Certain Investigations on Data Aggregation issues in Wireless Sensor Networks

S. Gopikrishnan and P. Priakanth

Karpagam College of Engineering, Coimbatore-641032, India
Kongu Engineering College, Perundurai-638052, India

Abstract: Wireless sensor network is a most important research area in wireless communications. Wireless sensor networks (WSNs) consist of sensor nodes. Data aggregation is an important functionality and very crucial technique in WSNs. Data aggregation in WSN faces many critical challenges like collision, delay, security, scalability, localization, energy efficiency. Different algorithms and protocols are proposed over the last decade to resolve the data aggregation issues of wireless sensor network. Even many review studies published in data aggregation algorithms; this work focuses on summarizing various approaches proposed based on the issues of data aggregation. This paper also provided a detailed analysis on existing models and summarized the unresolved issues in the existing protocols.

Key words: Collision • Data Aggregation • Energy Efficient • Security • Wireless Sensor Networks

INTRODUCTION

In the recent years, the rapid growth of the applications using Wireless Sensor Networks (WSN) shows the involvement of both the research community and actual users. As sensor nodes are generally battery-energized devices that are deployed in hostile environments. The network lifetime of WSN is fully based on non rechargeable batteries. Therefore, energy management is one of the most challenging problems in wireless sensor networks. Data aggregation algorithms are one of the most acceptable solutions in the past years which reduce the energy consumption of the sensor nodes.

Recent research efforts on the energy limitation problems deal with techniques during the process of node deployment, searching the target node, data collection and communication. Since the sensing is the main cause for the draining the energy of a sensor node, the energy wastage can be avoided only when sending the sensed data to the base station. To resolve these energy wastage many data aggregation algorithms has been proposed in recent years. Implementing a data aggregation algorithm in WSN is a difficult task due to many deployment and protocol issues.

Data Aggregation: In the sensor network, sensor nodes might generate redundant data; similar packets from multiple nodes can be aggregated so that the number of transmissions would be reduced. Data aggregation is the combination of data from different sources by using functions such as suppression (eliminating duplicates), min, max and average. Some of these functions can be performed by the aggregator sensor node, by allowing sensor nodes to conduct in network data reduction. Recognizing that computation would be less energy consuming than communication, substantial energy savings can be obtained through data aggregation.

The efficacy of data aggregation can be judged using different metrics:

- Accuracy: Perhaps the most important metric is accuracy, which is the difference between the resulting value at the sink and the true value since not all data is delivered to the sink any longer. Accuracy can be expressed as differences, ratios, statistics, or other values depending on the particular case.
- Completeness: Potentially an operational approximation of accuracy is completeness, the percentage of all readings that are included in the computation of the final aggregate at the sink.
Latency: Aggregation can also increase the latency of reporting as intermediate nodes might have to wait for data.

Message Overhead: The main advantage of aggregation lies, of course, in the reduced message overhead, which should result in an improved energy efficiency and network lifetime. Aggregation protocols can often deliberately tradeoff between accuracy, message overhead and latency and only provide estimates of the actual aggregated value.

Data Aggregation Algorithm Classification: To design a data aggregation protocol, we should take into account three main components which are,

- Routing scheme: This component describes routing procedure to send the aggregated data to the base-station by using or creating a network structure. Routing schemes plays a major role in all the data aggregation approaches. When applying an effective routing scheme on a sensor network, the data aggregation issues like energy, collision, delay and security are could be avoided. Depending on the routing schemes implemented in existing data aggregation, it can be classified into three categories: i) cluster based data aggregation ii) chain based data aggregation and iii) tree based data aggregation.

- Aggregation function: This component defines how data are aggregated by specifying the aggregation function used by the protocol towards the base-station through a specific network structure. Generally, tree or cluster structures are created as routing schemes. The existing solutions in these approaches try to create a specified structure, which can be exploited later by the data aggregation protocols. The use of such structure leads to reduce the broadcast in the network; each node sends packets to its upstream which will be routed towards the base station. Based on the routing schemes used, the data aggregation function can be classified into two categories i) centralized approach and ii) distributed approach.

- Data aggregation scheduling: This component defines the waiting time of each node should wait before aggregating and forwarding received data. The data aggregation scheduling has the strongest impact on data aggregation goals. An optimal scheduling of data aggregation should lead to save energy, increase data accuracy and ensure aggregation freshness. Depending on the waiting time (WT) nature, we classify the data aggregation scheduling into two categories:
  - Unslotted data aggregation scheduling
  - Slotted data aggregation scheduling

Even many literature surveys Fasolo et al. [1], Akkaya et al. [2], Rajagopalan et al. [3], Chen et al. [4] made on data aggregation algorithms in existing models, the detailed studied on all the data aggregation issues has not been described. Since the objective of the proposed thesis is to resolve many issues in a combined solution, the detailed literature has to be made on issues based data aggregation algorithm. From most of the existing surveys, the issues based existing solutions can be classified into four categories.

- Delay efficient data aggregation algorithms
- Collision aware data aggregation algorithms
- Security enabled data aggregation algorithms and
- Energy efficient data aggregation algorithms

The detailed survey on these categories has been made and the detailed background about the existing models has been described in this section.

Delay Efficient Data Aggregation Algorithms: Minimizing latency is of primary importance for data aggregation which is an essential application in wireless sensor networks. Many fast data aggregation algorithms under the protocol interference model have been proposed, but the model falls short of being an accurate abstraction of wireless interferences in reality. In contrast, the physical interference model has been shown to be more realistic and has the potential to increase the network capacity when adopted in a design. It is a challenge to derive a distributed solution to latency-minimizing data aggregation under the physical interference model because of the simple fact that global-scale information to compute the cumulative interference is needed at any node.

Some surveys published on data aggregation in WSN have been studied Fasolo et al. [1], Akkaya et al. [2], Rajagopalan et al. [3] and Chen et al. [4]. Based on the survey articles reviewed, most of the existing algorithms can be employed on the two basic structures i) centralized and ii) distributed. The investigation made on the existing algorithms in terms of delay issues are described here.
DAS (Distributed data Aggregation Scheduling in wireless sensor networks) Yu et al. [5]: DAS is a distributed algorithm for data aggregation scheduling in WSNs. This scheduling generates a collision-free schedule with a data latency bounded by $24D + 6\Delta + 16$ for a data aggregation tree; where $D$ is the network diameter and $\Delta$ is the maximum node degree. In the aggregation tree construction phase, the tree generated by DAS is similar to the one in NCA Huang et al. [6] and the distributed aggregation algorithm is employed in a distributed way using the approach presented in Wan et al. [7]. The use of a sequential approach and a connecting dominating set structure presented in DAS may increase the data latency. However, as the parent’s time slot is higher than the ones of its children, the aggregation freshness and data accuracy are guaranteed in DAS.

ISDA (Improved SDA) Jianming et al. [8]: A centralized algorithm proposed in this ISDA enhances the time latency of SDA Chen et al. [9]. This algorithm generates a collision-free scheduling with a data latency upper-bound of $(7\Delta / \log2 (\lfloor N \rfloor)) \times (R - 1)$ for an aggregation tree, where $R$ is the network radius, $N$ is the number of nodes and $\Delta$ is the maximum node degree. In ISDA the network is divided into equal cells. Each one of them can contain a set of nodes. By combining TDMA and ensuring that parent’s time slot is higher than the children’ ones, ISDA ensures the aggregation freshness and data accuracy. Even so, the data latency is affected in ISDA due to the scheduling process. In ISDA all the leaf nodes should be scheduled before the other ones.

CIAS (Centralized Improved data Aggregation Scheduling) Xu et al. [10]: CIAS is a centralized solution that aims to minimize the data aggregation latency by using CDS topology. The authors choose the network topology centre as the aggregation tree root instead of the base-station. This choice is made to reduce the upper-bound time latency of CIAS to a function of the network radius $R$, instead of the network diameter $D$. The time latency in CIAS is at most $16R+\Delta-14$. Due to the uses of TDMA scheduling and the parent’s time slot is always higher than the children’ ones the data accuracy and aggregation freshness are ensured in CIAS solution. However, this technique may increase the data latency and Bagaa et al. in Bagaa et al. [11] have also proved that the data latency upper-bound of CIAS is incorrect.

SAS: SAS (Sequential Aggregation Scheduling) Wan et al. [12]): The scheduling proposed in SAS is a centralized solution. This algorithm proposed a collision free schedule with a latency bound of $15R+\Delta-4$; where $R$ is the network radius and $\Delta$ is the maximum node degree. SAS is employed in two steps: the aggregation tree construction and the data aggregation scheduling. The first one is constructed using an existing approach Wan et al. [7]. The second one schedules all the dominate nodes using the same approach as Xu et al. [10]. In SAS, the use of CDS structure may increase the data latency. In order to ensure the data accuracy and aggregation freshness, SAS uses the same method proposed in Xu et al. [10]. Bagaa et al. in Bagaa et al. [11] have proved that the data latency upper-bound of SAS is also incorrect.

DAS-ST: DAS-ST (Data Aggregation Topology based on Structured Topology) Bagaa et al. [11]: DAS-ST: DAS-ST is a centralized algorithm that aims to minimize the data aggregation latency. DAS-ST is an improved version of DAS proposed in the same article. DAS-ST bypasses more the rules to select parents by excluding the use of the MIS. DAS-ST allows each node to select its candidate parents from its neighborhood, which belong to the three levels without restriction. DAS-ST allows simultaneous execution of aggregation tree construction and scheduling. In addition to selecting the parents from three levels, the parent can also be selected from the already scheduled nodes. Thus, the data latency in DAS-ST is significantly reduced. Moreover, the aggregation freshness and data accuracy are ensured due to the condition that parent’s time slot should higher than the children.

Collision Aware Data Aggregation Algorithms:
Communication collision is a primary reason for long latency in data aggregation. It occurs when two or more sensor nodes transmit data to a common node at the same time. Unfortunately, the collision problem is normally left to the MAC layer, which could incur a large amount of energy consumption and time latency during aggregation. Hence, we concentrate on the TDMA scheduling problem above the MAC layer, while the main objectives are to minimize aggregation latency and ensure the collision-free transmissions.

ATC: ATC (Adaptive Timing Control) Li et al. [13] is a distributed solution that considers either monitoring or event-driven applications. The authors assume that each node in the tree knows its hop count $h_{max}$ and the number of children $N_{data}$. The aim of ATC is to satisfy the delay required by the application $d_{max}$ and hence ensures that all the aggregated data reach the base-station before $d_{max}$ expires. To ensure the aggregation freshness in ATC,
any received data after WT is elapsed will be discarded. But, the use of weighted-path adjusted interval enhances the data latency in ATC. Note that in ATC, the clock drifts between nodes which can be significantly increased due to the absence of any mechanism to adjust the WT during the data gathering phase.

FAST Yousefi et al. [14]: In FAST, Yousefi et al. have proposed a balanced connected 3-hop dominating sets based structure data aggregation tree that is different from commonly used two hop tree construction method. This method uses simultaneous execution of tree construction method and follows a collision-free TDMA scheduling algorithm. The distributed nature of this algorithm minimizes the aggregation delay. The authors have theoretically proved the upper bound for data aggregation latency at $12R + D - 2$ time slots under the protocol interference model. The main objectives of the FAST are to minimize aggregation delay and ensure the collision-free transmissions.

CLSD Ma et al. [15]: Ma et al. has proposed a new energy model to avoid energy wastage when changing the state of the nodes. This paper identifies the link scheduling problem and proved that it is NP-complete. The main objective of the link scheduling is to find the number of time slots used in a period, by maximizing the spatial reuse of concurrent transmission without interferences. This increases the network throughput and improves the data aggregation speed. They have used centralized and distributed algorithms to improve the upper bound of data aggregation delay in homogeneous as well as heterogeneous networks.

SFEB Chao et al. [16]: Structure-Free and Energy-Balanced data aggregation have proposed an energy efficient structure free data aggregation based on the number of aggregators and the time spend on the data aggregation. This protocol consists of two phases. In the first phase, the data aggregators are dynamically selected to collect the data from leaf nodes. In the second phase all the aggregators send the data to sink node. They have proposed a light weight algorithm to select aggregator node by avoiding extensive pocket exchanges. Based on Data-Aware Anycast (DAA) and Randomized Waiting (RW) Protocol Fan et al. [17], they have proposed a structure-free data aggregation algorithm to aggregate the data from the leaf and send it to sink node in phase one and two. Through the analysis and simulation results, they have proved their algorithm efficiency by comparing DAA and RW protocol.

MLST Shan et al. [18]: In Maximum Lifetime Shortest path data aggregation Trees, Shanet et al. has proposed a shortest path aggregation tree for each sensor network to maximize the network lifetime. From the literature studies of this method, they have identified the general version of semi-matching problems and it can be solved by polynomial time. They also found that when the shortest path tree is restricted, the original NP-complete problem is in P. They have proposed a centralized algorithm that runs in $O(|E| \log N)$ time. For better implementation they have used distributed protocol. Since it is difficult to transform the centralized algorithm to a distributed one, they have proposed another algorithm for semi-matching problem. The simulation results are proved the improvement in lifetime of the network.

DICA Bagaa et al. [19]: Distributed Low-Latency Data Aggregation Scheduling algorithm proposes a distributed low-latency data aggregation algorithm that interlinks the tree construction and node scheduling to reduce the time latency. When constructing the data aggregation tree in this method, it uses the maximum choices for aggregator selection at every node based on the hop count to the sink node. This algorithm has three key design functionalities are as distributed, interlink formation of aggregation tree and node ordering on the tree. These functionalities allow more choices in choosing the parents and maximizing the reusability of time-slot. By enabling flexibility in aggregator node selection, this algorithm finds the shortest path to reduce aggregation latency. DICA uses compared simulation results with six existing models including CIAS to prove its superiority over the other models.

CIAS (Centralized Improved data Aggregation Scheduling) Xu et al. [10]: CIAS is used to minimize the data aggregation latency by using connected dominating set (CDS) topology. The authors choose the network topology center as the aggregation tree root instead of the base-station. This choice is made to reduce the upper-bound time latency of CIAS to a function of the network radius $R$, instead of the network diameter $D$. The time latency in CIAS is at most $16R + \Delta - 14$. Due to the uses of TDMA scheduling and the parent’s time slot is always higher than the children’s ones, the data accuracy and aggregation freshness are ensured in CIAS solution.

LODAT: A distributed solution has been published in Sivagami et al. [20] called LODAT (Latency optimized data aggregation timing model for wireless sensor networks) to manage the waiting time for monitoring.
applications. LODAT uses CTP (Collection Tree Protocol) proposed in this model ensures the false data detection. LODAT is greedy with respect to energy consumption. Nevertheless, the periodic update of waiting time and the use of adjusted weighted-path interval, have a positive impact on data latency, data accuracy and aggregation freshness.

DRFT: DRFT (Delay-ranged feedback timing control) is a distributed solution that aims to increase the data accuracy while taking into account the data latency and aggregation freshness. DRFT assumes that the data aggregation tree is already constructed. To distribute the waiting time over network nodes, DRFT uses the same technique as ATC Li et al. [13].

Security Aware Data Aggregation Algorithms: Even the popularity of WSN has been increasing for a wide variety of applications the security in WSN is a challenging task. Cryptography is one of the significant ways to provide security. It can be provided through by symmetric key techniques, asymmetric key techniques and hash function. Since the sensor networks are having more restrictions in terms of computing, communication and battery power, it is not possible to implement a complex cryptographic algorithm.

Due to the limitations of sensor nodes, the selection of cryptographic technique is vital in WSN. It is important to select the most appropriate cryptographic method because all the security requirements are ensured by cryptography Prasad et al. [23], Wei et al. [24]. Cryptographic methods used in WSNs should meet the constraints of sensor nodes and be evaluated by code size, data size, processing time and power consumption. Based on the existing cryptographic techniques, it can be classified into three classes: symmetric cryptographic techniques, asymmetric cryptographic techniques and hybrid cryptographic techniques. Asymmetric cryptographic techniques can further be classified into three classes: RSA based techniques, ECC based techniques and pairing based techniques.

DAA - Ozdemir et al. [25] proposed a hop-by-hop symmetric protocol which provides the data confidentiality through a public key that is kept secret within the network. The routing scheme proposed in this model ensures the false data detection and it integrates false data detection with data aggregation and confidentiality. The monitoring aggregator nodes are also performs data aggregation and computes the MACs for data verification. In these protocols, the computation overhead is increased for hop-by-hop encryption.

LDAI - In order to mitigate the drawbacks of the hop-by-hop schemes, some end-to-end protocols are proposed. Niu et al. [26] Niu et al. [26] proposed an end-to-end symmetric algorithm which is known as lossy data aggregation integrity scheme. A homomorphic hash algorithm and identity-based aggregate signature is used for encryption in LDAI. The mechanism described in this method involves, the sink node and other sensor nodes shares different keys for encryption and then it uses weights to verify the authenticity of the aggregated data. By implementing an end-to-end cryptography, the computation overhead is reduced. But this protocol has failed to avoid the data duplication due to sink node decryption.

IPHCD - Ozdemir et al. [27] proposed an asymmetric cryptography. In this method an elliptic curve cryptography-based homomorphic encryption is used to achieve hierarchical data aggregation. The data integrity is ensured because of end-to-end cryptography data aggregation and the data confidentiality is assurred by dual key encryption. Because of using end-to-end cryptography this method, the integrity is verified at the sink node. This scheme is employed on asymmetric cryptography to provide confidentiality which has a larger computation overhead. Therefore this method fails in reducing the energy consumption.

SCUR - Tahir et al. [28] proposed a Lightweight Encryption Mechanism based on the Rabbit stream cipher for providing confidentiality in Wireless Sensor Networks. It fulfils both requirements of security as well as energy efficiency. The method reduces the communication overhead during data aggregation. And it also provides the security with less computation overhead. Even it reduces the communication and computation overhead on data aggregation it fails in data integrity because of its end-to-end encryption behavior.

MASA - Based on the concept of virtual geographic grid Alzaid et al. [29] proposed a Hybrid Cryptographic security system to provide end-to-end data
security for wireless sensor networks. In this method the entire sensor network is broken down into smaller regions called cells. Each sensor carries two types of keys, asymmetric and symmetric. MASA uses the private key to sign a hashed event notification and public key to authenticate the event notification. Even this method provides confidentiality and authenticity; the data duplication cannot be avoided due to end-to-end hybrid cryptography.

SDA-PH - Zhou et al. [30] Zhou et al. [30] A secure data aggregation scheme based on homomorphic primitives is proposed in Yadav et al. [31]. It applies symmetric cryptography-based privacy homomorphism and homomorphic MAC to protecting data confidentiality and detecting data integrity. It also computes all packets during the process of integrity verification. Since this method uses end-to-end cryptography, this method provides less computation overhead. But this method uses homomorphic primitives at MAC layer which leads to collision problem on the network.

SecFleck- Hu et al. [32] Hu et al. [32] described the design and implementation of a public-key platform. It is based on a commodity Trusted Platform Module (TPM) chip that extends the capability of a standard node. SecFleck chooses XTEA symmetric key cryptography because of its small RAM footprint, which made it as a good candidate for tiny sensor devices that typically have less than 10 KB RAM. XTEA can be used in an output feedback mode to encrypt or decrypt variable length strings. The author’s had discussed the performance of the secFleck platform in terms of computation time, energy consumption and financial cost. But this cryptography algorithm is hardware dependent. So this method is not suitable for all the network models.

Energy Efficient Data Aggregation Algorithms: In WSN, the nodes are mostly operated using batteries. The output capability of a battery over a period of time is referred to as its capacity. It is measured in ampere-hours and is mostly proportional to the voltage. Over a period of time, the capacity of a battery can be established as a function of a continuous discharging process. As the current drawn increases, the corresponding voltage, remaining capacity, available energy and expected lifetime of the battery decreases. This necessitates the development of energy aware mechanisms that minimize the number of radio related activities, thereby reducing the energy consumption in the node and network.

A near-optimal scheduling (NOSA) algorithm Mpitzopoulos et al. [33] has been proposed by Mpitzopoulos et al. for finding an appropriate number of mobile sensor nodes and their near optimal paths using the Esau-Williams heuristic. In NOSA, the parallel deployment of multiple agents is suggested where each agent visits a subset of nodes. NOSA outperforms the single agent-based approaches (e.g., LCF, GCF and GA), in terms of data fusion cost and overall response time Mpitzopoulos et al. [33], but it endures high computational complexity in determining the agents’ itineraries.

In order to reduce the high computational complexity in Mpitzopoulos et al. [34], a multi-agent scheduling (MAS) algorithm has been proposed in Chen et al. [34], Chen et al. [35], Cai et al. [35]. To reduce the latency, the authors have proposed the scheme MAS that help in the collection of concurrent sensor data. These algorithms differ in cluster group formation methods. In Chen et al. [34], the authors used an angle gap for clustering all the sensor nodes in a particular direction as a single group. This approach does not describe how to find out an optimal angle gap threshold. A genetic algorithm has been proposed by the authors in Cai et al. [36] to form the clusters. The limitation of a genetic algorithm-based approach is its higher computational overhead Wang et al. [37]. These algorithms assume that the set of sensor nodes to be aggregated by the sink are predetermined, which limits the application scope of the network.

In Konstantopoulos et al. [38], the authors proposed a greedy tree-based scheduling (TBSA) algorithm to find near optimal paths for multiple agents. This algorithm is a centralized algorithm where the sink is statically determined the number of aggregators and their schedules. The main theme of the TBSA algorithm is to divide the area around the sink into concentric zones to construct the near optimal path tree from inner zones to the outer zones.

Qi Xu et al. [39] and Xu et al. [40] proposed a dynamic data aggregation algorithm for a target tracking application. Energy consumption, information gain and the remaining energy of a node are the cost function used in this method for selecting the next node. Once the cluster head aggregates the data from its cluster and return backs to the sink node. This algorithm is more time expensive and may face difficulty in returning back to the sink without additional forwarding information. In
Chen et al. [41], the authors proposed the mobile agent-based directed diffusion (MADD) where an aggregator node visits a subset of nodes. In MADD, the sink uses the first phase of the directed diffusion algorithm Intanagonwiwat et al. [42] to determine the clusters. However, the actual data aggregation is carried out by dispatching an aggregator that sequentially visits the subset of nodes Wang et al. [37].

A software agent based directed diffusion was presented in Shakshuki et al. [43] where the order of node visits is determined at the sink node, but the authors did not describe the procedure. This method takes the routing cost and the remaining energy of a node for selecting a next node to be visited by an aggregator. The main limitations of the schemes described in Chen et al. [41], Shakshuki et al. [43] are that they depend on a directed diffusion scheme, that they incur extra communication overhead for data aggregation and that they are only appropriate for request based data aggregation applications. Gupta et al. Gupta et al. [44] proposed a multiple mobile aggregator with dynamic scheduling-based data dissemination (MADD) protocol where nodes are organized in a set of the fixed regions and each aggregator is responsible for collecting aggregated data from each region. The route of an aggregator is dynamically routed at each node using cost function. MADD adapts to unexpected node failures during data aggregation from aggregator to the sink, but it consumes slightly more energy than TBID Konstantopoulos et al. [38].

Boudia et al. Boudia et al. [45] proposed an additive homomorphic encryption and aggregate MAC to provide the end-to-end confidentiality and the end-to-end integrity, respectively. SASPKC adopts state full Public Key Encryption (StPKE) for efficiency in terms of computation and communication costs. SASPKC aggregates not only cipher texts but also signatures, the end-to-end data confidentiality and integrity security services are provided using symmetric homomorphic encryption and aggregate Message Authentication Code (MAC), respectively. While considering new attacks such as selective forwarding, SASPKC does not support for nodes mobility. The main contribution of Sun et al. in Sun et al. [46] is to propose a combination of trust mechanism, data aggregation and fault tolerance to enhance data trustworthiness in Wireless Multimedia Sensor Networks (WMSNs) which considers both discrete and continuous data streams. Ho et al. in Ho et al. [47] proposed a framework to detect compromised nodes in WSN and then apply a software attestation for the detected nodes. They reported that the revocation of detected compromised nodes cannot be performed due to a high risk of false positive in the proposed scheme.

Rezvani et al proposed a novel collusion attack scenario against a number of existing IF algorithms in Rezvani et al. [48]. The authors have proposed an improvement for the IF algorithms by providing an initial approximation of the trustworthiness of sensor nodes. When compromised aggregators are involved in data aggregation, this model fails in protecting the data. Moreover this method is only suitable for new deployment of data aggregation. In Ghosh et al. [49], the authors dedicated their efforts on the aggregation scheduling problem in WSNs when multiple frequency channels are available. The authors then demonstrated that finding the minimum number of channels required in the network to alleviate all the interference is NP-hard. The NP hardness of minimizing the scheduling latency in an arbitrary network with respect to multiple channels has been proved. The work in Joo et al. [50] formulates the scheduling problem of minimizing the overall data transmission delay. The characteristics of optimal scheduling were studied, followed by the proof of the lower bound of optimal performance. Two scheduling policies were proposed in this paper Joo et al. [50]. Decision in one policy was made based on the current system state. In the other policy, prediction of the future system conditions was also taken into consideration.

The authors of Li et al. [51] studied the distributed aggregation scheduling problem in WSNs with respect to minimum latency. Compared with centralized solutions, a distributed scheduling plan has its own advantages. In this paper, an algorithm based on vertex coloring was proposed with a proved delay of $4R + 2D$. In Yu et al. [52], minimum latency aggregation scheduling in WSNs with multiple sinks were investigated. Differentiating from prior literatures, this model proposes a dynamic selection of sink node based on the shortest path by the sensor nodes for the purpose of minimizing the transmission latency. Two approximation algorithms with bounded latency were proposed. Zeydan et. al have proposed an adaptive and distributed routing algorithm for correlated data gathering and exploit the data correlation between nodes using a game theoretic framework in Zeydan et al. [53]. Routes are chosen to minimize the total energy
expended by the network using best response dynamics to local data. The cost function that is used for the proposed routing algorithm takes into account energy, interference and in-network data aggregation. This paper specifically addresses the problem of effective energy minimization but the quantitative analysis of delay minimization has not been resolved.

In Yao et al. [54], Yanjun Yao et.al proposed an Energy-efficient Delay-Aware Lifetime-balancing protocol for data collection in wireless sensor networks. The authors proved that the problem formulated by EDAL is NP hard, therefore, they have proposed a centralized and distributed heuristic to reduce its computational complexity. Even EDAL achieves a significant energy consumption and delay, this model proposed only to resolve energy issues only in heterogeneous networks. Chih-Min et al. proposed a structure-free and energy-balanced data aggregation protocol (SFEB) in Chao et al. [16]. SFEB consists of two phases. In first phase, we designate some nodes as aggregators to gather as many packets as possible. Then, these aggregators send the collected packets back to the sink in phase two. Sensor nodes that fail to send data to aggregators will also transmit their packets to the sink in phase two. This model requires location information of sensor nodes to avoid structure based data aggregation. Location information can be obtained by applying a localization protocol.

Dervis et al. proposed a novel energy efficient clustering mechanism, based on artificial bee colony algorithm, is presented to prolong the network life-time in Karaboga et al. [55]. Artificial bee colony algorithm, simulating the intelligent foraging behavior of honey bee swarms, has been successfully used in clustering techniques. The proposed protocol ICWAQ uses efficient and fast searching features of the ABC algorithm to optimize clustering of the nodes in the selection process of cluster-heads defining routing gateways. Since this algorithm uses random cluster head selection it leads to collision problem. This algorithm is not suitable for routing mobile networks due to the MAC layer issues.

**Analysis and Summary:** At the summary of data aggregation algorithms in wireless sensor networks, all the algorithms are focuses on optimizing important performance measures such as collision, data latency, data security and energy consumption. Efficient organization, routing and data aggregation tree construction are the three main focus areas of data aggregation algorithms. We have analyzed the main features, the advantages and disadvantages of each data aggregation algorithm.

**Delay Analysis:** All the protocols which are discussed in Section 2 are slotted-based data aggregation protocols are centralized, except DAS. The algorithms proposed in TDMA slotted category Xu et al. [10] are based on the cross-layer principle (routing and MAC layers). They are resilient to clock drifts, as long as it has assumed that nodes have synchronized clocks. All these solutions use only the interference protocol model and ensure data accuracy and aggregation freshness. To reduce the data latency many techniques are used in these solutions, where the most efficient are those which: (i) allow parents selection from three layers and (ii) execute tree construction and network nodes scheduling simultaneously.

By analyzing the existing models in terms of data aggregation latency the problems has been identified. Since all the existing models in section 2 are allow parents selection from three layers, it leads to collision on parent selection. When uses a TDMA slotted algorithm Xu et al. [10], the existing models Ma et al. [15] performs on/off mechanisms on transceiver to save the energy. But it leads to more computational overhead on all the nodes and also not suitable for event based applications. Even though all the algorithms are executed on tree based scheduling most of the existing models Yu et al. [5], Wan et al. [12], Xu et al. [10] followed a same tree construction method. The drawbacks of all the existing models will lead to find a better tree construction and delay efficient data aggregation algorithm that suitable for all the type of networks like static, dynamic, homogeneous and heterogeneous networks.

**Collision Analysis:** At the conclusion of the collision based existing models, there are nine protocols in this section 3, one of them is structure free (SFEB Chao et al. [16]) and the rest are distributed (ATC Li et al. [13], LODAT Sivagami et al. [20], FAST Yousefi et al. [14], DICA Bagaa et al. [19] and DRFT), centralized (CIAS) and hybrid (CLSD, MLST). The aim of DRFT is to increase the data accuracy, while the others aims is to reduce the energy utilization (CLSD, SFEB and MLST) and latency (ATC, LODAT, FAST, DICA and DRFT). Note that, only LODAT and FAST are simultaneous in this category. As LODAT and DRFT ensure the resilience to the clock
drifts, both of them suffer from high energy consumption and message collisions due to periodic update of waiting time (WT). In order to ensure the aggregation freshness in ATC, each node discards the received data from its children when its WT is elapsed. However, this strategy affects the data accuracy in ATC. Although FAST and CIAS uses TDMA scheduling to minimize the delay and collision, FAST uses 3-hop tree construction where as CIAS uses CDS Topology. Even FAST and CIAS reduce the aggregation delay; it uses time slot based data aggregation that is not preferable for event based applications and need to implement effective node state management. DICA uses cluster based data aggregation and it provides more flexibility in aggregator selection. But the aggregator selection is not unique for all the clusters and it leads to collision problem. Among the energy efficient algorithms, SFEB uses dynamic selection of aggregators. But it uses two different algorithms for leaf nodes and aggregator node which is not suitable for dynamic networks. CLSD and MLST uses hybrid algorithm, where CLSD can be implemented in homogeneous and heterogeneous networks. In order to save the energy MLST uses three algorithms for data aggregation. It increases the latency in data aggregation process. In CLSD, it uses state change mechanism to save the energy. So maximum time a node will be in off state which is not preferable for continuous monitoring applications. From the conclusion of existing models, all the models resolve only one issue in the data aggregation scenario.

**Security Analysis:** Based on the security issues in data aggregation there are seven protocols that are discussed in this section 4, one of them is hop-by-hop encryption (DAA Ozdemir et al. [25] and the others is end-to-end encryption technique. As Niu et al. [26] and Alzaid et al. [29] ensure the data integrity and confidentiality; both of them are unsuccessful in avoiding data duplication due to the end-to-end encryption techniques. Although Ozdemir et al. [25] provides high secure data aggregation, it does have high communication overhead which may not possible to implement in holistic environment based applications. The model proposed in Hu et al. [47] this designed by modifying the hardware architecture of the sensor nodes to provide the high security. But it makes this model is not suitable for the sensor networks which uses highly configured sensor nodes. Paper Tahir et al. [28] achieves data confidentiality and less energy consumption, but the data integrity cannot be achieved due to the direct forwarding happened at intermediate nodes. Confidentiality and integrity is achieved in the security algorithm proposed in Zhou et al. [30]. Even collision problem cannot be avoided since the algorithm does the integrity verification at MAC layer. Even all the related studies discussed here are provides data confidentiality, integrity, unsuccessful in achieving all the security requirements. Moreover when implementing the WSNs for Internet of Things (IoT) based applications, the energy efficient secure data aggregation protocol is highly significant Sheng et al. [56], Zhou et al. [57]. At the summary of related studies on security based data aggregation algorithms, the researchers should made contribution on a combined solution that resolves the data confidentiality, integrity, computation overhead and energy limitation.

**Energy Analysis:** The literature survey made in section 5 analysis energy issues in data aggregation and the best solutions has been presented from the existing models. EECD Zeydan et al. [53] presents an effective solution that resolves energy and collision problems. When the network density is more EECD fails in reducing the aggregation delay and produces less throughput. A centralized and distributed algorithm that resolves energy and delay issues has been proposed in EDAL Yao et al. [54]. But the authors in EDAL specifically presented their protocol only to heterogeneous networks. To avoid structure based data aggregation the authors have presented a modified aggregator based data aggregation known as SFEB Chao et al. [16]. Even energy efficient data aggregation has been achieved in SFEB; it fails in fast data aggregation due to mediatory implementation of localization algorithm. In ABC Karaboga et al. [55] authors have presented a cluster based data aggregation algorithm that mainly concentrates on reducing aggregation delay. The randomized selection of cluster heads leads ABC into an insecure data aggregation.

**CONCLUSION**

We have presented a comprehensive survey of data aggregation algorithms in wireless sensor networks. All the data aggregation algorithms focus on resolving a specific issue. In order to improve the network performance in data aggregation there is a need of efficient data aggregation algorithm which resolves the
maximum data aggregation algorithm design goals in a combined solution. By concluding all the drawbacks of existing models, the following problems have to be resolved in a combined unique solution:

- An effective data aggregation tree has to be constructed to implement a distributed approach for static as well as dynamic sensor networks.
- The tree construction should follow a shortest path algorithm to reduce the data aggregation delay.
- The parent and child node selection in tree structure should avoid collision. And an effective algorithm has to be implemented for collision-free data aggregation.
- Data integrity and confidentiality should be ensured through a lightweight secure data aggregation algorithm.
- All the above problems have to be resolved in an energy-efficient way. And the combined solution must be able to implement in all the type of sensor networks like homogeneous, heterogeneous, MANET etc.

REFERENCES


