IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 10, December 2014.
www.ijiset.com
ISSN 2348 – 7968

Data Gathering using New Clustering Algorithm

Jeevitha Vanmathi , Mrs. R.Kavitha M.E

Infant Jesus college of Engineering and Technology

Abstract:

The main objective of this project is big data gathering using clustering. Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process them using traditional data processing applications. This project proposes big data gathering in distributed sensor network. Data gathering is very complex. This project proposed the new clustering method to gather the data. This technique can reduce energy consumption of the sensor nodes**.** The mobile sink schemes aim to reduce wireless transmissions, the trajectory of the sink node is decided based on the sensor nodes' information.

1. Introduction:

Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications. The challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and privacy violations. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, prevent diseases, combat crime and so on Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers". Big Data is a moving target; what is considered to be "Big" today will not be so years ahead. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration." Big data is a buzzword, or catch-phrase, used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques. In most enterprise scenarios the data is too big or it moves too fast or it exceeds current processing capacity. Big data has the potential to help companies improve operations and make faster, more intelligent decisions.

Clustering is a popular strategy for implementing parallel processing applications because it enables companies to leverage the investment already made in PCs and workstations. In addition, it's relatively easy to add new CPUs simply by adding a new PC to the network. Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data. A loose

ISSN 2348 – 7968

definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called acluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning,pattern recognition, image analysis, information retrieval, and bioinformatics.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

2. Related Work:

Oracle [1] This paper presents the oracle approaches for big data. The Oracle approach has three key strengths. First, the customer is relieved of the integration involved in assembling a suitable set of hardware and software components to create a big data architecture. Second, commercial quality support is available and the entire system is supported by a single vendor. Third, and more significantly for most companies already using Oracle Database: the new big data environment connects to the existing Oracle database environment at the data management software level. For Oracle customers with an investment in Oracle software and hardware, it is a major benefit to be able to initiate work in big data, knowing that the data warehouse environment will be able to exchange and integrate data with the big data environment. Yuichi Kawamoto [2] This paper presents A Centralized Multiple Access Scheme for Data Gathering in Satellite-Routed Sensor System (SRSS). In our proposed method, the satellite partition sensor terminals into non overlapping groups and allocate time-slots to sensor terminals that have data to send by using a combination of random access control technique and fixed assignment scheme in an on-demand fashion. Moreover, the total number of groups is optimized to minimize the operation time by using the efficient allocation method along with mathematical expressions. C. Intanagonwiwat [3]This paper presents the directed diffusion for sensor network.

ISSN 2348 – 7968

Directed diffusion is data centric in that all communications is for named data. All nodes in a directed diffusion based network are application aware. This enables diffusion to achieve energy savings by selecting good paths and by catching the data in network. Angelika Herbold [4] This paper presents the resilent event detection in wireless sensor network. The contribution is the motivation, design, implementation and evaluation of ResTAG, which provides fault-tolerant aggregate queries in the popular, user-friendly TinyDB middleware These new query types return not only data values, but a quantified measure of confidence in those values. If a sensor becomes mis-calibrated or physically compromised, the fault-tolerant queries use the redundancy in the network to reduce the confidence in any report from that sensor. Varying properties such as the network density and epoch length of queries may have a significant impact on the results.

In existing, the Low-Energy Adaptive Clustering Hierarchy (LEACH) is one of the most famous clustering algorithms in WSNs using the static sink node. In clustering, the K-hop Overlapping Clustering Algorithm (KOCA) and k-hop connectivity ID (k-CONID) are used. Power-Efficient Gathering in Sensor Information Systems (PEGASIS) and KAT mobility (K-means And TSP mobility) are one of the centralized clustering algorithms are used in previous. The limitation of this is high energy. To overcome this drawbacks proposed the data gathering using a new clustering algorithm. an energy-efficient method to gather huge volume of data from a large number of sensors in the densely distributed WSNs. Based on EM algorithm, we proposed our clustering method and the procedure to gather data using the proposed method. The following chapters shows the methods and results.

