INTELLIGENT COMPRESSIVE DATA GATHERING USING DATA FERRIES FOR WIRELESS SENSOR NETWORKS

Siwang Zhou, Qian Zhong, Bo Ou

Hunan University
College of Computer Science
and Electrical Engineering
ChangSha, China

Yonghe Liu

the University of Texas at Arlington
Department of Computer Science
and Engineering
Arlington, USA

ABSTRACT

The latest research progress of the theory of compressed sensing (CS) over graphs makes it possible that the advantage of CS can be utilized by data ferries to gather data in WSNs. In this paper, we leverage the non-uniform distribution of the sensing data field to significantly reduce the required number of data ferries, yet ensuring the recovered data quality. Specially, we propose an intelligent compressive data gathering scheme consisting of an efficient stopping criterion and a novel learning strategy. The proposed stopping criterion is based only on the gathered data, without relying on the priori knowledge on the sparsity of unknown sensing data. Our strategy minimizes the number of data ferries while guaranteeing the data quality by learning the statistical distribution of gathered data. Simulation results show that the proposed scheme improves the reconstruction quality compared to the existing ones.

Index Terms— Compressed sensing, data ferry, wireless sensor network

1. INTRODUCTION

Data gathering is one of the most fundamental tasks performed within wireless sensor networks (WSNs), where sensing data have to be collected from sensors to the Sink node (Data center) [1]. Data gathering using data ferries, also termed as data mules, has recently emerged as a new alternative to the traditional multi-hop transmission paradigm that usually suffers from high energy consumption of forwarding nodes [2–4]. Software agents can also be used as data ferries that migrate from sensor to sensor performing data processing autonomously [5]. The use of data ferries, to some degree, overcomes the disadvantages of multi-hop transmission methods, and significantly increases network lifetime. However, existing ferries based data gathering approaches usually need complex scheduling algorithms to optimize the paths of data ferries for improving the data delivery efficiency, since these ferries have to travel through the whole WSN [2–5].

It is noted that compressed sensing (CS) theory over graphs has been developing recently [6]. For a sufficiently connected graph with N vertexes, it has been proved that a k-sparse signal (k nonzero entries, or k significant entries) can be recovered using $\mathcal{O}(k \log N)$ random path measurements [7]. This makes it possible that the advantage of CS is utilized by data ferries to gather data in WSNs. Sartipi et al first studied the data gathering for WSNs based on CS over graphs and designed rateless coding and decoding algorithms to gather sensing data [8]. Zheng et al further introduced a random walk based data gathering algorithm for WSNs [9]. As we know, randomness is essentially the guarantee of data recovery for the theory of CS over graphs. However, for WSNs, randomness means that sensory data field has to be uniform, or smooth. These existing schemes for data gathering in WSNs, including [8] and [9], actually work well under this uniform or smooth assumption. Unfortunately, in more realistic sensing scenarios, data field may have regional fluctuations, and the existing random data aggregation schemes will yield a poor gathering efficiency.

In this paper, we introduce a data gathering scheme for WSNs using data ferries based on the CS theory over graphs. In the proposed scheme, a number of data ferries are allocated and one data ferry only visits a subset of nodes in WSNs. They randomly move among sensors, and we need not do path planning for any mobile ferries. To accomplish the data ferries based compressive data gathering, we need to address the following two research challenges:

(1) How many data ferries should be assigned for data gathering? Intuitively, increasing the number of data ferries increases the quality of the recovered data field. However, this means that more resources need to be used. It is not trivial to lower the number of data ferries as much as possible while still guaranteeing the predefined data quality.

(2) How to allocate those data ferries to the WSNs? In order to find the minimum number of data ferries, we need a method to efficiently identify the regions or sensor nodes whose sensing values, if collected, can help improve recovery accuracy to the maximum extent.

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

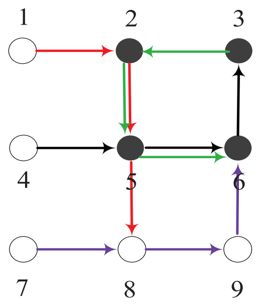


Fig. 1: A WSN with 9 nodes and 4 ferries (red, green, purple and black arrow, respectively)

The remainder of this paper is organized as follows. Section 2 introduces the idea of ferries based compressive data gathering over a WSN. Section 3 proposes a scheme, called ICDG, to overcome the above two research challenges. Section 4 analyzes the performance of the proposed scheme. Finally, we conclude the paper in section 5.

