other. Therefore, nodes that cannot communicate with each other belongs to different groups. To collect data from all nodes, the number of clusters must be set to more than the number of groups. At first, the mobile sink sets the cluster centroid, *μ*, to random locations. By using a random position vector of cluster centroid, communication distances of each node to cluster centroid, D_{nk}, are calculated. Thereafter, the mixing coefficient, *π*, and covariance matrix, Σ, are calculated. After the cluster initialization phase, our proposed method selects a group *g* that has the largest value of proportion of number of nodes to the number of clusters in group *g*, shown as follows,

$$v_g = \frac{K_g}{N_g}.$$

In the selected group that has the highest value of v_g, our proposed method picks up all nodes that belong to group g and updates these nodes responsibility value, γ_{nk}. This responsibility value reflects how much node n belongs to cluster k. By using the updated responsibility, γ_{nk}, cluster centroid, μ, and covariance matrix, Σ, are re-calculated, and the number of nodes which belongs to kth cluster is calculated as shown in the following equation,

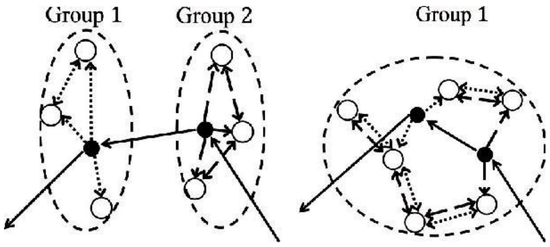
$$N_k = \sum_{x_n \in \boldsymbol{X}} \gamma_{nk}.$$

These calculations are repeatedly executed until the difference between the newly calculated P and previously calculated P becomes smaller than small number, _.

E. Data gathering procedure using the proposed clustering technique

After clustering, the mobile sink patrols every cluster centroid and collects the data from the nodes in the cluster. It is easy to see that delay is a main problem of using mobile sink in WSNs. This delay is the waiting time from data generation to data sending. Because the mobile sink moves relatively slow compared with electrical communication between nodes, the mobile sink scheme causes long delay. To shorten this delay, we need to minimize total patrolling path length. Thus, in our scheme the mobile sink patrols along Traveling Salesman Problem (TSP) path of all cluster centroid.

In our method, we consider using a typical example of them, i.e., One Phase Pull where the mobile sink node sends data request message at the cluster centroid. When a sensor node receives a data request message from cluster k, the node re-broadcasts the data request message and replies data to the neighboring node, which is the parent node in the data request tree of cluster k. Then, the node relays data messages to the sink. To minimize the total required energy to send data; all nodes send the sensed information according to the value of responsibility of the cluster. The responsibility value is calculated based on the given parameters, μ, π, and Σ, according to (3). These parameters are added to data request message and sent by the sink. Only after the sensor deployment, each node exchanges its own position vector, x, with sensor nodes belonging to same groups. Because the exchange of position vector is executed only one time after the sensor deployment, the energy consumption is not significant. As a result, when a node belongs to only one cluster, the node can send all data to the sink node. And when a node belongs to more than one cluster, the node sends data according to the responsibility of each cluster. In case of $\gamma_{n1} = 0.6$ and $\gamma_{n2} = 0.4$, if the nth node receives a data request message that is sent by the sink node at the centroid of cluster 1, the node replies 60% of data. And if the node receives data request message that is sent from cluster 2, the node sends 40% of data to the sink node at the centroid of cluster 2. By sending data using this cluster adapted Directed Diffusion scheme, we can minimize total required energy to send data.



(a) Low connectivity network (b) High connectivity network

Figure . 4. Data request flooding in low and high connectivity network

IV. DERIVING THE OPTIMAL NUMBER OF CLUSTERS IN THE PROPOSED CLUSTERING METHOD

The data gathering method presented in the previous section aims to minimize energy consumed by gathering data. However, it still has a remaining issue, which is to end the optimal number of clusters. Previous researches in literature often consider increasing the number of clusters lead to the decrease of energy consumption for data transmission. However, such researches do not take into consideration the energy consumption of data request message. In this section, we point out this problem, and show an analysis to derive the optimal number of clusters.

