A Proficient Processing of Probabilistic Top- k Queries in Distributed Wireless Sensor Networks

Shaik Salma Sultana[1], Shaik Khamar Zahan[2]

[1]M.Tech (CSE), Nimra College of Engineering and Technology, A.P., India.

[2]Asst. Professor, Dept. of Computer Science & Engineering, Nimra College of Engineering and Technology, A.P., India.

Abstract— Wireless sensor networks are revolutionizing the ways to collect and use information from the physical world. Wireless Sensor Networks (WSNs) are usually defined as large-scale, ad-hoc, multi-hop and wireless un-partitioned networks of homogeneous, small, static nodes deployed in an area of interest. Applications of sensor networks include monitoring volcano activity, building structures or natural habitat monitoring. In this paper, we present the problem of processing probabilistic top-k queries in a distributed wireless sensor networks. The basic problem in top-*k* query processing is that, a single method cannot be used as a solution to the problem of top-*k* query processing because there are many types of top-*k* query processing. The method has to be based on the situation, the classification and the type of database and the query model. Here we develop three algorithms, namely, sufficient set-based (SSB), necessary set-based (NSB), and boundary-based (BB), for inter- cluster query processing with bounded rounds of communications. Moreover, in responding to dynamic changes of data distribution in the overall network, we develop an adaptive algorithm that dynamically switches among the three proposed algorithms to minimize the transmission cost.

Keywords— BB, NSB, SSB, Top- k query, WSN.

I. Introduction

Wireless Sensor Networks (WSNs) are increasingly used in data-intensive applications such as microclimate monitoring, precision agriculture, and audio/video surveillance. A wireless Sensor Network(WSN) consists of number of nodes that is used in different applications such as military, health care, commerce, etc. Usually a sensor node is used for sensing precision to monitor environmental conditions. This will be varies in sensing precision. Every sensor node will be varying in the sensing quality. So, whatever the values i.e. raw sensor readings that are collected from sensor is of data uncertainty and energy consumption. In order to remove the data uncertainty many approaches has been used, but that gives inefficient results. A data uncertainty is removed by placing more sensor nodes and as well as by calculating the probability i.e. aggregate probability.

In many application domains, top-k query is a fundamental query to search for the most important objects according to the object ranking. Being different from those studies of top- k query in the centralized databases, in this paper we focus on the top-k query optimization in resource-constrained wireless sensor networks (WSNs). Technological advances have enabled the deployment of the large-scale sensor networks consisting of thousands of inexpensive sensor nodes in an ad-hoc fashion for a variety of environmental monitoring and surveillance purposes. During this course, a large volume of sensed data are needed to be aggregated within the sensor network to respond to user queries. The WSN thus is treated as a virtual database by the database community [1]. However, query processing in sensor networks is essentially different from it in traditional databases due to the unique characteristics imposed on sensors, e.g., slow processing capability, limited storage, and energy-limited batteries, etc. [2], which can be seen from several aspects. Firstly, to prolong network lifetime, the energy consumption is an optimization objective in sensor networks, because the battery-powered sensor nodes will quickly become inoperative due to the large quantity of energy consumption, and the network lifetime is closely tied to the energy consumption rate of the sensors. Secondly, a WSN that senses the data periodically can be viewed as a distributed stream system [3]. However, this special distributed stream system is different from the general distributed stream system because it is more expensive to obtain the sensed information from the sensors far away from the base station than those nearby. Finally, for query processing in sensor networks, minimizing not only the total energy consumption but also the maximum energy consumption among the sensors is the optimization objective. Hence, how to evaluate queries effectively and efficiently in sensor networks poses great challenges.