2. CS OVER A WSN USING DATA FERRIES

In this section, we make a connection between the ferries based data gathering and the CS theory over graphs.

Consider a wireless sensor network with N sensor nodes, represented by an undirected graph $G = (V, E)$, where V is the vertex set with cardinality $|V| = N$ and E is the edge set. Suppose that we have M data ferries who walk along M source-destination pairs over the WSN, respectively. Abstractly, let x be an $N \times 1$ vector whose s^{th} element represents the sensing datum over the sensor s, and let y be an $M \times 1$ dimensional vector whose l^{th} element is the sum of those data gathered by the l^{th} data ferry. In other word, the N-length vector y is the data gathered by N ferries. Then $y = \mathcal{A}x$, where $\mathcal{A}$ is an $M \times N$ matrix, whose element in the l^{th} row and s^{th} column is '1' if the s^{th} sensor node is in the route of the l^{th} data ferry and '0' otherwise.

For example, for a WSN with 9 nodes and 4 data ferries, as shown in Fig. 1, the CS measurement matrix $\mathcal{A}$ is:

$$\mathcal{A} = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

According to the theory of CS over graphs, x can be recovered through y and the measurement matrix $\mathcal{A}$. The ferries based data gathering is compressive since $M << N$.

3. THE PROPOSED SCHEME

In this section, an intelligent compressive data gathering (ICDG) scheme is proposed. We first present a stopping criterion to stop data delivery of ferries once recovered data meets the predefined quality requirement. Then, we introduce a learning strategy to adaptively allocate new data ferries, which can increase the reconstruction accuracy while minimizing the total number of data ferries. After that, we give the proof that the proposed scheme can reach the condition for l_1 minimization recovery according to the CS theory over graphs.

3.1. Stopping Criterion

In the process of data gathering scheme for WSNs, deciding when to stop data delivery for data ferries is a key issue, since we may not have any priori knowledge on the sparsity of the data field. If data ferries are stopped too early, the Sink might not collect enough data to achieve the predefined quality. If they are stopped too late, then the Sink might collect redundant data, which would lead to additional resource overheads.

We propose a simple but efficient stopping criterion. The data field is reconstructed in the Sink when $M+jT$ data from $M+jT$ data ferries are received sequentially, where M is the initial number of ferries, T is the number of the new ferries and $j \geq 0$. If the accuracy of the recovered data does not meet the required quality, then sensory data will continue to be collected; If meets, a feedback is formed to inform data ferries to stop gathering data.

However, the original sensory data is unknown, and we can not measure the accuracy of the recovered data directly by comparing them with the unknown ground truth. In the proposed criterion, we estimate the data quality via two successive data recovery. According to [10], for a k-sparse N-length signal x, the correct recovery can be declared if $\hat{x}^{M+(j+1)T} = \hat{x}^{M+jT}$, then $\hat{x}^{M+jT} = x$, where $\hat{x}^{M+(j+1)T}$ and $\hat{x}^{M+jT}$ denote the recovered data by $M + jT$ measurements and sequential $M + (j + 1)T$ measurements, respectively. If x is a compressible signal, then the recovery accuracy $error(\hat{x}, x)$ equals $C_{M,T} \cdot error(\hat{x}^{M+T}, \hat{x}^{M})$ where the mean $E(C_{M,T}) \approx \sqrt{\frac{N-M}{T}}$.

To implement the proposed stopping criterion, it intuitively needs to run the reconstruction algorithm for $j + 1$ times. Fortunately, there is some potentials of "memory" in the sequential $j + 1$ reconstructions, and the sequential solutions can be solved with low complexity [10].

3.2. Online learning

If the Sink receives $M + jT$ $(0 \leq j < i)$ data and still does not reach the stopping criterion, then additional T ferries have to be allocated. In this section, an online learning strategy is presented to adaptively assign data ferries. The number and the granularity of online learning depend on the range of j and the value of T, respectively.