A. Definition of connectivity

To analyze the correlation between energy consumption and connectivity, we formulate the connectivity of nodes. In this paper, we define the connectivity as the portion of nodes that can communicate with each other.

$$C = \frac{{}^{G}_{=1} N_g \, (N_g \qquad 1)}{{}^{g} N \, (N - 1)^{-}} \quad . \qquad (7)$$

This metric takes a value between 0 and 1. When all nodes can communicate with each other, the value of connectivity is 1. If every node is isolated, the value is 0. When the mobile sink starts computing the optimal number of clusters, the mobile sink node knows every sensor nodes location. Therefore, the mobile sink can calculate the connectivity value *C* based on nodes location.

B. Data request flooding problem

In WSN using mobile sink, the sink node sends data request message to invoke data transmission from sensor nodes when it arrives at the cluster centroid. The nodes that receive data request message send the data to the sink node and broadcast data request message to their neighboring nodes. That data request message is repeatedly broadcasted until all nodes that belong to the same group receive the message. Although some nodes may receive data request message more than 2 times, they only send data and broadcast the data request message once after the first time of receiving the message. These broadcasts of data request message cause high energy consumption because the network will be flooded with redundant wireless communication. Thus, reducing data request transmission is also important for mobile sink scheme.

The impact of data request flooding issues becomes significant when connectivity becomes larger as an example shown in Fig. 3. In Fig. 3(a) and Fig. 3(b), there are two groups and one group, respectively. Six sensor nodes are scattered on the ground. Furthermore, the sink nodes traverse the two cluster centroid, and broadcasts the data request message. In the case of Fig. 3(a), nodes can only communicate with the nodes that belong to the same group. Sink node broadcasts data request message to each node at cluster 1, and these nodes broadcast the data request message. Therefore, the sum of the transmission of data request message of both cluster 1 and cluster 2 is 6. On the other hand, in Fig. 3(b), where nodes can communicate with all nodes, the data request message sent at the cluster 1 is transferred to all nodes. Furthermore, all nodes broadcast the data request message. Therefore, sum of the transmission of data request message of both cluster 1 and cluster 2 is 12.Even if number of nodes and clusters stay the same, the data request flooding problem becomes more serious with higher connectivity. Moreover, it is clearly understood that the total number of transmitted data request messages increase when the numbers of clusters increases. Because of this problem, it is necessary to find the optimal number of clusters in terms of connectivity and energy consumption.

C. Computing the optimal number of clusters

To decide the optimal number of clusters, we need to define objective function. The objective function is defined as the sum of required energy of data and data request message transmissions. Thus, the objective function, *W* (*K*), can be deÞned as the sum of energy consumption in one cycle of mobile sink patrol as follows.

$$W(K) = D_{\mathrm{Req}} E_{\mathrm{Req}}(K) + D_{\mathrm{Dat}} E_{\mathrm{Dat}}(K), \quad (8)$$

where $E_{\mathrm{Req}}(K)$ and $E_{\mathrm{Dat}}(K)$ are the sums of the square of transmission distance of data requests and data messages, respectively. D_{Req} and D_{Dat} indicate the data size of data and data request messages, respectively. $E_{\mathrm{Dat}}(K)$ is evaluated according to the following equation:

$$E_{\mathrm{Dat}} = \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{h=1}^{H_{nk}} \gamma_{nk} \cdot l_h^2,$$

$$E_{\mathrm{Dat}} = \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{h=1}^{H_{nk}} \gamma_{nk} \cdot l_h^2,$$

Where H_{nk} is the hop count from *n*th node to *k*th cluster centroid and l_h is communication distance of each hop. When *n*th node cannot communicate with *k*th centroid, we set the value of H_{nk} to 0 and the value of required energy to 0. Moreover, each node re-broadcasts each data request message one time with the maximum transmission power. Since the data transmission energy, $E_{\mathrm{Dat}}(K)$, is a decreasing function of *K* while data request transmission energy, $E_{\mathrm{Req}}(K)$, is an increasing function of *K*, there is a trade-off relationship between the Þrst and second terms in the right side of (8). By considering the condition that the number of clusters, *K*, must be greater than the number of groups, *G*, the optimal number of clusters, K_{opt}, is defined by the following equation.

$$K_{opt} = \max(G, \arg\min_{K}(W(K))).$$

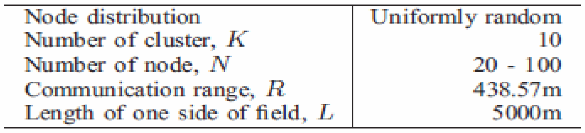
V. PERFORMANCE EVALUATION

We conducted performance evaluation by using a clustering simulator built by C++ programming language. In this section, we first evaluate the clustering efficiency. Then we evaluate total energy consumption to evaluate our proposed methd of optimizing number of clusters.

