Computer Networks 79 (2015) 103–119

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

A scalable approach for serial data fusion in Wireless Sensor Networks

A. Mostefaoui [a,*], A. Boukerche [b], M.A. Merzoug [a], M. Melkemi [c]

[a] FEMTO-ST/ DISC dep., Franche-Comte University, 1 Rue Engel Gros, Belfort, France
[b] PARADISE, University of Ottawa, 800 King Edward Avenue, Ottawa, Ontario, Canada
[c] LMIA, Haute Alsace University, 4 rue des freres Lumiere, Mulhouse, France

ARTICLE INFO

Article history:
Received 18 April 2014
Received in revised form 22 December 2014
Accepted 24 December 2014
Available online 7 January 2015

Keywords:
Localized algorithms
Serial data fusion
Wireless Sensor Networks

ABSTRACT

In Wireless Sensor Networks (WSNs), serial data fusion approaches, in which a parameter of interest is estimated through the serial communication of nodes, have shown their effectiveness over centralized and distributed ones. Nevertheless, they still suffer two major drawbacks: (i) they require the construction of a path passing through every node in the network exactly one time, which is known to be a NP-Complete problem and (ii) they experience poor scalability, which is an important concern in large scale WSNs. In this paper, we tackle these issues by proposing a novel localized serial algorithm, called Peeling Algorithm (PA). In the proposed algorithm, a packet travels serially from node to node, carrying with it the parameter estimate. Each visited node determines locally the next hop for the packet and does not need to store any information about the network topology. This unique feature allows a very good scalability of our approach. We also present a second algorithm, called Enhanced PA (EPA). We discuss their implementations, provide proof of correctness and report on their performance evaluation through an extensive set of simulation experiments using OMNET++ simulator. Our results indicate clearly that our proposed algorithms outperform previously known and existing ones.

1. Introduction

A Wireless Sensor Network (WSN) is composed of hundreds to several thousands of nodes deployed in an area of interest. Using their sensing, computing and wireless communication capabilities, the sensors respond to the application/end-user queries by collecting and forwarding data to a base station, also known as the sink node [1]. Usually, the objective behind this process, called *query processing* [2], is either to get a lower/upper value (e.g., minimum or maximum) or to derive an estimate of a parameter of interest [3] (e.g., average [4], target location [5–7], etc.), rather than collecting all raw data at the sink. This *in-network processing*, where data is collected and aggregated at the same time the query is spread (e.g., processed) in the network, has been established as a very efficient technique in WSNs [8].

In-network query processing in WSNs has attracted many research efforts that can be classified in three categories: (a) centralized approaches, (b) distributed approaches and, more recently, (c) serial approaches.

Centralized approaches: these approaches operate in two steps [2,8,9]: the first step is dedicated to constructing a tree-based structure, rooted at the sink, in order to disseminate the queries. In the second step, each node, upon receiving a query, will aggregate its own data with the data of its children before sending the result to its upper node.

* Corresponding author.
E-mail addresses: Ahmed.Mostefaoui@univ-fcomte.fr (A. Mostefaoui), boukerch@site.uottawa.ca (A. Boukerche), amine.merzoug@gmail.com (M.A. Merzoug), mahmoud.melkemi@uha.fr (M. Melkemi).

http://dx.doi.org/10.1016/j.comnet.2014.12.008
1389-1286/

Besides the fact that these approaches are not scalable, and, consequently, are not suitable for large-scale deployments, they present several additional drawbacks:

- Over-utilization of the network resources, mainly communications.
- Generation of a high degree of collisions, especially in large scale and dense networks.
- Low resilience to nodes and link failures as any link or node failure will trigger the re-construction of the entire tree-structure, in particular if the broken link or node is near the root of the tree. This could be very costly in large-scale networks.
- Raising the risk of the energy depletion hole problem [10,11] which occurs when some nodes, located near the root, are overused while relaying the traffic to the root.