3. Methodology:

3.1 System Architecture:

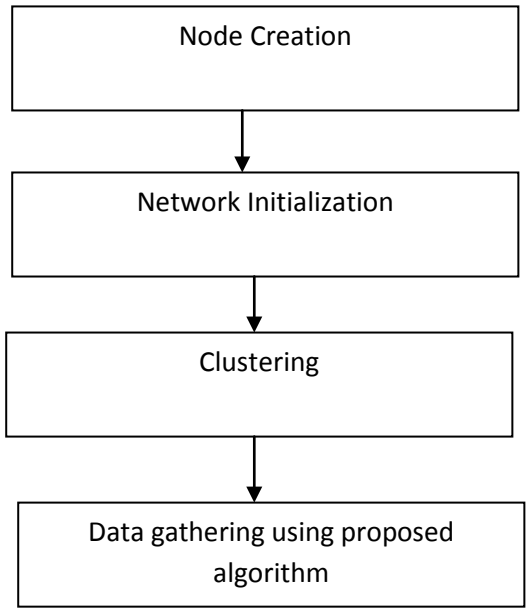


Fig: 1Block Diagram

3.2 Modules

3.2.1.Node Creation:

A node network is a series of two or more connected nodes. Once a connection between two or more nodes has been defined, all searches produce listings of configured users and resources from both local and remote nodes. This basic information is maintained on each

ISSN 2348 – 7968

computer in the node network. All calendaring data for each user and resource, however, resides only on that entity's local node, thus eliminating the space and consistency problems created by replicated databases. All exchanges of this information between nodes is done in real-time, making the scheduling of meetings with people or resources on remote nodes completely transparent to the user. When setting up a node it is important to note that the node-ID cannot be changed once the node has been created. The node can be created with its node type, size and properties. Created node can be located in the network.

3.2.2. Geographic area Model:

This model tries to realize division of geographical area in the form of cluster (sub-area) with higher node density and paths in between lower node density. The cluster is recognize as a vertices of the area graph while path as edges. The movement of cluster node could be managed with random way point model. An energy-aware key distribution scheme, which uses geographical location information.

3.2.3. Network Model:

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 centroids that are 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.

3.2.4. Clustering Algorithm:

At first, the mobile sink sets the cluster centroids, μ, to random locations. By using a random position vector of cluster centroids, communication distances of each node to cluster centroids, Dnk, 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 vg, our proposed method picks up all nodes that belong to group g and updates these node's responsibility value, γnk. This responsibility value reflects how much node n belongs to cluster k. By using the updated responsibility, γnk, cluster centroids, 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,

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

ISSN 2348 – 7968

3.2.5. 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.

Algorithm 1 Proposed clustering algorithm

Initialize cluster centroids, μ, to random locations.
Calculate clusters' parameters, π and Σ.
Calculate D_{nk} and $\mathcal{P}$.
while $|\mathcal{P} - \mathcal{P}^{\text{new}}| < \epsilon$ **do**
  Select a group g which has the biggest value v_g.
  for $k \in K_g$ **do**
    for $n \in N_g$ **do**
      Calculate nth node's responsibility value, γ_{nk}.
    end for
    Calculate number of nodes belong to cluster, N_k.
    Update the clusters' parameters, π, μ and Σ, by using N_k.
  end for
  Evaluate the log likelihood $\mathcal{P}^{\text{new}}$.
end while
Return cluster centroids, μ, covariance matrix, Σ, and the number of nodes that belongs to each cluster.

