Theoretical Computer Science 607 (2015) 381–390

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Approximate aggregation for tracking quantiles and range countings in wireless sensor networks

Zaobo He [a,1], Zhipeng Cai [a,*], Siyao Cheng [b,2], Xiaoming Wang [c,3]

[a] Department of Computing Science, Georgia State University, United States
[b] School of Computer Science and Technology, Harbin Institute of Technology, China
[c] School of Computer Science, Shaanxi Normal University, China

ARTICLE INFO

Article history:
Received 27 March 2015
Received in revised form 6 July 2015
Accepted 28 July 2015
Available online 31 July 2015

Keywords:
Quantiles
Range countings
Approximate aggregation

ABSTRACT

We consider the problem of tracking quantiles and range countings in wireless sensor networks. The quantiles and range countings are two important aggregations to characterize a data distribution. Let $S(t) = (d_1, \ldots, d_n)$ denote the multi-set of sensory data that have arrived until time t, which is a sequence of data orderly collected by nodes s_1, $s_2, \ldots, s_k$. One of our goals is to continuously track ϵ-approximate ϕ-quantiles ($0 \leq \phi \leq 1$) of $S(t)$ for all ϕ's with efficient total communication cost and balanced individual communication cost. The other goal is to track (ϵ, δ)-approximate range countings satisfying the requirement of arbitrary precision specified by different users. In this paper, a deterministic tracking algorithm based on a dynamic binary tree is proposed to track ϵ-approximate ϕ-quantiles, whose total communication cost is $O(k/\epsilon \cdot \log n \cdot \log^2(1/\epsilon))$, where k is the number of the nodes in a network, n is the total number of the data, and ϵ is the user-specified approximation error. For range countings, a Bernoulli sampling based algorithm is proposed to track (ϵ, δ)-approximate range countings, whose total communication cost is $O(\frac{2}{\epsilon^2} \ln \frac{2}{1-\sqrt{1-\delta}} + n_c)$, where δ is the user-specified error probability, n_c is the number of clusters.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Wireless Sensor Networks (WSNs) consist of many nodes which interact with each other through wireless channels. They are now being widely employed to collect physical information, such as temperature, pressure, light intensity and so forth [13]. With the development of wireless communication technologies, the scale of a WSN can be very large [11]. However, the most severe constraint imposed on their extensive applications is the power supply as the on-board power is still the main power source which is limited in most cases. Compared with computation, communications among nodes consume more energy. According to [16], the energy consumption for sending one bit is enough for executing 1000 instructions for one sensor. Thus, how to extract significant information from a huge amount of sensory data with a reasonable communication cost becomes a crucial problem.

* Corresponding author. Tel.: +1 (404) 413 5721.
E-mail address: zcai@gsu.edu (Z. Cai).

[1] Supported by NSF under grant No. CNS-1252292.
[2] Supported by the NSFC of China under grant No. 61370217.
[3] Supported by the NSFC of China under grants Nos. 61173094 and 61373083.

http://dx.doi.org/10.1016/j.tcs.2015.07.056
0304-3975/© 2015 Elsevier B.V. All rights reserved.

Much effort has been spent on studying various aggregations, including algebraic aggregations such as Sum, Count, and Average [4,9,10], or holistic aggregations such as Heavy hitters [12] and Quantiles [5,7,8,15]. In-network aggregation techniques are very meaningful for algebraic aggregations through computing partial results at intermediate nodes during the process of transmitting data to the sink [4,7,8,10,15]. By preventing nodes from forwarding all the data to the sink, in-network aggregation techniques significantly reduce energy consumption. They are meaningful for algebraic aggregations since these aggregations are decomposable [2], however, holistical aggregations are not decomposable so that the in-network aggregation techniques cannot be employed to track quantiles directly [5]. The ϕ-quantile ($0 \leq \phi \leq 1$) of an ordered dataset S is the data x such that $\phi|S|$ elements of S are less than or equal to x and no more than $(1-\phi)|S|$ elements are larger than x, particularly, the $\frac{1}{2}$-quantile is the median of S. Another fundamental aggregation is range counting. For any query interval Q, the Q-range counting over an ordered dataset S is the value of $|Q \cap S|$.

Since exact aggregations incur large communication costs and storage spaces, approximate results are usually expected. The work in [17] shows that a random sample of size $\Theta(1/\epsilon^2)$ needs to be drawn from a dataset to compute ϵ-approximate quantiles with a constant probability. Moreover, practical applications involve various precision requirements so that aggregation results that can satisfy an arbitrary precision is expected. ϵ-approximate ϕ-quantiles and (ϵ, δ)-approximate Q-range countings are defined respectively as follows:

Definition 1 *(ϵ-Approximate ϕ-quantiles).* The ϵ-approximate ϕ-quantiles are those elements in dataset S such as element x satisfying $(\phi-\epsilon)n \leq r(x) \leq (\phi+\epsilon)n$ where $r(x)$ is the rank of x in S and n is the total number of the elements in S.