II. Related work

In recent years, many works have been done. Here we review representative work in the areas of 1) top-k Query processing in WSNs, and 2) top-k query processing on the uncertain data. An extensive number of research works in this area has appeared in the literature [4], [5], [6]. Due to the limited energy budget available at sensors, the primary issue is how to develop energy-efficient methods to reduce communication and energy costs in the networks. TAG [4] is one of the first studies in this research area. By exploring the semantics of aggregate operators (e.g., sum, avg, and top-k), in-network processing approach is adopted to suppress the redundant data transmissions in wireless sensor networks. Moreover, continuous top-k queries for sensor networks have been studied in [7] and [8]. In addition, a distributed threshold join algorithm has been developed for the top- k queries [5]. These studies, considering no uncertain data, have a different focus from our present study.

For uncertain databases, two interesting top-k definitions (i.e., U-Topk and U-kRanks) and like methods are proposed [9]. U-Topk returns a list of k-tuples that has the highest probability to be in the top-k list over all possible worlds. U-k- Ranks returns a list of k tuples such that the ith record has the highest probability to be the ith best record in all possible worlds. In [10], PT-Topk query, which returns the set of the tuples with a probability of at least p to be in the top-k lists in the possible worlds, is studied. Inspired by the concept of dominate set in the top-k query, a method which avoids unfolding all possible worlds is given. Besides, a sampling method is developed to quickly compute an approximation with quality guarantee to the answer set by drawing a small sample of the uncertain data. In [11], the expected rank of each tuple across all possible worlds serves as the ranking function for finding the final result. In [12], U-Topk and U-kRank queries are improved by exploiting their stop conditions. In [13], all existing top-k semantics have been unified by using some generating functions. Recently, a study on processing top-k queries over a distributed uncertain database is reported in [14].

III. Proposed work

A. Sufficient and Necessary sets

In this sectionWe introduce the notion of sufficient set and necessary set for distributed processing of probabilistic top-k queries in cluster-based wireless sensor networks. These two concepts have very nice properties and can facilitate localized data pruning in clusters.

Given an uncertain data set Ti in the cluster Ci, if there exists a tuple tsb€Ti (called sufficient boundary) such that the tuples ranked lower than tsb are useless for the query processing at the base station, then the sufficient set of Ti, denoted as S(T), is a subset of Ti as specified below:

S(Ti)={t|t=ftsbort<ftsb}

where f is a given scoring function for ranking. Note that a sufficient boundary may not exist for a given data set

Given a local data set Ti in the cluster Ci, assume that Ai is the set of locally known candidate tuples for the final answer and tnb (called necessary boundary) is the lowest ranked tuple in Ai. The necessary set of Ti, denoted as N(Ti), is

N(Ti)= {t|t€ Ti, t <f <tnb}

Using the notion of sufficient and necessary sets as a basis, we propose 3 distributed algorithms for processing probabilistic top-k queries in wireless sensor networks, namely 1) Sufficient Set-based method; 2) Necessary Set-based method; and 3) Boundary-based method.

B. Sufficient Set-Based (SSB) Algorithm

After collecting data tuples from its cluster, ci computes the S(Ti) from the locally collected tuples and sends it to the base station. If a sufficient set cannot be obtained, then all the tuples are transmitted to the base station. After receiving the transmitted data tuples from all the cluster heads, they compute final answer.

Algorithm 1: SSB ALGORITHM

```
AT CLUSTER HEAD (ci):
    1.  if SB(Ti) exits
          $S(Ti) \leftarrow \{x|x \leq f\ SB(Ti)\ \Lambda\ x \in Ti\}$
          $Yi \leftarrow S(Ti)$
        Else
              $Yi \leftarrow Ti$
    2.  Now, Yi is delivered to the base station.
```

```
AT BASESTATION:
  1. It receive the tuples Yi from the cluster head.($1 \leq i \leq N$)

  2.$T' \leftarrow U1 \leq i \leq N\ Yi$
```

Where, x is the tuples ci is the cluster head S(Ti) is the sufficient set Ti is the records collected from the sensor N is the number of clusters in the zone Ci is the cluster Yi is the sufficient boundary for SSB. T′ is the aggregation of data sets received from the clusters

C. Necessary Set-Based (NSB) Algorithm

After receiving all the necessary sets, the received tuples are merged into a table in a base station and finds the necessary boundary called the global boundary (GB)). If GB is ranked higher than the highest ranked necessary boundary, all the necessary data have delivered to the base station. Otherwise, it entering the second phase, it sends the GB back to the ci, which return the supplementary data tuples ranked between its local necessary boundary and GB. Then, the base station computes the final answer.