Distributed approaches: in large-scale WSNs, the two last limitations of centralized approaches can have a deeply negative impact on the overall query processing performances. To overcome these two drawbacks, distributed approaches, based on in-network processing, were proposed as interesting alternatives [2,8,12–14]. In such approaches, each node maintains a local estimate of the unknown parameter. This estimate is *refined* successively (i.e., iteratively), through communications with its immediate neighbors, until convergence to the right value is reached (i.e., the difference between two consecutive estimates is less than a convergence threshold). In these approaches, nodes do not need to have an overview of the current network topology, since all the communications are 1-hop. Furthermore, since each node holds locally the estimate, there is no need to store any routing data as there is no need for a central base station. Finally, these approaches are more robust against link and node failures [15,16].

Even though distributed approaches are independent of any rooting structure, they remain dependent on the network topology; i.e., in sparse deployments (linear topologies for instance), the convergence of the unknown parameter can require many iterations, leading thus to a huge communication overhead, especially in large-scale networks. Knowing that communication is the most energy consuming task in WSNs [1], this limitation can seriously compromise the whole network lifetime. What is worse, concurrent communications between nodes, at the same iteration, could not only increase collisions, but can also augment considerably the query response time, as observed in our performance evaluation, presented in Section 6.

Serial approaches: in these approaches, the query is processed serially from node to node until all nodes in the network have been visited (i.e., contributed to the query). The final node holds the right value of the query [17,18]. Serial algorithms achieve superior performance, under certain conditions viewed below, in comparison to centralized and iterative approaches. Furthermore, they do not generate any communication collisions, since all the communications in the network are serial; i.e., at each step, only one node transmits. Consequently, the query response time is noticeably improved, as confirmed by our simulation study presented in Section 6.

Apart from the theoretical foundations of serial approaches (i.e., convergence proofs in the case of a parameter/function estimation) discussed and provided in [17–19], to the extent of our knowledge, some practical issues of serial approaches, exposed in Section 1.1 below, remain open and deserve additional research efforts, especially in large-scale and randomly deployed network topologies.

1.1. Motivations

As a matter of fact, many previous serial approaches required a Hamiltonian path (i.e., a path which passes through all nodes in the network and visits each node just once) [17]. Yet, previous research works make implicit the assumption that such a path exists, which is not true for any network configuration. Furthermore, even when such a Hamiltonian path exists, finding it is known to be an NP-Complete problem [20]. Accordingly, the construction of such a path in a decentralized fashion, to ensure scalability, could generate a prohibitive overhead, particularly in sparse and large-scale networks.

To overcome this obstacle, the authors in [21] use space filling curves to derive a path that is not necessarily Hamiltonian (i.e., nodes could be visited more than once). This approach, although it performs well in dense and regular networks topologies, cannot handle all network topologies, specifically those with holes. Indeed, it is not able to ensure query completeness (i.e., make sure that all connected nodes in the network contribute to the query). This is particularly harmful for sensitive applications, in which query accuracy is essential.

The motivation that drives our work is the design of a novel serial algorithm with the following requirements:

- *Network topology*: we want the proposed approach to be able to handle any network topology, provided that the latter is connected.
- *Localized nature*: to be scalable, which is a fundamental requirement in large-scale network deployments, our approach has to be of a localized nature. This means that the current node, when selecting its next query hop, has to consider only local available information and no additional information is requested.
- *Query completeness*: the proposed approach has to ensure that all nodes will be queried.
- *Rooting structure*: in some serial approaches, as those using space filling curves [21], the visiting order of nodes, named *visiting path*, is constructed/known in advance. By doing so, such approaches increase their *vulnerability* since the current node, when selecting the next hop, has no choice other than following the predefined path. Hence, it cannot easily handle topology changes; e.g., in the case where the next hop is not available. On the contrary, in our proposal, we want the path to be constructed *gradually* at each hop.

1.2. Contributions

In this paper, we extend the preliminary work initiated in [22], by improving our previously proposed algorithm, called Peeling Algorithm (PA), and by proposing a novel