Definition 2 *((ϵ, δ)-Approximate Q-range countings).* Let Υ and Υ' denote the exact and approximate Q-range countings respectively, *i.e.*, $\Upsilon = |Q \cap S|$ and $\Upsilon' = |Q \cap U|$ where U is a random sample set of S. An (ϵ, δ)-approximate Q-range counting is such Υ' that satisfies $\Pr\{\Upsilon - \epsilon|S| \leq \Upsilon' \leq \Upsilon + \epsilon|S|\} \geq 1-\delta$, where $\Pr\{X\}$ is the probability of random event X.

Various data models have been studied in the database communities, such as the static model, single-stream model and multi-stream model. For the static model, data is predetermined and stored at nodes and aggregation function f is computed over the union of these multiple datasets. For the single-stream model, there is only one node, and data stream into it in an online fashion. Nowadays, the multi-distributed streaming model attracts a lot of attention since it is more general in the physical environment. In this model, data stream into each node in a distributed way and the tracking results are returned in a logical coordinator. Moreover, for the quantile tracking results, the querying mode can be divided into two classes: single ϕ-quantile and all ϕ-quantile. For the single ϕ-quantile tracking, a certain summary is always maintained by a coordinator to compute a certain ϕ-quantile. Comparatively, the data structure or summary preserved by a coordinator for all ϕ-quantiles can be used to compute any ϕ simultaneously.

We want to track quantiles in a general manner, where the sensor nodes are organized into a spanning tree and sensory data stream into each node in an online fashion. $S(t) = (d_1, \ldots, d_n)$ is the multi-set of items of the entire network that have arrived until time t, which is a sequence of data orderly collected by nodes $s_1, s_2, \ldots, s_k$. The goal is to continuously track ϵ-approximate ϕ-quantiles ($0 \leq \phi \leq 1$) of $S(t)$ at the sink for all ϕ's. Moreover, we design a Bernoulli sampling based algorithm for tracking range countings to satisfy arbitrary precision requirements specified by different users.

The main contributions of this work can be summarized as follows. First, quantiles can be tracked over the arrived data at any time t rather than through a one-time computation over a predetermined dataset. Second, quantiles are computed based on an arbitrary topological spanning tree rather than the centralized flat model. Third, a data structure can be maintained in the tree from which all the ϕ-quantiles can be tracked simultaneously rather than for just a specific ϕ. Fourth, the proposed range countings tracking algorithm can satisfy arbitrary precision requirements.

2. Related works

The previous quantile tracking techniques can be divided into three categories, which are the exact algorithms, deterministic algorithms and probabilistic algorithms. For a given ϕ, the exact algorithms return the exact ϕ-quantile result. According to [14], the space complexity for computing the exact median with p passes is $\Omega(n^{1/p})$. Clearly, the space complexity of the exact algorithms is high, especially when the number of the passes p is small.

To further reduce time and space complexities during tracking quantiles, some deterministic algorithms are proposed, such as the recent works in [6,8,15,18]. Unlike the exact algorithms, the deterministic ones return ϵ-approximate ϕ-quantiles of a dataset. Since the deterministic algorithms just require approximate results, they have smaller space and communication complexities. In 2005, Cormode *et al.* [5] proposed an all-tracking algorithm with the cost of $O\left(\frac{k}{\epsilon^2}\log n\right)$. The work in [18] improves this result by a $\Theta(\frac{1}{\epsilon})$ factor, whose result has an upper bound $O\left(\frac{k}{\epsilon}\log n\right)$. Note that the work in [18] discusses the all ϕ-quantiles tracking problem under the flat model. However, it is unclear how to track quantiles in the tree model.

Considering that the approximate quantile with a probability guarantee can be accepted in most cases, the complexity of tracking the quantile can be further reduced. Thus, a group of probabilistic algorithms [4,8,10] were proposed. Different from the above two types of algorithms, the probabilistic algorithms require that the ϵ-approximate ϕ-quantile result is

دریافت فوری ← متن کامل مقاله

ISIArticles
مرجع مقالات تخصصی ایران

✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
✓ امکان دانلود نسخه ترجمه شده مقالات
✓ پذیرش سفارش ترجمه تخصصی
✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
✓ دانلود فوری مقاله پس از پرداخت آنلاین
✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات