ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com
International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 2, Issue 8, August 2016.

# Towards Energy Efficient Big Data Gathering In Densely Distributed Sensor Networks

**J. Christy Grace**
**Research scholar,**
**VELSUNIVERSITY,**
**Chennai- 117**

**Dr. S. Prasanna**
**Associate Professor**
**Head of the department**
**Department of MCA,**
**VELS UNIVERSITY**
**Chennai-117**

## Abstract

Recently, the "big data" emerged as a hot topic because of the tremendous growth of the Information and Communication Technology (ICT). One of the highly anticipated key contributors of the big data in the future networks is the distributed Wireless Sensor Networks (WSNs). Although the data generated by an individual sensor may not appear to be significant, the overall data generated across numerous sensors in the densely distributed WSNs can produce a significant portion of the big data. Energy-efficient big data gathering in the densely distributed sensor networks is, therefore, a challenging research area. One of the most effective solutions to address this challenge is to utilize the sink node's mobility to facilitate the data gathering. While this technique can reduce energy consumption of the sensor nodes, the use of mobile sink presents additional challenges such as determining the sink node's trajectory and cluster formation prior to data collection. In this paper, we propose a new mobile sink routing and data gathering method through network clustering based on modified Expectation- Maximization (EM) technique. In addition, we derive an optimal number of clusters to minimize the energy consumption. The effectiveness of our proposal is verified through numerical results.

**Index Terms—Big data, Wireless Sensor Networks (WSNs), clustering, optimization, data gathering, and energy efficiency.**

# 1. INTRODUCTION

Recent development of various areas of Information and Communication Technology (ICT) has contributed to an explosive growth in the volume of data. According to a report published by IBM. 90 percent of the data in the world was generated in the previous two years. As a consequence, the concept of the big data has emerged as a widely recognized trend, which is currently attracting much attention from government, industry, and academia. As shown in 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. Although currently used services (e.g. social networks, cloud storage, network switches, and so forth) are already generating much volume of the big data it is anticipated that more and more data will be generated by sensors/RFID devices such as thermometric sensors, atmospheric sensors, motion sensors, accelerometers, and so on. In fact, according to a report by ORACLE [4], the volume of data generated by sensors and RFID devices is expected to reach the order of pet bytes. Interestingly as shown in Fig. 1, the sensors are responsible for generation of big data in big volume and also in a wide variety

# 2. RELATED WORKS

## Cluster Creation:

WSN are autonomous systems consisting of mobile hosts that are connected by multi hop wireless links. In this cluster head (CH) is elected according to its weight computed by combining a set of system parameters (Mobility). Sensor nodes are equipped with store sensed information until mobile sink approaches the cluster centred.

The 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 the region.

## Twitter Data Generation:

Twitter is a highly popular platform for information exchange, can be used as a data-mining source which could aid in the aforementioned challenges which is collected by sensor nodes. Specifically, using a large data set of harvested tweets, sensor nodes connect with sink to transfer the dataset to HDFS system.

The REST APIs provides programmatic access to read and write Twitter data. Author a new Tweet, read author profile and follower data, and more. The REST API identifies Twitter applications and users using OAuth, responses are available in JSON

## EM computation:

The sink node sends data request message to cluster head to invoke data transmission from sensor nodes when it arrives at the cluster centroids. The nodes that receive data request message send the data to the sink node and broadcast data request message to their neighbouring nodes using multi hop traversal. It was realized that clustering can be based on probability models to cover the missing values. This provides insights into when the data should conform to the model and has led to the development of new clustering methods such as Expectation Maximization (EM) that is based on the principle of Maximum Likelihood of unobserved variables in finite mixture models.

## Data collection:

Once, the mobile sink patrols every cluster centroid and collects the data from the nodes in the cluster. This leads to transfer the sensor data to HDFS system with less energy consumption. The spectral clustering is performed to perform data analytics based on the Hash tag, Location and rewet count. As Big

Data applications are featured with autonomous sources and decentralized controls, aggregating distributed data sources to a centralized site for mining is systematically prohibitive due to the potential transmission cost and privacy concerns. On the other hand, although we can always carry out mining activities at each distributed site, the biased view of the data collected at each site often leads to biased decisions or models.