Algorithm 2: NSB ALGORITHM

```
 AT CLUSTER HEAD:
1.Compute the necessary boundary NB(Ti),
  $N(Ti) \leftarrow \{x|x \leq f\ NB(Ti)\ \Lambda\ x \in Ti\}$

2. Deliver N(Ti) to the base station

3.if cluster head receive GB from the base station then
     $N'(Ti) \leftarrow \{x|x \leq f\ GB\ \Lambda\ x \in [Ti - N(Ti)]\}$ Now,
     N′(Ti) is send to the base station.
     end if

AT BASESTATION:

1. It receives the tuples N (Ti) from the cluster head.
  ($1 \leq i \leq N$) $T' \leftarrow U1 \leq i \leq N\ N(Ti)$
2. Now, it will calculate the global boundary.
3. if global boundary GB is less than that of NB(Ti), then
    It calculate the final necessary boundary
else
    It will broadcast GB to ci and once again it collects necessary tuples
        $T' \leftarrow U1 \leq i \leq N\ N'(Ti)$
end if
```

Where, x is the tuples ci is the cluster head N(Ti) is the necessary set NB(Ti) is the necessary boundary Ti is the records collected from the sensor N is the number of clusters in the zone T′ is the aggregation of data sets received from the clusters

D. Boundary-Based(BB) Algorithm

The boundary-based method first delivers the local knowledge in clusters, in the form of NB and SB, to the base station in order to provide a refined global data pruning among clusters. It is done instead of directly delivering data tuples to the base station.

Algorithm 3: BB Algorithm

```
 AT CLUSTER HEAD:
1. Calculate the Necessary Boundary (NB) and Sufficient Boundary (SB) and send it to the base station.
2. Base station receive Global Boundary (GB)
3. $Yi \leftarrow \{x|x \leq f\ GB\ x \in [Ti - N(Ti)]\}$
4. Now, Yi is delivered to the base station.

AT BASESTATION:
1. It will receive the NB and SB from cluster heads (ci),
 2. Now, base station computes the (Sufficient Boundaryhigh and Necessary Boundarylow ).
3. if SBhigh < NBlow , then SBhigh → GB
else
   NBlow → GB
end if
4.Now, broadcast the global boundary to each
        Ci $T' \leftarrow U1 \leq i \leq N\ Y(Ti)$
```

Where, x is the tuple ci is the cluster head S(Ti) is the sufficient set N(Ti) is the necessary set Ti is the records collected from the sensor N is the number of clusters in the zone Yi is the sufficient boundary for SSB T′ is the aggregation of data sets received from the clusters

E. Cost Analysis

We perform a cost analysis on data transmission of the three proposed methods by using adaptive algorithm. Adaptive Algorithm: The performance of the data transmission using proposed method is affected by factors such as the skewness of data distribution among clusters which may change continuously over time. A cost-based adaptive algorithm that is used dynamically Sufficient Set Based, Necessary Set Based, and Boundary Based as the data distribution within the network changes.

```
Algorithm 4: Adaptive Algorithm
Count=0 ;
 ZSSB , ZNSB , ZBB =0 Where R is varied window size.
Then estimate the cost of CSSB, CNSB, CBB
ZSSB ← ZSSB + CSSB
ZNSB ← ZNSB + CNSB
 ZBB ← ZBB + CBB
if count ≥ R then
    if ZSSB = min{ ZSSB , ZNSB , ZBB} then
        switch to SSB
    end if
    if ZNSB = min{ ZSSB , ZNSB , ZBB} then
        switch to NSB
    end if
    if ZBB = min{ ZSSB , ZNSB , ZBB} then
        switch to SSB
    end if
end if
```

