Computer Networks 126 (2017) 100–113

Contents lists available at ScienceDirect

Computer Networks

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

On sensor selection in linked information networks[☆]

Charu C. Aggarwal[a], Amotz Bar-Noy[b], Simon Shamoun[c,1,*]

[a] IBM T. J. Watson Research Center,1101 Kitchawan Rd, Yorktown Heights, NY 10598, USA
[b] Computer and Information Science Department, Brooklyn College, 2900 Bedford Avenue, Brooklyn, NY 11210, USA
[c] Faculty of Engineering, Bar Ilan University, Ramat Gan, 5290002, Israel

ARTICLE INFO

Article history:
Received 21 October 2016
Revised 24 February 2017
Accepted 24 May 2017
Available online 25 May 2017

Keywords:
Sensor selection
Stream prediction
Information networks
Greedy algorithm

ABSTRACT

Sensor networks are often redundant by design in order to achieve reliability in information processing. In many cases, the relationships between the different sensors are known *a-priori*, and can be represented as virtual linkages among the different sensors. These virtual linkages correspond to an information network of sensors, which provides useful external input to the problem of sensor selection. In this paper, we propose the unique approach of using external linkage information in order to improve the efficiency of very large scale sensor selection. We design efficient theoretical models, including a greedy approximation algorithm and an integer programming formulation for sensor selection. Our greedy selection algorithm provides an approximation bound of $1 - 1/\sqrt{e}$, where e is the base of the natural logarithm. We show that our approach is much more effective than baseline sampling strategies. We present experimental results that illustrate the effectiveness and efficiency of our approach.

1. Introduction

In recent years, sensor networks have become very popular because of their numerous military and industrial applications. In many cases, a network may contain tens of thousands of sensors producing continuous streams of data. The large amount of data produced and transmitted is a challenge from a variety of cost-driven perspectives. For example, sensors have limited battery supplies, so it is often desirable to minimize the total power consumption by the sensors. Therefore, a lot of recent research has been devoted to efficient data collection and processing in sensor networks [2–4].

While a sensor network may be inherently redundant, it is often the case that not all sensors are used in the data collection process. A user may wish to approximate as many streams as accurately as possible with the use of only a subset of the sensors. In many cases, sensors have predictability relationships with each other that can be used in deciding which sensors to use to approximate the streams of others. For example, a bird call heard by one sensor will be replicated by another sensor nearby, albeit with some lag. Similarly, visual sensors located within a few meters of one another often produce almost the same data. Such relationships can be represented in the form of *logical*, *information-network*, or *virtual links* between sensors, some examples of which are as follows:

- Two sensors of a particular type may be located at a particular distance apart (which is typically small) and facing the same direction. Depending upon the distance between the two sensors, a virtual link can be created between the two sensors that includes information about the type and the distance between the two sensors.
- Two sensors of different types may carry highly correlated information and may be placed at the same location. For example, an acoustic sensor and a vibration sensor may often collect closely correlated information. While the sensors may be of different types with different kinds of output, the correlation can be used in order to make inferences.
- In fluid tracking applications, the sensors may have similar readings (after a small time lag), depending upon the direction of fluid flow or the topology of the pipe network which enables such flow. In some cases, such as wind direction in open spaces, the topology may be defined implicitly at a given time. In some cases, the topology may even dynamically change over time as the direction of the flow changes. Therefore, it is critical to be able to make quick decisions for inference purposes.

When a sensor stream is not available, one or more of its linked sensors can be used to predict it with the use of numerical regression analysis [5–7]. The resulting loss of information is essentially the *error* of not using all the sensors. We note that the best

☆ Preliminary results appeared in DCOSS 2011 [1].

* Corresponding author.
E-mail addresses: xlf07221@gmail.com (C.C. Aggarwal), amotz@sci.brooklyn.cuny.edu (A. Bar-Noy), srshamoun@yahoo.com (S. Shamoun).

1 Simon Shamoun began this work at the CUNY Graduate Center.

http://dx.doi.org/10.1016/j.comnet.2017.05.024
1389-1286/© 2017 Published by Elsevier B.V.

choice of sensors depends on the strength of the relationships between the sensors and the topology of the corresponding relationships. The graph representations of such informative relationships between different data objects are referred to as *information networks* [8]. Information networks have proven to be a useful logical construct to analyze a wide variety of web and social network scenarios, though the approach has not been used in the context of sensor networks to the best of our knowledge. This paper takes a first approach to treating sensors as data objects and the corresponding relationships among them as information network relationships. We will see that such a systematic approach provides us the ability to model generic relationships between sensor nodes in the context of selection. Such a model also allows us to design a general approximation algorithm for the problem.