The latter usually is used for applications in specific areas, and thus is not able to meet the universality and readability requirements of electric power big data knowledge models [5]-[7]. Since the beginning of this century, artificial intelligence technologies have made great progress in terms of knowledge modeling. In all of these achievements, the ontology theory [8]-[10] and the semantic web technology [11]-[1 3] provide a new path to solve the problem of electric power knowledge modeling for big data. Ontology theory is based on description logics. Domain ontology, specifically, means "an explicit specification of a conceptualization," and it focuses on defining the relation between important concepts in a particular domain.

## 3. ENHANCEMENT AND RESULTS

In this paper, we propose energy minimized clustering algorithm by using the Expectation-Maximization (EM) algorithm for 2-dimensional Gaussian mixture distribution. Our proposal aims to minimize the sum of square of wireless communication distance since the energy consumption is proportional to the square of the wireless communication distance. Moreover, we first focus on the “data request flooding problem” to decide the optimal number of clusters. The data request flooding problem refers to the energy inefficiency that occurs when all the nodes broadcast data request messages to their respective neighbouring nodes. This problem wastes energy, particularly in the high density WSNs. Previous research work advocates increasing the number of clusters to reduce the data transmission energy. However, in this paper, we point out that an excessive number of clusters can result in performance degradation, and therefore, we propose an adequate method for deriving the optimal number of clusters.

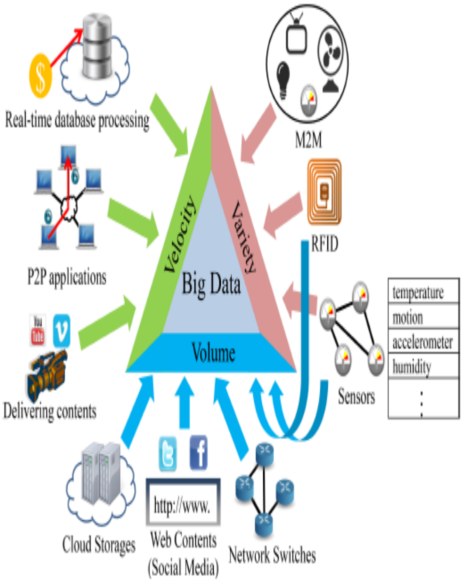

Fig 1 **Major trends of big data gathering**

The review conducted by Sagiroglu et al. [3] highlighted that big data and its analysis are at the core of modern science and business. Sagiroglu et al. 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.

## 4. CONCLUSION

In this paper, we investigated the challenging issues pertaining to the collection of the “big data” generated by densely distributed WSNs. Our investigation suggested that energy efficient big data gathering in such networks is, indeed, necessary. While the conventional mobile sink schemes can reduce energy consumption of the sensor nodes, they lead to a number of additional challenges such as determining the sink node’s trajectory and cluster formation prior to data collection. To address these challenges, we proposed a mobile sink based data collection method by introducing a new clustering method. Our clustering method is based upon a modified

Expectation- Maximization technique. Furthermore, an optimal number of clusters to minimize the energy consumption were evaluated. Numerical results were presented to verify the effectiveness of our proposal.

## 5.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/bigdatawhitepaper.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] D. Baum and CIO Information Matters, “Big Data, big opportunity,” http://www.oracle.com/us/c-central/cio-solutions/informationmatters/ big-data-big-opportunity/index.html.

[13] 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.

[14] 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.

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

[16] S. Wen-Zhan, H. Renjie, X. Mingsen, B. Shirazi, and R. LaHusen, “Design and deployment of sensor network for real-time high-fidelity volcano monitoring,” IEEE Transactions on Parallel and Distributed Systems, vol. 21, no. 11, pp. 1658–1674, Nov. 2010.