Addressing Data Redundancy in IoT Networks

Abhinav Garg, Manisha Jailia

***Abstract*: With the proliferation of ubiquitous information across the globe, Wireless sensor networks have been playing a crucial rule in availability of data at our finger tips. A large portion of this inundated data generated by these networks is repeated and of very little use. This paper aims to explore the work done so far by the researchers in field of handling replicated data produced by thousands of miniature devices spread around us.**

***Keywords:* IoT (Internet of Things), Redundancy, Wireless Sensor Networks (WSN's).**

I. INTRODUCTION

A Wireless Sensor network is made of numerous tiny IOT devices that are connected together to sense the physical environment variables like pressure, humidity, temperature, motion etc. and relay the information to sink or cloud. Internet of Things (IoTs) refers to a network of interconnected computing/mechanical devices, sensors and actuators that aid real time analytics.

With the rise in Wireless Sensor Networks and miniature devices all around us collecting GB's and TB's of data every second, the world today is completely inundated by Big Data. According to Gartner, by 2020 approximately 20.8 billion connected things will be continuously monitoring and reporting everything from daily activities, health records, temperature, proximity, quality, stocks and many more. Industries such as manufacturing, transportation, retail, automobiles etc. can harness the powers of these networks and data generated by them to reduce sustentation costs, prognosticate device malfunction, and evolve their business operations.

In legacy wireless sensor networks the end nodes or sensors sense the physical phenomenon, collect all the readings and transmit it to the base station/gateway. Gateways usually consolidate the data and send it to the servers. All the task of knowledge discovery is usually done at the powerful end station which is normally a cloud.

Some key issues concerning the data transmission in sensor networks are:-

1. Latency: Significant work has been done in the field of online and offline data mining so far, however it has given rise to the problems of cost and latency as every data transaction takes place to or from cloud for further processing which incurs high bandwidth and latency. Mission critical IoT applications, such as autonomous public transit, disaster monitoring, drone or flight control applications, remote surgery, and industrial control systems expect send to end latencies to be below a few tens of milliseconds. It is very difficult for mainstream cloud services to achieve this [1][2].
2. Redundancy: Owing to the high mobility of nodes in a single IoT network, or multiple sensors deployed in a common spatial region, sensors end up sending highly duplicated data at the resolution of seconds to the sink which results in wasting the energy, bandwidth and efficiency of the resources.

This paper aims to study some of the techniques used by researchers so far to address and resolve these issues. Rest of the paper is divided into five sections.

1) Sleep Scheduling Based Algorithms.
2) Data Encoding/Compression
3) Data Aggregation Techniques
4) Proposed Study
5) Conclusion

II. SLEEP SCHEDULING BASED DATA TRANSMISSION

One of the most frequently used approaches in Wireless Sensor Networks is Sleep Scheduling. It is also known as Duty cycling approach or sleep/awake scheduling. Here sensors are put to sleep for a specific period of time. Sensors sense the environment periodically only when a predefined event has occurred. These algorithms use multiplexing and various scheduling techniques to increase throughput, save energy and reduce redundant data.

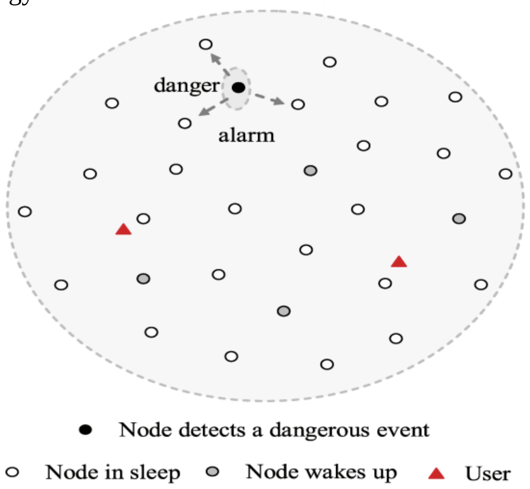

Fig. 1. Critical event monitoring with a WSN. [26]

Manuscript published on November 30, 2019.
* Correspondence Author

Abhinav Garg*, Assistant Professor, NIFT, Hyderabad, India. Email: abhinavarungarg@gmail.com

Dr. Manisha Jailia, Associate Professor, Banasthali Vidyapith, Rajasthan , India.. Email: manishajailia@yahoo.co.in

D. Goldsmith and J. Brusey have developed Spanish Inquisition Protocol (SIP) [3] in which a node transmits only the unexpected events. A state vector is transmitted in place of individual readings that too only when the difference between estimated new state and predicted sink state is more than a predefined threshold.

