Energy-Efficient Data Collection Using Lossless Compression for Industrial Wireless Sensor Networks

Xiaolan Tang[1], Hua Xie[1], Wenlong Chen[1(✉)], and Jianwei Niu[2]

[1] College of Information Engineering, Capital Normal University,
Beijing 100048, China
chenwenlong@cnu.edu.cn

[2] State Key Laboratory of Virtual Reality Technology and Systems,
School of Computer Science and Engineering,
Beihang University, Beijing 100191, China

Abstract. Industrial wireless sensor network is an important technology for precise monitoring in industrial systems. Sensors are deployed densely in various industry applications, where the high density of sensors results in large amounts of redundant data. Therefore, information aggregation is used to avoid forwarding redundant data and thus save limited resources. However, when decreasing transmission cost, existing aggregation schemes lead to low data accuracy and long delivery latency. In this paper, we propose an energy-efficient data collection solution using lossless compression for industrial wireless sensor networks, namely ECL, aiming for high energy efficiency and high information entropy. According to three aggregation rules, aggregation regions are constructed in a distributed way based on a preset threshold of sensing duplication rate. Therefore, the aggregated data are probably similar, and ECL has the original entropy through removing only the redundant data. Experiment results show that compared with other schemes, ECL keeps about 38% and 48% higher data accuracy and 12% and 25% shorter maximum end-to-end delay than EEUC and HEER, respectively, with a similar lifetime.

Keywords: Industrial wireless sensor networks
Sensing duplication rate · Data aggregation · Redundant data

1 Introduction

Industrial Wireless Sensor Network (IWSN) plays a vital role in creating a highly reliable and self-healing industrial system that rapidly responds to real-time events with appropriate actions. Machines are automatically controlled by using the obtained information to make an efficient production line [1]. In order to improve the stability and robustness of the entire network and the accuracy of gathered information, sensor nodes are usually densely deployed in industrial

Y. Chen and T. Q. Duong (Eds.): INISCOM 2017, LNICST 221, pp. 14–27, 2018.
https://doi.org/10.1007/978-3-319-74176-5_2

areas [2]. However, more than enough sensors lead to a large amount of redundant data. Forwarding these redundancy further wastes the nodes' energies and bandwidths, which reduces energy efficiency and shortens network lifetime.

Data aggregation is widely utilized in IWSNs, aiming for reducing the transmissions of redundancy [3]. Specifically, in most aggregation protocols, the general network is divided into some areas/cells based on the geographical locations. A selected node gathers and aggregates all the obtained data in each area [4]. Since the data might be delivered to an aggregation node through several hops, the redundant information wastes a large amount of communication resources due to multi-hop relays. Meanwhile, multiple forwarding brings in a long latency for data aggregation [5,6]. In addition, because of the possible random distribution of nodes and several times of aggregation of the raw data, data accuracy is reduced, which affects the performance of data collection.

In this paper, we propose an energy-efficient data collection scheme using lossless compression for IWSNs, named ECL, to keep high data accuracy when aggregating data. Several sensors construct an aggregation region if their sensing duplication rate equals or is larger than a preset threshold. In this way, the data in an aggregation region are spatial correlated, and the contents are similar, which would be aggregated efficiently and accurately. Moreover, since there are only two levels of nodes in an aggregation region, it limits the hops of redundancy forwarding and decreases the transmission overhead as much as possible. Additionally, for further energy efficiency, a proper neighbor node could also be selected as an aggregation node. Through lossless compression algorithms, ECL keeps the original entropy by only deleting the repeated information in the collected data.

The contributions of ECL scheme are as follows. (1) Utilize the sensing duplication rates to construct aggregation regions. It helps to improve the accuracy of information and diminish the energy consumption. (2) Three aggregation rules are designed to support aggregating node selection from available parent nodes or appropriate neighbor nodes. Thus, it establishes an efficient and accurate data route for each sensor. (3) From lots of simulation experiments, ECL scheme shows high information accuracy, energy efficiency and fast collection of data.

The reminder of this paper is organized as follows. Section 2 introduces related work on data aggregation. After discussing on aggregation rules used in ECL in Sect. 3, the implementation of ECL is illustrated in Sect. 4. Experimental results are analyzed in Sect. 5, and Sect. 6 concludes this paper.

2 Related Work

In IWSNs, data aggregation helps to remove the same information from several collected packets and thus reduce the resource consumption when delivering information from sensor nodes to the sink. In specific, data aggregation schemes are classified into three categories, i.e., tree-based aggregation [7,8], hybrid aggregation [9], and cluster-based aggregation [10–13].

Cluster-based aggregation usually has better scalability and higher energy efficiency than tree-based and hybrid aggregation [10,11]. Since the cluster heads which are closer to the sink, relay more data than others, they are easy to run out of power. In an energy-efficient unequal clustering scheme (EEUC) [12], cluster heads are elected by localized competition, and the competition range becomes small when it is near the base station. Therefore, those clusters closer to sink have smaller cluster sizes than others, and the energy consumption of cluster heads is balanced. Even though, the cluster maintenance is somewhat difficult.

In order to deal with the cluster update problem, Yi and Yang design Hamilton energy-efficient routing protocol (HEER) [13]. A Hamilton path consists of members in a cluster, which take turns to be the unique cluster head. The first round of cluster construction is like LEACH and may result in energy hole problem. The assumption that all members in a cluster can communicate with each other is too strict, and data transmissions in turn along the Hamilton Path lead to a long delivery latency.

IWSNs are required to provide highly reliable and real-time transmissions. Shu et al. investigate the routing performance of TPGF in CKN-based duty-cycled IWSNs with radio irregularity, in terms of the number of explored routing paths as well as the lengths of the average and shortest routing paths. They prove that the cross-layer optimized version of TPGF finds reliable transmission paths with low end-to-end delay [14]. Considering the resource constraint in IWSNs, a cross-layer optimization scheme named Adjusting the Transmission Radius (ATR) is proposed. In EC-CKN-based WSNs, it solves two important problems, namely, the death acceleration problem and the network isolation problem [15].

Present data aggregation researches construct clusters based on preset geographical scale. If the preset scale is too large, the collected data from member nodes have low similarity, which results in low accuracy after aggregation; if the preset area is too small, data aggregation does not function well. To combine the advantages of data aggregation and high accuracy, we attempt to remove redundant data efficiently by using the threshold of sensing duplication rate to construct aggregation regions.

3 Network Model and Aggregation Rules

In an IWSN we focus on, all sensor nodes have the same sensing radius, denoted by R_S; the communication radius is the same, denoted by R_C, and $R_C > 2R_S$ [16]; the size of data collected from each sensor is d. Table 1 lists the main symbols used in ECL.

Definition 1. *Transfer topology L: The topology is a directed acyclic diagram of all the sensors, which indicates communication relations and levels.*

For the adjacent levels in the topology, the level with a smaller value is called upper-level (L_i) and that with a higher value is called lower-level ($L_i + 1$). Since a node may link with more than one up-level nodes, L is similar to but not the same with a tree structure. In a dense network we focus on, it is assumed that all the sensor nodes can communicate with the sink by one hop or multi-hop transfers. Therefore, all the nodes are included in the transfer topology.

Table 1. Symbols.

Symbol	Description
R_S	Sensing radius of a sensor node
R_C	Communication radius of a sensor node
SDR	Sensing duplication rate, which indicates the ratio of two nodes' sharing sensing area to each one's own sensing range
SDR_T	Threshold of sensing duplication rate, which is utilized to select the aggregation nodes
AR_i	Aggregation region with the aggregation node v_i, in which the collected information are aggregated by the node v_i
AN	Aggregation node, which aggregates data collected in an aggregation area and transfers the aggregated to sink
AVL_AN_i	Available aggregation node set of v_i, where the nodes have the privilege to be aggregation node
MN	Member node, which sends its data to the aggregation node in its aggregation region
AVL_P_i	Available parent node set of v_i, which is comprised of all the upper-level nodes that could communicate with v_i directly
N_i	Neighbor node set of v_i, including all the nodes in the same level that could communicate with v_i directly
IN	Independent node, which has a low sensing duplication rate with its available parent nodes and neighbor nodes, and sends its data to sink without data aggregation
L	Level, which represents the hierarchy value of the sensor node and $L \geq 0$
S_C	The overlapping sensing area of two nodes

An example of transfer topology is illustrated in Fig. 1. There are m nodes in the network. Node 0 (v_0) is the sink, and the edges show the possible communication probabilities among nodes. In particular, the parent-child relations are shown by solid lines, while the neighbor relations are indicated by dotted lines.

Definition 2. *Sensing duplication rate, SDR: The ratio of the sensing duplication area to the sensing range of each node ($\pi {R_S}^2$). Thus the nodes sharing the sensing duplication area have the same sensing duplication rate.*

Since each sensor monitors its whole sensing area, we assume that the amount of collected data from a sensor is proportional to the area of sensing range. Therefore, a larger SDR implies a larger amount of duplicated information. ECL scheme only removes duplicated data, and retains the entropy of all the original data. It works as a lossless compression algorithm with the sensing duplication rate as its compression ratio.

For the sake of data accuracy, a threshold of sensing duplication rate, denoted by SDR_T, is introduced for aggregation node selection. The specific value of

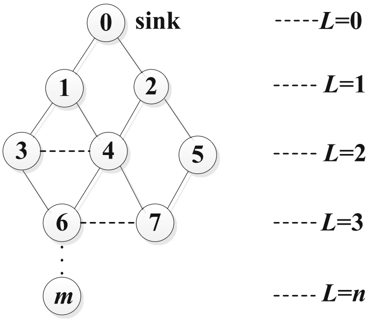


Fig. 1. Transfer topology.

SDR_T influences the performance of data aggregation. A small SDR_T may result in large aggregation regions, where the collected data from members have low similarity; otherwise, the larger SDR_T is, the smaller aggregation regions leads to low energy efficiency. Selecting a suitable value for SDR_T requires comprehensive analysis.

Definition 3. *Aggregation region, AR: An aggregation region is an area consisting of an aggregation node and several member nodes. A aggregation node gathers and aggregates all the data in a region and then forwards the results to the sink, while a member node only sends its data to its aggregation node.*

An aggregation region with node v_i as aggregation node and $v_j, \ldots, v_k$ as member nodes is presented with $AR_i = (v_i, \{v_j, \ldots, v_k\})$. Note that one node is included in at most one aggregation region; $AR_i \bigcap AR_j = \emptyset (i \neq j)$. If node v_k does not join in an aggregation region, v_k works as an independent node, $AR_k = (v_k, \emptyset)$. If a node v_m is not an aggregation node, we call v_m non-aggregation node. In the initial phrase, no aggregation regions exist, and all the nodes are free nodes.

Definition 4. *Available aggregation node set of v_i, AVL_AN_i: A set consists of all the nodes having the privilege to be the aggregation node for v_i.*

Definition 5. *Available parent node set of v_i, AVL_P_i: A set consists of those nodes at level $L_i - 1$ which could directly communicate with v_i (level L_i).*

Definition 6. *Neighbor nodes set of v_i, N_i: A set is composed of those nodes at level L_i which could directly communicate with v_i (level L_i).*

Figure 2 shows a transfer topology, which has three aggregation regions. v_3, v_6 and v_7, as the member nodes, transmit their data to the aggregation node v_4 directly, therefore $AR_4 = (v_4, \{v_3, v_6, v_7\})$. Similarly, aggregation node v_2 receives the data from v_5 and aggregates these data, $AR_2 = (v_2, \{v_5\})$. v_1 transmits its data to sink directly, expressed as $AR_1 = (v_1, \emptyset)$.

There are three rules to guide aggregation node selection and data routing in ECL.

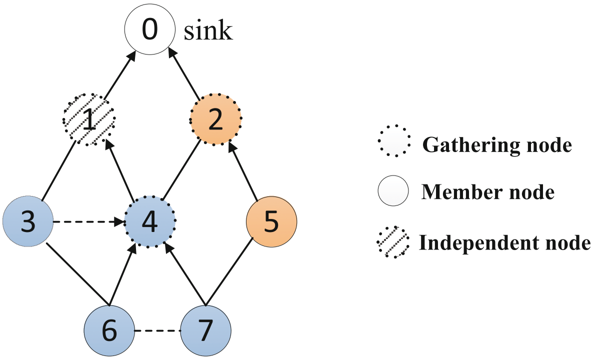


Fig. 2. An instance with three aggregation region.

Rule 1. *In an aggregation region, the available parent nodes have priorities to be chose as aggregation nodes.*
Rule 2. *After all the lower-level nodes joined an aggregation region, the free nodes in the upper-level prefer to find their aggregation nodes from the available parent nodes, and then the neighbor nodes are taken into account.*
Rule 3. *Sensor data are aggregated at most once. In the phase of relay node selection, the non-aggregation nodes are first choice and then the available parent node with most power is considered.*

4 Implementation of ECL Scheme

The transfer topology of sensor network is established through broadcasting hello messages among nearby nodes. Hello message has the sender's identifier, location, remaining power and the level in transfer topology. The nodes that could communicate directly at the same level are included in the neighbor node set, while the upper-level nodes communicating directly are added to the available parent node set. In initialization phase, the level of sink is set to 0, and that of sensors is infinity. Sink sends its hello message, and then other nodes update their levels with the small values compared with receiving hello messages. In detail, upon receiving a hello message, v_i checks whether its current level value L_i is 1 plus the level in received hello message. If true, v_i updates L_i with the level in hello message plus 1, and then propagates hello message with the new level value; otherwise, if the level in received message is larger than L_i by 1, v_i adds the identifier in hello message to AVL_P_i; if the difference between the local level and the received level equals 0, v_i stores the identifier in hello message to N_i.

4.1 Aggregation Region Construction

After transfer topology is finished, sensor nodes begin to construct the aggregation region distributively. Aggregation region construction is sequencing activity which starts from free nodes in the lowest level and then to the upper-level layer by layer. The construction phase follows three rules in Sect. 3. A free node might be picked as an aggregation node to construct an new aggregation region, or join in an aggregation region as a member node, or work as an independent node to deliver data to sink directly. Take v_i as an example, Algorithm 1 shows the aggregation region construction phase, and finding available aggregation nodes is described in Algorithm 2.

Algorithm 1. Aggregation region construction.

```
Input: the transfer topology, a free node $v_i$
Output: $AR_i$
if several nodes $M$ request to be members of $AR_i$ then
    construct aggregation region $AR_i = (v_i, M)$, and reply to $M$;
else
    if $L_i \neq 1$ then
        $AVL\_AN_i \leftarrow$ Algorithm 2;
        if $AVL\_AN_i = \emptyset$ then
            $v_i$ is an independent node, $AR_i = (v_i, \emptyset)$;
        else
            select $v_t \in AVL\_AN_i$ with max $E_t$ to be the aggregation node of $v_i$;
            send the request to be a member of $AR_t$;
        end
    else
        $v_i$ is an independent node, $AR_i = (v_i, \emptyset)$;
    end
end
return $AR_i$;
```

In Algorithm 1, the residual energy of node v_i is denoted by E_i. At the first, complete transfer topology construction with $(n+1)$ levels for all the free nodes, and thus the nodes at level n begin aggregation region construction. Any node v_i with $L_i > 1$ calculates its sensing duplication rates related to nearby nodes, and updates its available aggregation node set by executing Algorithm 2 (finding available aggregation node set algorithm). If the available aggregation node set AVL_AN_i is empty, v_i turns into an independent node; otherwise, v_i selects the node v_t with the most remaining power in CG_i as its aggregation node, and sends a request to join the aggregation region AR_t. Regarding a free node v_j in level 1, it is invalid to take the sink as aggregation node. Meanwhile, its sensing duplication rate with any neighbor cannot be bigger than 2 (Rule 2). Thus, v_j becomes an independent node, and sends its data to the sink without aggregation.

Algorithm 2 returns the available aggregation node set of v_i. For every available parent node v_j of v_i, if $SDR_{i,j} \geq SDR_T$, v_j is inserted to the set AVL_AN_i. Only if no available aggregation node is picked from the available parent node

Algorithm 2. Finding available aggregation node set.

```
Input: $AVL\_P_i$, $N_i$
Output: $AVL\_AN_i$
initialize:$AVL\_AN_i \leftarrow \emptyset$;
for $\forall\ v_j \in AVL\_P_i$ do
    if $\exists\ SDR_{i,j} \geq SDR_T$ then
        $AVL\_AN_i \leftarrow AVL\_AN_i \cup \{v_j\}$;
    end
end
if $AVL\_AN_i = \emptyset$ then
    for $\forall\ v_k \in N_i$ do
        if $SDR_{i,k} \geq SDR_T$ and $SDR_{i,k}(2L_i - 1) > 2$ then
            $AVL\_AN_i \leftarrow AVL\_AN_i \cup \{v_k\}$;
        end
    end
end
return $AVL\_AN_i$;
```

set, v_i considers the sensing duplication rate with its neighbors. If a neighbor v_k has $SDR_{i,k}(2L_i - 1) > 2$ and $SDR_{i,k} \geq SDR_T$, v_i stores v_k into AVL_AN_i.

4.2 Data Routing

In an aggregation region, member nodes only transfer their own data to the aggregation node by one hop transfer; the aggregation node collects and aggregates data from all the members in this region and transmits them to sink along energy-efficient paths; sensing data of independent nodes are sent to sink without any aggregation. Energy-efficient paths are selected according to Rule 3, in which the non-aggregation nodes have priority to forward data and then powerful available parent nodes are picked to be relays.

Take Fig. 1 as an instance. Its construction of aggregation regions and data routing are depicted in Fig. 3. Suppose that v_4 has more energy left than v_3, $SDR_{3,4} = 0.7$. At the highest level, there are two nodes, v_6 and v_7, which obtain the sensing duplication rates with their available parent nodes v_4, v_5 and v_3. Suppose that $SDR_{4,6} > SDR_{3,6} > SDR_T$, $SDR_{4,7} > SDR_T > SDR_{5,7}$, thus $CG_6 = \{v_3, v_4\}$, $CG_7 = \{v_4\}$. Due to $E_4 > E_3$, v_6 and v_7 both choose v_4 as their aggregation node, $GA_4 = (v_4, \{v_6, v_7\})$. Since no free nodes exist in level 3, v_3 and v_5 in level 2 start to find their aggregation regions. Considering the available parent nodes, suppose that $SDR_{1,3} < SDR_T$, $SDR_{2,3} < SDR_T$ and $SDR_{2,5} = SDR_T$. Thus, at the moment, $CG_3 = \emptyset$, and $CG_5 = v_2$. In this way, v_5 chooses v_2 as it aggregation node, $GA_2 = (v_2, \{v_5\})$. Since $CG_3 = \emptyset$, v_3 further considers its neighbor node v_4, and gets $SDR_{3,4} \times (2L_3 - 1) = 0.7 \times 3 > 2$. Since v_4 is the only neighbor of v_3, v_3 selects v_4 to be its aggregation node ($AVL\_AN_3 = \{v_4\}$). Until then, $AR_4 = (v_4, \{v_3, v_6, v_7\})$, and no free nodes are in level 2. According to previous analysis, all free nodes in level 1 should be independent nodes. Thus, v_1 turns into independent node. The aggregation region construction is completed here. For the data routing, nodes v_3, v_6 and v_7 transfer their own data to v_4, which aggregates all the data in AR_4. Then v_4 selects v_1 (non-aggregation node) as relay node to sink. Node v_5 sends data to its aggregation node v_2, which transfers data to sink directly, while v_1 delivers its data to sink without aggregation.

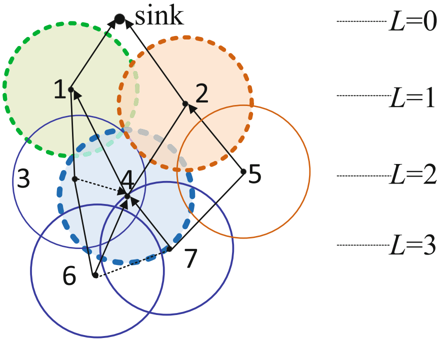


Fig. 3. An example of ECL implementation.

4.3 Complexity Analysis

In the Algorithm 2 (finding available aggregation node set), if each node v_i visits all the nodes in AVL_P_i and N_i, the computation complexity is $O(\max_{\forall i \in [1,m]} (|AVL\_P_i| + |N_i|))$ in the process of finding the available aggregation node set AVL_AN_i. Correspondingly, the computation complexity of Algorithm 1 (aggregation region construction) is $O(\max_{\forall i \in [1,m]} |AVL\_AN_i|)$. Because the number of nodes in AVL_P_i, N_i and AVL_AN_i are all smaller than m (the number of sensors), the computing complexity of algorithms is $O(m)$.

With regard to transfer topology establishment, sink broadcasts the control message firstly and other nodes update and resend it out after reception, in which the control message complexity is m. In order to set aggregation regions, node v_i sends requests to its potential aggregation nodes, and after receiving acceptance messages, v_i replies with a response message to join in an aggregation region. In this phase, control message cost is $3m$. In conclusion, message complexity of ECL is $O(m)$.

With respect to additional information, every node only carries its level, remaining energy, available parent set and neighbor set, and thus the storage complexity is $O(m)$. The message complexities of HEER and EEUC both are $O(m)$, while HEER has exponential computation complexity. Overall, ECL has relatively low computation complexity and requires small storage spaces.

5 Performance Evaluation

5.1 Network Configurations

For accurate analysis, we analyze the performance of ECL scheme on OPNET Modeler [17] network simulation platform. Table 2 lists the network configurations. Note that sensor nodes are distributed in the target field with common density. In scenario, the nodes are deployed in a pyramid field, of which the top is the sink. In the experiments, for the adequate power of sink, only the energy of sensor nodes are taken into account. Since data transmission and reception cost most of the energy, small consumptions in data processing are ignored. Besides, the energy consumption of sensing is fixed, because every node collects the same size of data. Therefore only the transmission overhead is focused on.

Table 2. Simulation parameters.

Parameter	Value
Scenario (m^2)	100 × 100
Number of sink	1
Number of sensors	40, 80, 120, 160 and 200 (nodes)
Sensing radius (m)	25
Communication radius (m)	52
Data collection cycle (s)	60
SDR_T	0.5

In the simulation experiments, we compare the proposed scheme ECL with a latest data collection scheme HEER and a typical aggregation scheme EEUC. HEER uses Hamilton Path to realize data routing and forms clusters in a similar way with LEACH. Specifically, Hamilton Path is designed for the sequence of data transfers, as analyzed in Sect. 2. In another compared scheme EEUC, cluster heads are elected by localized competition in a distributed way. In the cluster construction phase, a competition range is calculated according to the distance to the base station, and is used to control the sizes of clusters. The final cluster heads are elected by several tentative nodes. After the cluster heads have been decided, accordingly ordinary nodes request to be members of their closest cluster heads. In order to see the different performances of aggregation node selection algorithms, we also regard a variation of ECL, ECL-CP, as a compared scheme. In ECL-CP, the aggregation nodes are only selected from available parent nodes. In other words, neighbor nodes cannot be picked as aggregation nodes.

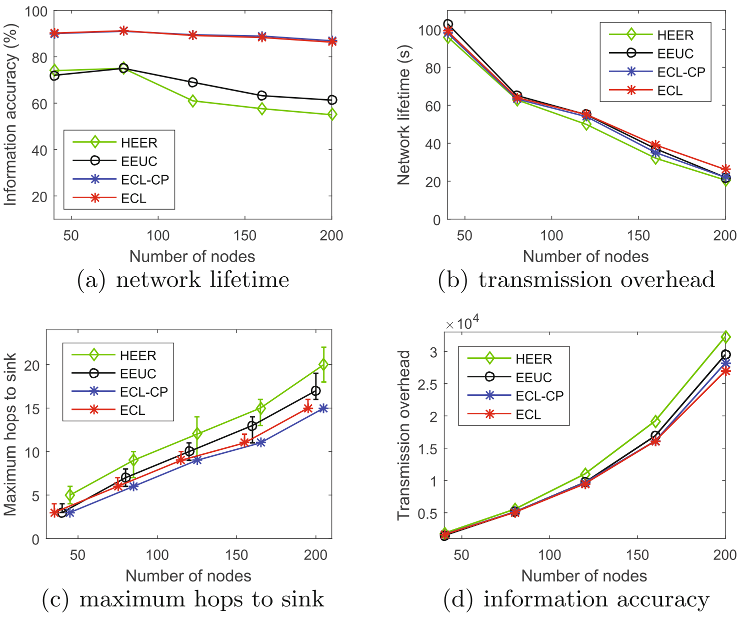
It is noteworthy that, in different data collection schemes, the sizes of aggregation output packets are not the same. In ECL, when an aggregation node aggregates x data packets, it gets $p(0 < p < x \cdot d)$ amount of data thereafter, where p is decided by sensing duplication rate in this aggregation region. Because

ECL uses lossless compression algorithm, which only deletes repetition of sensing duplication areas, the general information received by sink is correct and lossless. While in compared schemes, aggregation function is compressing x data packets into a fixed packet whose size is d.

For the performance analysis, four following metrics are introduced. First, information accuracy [18] is the proportion of the information entropy gathered by sink in all the generated data in the scenario. Second, network lifetime is the period of time from the start of data collection to the time when a sensor runs out of battery. Third, the longest path to sink is the longest path from sensors to sink in the network lifetime. A larger value for the longest path to sink indicates a longer delay for the sink to gather the sensing data from all the sensors deployed in the target monitoring field, and thus it represents the data collection latency. Fourth, transmission overhead is the number of data transmitted in one round of data transfer. It implies the energy consumption of data transfer and reception by all the sensors.

5.2 Experiment Results

In this subsection, we discuss the simulation results of all the schemes mentioned above, i.e., ECL, ECL-CP, EEUC and HEER, and the results are illustrated in Fig. 4.



(a) network lifetime

(b) transmission overhead

(c) maximum hops to sink

(d) information accuracy

Fig. 4. Experiment results.

As Fig. 4(a) depicts, the data correctness in ECL is about 88%, which is the highest compared with other three comparisons. It is very similar to ECL-CP and has 38% and 48% higher ratio than EEUC and HEER respectively. It is obvious that aggregating a large amount of data into a small package easily introduces information loss. Moreover, regarding an original data packet generated by a sensor, multiple times of data aggregation further aggravate the information inaccuracy. On a Hamilton Path in HEER, the end node begins data transmission to its neighbor which is closer to cluster head. After that, the neighbor aggregates its obtained data into one package with a fixed size, and then send it to the next node on the path until arriving at the cluster head. As a result, the raw data may be aggregated multiple times at different forwarders. Every cluster, in EEUC, aggregates data from all members into one packet with a constant size, not considering different duplication ratios. In comparison, sensor data of ECL are required to be aggregated once, and keep the information entropy through lossless compression. Therefore, ECL keeps the most information compared with others. Because some nodes are not included in the transfer topology (in other words, they can not communicate with others), the information accuracy is not 100% in ECL.

In the Fig. 4(b), EEUC has the longest lifetime with 40 sensors deployed in the scenario, and the lifetime of ECL is the second longest. As the number of nodes increases, the lifetime of ECL which is longer than ECL-CP, gradually becomes longer than EEUC when there are over 120 nodes. Meanwhile, HEER has the shortest lifetime. EEUC has a single length-fixed packet as aggregation output in each cluster, and the fixed size is smaller than that in ECL. There are two primary reasons why ECL gradually has long lifetime. (1) The duplicated data are transferred for multiple hops in EEUC, but ECL only forwards the redundant information once. (2) The cluster heads of EEUC may be in lower levels of the cluster, while in ECL, almost all the aggregation nodes are from upper levels, which ensures that the data always go up along the transfer topology without loops.

Since the longest paths in several rounds of experiments for HEER are not stable, median values are calculated in Fig. 4(c). As the figure illustrates, ECL-CP has the shortest paths and the numbers of hops are relatively stable. In addition, ECL only has few cases with long paths to sink. The reason for that is all the aggregation nodes in ECL-CP are the upper-level nodes, and hence data are only transferred upwards; in ECL, a part of aggregation nodes are picked from neighbor nodes, which slightly prolongs the paths for data collection. When there are more sensors, the numbers of hops along the longest paths in EEUC and HEED have sharp rises and fluctuate. The paths are longer than those in ECL, especially for HEED. With the scale of 200 nodes, the longest paths in ECL are shorter than those in EEUC and HEER by 12% and 25%, respectively. Because members transfer data to their aggregation nodes directly in ECL and ECL-CP, which ensures the shortest distances to sink. However, multi-hop routing inside a cluster is common in EEUC, and the members, in HEER, forward their data to the cluster head according to the order of nodes on Hamilton Paths, which leads to a longer path to sink and a longer end-to-end latency.

As Fig. 4(d) shows, when the network scale is 80 nodes, the transmission overhead of one data collection round in ECL is a little smaller than those in ECL-CP and EEUC schemes. When the number of nodes increasing, the transmission overhead of HEER, EEUC and ECL-CP are increasing quicker than ECL. In the 200 nodes scenario, ECL has a smaller transmission overhead than ECL-CP, EEUC and HEER by about 4%, 11% and 18% respectively. However, HEER always has the biggest overhead among four comparisons.

To sum up, the proposed scheme ECL reaches a high data accuracy, and meanwhile maintains energy-efficient and fast data collection.

6 Conclusion and Future Work

In order to support precise control in industrial systems, IWSNs are highly required in modern industry, and require an energy-efficient and lossless data aggregation protocol. In this paper, we propose an energy-efficient data collection scheme using lossless compression for IWSNs, named ECL. In the transfer topology, a threshold SDR_T is introduced for aggregation node selection from available parent nodes and neighbor nodes, which guarantees that sensors in an aggregation region keep a high correlation of collected information. Only cleaning redundant data does not reduce the data accuracy and ensures energy efficiency. Member nodes only forward data to their aggregation nodes by one hop transmission, and then the aggregation nodes aggregate sensor data immediately, which maintains a short collection latency. Simulation experiments on OPNET platform show that ECL scheme achieves a much higher data accuracy and a shorter latency than EEUC and HEER schemes, when working for a similar lifetime.

Moreover, network density is a significant factor for the proper assignment of SDR_T, which requires further study. Meanwhile, the energy-efficient data collection solution with a high accuracy for several concurrent events [19] also needs indepth explorations in the future.