4. Results:

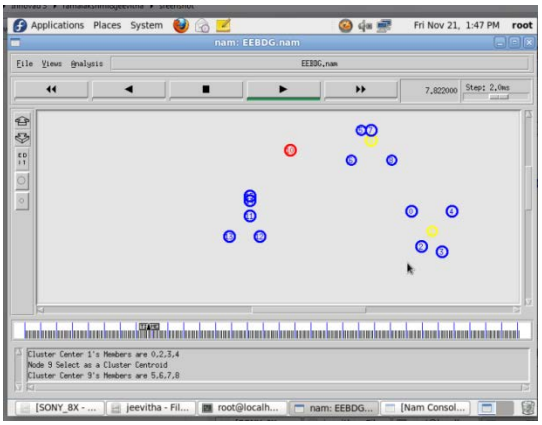


Fig: 4.1 Cluster centre

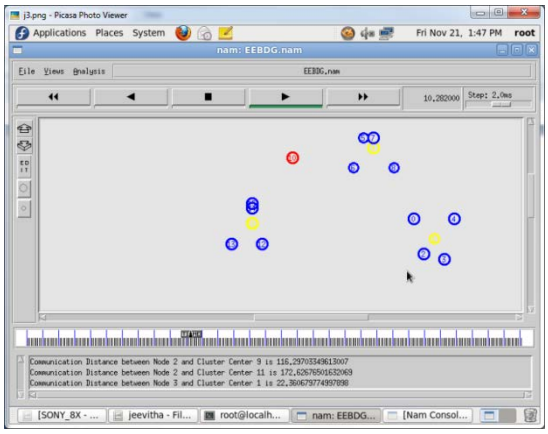


Fig: 4.2 Communication Distance

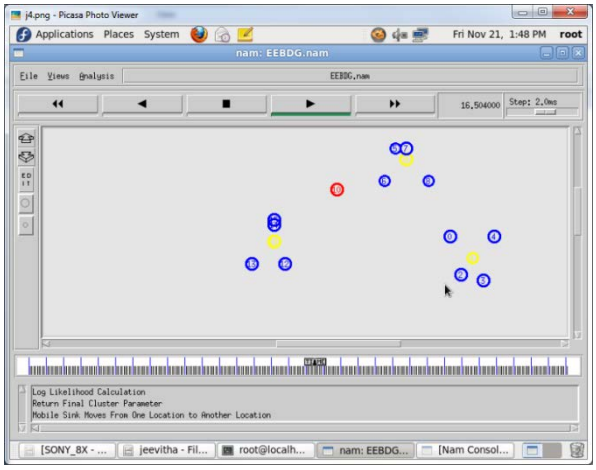


Fig: 4.3 Mobile sink move another location

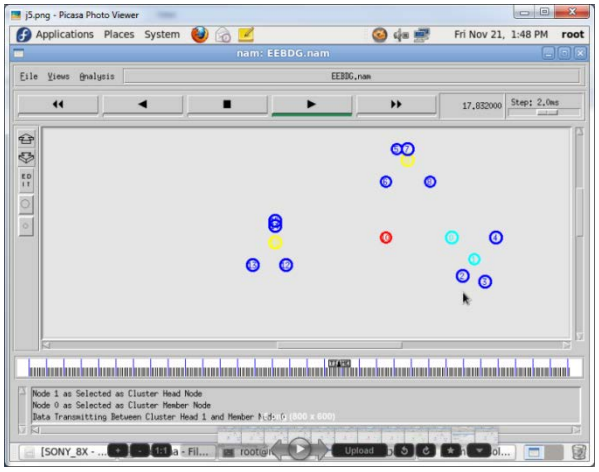


Fig: 4.4 Cluster Head selection

ISSN 2348 – 7968

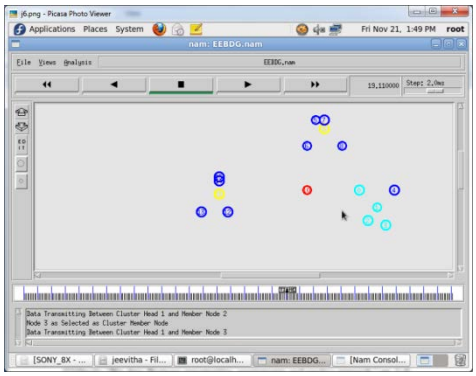
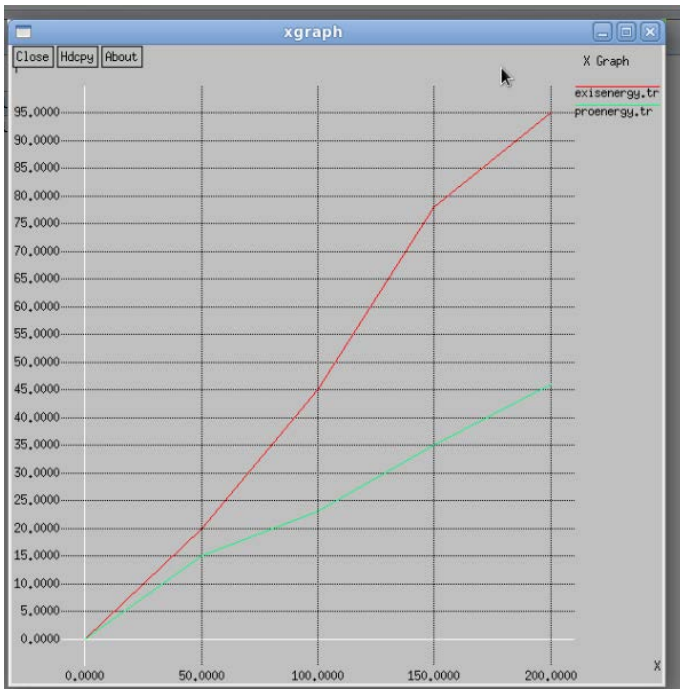