Another specific algorithm by J.Bruseyet. al. [4] is a human posture recognition approach that records and share posture and the timing of postural changes in place of the original signal. Bare Necessities [5] discards even timing and is appropriate where only a summary of relative time spent in different states is needed. These algorithms save energy and storage space by minimizing number of transmissions to certain extent.

Arora et al.[6] have designed a distributed sensor based surveillance system that works on dense, resource constrained networks. Here each node processes intrusion data locally and informs its peers only when an unexpected event has occurred.

III. DATA ENCODING/COMPRESSION

There tends to be a high correlation in the consecutive samples recorded by a sensor node in the network. Data compression and encoding schemes like delta encoding, S-LZW, Entropy coding, Huffman coding have been applied and tested so far to provide an in-network solution to the problem of huge data in transit.

S. Bhattacharya et al. [7] have also used compressed sensing scheme to design high resolution imaging system named as SAR (Synthetic Aperture Radar). They advocate that sampling at higher rate and then processing the large quantity of data to reduce redundancy increases the computational complexity. So SAR samples the signal below Nyquist rate.

A. Scaglione and S. Servetto [8] have used scalar quantizers locally to compress the analog sequences from sensor using Lempel-Ziv compression scheme in contrary to vector quantizers. They have come up with an idea of combining routing with source coding to eliminate correlations.

F. Chen et. al. [9] proposed compressed sensing scheme in which data sample rate is proportional to the information content rather than frequency to address energy and bandwidth limitations. However, generally data sampling is based on the theory of Nyquist-Shannon's sampling theorem which states that sampling rate should be greater than twice the maximum frequency of the signal i.e. fs>2B where B is band limit or frequency and fs is sampling rate.

Bhavish Aggarwal et al [10] have proposed a novel redundancy elimination scheme called EndRE. It is divided into 2 phases- 1. Fingerprinting and Marking- Here authors have illustrated a new mechanism named as Sample Byte based on Rabin Fingerprinting. They have compared 4 fingerprinting techniques- MODP, MAXP, Fixed and their Sample Byte. They claim their technique to be adaptive and faster in comparison to the other approaches. 2. Matching and Encoding- Here authors have introduced two approaches. "Chunk match" approach uses chunk hashes to identify full chunk matches whereas Max Match approach uses fingerprinting to identify maximal matches in the data.

Neil T. Spring and David Wetherall [11] had proposed a technique to identify duplicates in network traffic. A cache with most recent packets is indexed using representative fingerprinting. Then fingerprint of every incoming packet is compared with that of the cache's to find a hit or miss.

Biljana Risteska Stojkoska [12] has developed a lightweight, computationally cheap and optimal data compression algorithm based on delta coding scheme. Author encoded the most probable difference between the present and previous values using four basic two bit variable length codes. Remaining values were derived using these four values in suffix or prefix.

Christopher M. Sadler and Margaret Martonosi [13] have designed a LZW based compression scheme known as S-LZW. It divides data into small individual blocks which are then compressed using dictionary schemes. A 512 entry dictionary is used to compress data in a block of two flash pages for a significant gain in compression ratio. A variant of S-LZW ie. S-LZW MC uses a hash indexed mini cache to hold recently used dictionary entries. The technique results well in terms of energy saving as well as compression ratio.

IV. DATA AGGREGATION TECHNIQUES

Nodes in a WSN are distributed densely across the area under surveillance. If all these nodes send all their data to the sink node, the process will consume lot of energy, bandwidth and time. It is not practical for a resource constrained mobile node, so data from various end sensors is aggregated by an extra node which acts as a cluster head or aggregator and then aggregated data is forwarded to the sink or border gateway.

