An Efficient Data Aggregation Approach for Underwater Wireless Sensor Networks

Khoa Thi-Minh Tran¹, Seung-Hyun Oh², Jeong-Yong Byun³

¹, ² Department of Computer Science, Dongguk University, South Korea
³{ttmk84, shoh, byunjy}@dongguk.ac.kr

Abstract. The main goal of data aggregation technique is to gather data in energy manner for a long-term network monitoring. In data aggregation technique, the role of an aggregator is to collect sensed data from surrounding environment and transmit the collected data to base station. One part of data aggregation process is applying similarity function in order to minimize redundancy from the raw data and reduce the packet size is being sent to the base station. Our main research focuses on long-term underwater wireless sensor networks (UWSNs), especially in cluster-based UWSNs. In this paper, we evidence the effectiveness of similarity functions on reducing the packet size and minimizing data redundancy of cluster-based UWSNs. We show through the results that Euclidean distance and Cosine distance growth the efficiency of the network in both theory and simulation.

Keywords: data aggregation, similarity function, UWSNs

1 Introduction

Data aggregation has been inspected as an essential technique for reducing the energy consumption in wireless sensor networks by minimizing redundancy from the raw data sensed by the multiple sensor nodes as well as the number of transmissions to the sink or base station (BS) [1]. Not only the data aggregation process help to enhance the accuracy of information which is obtained by entire networks but also reduce the redundant information, the traffic load and the prolong the network lifetime [2]. However, the sensor nodes can be positioned randomly in the networks. Hence, neighboring sensor nodes may collect very similar data if they are positioned close to others. At that point, the data aggregator node may collect inconsistency data that need to be resolved before the data can be used for accurate data analysis [3]

The crucial issue for efficient deployment of UWSNs is to maximize the network lifetime. In UWSN data aggregation, the aggregator node aggregates the sensed data from surrounding nodes, process, and transmit the data to the sink or BS. So, the major challenge of data aggregation in UWSNs is to minimize data redundancy while ensure high data accuracy. A promising approach to compare between two set of captured data is using similarity functions.

The remainder of this paper is organized as follows. Section 2 provides detail information about our working on similarity functions in UWSN data aggregation. Section 3 shows the simulation results. Section 4 concludes the paper.
2 Similarity Functions in Data Aggregation

Data aggregation is one of key process of aggregator in order to help decreasing the network consumption by eliminating the redundancy as well as to reducing the packet size being transmit to the sink/BS. An aggregator has responsibility to collect sensed data from neighbor nodes and transmit collected data to the sink/based station node in order to save energy as well as reduce the packet size. The idea here is, at a certain time, each aggregator collects and stores a set of measured data as a vector. Then, the aggregator does identifying two pairs of sets whose similarities are above a given threshold. Hence, applying similarity function is one of the promising methods for the aggregator. Similarity function is a method which uses threshold to decide how much similar between two compared data. To reach the goal of minimizing network consumption and size of data packet, we work on applying similarity function for aggregators. If the compared data are considered similar each other, the aggregator no need to transmit all sets of data to the sink/BS node. There are some similarity functions can be used in set comparison, such as edit distance, Euclidean distance, Cosine similarity, Jaccard similarity, and generalized edit distance, etc. Paper [1] has proved that Euclidean distance and Cosine distance are two appropriate similarity functions for UWSNs. In this section, we describe in details how Euclidean distance and Cosine distance work and how each of them affect to the network.

The Euclidean distance is the dissimilarity between each pair of data in the data set, calculated by (1). Thus, u and v are said to be similar if \( E_d \leq t_d \).

\[
E_d = \sum_{i=1}^{n} \sqrt{(u_i - v_i)^2}
\]  

(1)

The Cosine distance equals one minus the Cosine of the included angle between two vectors, represented by (2). The Cosine of the included angle between two vectors is one kind of similarity. Thus, u and v are said to be similar if \( C_d \leq t_d \).

\[
C_d = 1 - \frac{\sum_{i=1}^{n} (u_i \times v_i)}{\sum_{i=1}^{n} (u_i)^2 \times \sum_{i=1}^{n} (v_i)^2}
\]  

(2)

Cosine and Euclidean distances directly use the collected value to compute the dissimilarity between pairs of data. However, the Cosine distance computes the distance based mainly on the included angle between two vectors, whereas Euclidean distance calculates the straight-line distance between two vectors. This generates different values that must be made as the same range for comparison. Normalization is a basic method for doing this. In (3), normalization formula [4], \( t_d \) is the original distance threshold value, \( t'_d \) is the normalized distance value, \( m \) denotes the mean value, and \( \sigma \) indicates the standard deviation of pairwise distance over all data.

\[
t'_d = \frac{t_d - m}{6\sigma} + \frac{1}{2}, \quad \text{(with} \quad \sigma = \sqrt{\frac{\sum_{i=1}^{n} (t'_i - m)^2}{n}})\]

(3)

3 Simulation Results

Figure 1 shows the percentage of data sent to sink node (left side) and percentage of deleted data before it is sent to sink node (right side). The two columns indicate the results from analysis, and the two lines indicate the results from simulation. The graph in the left side shows the results of simulations are lower than analysis. This is
because of the impact of many factors in the real environment. However, it can be solved by using transmission protocol. The graph in the right side shows the deletion of duplicated data that result in eliminating data redundancy. This proves that Euclidean and Cosine distances are well used for the purpose of reducing the redundancy of data which is being transmitted to the sink/ based station.

4 Conclusion

In this paper, we evaluated the percentage of data sent to the sink and the percentage of deleted data sets in both analysis as well as simulation. Those metrics are used to prove the effectiveness of Euclidean distance and Cosine distance on reducing the packet size and minimizing data redundancy sent to the BS.

Acknowledgment: This work was supported by the MKE (Ministry of Knowledge Economy) [A004700008], Development of realistic sense transmission system with media gateway supporting multi-media and multi-device.

References