Information Fusion 22 (2015) 39–49

Contents lists available at ScienceDirect

Information Fusion

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

Information extraction from sensor networks using the Watershed transform algorithm

Mohammad Hammoudeh [a,*], Robert Newman [b]

[a] Manchester Metropolitan University, Chester Street, Manchester M1 5GD, UK
[b] University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK

ARTICLE INFO

Article history:
Available online 29 July 2013

Keywords:
Information extraction
Watershed segmentation
Query scoping
Macroprogramming
Wireless sensor networks

ABSTRACT

Wireless sensor networks are an effective tool to provide fine resolution monitoring of the physical environment. Sensors generate continuous streams of data, which leads to several computational challenges. As sensor nodes become increasingly active devices, with more processing and communication resources, various methods of distributed data processing and sharing become feasible. The challenge is to extract information from the gathered sensory data with a specified level of accuracy in a timely and power-efficient approach. This paper presents a new solution to distributed information extraction that makes use of the morphological Watershed algorithm. The Watershed algorithm dynamically groups sensor nodes into homogeneous network segments with respect to their topological relationships and their sensing-states. This setting allows network programmers to manipulate groups of spatially distributed data streams instead of individual nodes. This is achieved by using network segments as programming abstractions on which various query processes can be executed. Aiming at this purpose, we present a reformulation of the global Watershed algorithm. The modified Watershed algorithm is fully asynchronous, where sensor nodes can autonomously process their local data in parallel and in collaboration with neighbouring nodes. Experimental evaluation shows that the presented solution is able to considerably reduce query resolution cost without scarifying the quality of the returned results. When compared to similar purpose schemes, such as "Logical Neighborhood", the proposed approach reduces the total query resolution overhead by up to 57.5%, reduces the number of nodes involved in query resolution by up to 59%, and reduces the setup convergence time by up to 65.1%.

1. Introduction

Wireless Sensor Networks (WSNs) are enabling the production of applications that previously were not practical. They are currently being applied to a variety of use domains ranging from habitat monitoring to space exploration, and from scientific to military. Such applications share several aspects: (1) the demand for information; (2) the response to this demand generally exists in multiple unstructured, potentially unbounded sequences of data points; (3) the generation of great volume of data that is 'imperfect' in nature and is characterised by significant redundancy. The characteristics of the sense data coupled with the resource constraints on sensor nodes necessitate the development of resource-efficient WSN applications. In-network information extraction is one method to minimise resource utilisation while achieving applications objectives. This is due to the fact that local processing of sensed data is more energy and bandwidth efficient than the transfer of raw data to a central location for processing.

Information extraction is the sub-discipline of artificial intelligence [1] that selectively structures, identifies, filters, classifies, and merges multi-modal data produced by multiple sensor nodes to discover recurring patterns that form a coherent and meaningful information. Initially, information extraction converts, possibly unbounded, sequences of raw data into a more uniform and convenient structural format preparing it for further processing and analysis. It exploits domain-specific knowledge and data structural properties to harvest information by finding and combining relevant data while excluding irrelevant and erroneous ones. This definition will be used throughout this work to describe the term information extraction.

This work addresses distributed query-based information extraction systems that use in-network computation to provide cost effective query responses. Nevertheless, the proposed solution is general enough to be applicable to other information extraction models, e.g., periodic or threshold-based. The query-based model of information extraction is widely used in data intensive applications, where data is stored at multiple locations. Such systems are user controlled where users specify and inject a request for the information they require through simple queries; then the

* Corresponding author. Tel.: +44 0 161 247 2845.
E-mail addresses: M.Hammoudeh@mmu.ac.uk (M. Hammoudeh), R.Newman@wlv.ac.uk (R. Newman).

1566-2535/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.inffus.2013.07.001

information extraction infrastructure efficiently retrieves and processes the data within the network. Generally, query-based systems support high-level declarative languages for incorporating data retrieval and analysis. In this setting, most query languages provide an interface to the WSN user whilst concealing low level details such as the network topology or communication paths. Often, the query issuer is unaware of how the query is disseminated in the network and how the data is retrieved or analysed.