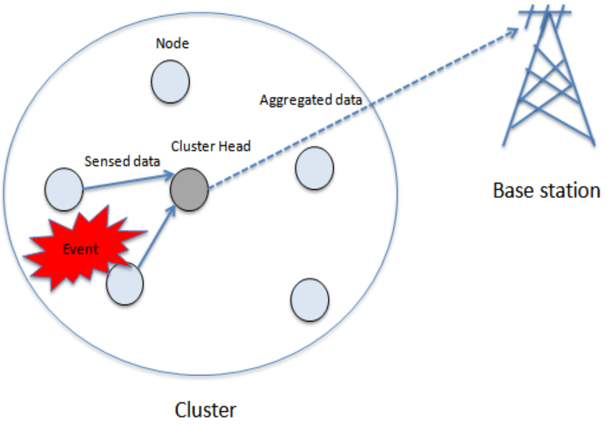

Fig. 2. Event detection at the cluster head [27]

Fusion approach:

G. Simon et. al. [14] have developed a counter Sniper approach named as PinPtr that advocates the use of multiple sensors to detect the shooter. A sensor fusion algorithm at the base station can then estimate the location of shooter. Increased number of sensors increases the probability of high coverage.

Clustering approach:

W. Heinzelman et al. [15] have developed and analyzed an application specific protocol LEACH(low- energy adaptive clustering hierarchy). This is an energy efficient distributed cluster formation approach. Here every cluster processes the data locally which can then be aggregated.

It assumes that nodes are close to each other, have equal processing-energy capabilities and always have data to send. All the nodes in the cluster sense the same event and produce redundant and correlated data.

Data redundancy exhausts network resources and increase latency.

M. Ye et al. [16] have proposed EECS(An Energy Efficient Clustering Scheme in Wireless Sensor Networks) for single hop wireless networks which works well in periodical data gathering applications. It elects cluster heads which have more residual energy through local radio communication in an autonomous way. It also balances load among the cluster heads using distance based parameters.

K. Maraiyaet. al. [17] have designed a data aggregation algorithm called Efficient cluster head selection scheme for data aggregation in wireless sensor network (ECHSSDA) which suggests replacing the worn out cluster head by an associate cluster head. This algorithm uses a localizes clustering technique LEACH(Low Energy Adaptive Clustering Hierarchy).

I. Gupta et. al. [18] states that efficient cluster head can be selected on the basis of 3 parameters viz. Node Energy (energy level of each node calculated by a fuzzy variable), node concentration (no. of nodes present in the vicinity) and node centrality (how easy it is for other nodes in cluster to transmit through cluster head candidate).

Wenwei Xue et al. [19] have proposed an event detection scheme in which sensor events are mapped into spatio-temporal patterns. Contour map of a sub network rooted at the node is transmitted as data aggregate for pattern matching.

Pasternak et al [20] have suggested using multiple gateways to connect different WSN's. They have recommended two topologies to configure wireless network- using 6LBR software on a border router as a gateway or to use a simple transparent bridge which simply forwards the traffic to the internet via an Ethernet backbone.

Fei Yuan et al [21] have proposed an accurate and energy efficient clustering scheme based on data density correlation degree (DDCD) to quantify the spatial correlation of data. It exploits the fact that nodes present in a cluster will generate highly correlated data, in comparison to the nodes present in different clusters. DDCD approach first identifies the type of sensor node. A data density correlation degree helps in identifying representative, isolated and member sensor nodes. Then local clusters are constructed in Local Cluster construction phase. Further local clusters are merged based on the DDCD parameter of every sensor node.

Tree based approach:

Prakash goud Patil et al [22] have developed a Delay Efficient Distributed data aggregation algorithm (DEDA) for WSN's. First a data aggregation tree is build using in-network aggregation and then DEDA scheduling approach is applied from child node to parent node and finally to the root node for aggregating data. A time slot is given for each node to transmit to avoid collisions. Side edges and back up edges are used along with the primary edges, which helps in recovery and makes the whole approach fault tolerant.

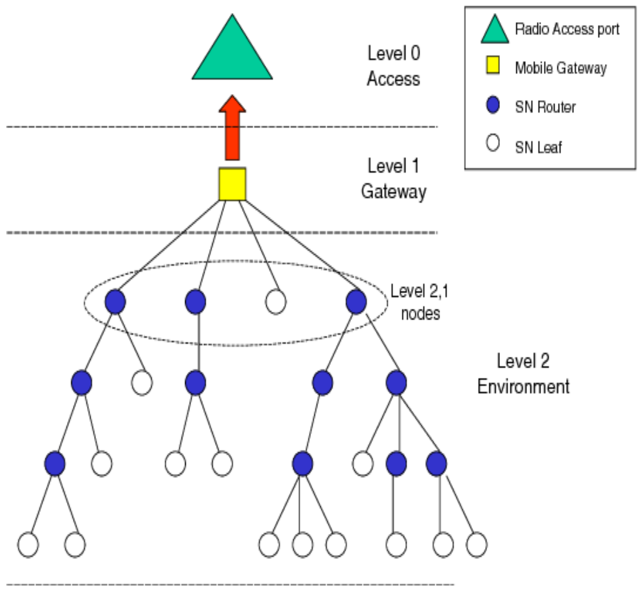

Fig. 3. Hierarchical tree-based network topology [28]

Charalambos Sergiou et. al. [22] have developed a congestion control scheme named as Hierarchical Tree Alternate Path (HTAP) algorithm. It consists of two phases APC (Alternate Path Creation) and HTC (Hierarchical Tree Creation) where congested nodes are by passed and unutilized idle nodes are also put to use.

Cheng et al. [23] have used Decision tree algorithm to classify hierarchically organized sensor nodes in a localized and iterative manner. In Hierarchical Distributed data classification utilizes unused node, eliminates dead nodes and create alternate path for transmission. A leaf node senses local data, builds a decision tree and sends to the parent node. Intermediate parent node combines its local data with the set of classifiers received from its children, builds a new decision tree and forwards the classifier to the parent node or base station. Base station then builds a global classifier. However disadvantage of using this approach is that the size of decision tree becomes too long.

CHUAN ZHU et al [24] proposed tree-cluster based data gathering algorithm (TCBDGA). The sensor nodes in TCBDGA can be divided into three categories: root nodes, leaf nodes, and normal nodes.

The whole algorithm consists of three stages, tree construction, Rendezvous Points (RPs) and Sub-Rendezvous Points (SRPs) selection, and data collection. The process starts with the construction phase where data gathering trees are constructed and every node is weighted and marked as root (Rendezvous point), parent or leaf. In RPs and SRPs selection stage, every tree is divided into sub-trees for load balancing. A mobile sink starts the data collection tour periodically from the BS, visits each rendezvous point, collects data from the sensors in its one-hop range directly, and finally, return back to the BS for one round. Sink finally decides whether to reconstruct the tree based on residual energy of each node, distance of nodes from BS and local densities of nodes.

TCBDGA claims to achieve significant advantage in terms of load balancing and low energy consumption.

V. PROPOSED STUDY

Traditionally, IoT edge analytics meant pushing IoT data to the cloud and dumping it into a data lake for big data analytics but unprecedented growth of smart city phenomenon has led to the requirement of real time analytics with time critical insight. In an environment where sensors are geographically distributed and cost/bandwidth incurred in transferring data to cloud is too large and chances of data being replicated increases manifold. Several techniques (as discussed above) have been considered by previous authors to address these issues. Although these approaches works well where events are both infrequent and easily traceable, it is difficult to specify event thresholds as a state of system varies with time (e.g. because of periodic variations). Predefined triggers may lose importance over real time events, eventually making the approach less useful.

However there is need to focus on amalgamation of cloud and Edge Analytics to make the best out of the data deluge produced by the sensor networks. Bringing the intelligence and analytics on smart sensing device in an energy efficient manner can help make the best use of data.

Every edge node in a cluster sense the environment over a certain period of time, generates a local classifier and sends it to Cluster Head (CH) along with the Time Stamp (Tn). CH will then generate congregated classifier based on the information from its children's classifier and its local data. It then updates its child nodes with the current Timestamp (Tc). Thereafter, edge nodes will periodically check the timestamp from the sink node using handshake protocol. If Ti>Tc edge node will transmit its current classifier. Eventually CH will update its classifier and time stamp also.