IV. CONCLUSION

Motivated by many applications, top-*k* query is a fundamental operation in the modern database systems. Technological advances have enabled the deployment of several large-scale sensor networks for environmental monitoring and surveillance purposes, efficient processing of top-*k* query in such networks poses great challenges due to the unique characteristics of sensor nodes and a vast amount of data generated by sensor networks. This work supports in-network top-k query process over uncertain data in the distributed wireless sensor network. We develop the notion of the sufficient set and necessary set for efficient in-network pruning of uncertain data in a distributed setting. This notion, along with its nice properties, provides a theoretical basis for the distributed query processing methods. Based on the notion of sufficient sets and necessary sets, we propose a suite of algorithms for in-network processing of PT-Topk queries in a two-tier hierarchical sensor network. These methods exploit individual and combined strengths of sufficient and necessary sets in query processing. We propose a cost-based adaptive algorithm that dynamically switches among the three proposed algorithms based on their estimated costs.

REFERENCES

[1] S. Madden, M. J. Franklin, J. M. Hellerstein, W. Hong. TAG: a tiny aggregation service for ad hoc sensor networks. ACM SIGOPS Operating Systems Review, Vol.36, pp.131–146, 2002.

[2] G. J. Pottie, W. J. Kaiser Wireless Integrated Network Sensors. Communication of ACM, Vol.43 No.5, pp.51–58, 2000.

[3] B. Babcock and C. Olston. Distributed top-k monitoring. Proc. of ACM SIGMOD, ACM, pp.28–39, 2003.

[4] P. Cao and Z. Wang, "Efficient Top-k Query Calculation in Distributed Networks," Proc. 23rd Ann. ACM Symp.Principles of Distributed Computing (PODC), pp. 206-215, 2004.

[5] M. Ye, X. Liu, W.-C. Lee, and D.L. Lee, "Probabilistic Top-k Query Processing in Distributed Sensor Networks," Proc. IEEE Int'l Conf. Data Eng. (ICDE '10), 2010.

[6] D. Zeinalipour-Yazti, Z. Vagena, D. Gunopulos, V. Kalogeraki, V.Tsotras, M. Vlachos, N. Koudas, and D. Srivastava, "The Threshold Join Algorithm for Top-k Queries in Distributed Sensor Networks," Proc. Second Int'l Workshop Data Management for SensorNetworks (DMSN '05), pp. 61-66, 2005

[7] Q. Han, S. Mehrotra, and N. Venkatasubramanian, "Energy Efficient Data Collection in Distributed Sensor Environments," Proc. 24th Int'lConf.Distributed Computing Systems (ICDCS'04), pp. 590-597, 2004.

[8] M.Wu, J.Xu, X. Tang, and W.-C. Lee, "Top-k Monitoring in Wireless Sensor Networks," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 7, pp. 962-976, July 2007.

[9] M.A. Soliman, I.F. Ilyas, and K.C. Chang, "Top-k Query Processing in Uncertain Databases," Proc. Int'l Conf.

[10] M. Hua, J. Pei, W. Zhang, and X. Lin, "Ranking Queries on Uncertain Data: A Probabilistic Threshold Approach," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '08), 2008.

[11] C. Jin, K. Yi, L. Chen, J.X. Yu, and X. Lin, "Sliding-Window Top-k Queries on Uncertain Streams," Proc. Int'l Conf. Very Large Data Bases (VLDB '08), 2008.

[12] D. Wang, J. Xu, J. Liu, and F. Wang, "Mobile Filtering for Error-Bounded Data Collection in Sensor Networks," Proc. 28th Int'l Conf. Distributed Computing Systems (ICDCS '08), pp. 530-537, 2008.

[13]K. Yi, F. Li, G. Kollios, and D. Srivastava, "Efficient Processing of Top-k Queries in Uncertain Databases with X-Relations," IEEE pp. 1669 1682, Dec.2008.

[14] F. Li, K. Yi, and J. Jestes, "Ranking Distributed Probabilistic Data," Proc. 35th SIGMOD Int'l Conf. Management of Data (SIGMOD '09), 2009.