A fundamental requirement to effectively use the query-based information extraction model is the user' s prior knowledge of the semantic data model and its contextual meaning. This allows network users to better understand, query, and control the extracted information. For instance, it could be essential to recognise what kind of events are occurring in a specific segment of the monitored area and when these events occurred. Depending on the type of the application, sensor data can be labelled with temporal, spatial, or other thematic semantic metadata. Accordingly, based on the user information requirements, queries can be built to distinguish a variety of events at various levels of semantic detail. In time critical applications, for example, it may be sufficient to state that a portion of a query is a temporal operator; whereas in other applications, it may be essential to distinguish among diverse temporal operators, e.g., operators specifying past, present and future. In event-driven applications, the semantic nature, scope, and unit of the requested information could be predefined. The unit of extraction defines the granularity of individual information chunks that are extracted from the sensor nodes. The scope of extraction is the process of identifying a subset of data or nodes holding data that is relevant to the requested information. To choose the chunks of information that should be considered in preparing the query response, an information extraction system uses a group of conditions. These conditions specify the formal attributes a specific chunk of information must have to belong to a certain semantic group.

As an example of this kind of requirement, from a real-world and practical application, in the recently completed RFID from Farm to Fork project [2], undertaken by one of the present authors, a system was built to trace provenance and condition of food products. The general strategy was that RFID systems were used to localise food products, while WSNs were used to monitor the environment through which they passed and condition information as temperature and humidity. It is required that the condition information be associated with an individual food item, so the resolution of a query requires determination of the location of an item at a particular time and estimation of the environmental parameters at that location and time. Sensor information not associated with any food product (the majority of it) is surplus to requirements. In-network information extraction allows only the required information to be collected, avoiding the swamping of the systems with unnecessary data.

The focus in this work is on large-scale sensor networks where multiple sensor types are fitted to each node. In such networks, many low-observability events require the retrieval of multi-modal data. To cope with the high-volume sensor data, we explore network segmentation techniques to provide effective mechanism for multi-modal query dissemination and local collaboration among nodes. Network segmentation provides high level abstractions for local and spatial data processing. The aggregation of sensor data and informed selective collaboration of sensor nodes can improve extracted information accuracy, reduce latency, and minimise bandwidth consumption. However, every node in the network performs sensing operations on every sensor without in-node preprocessing.

This paper proposes a distributed information extraction approach that retrieves data and accurately process it to respond to queries that users pose. This approach focuses on reducing the amount of data that has to be collected and analysed to obtain the desired information. It only query nodes that are closely linked to the user requested information. The ability to limit a query scope to arbitrary subsets of nodes results in considerable energy savings, lower bandwidth utilisation, and better information quality. The absolute result of the information extraction system depends on the nature of the query; however, it usually entails identifying entities sharing a common sensing state and the relations between those entities to provide accurate information access. Our proposed approach makes use of the Watershed transform algorithm [3] to assign homogenous nodes' to well delimited segments using nodes physical location and their sensors measurements. The purpose of the network segmentation process is to provide high-level primitives that abstracts away low-level system details. Unlike traditional information extraction approaches where each node is treated as an isolated computational unit, network segmentation allows efficient processing of high-level queries of logical interest over the entire network. Network segmentation is especially useful to dynamically build information rules and frames to quickly provide accurate answers to information queries at a low cost.

The Watershed algorithm is able to cope with the irregular and data dependent information requests by using the most suitable option at every processing step instead of using predefined threshold values. To comply with application constraints, typically the requirement for real-time processing and limited energy consumption, we propose a new asynchronous and parallel reformulation of the Watershed transform algorithm. The presented reformulation is able to calculate the Watershed transform based on the node's local data. It achieves similar outcome with just single point of synchronisation, hence reducing the running time without scarifying the segmentation effectiveness. These two claims are discussed in this paper and verified in experimental evaluations.

The remainder of this paper is organised as follows. The next section discusses the differences of our work from other closely related work. Section 3 introduces the Watershed algorithm. Section 4 explains the suitability of this algorithm for WSNs. Section 5 describes our modifications to the Watershed algorithm. Section 6 explain the use of segments as a query scoping mechanism. Section 7 presents the parallel asynchronous implementation of the Watershed algorithm. Section 8 explains the mechanism for resolving queries. The effectiveness of the presented approach is studied by simulation in Section 9. Section 10 concludes the work. Section 11 gives a number of interesting directions for future work.

2. Related work

The information extraction process begins by specifying the information needed by a user/application. Often, this takes the form of a query or event trap. Then, an information extraction approach selects a relevant subset of data or nodes carrying data that is relevant to the needed information from a larger set. To improve the accuracy of the extracted information and to minimise the cost of collecting raw data, high-level programming abstractions can be used [4–9]. Programming abstractions allow the user to specify how and where a query is disseminated. A large and growing body of literature has investigated micro- and macro-programming abstractions in the context of WSNs. Many methods and formalisms for abstraction have been proposed, but their design has often been focussed on synthetic problems, rather than requirements of working systems, such as the example given above. In this section different mechanisms to achieve abstraction are presented. We refer the interested reader to the recent survey [10] and the references therein for a comprehensive review of WSNs programming approaches.