In this approach every node (ie. origin, intermediate, cluster head, base station) involved in the network before cloud storage will have the capability of producing local classifiers, which can be tapped for gaining insights at any level of the network.

Algorithm	**Technique**	
SIP [3]	**Sleep Scheduling**	**State vector**
[4]		**Transmit posture and timing only**
Bare Necessities [5]		**Relative difference**
[6]		**Transmit Unexpected Events Only**
SAR [7]	**Data Compression**	**Compressed Sampling**
[8]		**LZW**
[9]		**Sampling rate proportional to Informational content**
EndRE [10]		**Finger Printing**
[11]		**Finger Printing**
[12]		**Delta coding**
S-LZW [13]		**LZW**
PinPtr [14]		**Fusion**
LEACH [15]		
EECS [16]		
ECHSSDA [17]	**Data Aggregation**	**Clustering**
[18],[19],[20]		
DDCD [21]		
DEDA [22]		**Tree based data aggregation**
HTAP [23]		
HDDC [24]		
TCBDGA [25]		

VI. CONCLUSIONS

Smart cities engulfed with innumerous miniature IoT devices have led to an era of digital transformation across the globe. The huge amount of data collected from the various sources can be structured, semi structured, unstructured or streamed data. This heterogeneity of data demands for smart and coherent data mining techniques that can deal with all types of data. Already existing solutions like data aggregation techniques or other probabilistic approaches focus on reducing redundancy at the base station, however new theories should address this problem wherever needed from origin to cloud.

REFERENCES

1. M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," IEEE Internet Things J., vol. PP, no. 99, pp. 1–1, 2016.
2. Fan X, Wei W, Wozniak M, et al., "Low energy consumption and data redundancy approach of wireless sensor networks with big data". Inf Technol Control 2018; 47(3): 406–418.
3. D. Goldsmith and J. Brusey, ''The Spanish inquisition protocol—Model based transmission reduction for wireless sensor networks,'' in Proc. IEEE Sensors, Nov. 2010, pp. 2043–2048.
4. J. Brusey, R. Rednic, E. I. Gaura, J. Kemp, and N. Poole, "Postural activity monitoring for increasing safety in bomb disposal missions," Meas. Sci. Technol., vol. 20, no. 7, p. 075204 (11pp), 2009.
5. E. I. Gaura, J. Brusey, and R. Wilkins, ''Bare necessities—Knowledgedriven WSN design,'' in Proc. IEEE SENSORS, Oct. 2011, pp. 66–70.
6. A. Arora et al. A line in the sand: A wireless sensor network for target detection, classification, and tracking. Computer Networks Journal, 46(5):605–634, 2004.
7. S. Bhattacharya, T. Blumensath, B. Mulgrew, and M. Davis, "Fast encoding of synthetic aperture radar raw data using compressed sensing," in Proc. IEEE Workshop on SSP, Aug. 26–29, 2007, pp. 448–452.
8. A. Scaglione and S. D. Servetto, "On the interdependence of routing and data compression in multi-hop sensor networks," in Proc. ACM MobiCom, Atlanta, GA, Sep. 2002, pp. 140–147.
9. F. Chen, A. P. Chandrakasan, and V. Stojanovic, "Design and analysis of a hardware-efficient compressed sensing architecture for data compression in wireless sensors," IEEE J. Solid-State Circuits, vol. 47, no. 3, pp. 744–756, Mar. 2012.
10. B. Agarwal, A. Akella, A. Anand, A. Balachandran, P. Chitnis, C.Muthukrishnan, R. Ramjee and G. Varghese. "EndRE: An End-System Redundancy Elimination Service for Enterprises," in Proceedings of the Second Symposium on Networked Systems Design and Implementation (NSDI '10), 2010, pp. 419-432.
11. N.T. Spring and D. Wetherall, "A Protocol-Independent Technique for Eliminating Redundant Network Traffic," Proc. SIGCOMM, pp. 87-95, 2000.
12. B. R. Stojkoska, and Z. Nikolovski, "Data compression for energy efficient IoT solutions", Telecommunications Forum, pp. 1–4, 2017.
13. C. M. Sadler and M. Martonosi, "Data compression algorithms for energy-constrained devices in delay tolerant networks," in Proc. SenSys '06: 4th Int. Conference on Embedded networked sensor systems, 2006, pp. 265–278.