Fig: 4.5 Data Transfer



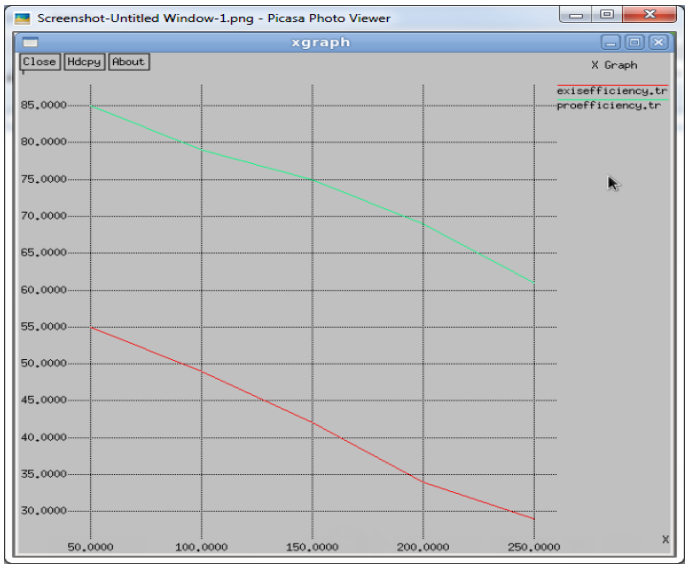
Performance Evaluation:

Performance Metrics:

Efficiency:

Accuracy of a measurement system is the degree of closeness of measurements of a quantity to that quantity's actual value

$$Detection\ Accuracy = \frac{Number\ of\ Nodes\ Detected}{Total\ Number\ of\ Nodes}\ X\ 100$$



Energy:

To define this amount of power consumption when the maximum transmission range is utilized as *P*max. Thus, the energy expanded ratio, denoted by *EER*, is expressed as follows.

$$EER = \frac{P_{\text{opt}}}{P_{\text{max}}}.$$

ISSN 2348 – 7968

5. CONCLUSION:

This project proposed big data gathering in distributed sensor network. Data gathering is very complex. This project proposed the new clustering method to gather the data. This technique can reduce energy consumption of the sensor nodes**.** The mobile sink schemes aim to reduce wireless transmissions, the trajectory of the sink node is decided based on the sensor nodes' information.

Reference:

[1] Oracle, "Big data: Business opportunities, requirements and oracle's approach," pp. 1–8, 2011.

[2] Yuichi Kawamoto, Hiroki Nishiyama, Nei Kato, Shinichi Yamamoto, Naoko Yoshimura, and Naoto Kadowaki, "A Centralized Multiple Access Scheme for Data Gathering in Satellite-Routed Sensor System (SRSS)," IEEE Global Communications Conference (GLOBECOM) 2013, Atlanta, Georgia, USA, Accepted.

[3] C. Intanagonwiwat, R. Govindan, and D. Estrin, "Directed Diffusion: a scalable and robust communication paradigm for sensor networks," in MobiCom'00 Proceedings of the 6th annual international conference on Mobile computing and networking, 2000.

[4] Angelika Herbold, Thierry Lamarr , Nirupama Bulusu and Sanjay Jha 'Resilient Event Detection in Wireless Sensor Networks"