3.2.1. *Compute weight map*

We first present a weight map approach to estimate the distribution of the data field. Suppose a WSN with N nodes is divided into r non-overlapped regions in advance, mathematically $R_1, \ldots, R_l, \ldots, R_r$, where $R_l = (n_1, \ldots, n_j)$. That is, the l^{th} region consists of n_j nodes. Let the weight of the region R_l be w_l. Then the weight map $\mathcal{W} = (w_1, \ldots, w_l, \ldots, w_r)$.

In the Sink, a sample data field is recovered once the new $jT(j = 0, 1, \cdots, i)$ measurements, denoted as $y_{M+(j-1)T}$, $\cdots$, y_{M+jT-1}, are received sequentially. Suppose the data field is sparse. It is easy to find that, if the k^{th} data ferry visits more sensor nodes with significant data than y_r does, then the value of y_k is bigger than y_r. In turn, if y_k is a big number, then the regions that k^{th} data ferry visits contain more sensor nodes with significant information, and thus these regions should be assigned higher weight. That is the intuitive idea upon which the proposed approach is built. For a walk $route_r$, if its measurement is y_r, and the number of regions that it visits is b_r, then each region associated with the walk $route_r$ will be allocated sub-weight $\frac{y_r}{b_r}$. So, for some region R_l, the weight is $\omega_l = \sum_{r=1}^{M} \frac{y^l}{b_r}$, where $y^l = 0$ if $R_l \cap route_r = \emptyset$, otherwise, $y^l = y_r$.

In many realistic scenarios, the data fields in WSNs are not sparse. However, their representations in some transform domain only have a small number of significant entries. Those data fields are generally called compressible signals. For a compressible signal x, the entropy $H(x)$ can be utilized to calculate its weight. That is, $H(x) = -\sum_l p(x_l) \cdot log(2, p(x_l))$, where $p(x_l)$ is the appearing probability of x_l in x. Suppose x is the signal in the region R_l, and for convenience let h_l represents its entropy. Then the weight of the region R_l is $\omega_l = h_l$

Although the weight map is computed only using partial measurements, it can reveal the distribution of significant information in sparse or compressible signals, which will be validated in the simulation.

3.2.2. *Allocate data ferry using weight map*

The number of starting nodes assigned to the region R_l with weight ω_l is computed as $m_l = \frac{\omega_l}{\omega} \cdot M$, where M is the initial number of ferries, and $\omega = \sum_l^r \omega_l$. According to this assignment, the region that contains more significant data will be allocated more data ferries. On the contrary, regions with more non-significant information will be less sampled.

If the total number of data ferries is staying constant, the data construction quality can be improved, since the number of data ferries for each region is periodically re-adjusted consciously when j in $(M + jT)$ changes from 0 to i. The value of T is capable of being used as granular control of online learning. If T is a relatively small value, then the proposed online learning algorithm runs with fine grained ability but leads to higher reconstruction complexity.

It can be seen that the proposed allocation strategy will be simplified into a general data gathering scheme, as described in [9], if the data field is completely smooth or uniform.

3.3. Guarantees for Signal Recovery

It has been proved that the whole data field can be reconstructed by the samples using CS over graphs with a high probability [6, 9]. In this section, we will analyze and prove that the measurement matrix formed by the proposed online learning strategy can still reach the condition for l_1 minimization decoding algorithm even if the starting nodes are chosen deliberately.

Suppose that the matrix of transition probability for a Markov chain is $P = \|P_{ij}\|$. If a probability distribution $\{\pi_i, i \geq 0\}$ satisfies $\pi_j = \sum_{i=0}^{\infty} \pi_i P_{ij}$ and $\sum_j \pi_j = 1, \pi_j \geq 0$, then we call it a stationary distribution of the Markov chain. As we know, if the degree of each vertex in a undirected graph $G = (V, E)$ with vertex set V and edge set E is between D and cD where constant $c \geq 1$, then the graph $G(V, E)$ is a (D, c) uniform graph.

Definition 3.1. (δ-stationary time [7]) Let $G = (V, E)$ be a (D, c) uniform graph, and π be the stationary distribution of a random source-destination path over the graph G. The stationary time $T(G)$ of G is defined as the smallest t' over which the path reaches a stationary distribution. That is, $\|\pi - \pi'\|_\infty \leq \delta$ where π' is the distribution that a path with length t' starting at a random vertex.

