Information Systems 38 (2013) 1285–1308

Contents lists available at ScienceDirect

Information Systems

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

In-network approximate computation of outliers with quality guarantees

Nikos Giatrakos [a,*,1,2], Yannis Kotidis [b,3], Antonios Deligiannakis [c,1], Vasilis Vassalos [b], Yannis Theodoridis [a,2]

[a] Department of Informatics, University of Piraeus, Central Building, 80 Karaoli & Dimitriou St., GR-18534 Piraeus, Greece
[b] Department of Informatics, Athens University of Economics and Business, 76 Patission St., GR-10434 Athens, Greece
[c] Department of Electronic and Computer Engineering, Technical University of Crete, University Campus., GR-73100 Chania, Greece

ARTICLE INFO

Available online 22 September 2011

Keywords:
Sensor network
Outlier
Locality sensitive hashing
Similarity

ABSTRACT

Wireless sensor networks are becoming increasingly popular for a variety of applications. Users are frequently faced with the surprising discovery that readings produced by the sensing elements of their motes are often contaminated with outliers. Outlier readings can severely affect applications that rely on timely and reliable sensory data in order to provide the desired functionality. As a consequence, there is a recent trend to explore how techniques that identify outlier values based on their similarity to other readings in the network can be applied to sensory data cleaning. Unfortunately, most of these approaches incur an overwhelming communication overhead, which limits their practicality. In this paper we introduce an in-network outlier detection framework, based on locality sensitive hashing, extended with a novel boosting process as well as efficient load balancing and comparison pruning mechanisms. Our method trades off bandwidth for accuracy in a straightforward manner and supports many intuitive similarity metrics. Our experiments demonstrate that our framework can reliably identify outlier readings using a fraction of the bandwidth and energy that would otherwise be required.

1. Introduction

Pervasive applications are increasingly supported by networked sensory devices that interact with people and themselves in order to provide the desired services and functionality. Because of the unattended nature of many applications and the inexpensive hardware used in the construction of the sensors, sensor nodes often generate imprecise individual readings due to interference or failures [1]. Sensors are also often exposed to severe conditions that adversely affect their sensing elements, thus yielding readings of low quality. For example, the humidity sensor on the popular MICA mote is very sensitive to rain drops [2].

The development of a flexible layer that will be able to detect and flag outlier readings, so that proper actions can be taken, constitutes a challenging task. Conventional outlier detection algorithms [3,4] are not suited for our distributed, resource-constrained environment of study. First, due to the limited memory capabilities of sensor nodes, in most sensor network applications, data is continuously collected by motes and maintained in memory for a limited amount of time. Moreover, due to the

* Corresponding author. Tel.: +30 210 4142449; fax: +30 210 4142264.
E-mail addresses: ngiatrak@unipi.gr (N. Giatrakos), kotidis@aueb.gr (Y. Kotidis), adeli@softnet.tuc.gr (A. Deligiannakis), vassalos@aueb.gr (V. Vassalos), ytheod@unipi.gr (Y. Theodoridis).

[1] These authors were partially supported by the European Commission under ICT-FP7-LIFT-255951 (Local Inference in Massively Distributed Systems).
[2] These authors were partially supported by the EU FP7/ICT/FET Project MODAP.
[3] Yannis Kotidis was partially supported by the Basic Research Funding Program, Athens University of Economics and Business.

0306-4379/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.is.2011.08.005

frequent change of the data distribution, results need to be generated continuously and computed based on recently collected measurements. Furthermore, a central collection of sensor data is not feasible nor desired, since it results in high energy drain, due to the large amounts of transmitted data. Hence, what is required are continuous, distributed and in-network approaches that reduce the communication cost and manage to prolong the network lifetime.

One can provide several definitions of what constitutes an outlier, depending on the application. For example in [5], an outlier is defined as an observation that is sufficiently far from most other observations in the data set. However, such a definition is inappropriate for physical measurements (like noise or temperature) whose absolute values depend on the distance of the sensor from the source of the event that triggers the measurements. Moreover, in many applications, one cannot reliably infer whether a reading should be classified as an outlier without considering the recent history of values obtained by the nodes. Thus, in our framework we propose a more general method that detects outlier readings taking into account the recent measurements of a node, as well as spatial correlations with measurements of other nodes.