A. Efficient data collection

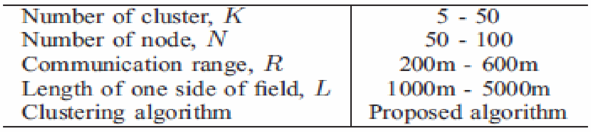
In this experiment, we measure the energy consumption for data transmissions, *E*Dat, and the efficiency of our proposed clustering algorithm by varying the number of nodes. Table I shows simulation parameters used in the first experiment. Sensors are uniformly deployed in a 5000 × 5000 square meters area. The nodes' communication range is set to 438.57 meters, and we measure *E*Dat and efficiency of our proposal Clustering by varying the number of sensor nodes. *E*Dat represented in (9) simply shows how much energy is needed for data transmissions from sensor nodes to the mobile sink.

TABLE I
ENVIRONMENTS OF 1ST EXPERIMENT



Node distribution	Uniformly random
Number of cluster, K	10
Number of node, N	20 - 100
Communication range, R	438.57m
Length of one side of field, L	5000m

TABLE II
ENVIRONMENT OF 2ND EXPERIMENT



Number of cluster, K	5 - 50
Number of node, N	50 - 100
Communication range, R	200m - 600m
Length of one side of field, L	1000m - 5000m
Clustering algorithm	Proposed algorithm

However, if locations of every centroid is far away from nodes and they cannot establish connection, *E*Dat value is calculated as 0 according to (9). This value does not reflect energy saving, but it only indicates failure of clustering. The clustering algorithm that do not consider connectivity suffer from this failure (e.g. pure EM algorithm). Thus, we also use a second metric, referred to as efficiency, which combines *E*Dat value and number of connected nodes.

$$\text{Efficiency} = \frac{(\text{Number of connected node})}{E_{\text{Dat}}}.$$

We compare our clustering algorithm with EM algorithm and *k*-COIND algorithm, which are centralized clustering algorithm and distributed clustering algorithm, respectively. Our clustering algorithm is a centralized algorithm considering connectivity, unlike EM algorithm. Figure 4(a) and 4(b) are experimental results of required energy and efficiency respectively. As can be seen from (9), (11), the energy consumption is proportional to the square of data transmission range, we measure energy in units of

m2, i.e., omitting constant variables. Figure 4(a) shows the proposed scheme and pure EM algorithm can reduce required energy significantly compared with *k*-CONID algorithm. The reason of this difference is based on difference between centralized and distributed cluster establishment. The centralized algorithm can calculate more efficient clustering than the distributed one. Our proposed scheme behaves similar to the EM algorithm, but has less energy consumption. This improvement occurs from considering connectivity and communication distance. Fig. 4(b) shows that EM algorithm is the worst clustering algorithm when node density is low.Since EM algorithm does not consider node connectivity and is centralized algorithm, when the number of nodes is low and node density is small, centroids of EM algorithm can connect only to a small number of nodes. Our proposed scheme succeeds to adapt to node density variation and minimizes transmission energy.

B. OPTIMAL NUMBER OF CLUSTERS

To evaluate our proposed method of optimizing number of clusters, we measure the energy consumption by varying the number of clusters. Energy consumption is defined as the sum of energy consumption of data transmissions and data requests. Given parameters are enumerated in Table II.

The required energy for data transmissions and data request transmissions, objective function as described in the previous subsection, energy is measured in units of m^2. Black dots are the optimal number of cluster computed by using our method. Dash lines are the area where the number of clusters is smaller than the number of groups. In those areas, a mobile sink cannot collect all data. By using our method, the optimal number of cluster is decided as 17, 23, 32, 47, and 50 when connectivity is 1.0, 0.8, 0.6, 0.4, and 0.2, respectively. These results show that traditional method which increases the number of clusters is not always the best solution to reduce energy consumption.

VI. CONCLUSION

In this project, we have presented an efficient distributed algorithm for the gathering problem in sensor networks. We design optimization formulation for the data gathering problem, and show that the optimal transmission structure depends on the data correlation and the wireless link capacities. The proposed algorithm considers both factors, while minimizing the total energy consumed in the network. Moreover, the algorithm is asynchronous and amenable to fully distributed implementations, making it feasible for practical deployment in large-scale sensor networks. To the best of our knowledge, there does not exist any previous work that addresses the gathering problem with data aggregation and wireless channel interference simultaneously, especially when a price-based strategy is employed to obtain a distributed algorithm to solve the problem. Finally, our system extends the proposed optimization framework to accommodate sensor networks with multiple sinks and arbitrary amount of data correlation.

ACKNOWLEDGMENT

Part of this work was conducted under the Research centre of Bharathiar University and additional support given by Jamal Mohamed college network laboratory manners.

REFERENCES

[1]. IBM, “Four vendor views on big Data and big data analytics: IBM,”http://www-01.ibm.com/software/in/data/bigdata/, Jan. 2012.

[2]. A. Divyakant, B. Philip, and *et al.*, “Challenges and opportunities with Big Data,” *2012, a community white paper developed by leading researchers across the United States. [Online]. Available:http://cra.org/ccc/docs/init/bigdata whitepaper.pdf*.

[3]. S. Sagiroglu and D. Sinanc, “Big data: A review,” in *International Conference on Collaboration Technologies and Systems (CTS)*, 2013

[4]. Oracle, “Big data: Business opportunities, requirements and oracle’s approach,” pp. 1–8, 2011.

[5]. I. Bisio and M. Marchese, “Efficient satellite-based sensor networks for Information retrieval,” *IEEE Systems Journal*, vol. 2, no. 4, pp. 464–475, Dec. 2008.

[6]. I. Bisio, M. Cello, M. Davoil, and et al, “A survey of architectures and scenarios in satellite-based wireless sensor networks: System design aspects,” *International Journal of Satellite Communications and Networking (IJSC)*, vol. 30, no. 6, 2012.

[7]. S. Katti, H. Rahul, W. Hu, D. Katabi, M. Medard, and J. Crowcroft, “XORs in the air: Practical wireless network coding,” *IEEE/ACM Transactions on Networking*, vol. 16, no. 3, pp. 497–510, Jun. 2008.

[8]. K. Miyao, H. Nakayama, N. Ansari, and N. Kato, “LTRT: An efficient And reliable topology control algorithm for ad-hoc networks,” *IEEE Transactions on Wireless Communications*, vol. 8, no. 12, pp. 6050– 6058, Dec. 2009.

[9]. N. Li, J. Hou, and L. Sha, “Design and analysis of an MST-based topology control algorithm,” *INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications*, vol. 4,no. 3, pp. 1195–1206, May 2005.

[10]. S. He, J. Chen, D. Yau, and Y. Sun, “Cross-Layer optimization of correlated data gathering in wireless sensor networks,” in *IEEE Communications Society Conference on Sensor Mesh and Ad Hoc Communications and Networks (SECON)*, Jun. 2010, pp. 1–9.

[11]. C. Jiming, X. Weiqiang, H. Shibo, S. Youxian, P. Thulasiraman, and S. Xuemin, “Utility-based asynchronous flow control algorithm for wireless sensor networks,” *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 7, pp. 1116–1126, Sep. 2010.

[12]. L. Ramaswamy, V. Lawson, and S. Gogineni, “Towards a qualitycentric big Data architecture for federated sensor services,” in *IEEE International Congress on Big Data (BigData Congress)*, 2013

[13]. C.-C. Lin, M.-J. Chiu, C.-C. Hsiao, R.-G. Lee, and Y.-S. Tsai, “Wireless health care service system for elderly with dementia,” *IEEE Transactions*

on Information Technology in Biomedicine, vol. 10, no. 4, pp. 696–704, 2006.

[14]. P. Ross, “Managing care through the air [remote health monitoring],” *IEEE Spectrum*, vol. 41, no. 12, pp. 26–31, 2004.

Authors

M.CHITRA GANESH, received B.E., computer science and engineering degree from Anna university Chennai and now currently pursuing M.E computer science and engineering degree from Arasu engineering college, Kumbakonam.

K.Kalaivani, received B.E., computer science and engineering degree from Oxford Engineering college and currently M.E computer science and engineering degree from Oxford engineering college, Trichy.Currently working as a assistant professor in Arasu Engineering college Kumbakonam.