The goal of this paper is to design a *link- and data-driven* technique to determine which sensors are most critical for overall sensing quality, especially when the real-time budget constraints are tight. Such budget constraints can be resource constraints, such as power consumption, storage capacity, and bandwidth, though the model is general enough to accommodate any kind of cost in the context of a sensor network. Specifically, the goals of this paper are as follows:

- We design a link-based error based sensor information network model.
- We show how to formulate an optimization problem for sensor selection using this model.
- We design an effective approximation algorithm for this optimization problem and show how this solution can be used to make effective predictions about the data while using cost-based budget constraints.

This paper is organized as follows. In Section 2, we review related work. In Section 3, we introduce the overall sensor selection problem in the context of linked information networks. In Section 4, we study the problem of link error modeling, which is required in order to assign values to the variables of the optimization problem. In Section 5, we show that the problem is NP-hard, and design effective integer programming and approximation algorithms. Section 6 presents the experimental results. Finally, Section 7 presents the conclusions.

2. Related work

The sensor selection problem has been addressed independently by the sensor network, signal processing, and stream mining communities.

The sensor network community has focused mostly on predicting sensor values using Gaussian processes. Krause et al. [9] take an information theoretic approach, setting the optimization goal to the maximization of the mutual information between selected and unselected sensors. Since mutual information is submodular, they can guarantee a constant factor approximation by greedy selection. This solution was directly applied to positioning cameras in scene observation [10], while [11] compared various optimization criteria for soil moisture prediction. The later included a way of ordering and clustering the sensors in order to simplify selection. In a sensornet-as-database [12], queries are answered using the posterior probability density given observations by selected sensors. Unlike these methods, the method in [13] does not assume prior knowledge of the probability model. Rather, it uses global optimization by sampling sets of sensors over time, using the earth mover's distance as a metric between sets. Golovin et al. [14] address the same problem for arbitrary submodular utility functions, using multiple bandit algorithms corresponding to the incremental error reduction when adding a sensor. By proving submodular properties of this method, it guarantees a constant approximation bound.

Instead of the prediction of all sensor values, Bian et al. [15] and Byers and Nasser [16] present frameworks for selecting sensors to estimate an arbitrary function of sensor values. Das and Kempe [17] present approximation and exact selection algorithms under specific conditions. They, too, use linear regression for prediction and, in certain instances, a graph model, but based on the covariance between sensors and not on information network linkage. Das and Kempe [18] analyze how to sample sensors to predict the maximum and average sensor value. By assuming sensors are embedded in a metric space, they reduce the problem to either the k-center or k-median problem, and are thus able to guarantee constant factor approximations of the worst-case prediction error.

The signal processing community defined the problem as estimating an unknown vector from sensor readings that are known linear functions of it with additive noise. Joshi and Boyd [19] use convex optimization to minimize the estimation error of the maximum likelihood estimators. Shamaiah et al. [20] map the same problem to maximum *a posteriori* estimation in order to apply greedy selection with guaranteed approximation bounds. Ranieri et al. [21] use greedy selection to minimize the frame potential–which is related to the orthogonality of the model vectors–as an approximation for minimizing the mean squared error of the least squares solution. Similar techniques are used to minimize the mean square error estimator when the noise is correlated [22] and the estimation error for non-linear models after transformation to linear models [23].

Of these solutions, only the soil moisture application uses meta-information about predictability relationships to reduce computation as in ours–specifically, by clustering sensors according to soil features. Marjovi et al. [24] highlighted the significance of logical relationships between sensors by showing that relating city streets by their logical relationships rather than physical proximity was more effective in urban pollution monitoring. [25] is somewhat similar to our method, in that they use linear regression for prediction and a correlation hypergraph for selection. However, the hyperedges are limited by the distance between sensors in the communication graph, and they are not differentiated by prediction quality. Note that all these solutions optimize the maximization of error reduction in order to apply solutions with provable approximation bounds. Only Das and Kempe [18] and Garnett et al. [13] minimize prediction error directly.

The problem of data driven stream selection has been studied extensively by the database community [2,3]. The work in the database community is somewhat distinct from that in the sensor community in that the problem of selection is studied from the perspective of *storage constraints* rather than *stream collection costs*. Nevertheless, the two ways of looking at this problem are quite closely related. The work in [6,7] examines the cross-predictability of streams with the use of online latent variable detection and its use for predictive purposes. Considerable work [26] has been done in using regression modeling for stream prediction and selection. However, none of this work studies the advantages of studying the stream selection problem in the context of a large information network of linkages, as is studied in this paper.

The model presented here was used in several other papers. Preliminary results of this paper were presented in [1]. Aggarwal et al. [27] applied the same model to dynamic sensor selection, updating the link weights according to reduced readings by unselected sensors. Gu et al. [28] designed a hybrid spatio-temporal prediction method for dynamic environments with sparse data, choosing between prediction methods based on the link quality. They also studied various measures of link quality for their prediction methods. Shamoun et al. [29] studied how to select sensors based on various measures of link quality. Bar-Noy et al. [30]