Similar to recent proposals for processing declarative queries in wireless sensor networks, our techniques employ an *in-network processing* paradigm that fuses individual sensor readings as they are transmitted towards a *base station*. This fusion, dramatically reduces the communication cost, often by orders of magnitude, resulting in prolonged network lifetime. While such an in-network paradigm is also used in proposed methods that address the issue of data cleaning of sensor readings by identifying and, possibly, removing outliers [6,2,1,7], none of these existing techniques provides a straightforward mechanism for controlling the burden of the nodes that are assigned to the task of outlier detection.

An important observation that we make in this paper is that existing in-network processing techniques cannot reduce the volume of data transmitted in the network to a satisfactory level and lack the ability of tuning the resulting overhead according to the application needs and the accuracy levels required for outlier detection. Note that it is desirable to reduce the amount of transmitted data in order to also significantly reduce the energy drain of sensor nodes. This occurs not only because radio operation is by far the biggest culprit in energy drain [8], but also because fewer data transmissions also result in fewer collisions and, thus, fewer retransmissions by the sensor nodes.

In this paper we present a novel outlier detection scheme termed TACO (TACO stands for Tunable Approximate Computation of Outliers). TACO [9] adopts two levels of hashing mechanisms. The first is based on locality sensitive hashing (LSH) [10], which is a powerful method for dimensionality reduction [10–12]. We first utilize LSH in order to encode the latest W measurements collected by each sensor node as a bitmap of $d \ll W$ bits. This encoding is performed locally at each node. The encoding that we utilize trades accuracy (i.e., probability of correctly determining whether a node is an outlier or not) for bandwidth, by simply varying the desired level of dimensionality reduction and provides tunable accuracy guarantees based on the d parameter mentioned above. Assuming a clustered network organization [13–16], motes communicate their bitmaps to their clusterhead, which can estimate the similarity amongst the latest values of any pair of sensors within its cluster by comparing their bitmaps, and for a variety of similarity metrics that are useful for the applications we consider. Based on the performed similarity tests, and a desired minimum support specified by the posed query, each clusterhead generates a list of *potential* outlier nodes within its cluster. At a second (inter-cluster) phase of the algorithm, this list is then communicated among the clusterheads, in order to allow potential outliers to gain support from measurements of nodes that lie within other clusters. This process is sketched in Fig. 1.

The second level of hashing (omitted in Fig. 1) adopted in TACO's framework comes during the intra-cluster communication phase. It is based on the hamming weight of sensor bitmaps and provides a pruning technique

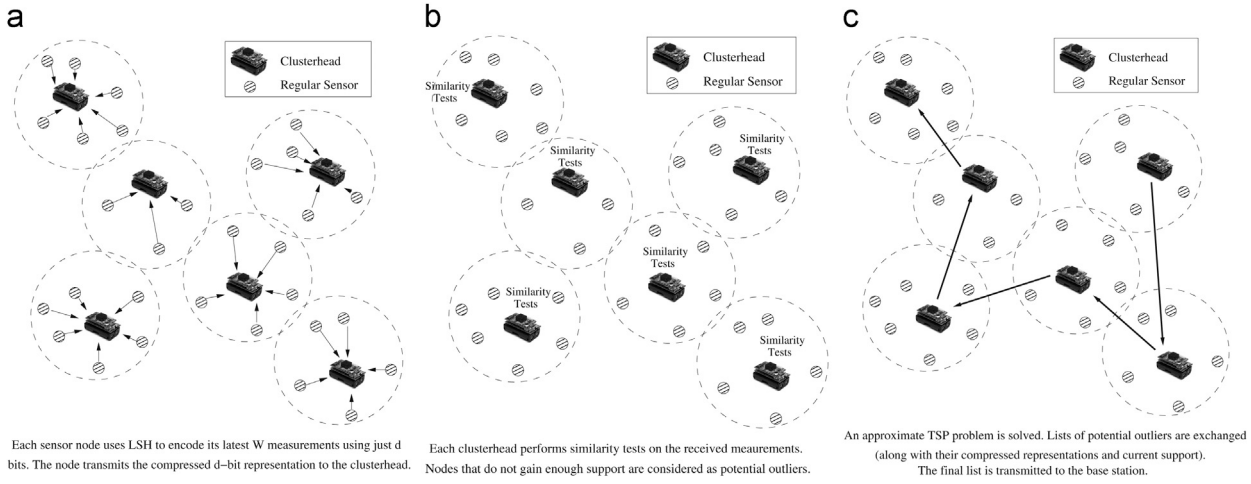

Fig. 1. Main stages of the TACO framework.