Acknowledgments. We gratefully acknowledge the support from the National Natural Science Foundation of China (61502320, 61373161, 61173009 & 61572060), Science & Technology Project of Beijing Municipal Commission of Education in China (KM201410028015), Youth Backbone Project of Beijing Outstanding Talent Training Project (2014000020124G133), 973 Program (2013CB035503), CERNET Innovation Project (NGII20151004, NGII20160316), Beijing Advanced Innovation Center for Imaging Technology, and Cultivation Object of Young Yanjing Scholar of Capital Normal University.

References

1. Gungor, V.C., Hancke, G.P.: Industrial wireless sensor networks: challenges, design principles, and technical approaches. IEEE Trans. Industr. Electron. **56**(10), 4258–4265 (2009)
2. Tang, X., Juhua, P., Gao, Y., Xiong, Z., Weng, Y.: Energy-efficient multicast routing scheme for wireless sensor networks. Trans. Emerg. Telecommun. Technol. **25**(10), 965–980 (2014)

3. Zhang, X., Liang, W., Haibin, Y., Feng, X.: Optimal convergecast scheduling limits for clustered industrial wireless sensor networks. Int. J. Distrib. Sens. Netw. **2012**, 1319–1322 (2012)
4. Kumar, A., Baksh, R., Thakur, R.K., Singh, A.P.: Data aggregation in wireless sensor networks. Int. J. Sci. Res. **3**(3), 249–251 (2014)
5. Singh, S.P., Sharma, S.C.: A survey on cluster based routing protocols in wireless sensor networks. Comput. Sci. **45**(18), 687–695 (2015)
6. Zeb, A., Muzahidul Islam, A.K.M., Zareei, M., Al Mamoon, I., Mansoor, N., Baharun, S., Katayama, Y., Komaki, S.: Cluster analysis in wireless sensor networks: the ambit of performance metrics and schemes taxonomy. Int. J. Distrib. Sensor Netw. **12**(7) (2016)
7. Fasolo, E., Rossi, M., Widmer, J., Zorzi, M.: In-network aggregation techniques for wireless sensor networks: a survey. IEEE Wirel. Commun. **14**(2), 70–87 (2007)
8. Zhang, Y., Juhua, P., Liu, X., Chen, Z.: An adaptive spanning tree-based data collection scheme in wireless sensor networks. Int. J. Distrib. Sens. Netw. **2** (2015)
9. Manjhi, A., Nath, S., Gibbons, P.B.: Tributaries and deltas: efficient and robust aggregation in sensor network stream. Int. J. Distrib. Sens. Netw. **17**(1), 287–298 (2005)
10. Shukla, K.V.: Research on energy efficient routing protocol leach for wireless sensor networks. Int. J. Eng. **2**(3), 1–5 (2013)
11. Younis, O., Fahmy, S.: HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks. IEEE Trans. Mob. Comput. **3**(4), 660–669 (2004)
12. Li, C., Ye, M., Chen, G., Wu, J.: An energy-efficient unequal clustering mechanism for wireless sensor networks. In: IEEE International Conference on Mobile Adhoc and Sensor Systems Conference, pp. 597–604 (2005)
13. Yi, D., Yang, H.: HEER-a delay-aware and energy efficient routing protocol for wireless sensor networks. Comput. Netw. **104**(C), 155–173 (2016)
14. Shu, L., Mukherjee, M., Hu, L., Bergmann, N., Zhu, C.: Geographic routing in duty-cycled industrial wireless sensor networks with radio irregularity. IEEE Access **4**, 9043–9052 (2016)
15. Shu, L., Wang, L., Niu, J., Zhu, C., Mukherjee, M.: Releasing network isolation problem in group-based industrial wireless sensor networks. IEEE Syst. J. **PP**(99), 1–11 (2015)
16. Juhua, P., Yu, G., Zhang, Y., Chen, J., Xiong, Z.: A hole-tolerant redundancy scheme for wireless sensor networks. Int. J. Distrib. Sens. Netw. **1550–1329**, 184–195 (2012)
17. George, T., Trevor, C.: Simulation tools for multilayer fault restoration. IEEE Commun. Mag. **47**(3), 128–134 (2009)
18. Villas, L.A., Boukerche, A., de Oliveira, H.A.B.F., de Araujo, R.B., Loureiro, A.A.F.: A special correlation aware algorithm to perform efficient data collection in wireless sensor networks. Ad Hoc Netw. **12**(1), 10–30 (2011)
19. Dong, M., Ota, K., Liu, A.: RMER: reliable and energy-efficient data collection for large-scale wireless sensor networks. IEEE IoT J. **3**(4), 511–519 (2016)