14. G. Simon, A. Ledezczi, and M. Maroti. Sensor NetworkBased Countersniper System. In Proc. SenSys, Baltimore, USA, November 2004.
15. W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "An Application-Specific Protocol Architecture for Wireless Microsensor Networks," IEEE Transactions on Wireless Communications, vol. 1, no. 4, pp. 660–670, October 2002.
16. M. Ye, C. F. Li, G. H. Chen, and J. Wu, "EECS: An Energy Efficient Clustering Scheme in Wireless Sensor Networks", in Proceedings of IEEE Int'l Performance Computing and Communications Conference (IPCCC), 2005, pp. 535-540.
17. K. Maraiya, K. Kant, and N. Gupta, "Efficient cluster head selection scheme for data aggregation in wireless sensor network," Int. J. Comput. Appl., vol. 23, no. 9, pp. 10–18, 2011.
18. I. Gupta, D. Riordan, and S. Sampalli, "Cluster-head election using fuzzy logic for wireless sensor networks," in Proc. Annu. Conf. Commun. Netw. Services Res., 2005, pp. 255–260.
19. W. Xue, Q. Luo, L. Chen, and Y. Liu, "Contour map matching for event detection in sensor networks," in Proc. ACM SIGMOD'06, June 2006, pp. 145–156.
20. M. Pasternak, M. Nycz, and P. Hajder, "Redundancy, Load Balancing and High Availability Mechanisms in 6lowpan based Sensor Networks," in Proc .of CSIS, Severodonetsk, Dec. 2014, vol. 5.
21. F. Yuan, Y. Zhan, and Y. Wang, "Data density correlation degree clustering method for data aggregation in wsn," IEEE Sensors Journal, vol. 14, no. 4, pp. 1089–1098, 2014.
22. P. Patil and U. Kulkarni, "Delay Efficient Distributed Data Aggregation Algorithm in Wireless Sensor Networks," International Journal of Computer Applications, vol. 69, 2013.
23. X. Cheng, J. Xu, J. Pei, and J. Liu, "Hierarchical distributed data classification in wireless sensor networks," Computer Communications, vol. 33, no. 12, pp. 1404–1413, 2010.
    C. Zhu et al., "A Tree-Cluster-Based Data-Gathering Algorithm for Industrial WSNs with a Mobile Sink," IEEE Access, vol. 3, May 2015, pp. 381–96.
24. P. Guo, T. Jiang, Q. Zhang, and K. Zhang, "Sleep scheduling for critical event monitoring in wireless sensor networks," IEEE Trans. Parallel Distrib. Syst., vol. 23, no. 2, pp. 345–352, Feb. 2012.
25. Ouadoudi Zytoune1, Driss Aboutajdine, "A Low Energy Time Based Clustering Technique for Routing in Wireless Sensor Networks" American Journal of Sensor Technology, 2014, Vol. 2, No. 1, 1-6.
26. C. Buratti, R. Verdone, "A Hybrid Hierarchical Architecture: From a Wireless Sensor Network to the Fixed Infrastructure," IEEE EW2008, 22-25 June 2008, Prague, Czech Republic

AUTHORS PROFILE

Abhinav Garg is Assistant Professor in Fashion Communication Deptt at NIFT Hyderabad.. He has done B.Tech. (C.S.E.) and M.Tech. (I.T.) His areas of Interest Include Data Mining, Intellectual Property, Design Management and CAD

Dr. Manisha Jailia is Associate Professor in Computer Science Deptt. at Banasthali Vidyapith, Rajasthan India. She has done MCA, UGC-NET, and Ph.D. Her research areas are Data Mining, Distributed Databases, Web technologies, Big Data, Clou Computing and High-performance computing.