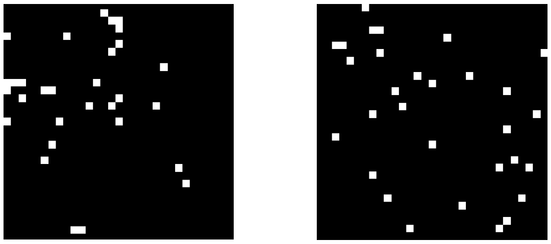
If a network meets the requirements of a uniform graph, a random walk over it forms a Markov chain. In the proposed scheme, although the nodes that data ferries start to walk are limited to some specified region by the weight map, the process of walks of data ferries still satisfy the property of Markov chain. So, a stationary distribution will be reached once data ferries walk over a stationary time.

Proposition 3.1. *Let $B_t(v)$ be an event that a random walk starting at node u visits node v over a period of t. Then the probability $Pr(B_t(v))$ satisfies $\frac{(1-\mu)t}{2(\eta+2)cn} < Pr(B_t(v)) < \frac{(1+\mu)t}{cn}$, where $\eta = \frac{1}{(1-\mu)\kappa}$, and c is a constant. [9]*

From the above proposition, starting from any node $u \in V$ for a random walk, the probability that a node is visited is $Pr(B_t(v))$ when a stationary distribution is reached. In other words, after $T(G)$, the walk in the proposed strategy will reach the stationary distribution as the general random walk does. Therefore, the measurement matrix $\mathcal{A}$ conforms to the requirement of a (D, c) uniform graph, and the signal can be recovered using an l_1 algorithm.

4. NUMERICAL SIMULATIONS

In this section, we evaluate the performance of the proposed ICDG scheme and compare it with the existing random



(a) (b)

Fig. 2: Two sparse signals with different distributions

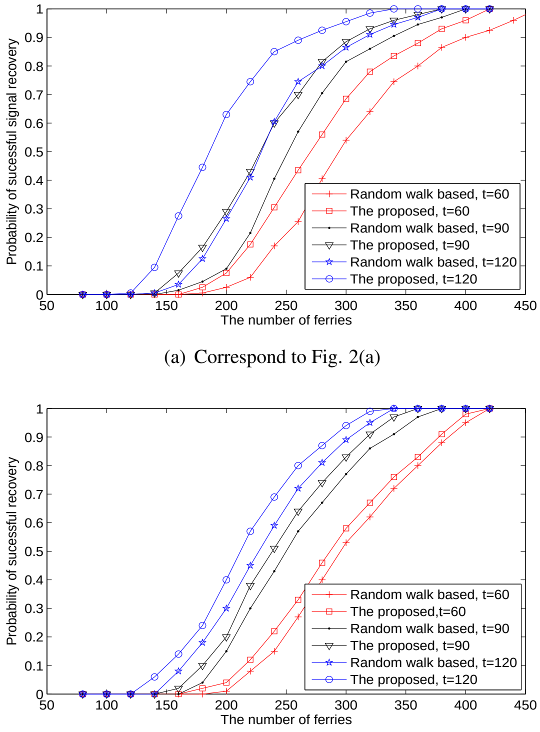
walk based ones [9]. All simulations are implemented on MATLAB and BP algorithm [11] is used to recover the original signal. In the simulation, a WSN with 1024 nodes is considered, in which sensor nodes are deployed in a two-dimensional grid. The data fields, generated by the WSN, is represented as 32×32 two-dimensional signals.

Two sparse signals with different distributions are first chosen to perform the simulation where the sparsity $k = 30$, as shown in Fig.2(a) and (b), respectively, and white dots represent significant value (nonzero data). The initial number of data ferries M is set to 40 and the length of each walk is $t = 60$. Suppose that the WSN is divided into 4 equal regions, and assume that a k-sparse signal is recovered successfully if the decoding error is not larger than 10^{-3} in l_2 norm. Fig.3 gives the probability of successful recovery over 200 realizations. The probability of successful signal recovery using the proposed online learning strategy is higher than other schemes when the significant data are distributed non-uniformly, as shown in Fig. 3(a). The advantage of our scheme is not so significant as the distribution of the data field is becoming uniform, as illustrated in Fig. 3(b).

We also apply a compressible signal to evaluate the performance of the proposed scheme. A temperature data set, which is from 1024 sensor nodes in a monitoring project of ocean environment [12], is chosen to do simulation. Initial sampling ratio is set to 0.1. That is, the initial number of data ferries is about 10 percent of the data size of a data set. As shown in Fig. 4, our scheme has less relative reconstruction error than the existing ones along with the increase of data ferries thanks to the proposed learning strategy.

5. CONCLUSION

In this paper, we present an intelligent data compressive gathering scheme, called ICDG, for WSNs on the basis of CS theory over graphs using data ferries. Our idea is that more data ferries should walk in the regions with significant information and data gathering can stop as soon as the recovered data quality meets the predefined bound. In ICDG, data ferries are adaptively allocated according to the proposed online learning strategy, but the walkable area of them is not limited. In this way, the non-smoothness or non-uniformness of data field in WSNs is explored while the randomness of data ferries is also guaranteed to sure that data field is capable of being recovered with high accuracy. Theoretically and experimentally, the proposed scheme is proved to be correct and has better reconstruction accuracy than the existing ones.



(a) Correspond to Fig. 2(a)

(b) Correspond to Fig. 2(b)

Fig. 3: The probabilities of successful recovery. For the random walk based approaches, the abscissa stands for the number of CS measurements.

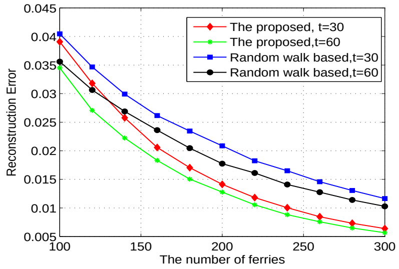


Fig. 4: The comparison of construction accuracy

References

[1] R.Rajagopalan and P.K.Varshney, *Data aggregation techniques in sensor networks: A survey*, IEEE Communications Surveys and Tutorials, vol.8, no.4, 2006.

[2] L. Chen, W. Wang, H. Huang and S. Lin. *Time-constrained Data Harvesting in WSNs: Theoretical Foundation and Algorithm Design*, in Proc. IEEE INFOCOM, 2015.

[3] R. Sugihara, R.K. Gupta, *Speed Control and Scheduling of Data Mules in Sensor Networks*, ACM Transactions on Sensor Networks, Vol. 7, No. 1, Article 4, 2010.

[4] L. He, J. Pan and J. Xu, *Progressive Approach to Reducing Data Collection Latency in Wireless Sensor Networks with Mobile Elements*, IEEE Transactions on Mobile Computing, vol.12, no.7, 1308-1320, 2013.

[5] Qi H., Xu Y., Wang X.. *Mobile agent based collaborative signal and information processing in sensor networks*. Proceedings of the IEEE, 91(8): 1172-1183, 2003.

[6] M. Wang, W. Xu, E. Mallada, and A. Tang, *Sparse Recovery With Graph Constraints*, IEEE Transactions on Information Theory, vol.61, no.2, 2015.

[7] W. Xu, E. Mallada, and A. Tang, *Compressive sensing over graphs*, in Proc. IEEE INFOCOM, pp. 2087-2095, Apr. 2011.

[8] M. Sartipi and R. Fletcher, *Energy-Efficient Data Acquisition in Wireless Sensor Networks Using Compressed Sensing*, Proc. IEEE Data Compression Conference (D-CC11), pp. 223-232, Mar. 2011.

[9] H. Zheng, F. Yang, X. Tian, X. Gan, X. Wang, and S. Xiao, *Data Gathering with Compressive Sensing in Wireless Sensor Networks: A Random Walk Based Approach*, IEEE Transactions on Parallel and Distributed Systems, VOL. 26, NO. 1, JANUARY 2015

[10] D. M. Malioutov, S. R. Sanghavi and A. S. Willsky, Sequential compressed sensing, IEEE Journal of Selected Topics in Signal Processing, 4(2): 435-444, 2010

[11] S. Chen, D.L. Donoho, M.A. Saljnders. *Atomic DecomPosition by Basis pursuit*, Siam Journal on Scientific computing, 20(l):33-61, 1999

[12] *National Oceanic and Atmospheric Administrations National Data Buoy Center.* `http://tao.ndbc.noaa.gov/refreshed/ctd_delivery.php`