IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. VIII (Feb. 2014), PP 48-53
www.iosrjournals.org

Distributed Processing of Probabilistic Top-kQueries with Efficient Query Evaluation

B.Prakash[1], T.S.Murunya[2]
[1]*(Department of CSE, Prist University, India)*
[2]*(Head of the CSE Department, Prist University, India)*

Abstract: *Uncertain data arises in a number of domains, including data integration and sensor networks. Top-k queries that rank results according to some user-defined score are an important tool for exploring large uncertain data sets. So, we introduce the efficient query evaluation of the sufficient set-based (SSB), necessary set-based (NSB), and boundary-based (BB) algorithm for distributed processing in Top-K queries in wireless sensor networks, for inter cluster query processing with bounded rounds of communications and in responding to dynamic changes of data distribution in the network, we develop an adaptive algorithm that dynamically switches among the three proposed algorithms to minimize the transmission cost. The generic method to evaluate the reliability of a data automatically retrieved from the web. Finally results given that the proposed algorithms reduce data transmissions significantly and incur only small constant rounds of data communications for reliability. The experimental results also demonstrate the superiority of the adaptive algorithm, which achieves a near-optimal performance under various conditions.*
Index Terms: *Distributed data management, network topologies, probabilistic databases, top K- Queries, Wireless sensor networks,*

I. Introduction

In Wireless sensor networks are revolutionizing the ways to collect and use information from the physical world. This new technology has resulted in significant impacts on a wide array of applications in various fields, including military, science, industry, commerce, transportation, and health-care. However, the quality of sensors varies significantly in terms of their sensing precision, accuracy, tolerance to hardware/external noise, and so on. For example, studies show that the distribution of noise varies widely in different photo voltic sensors, precision and accuracy of readings usually vary significantly in humidity sensors, and the errors in GPS devices can be up to several meters. Thus, sensor readings are inherently uncertain. On the contrary, our proposal is a general approach which is applicable to probabilistic top-k queries with any semantic. Furthermore, instead of repeatedly requesting data which may last for several rounds, our protocols are guaranteed to be completed within no more than two rounds. These differences uniquely differentiate our effort from. Our previous work as the initial attempt only includes the concept of sufficient set. In this paper, besides of sufficient set, we propose another important concept of necessary set. With the aid of these two concepts, we further develop a suite of algorithms, which show much better performance than the one. Probabilistic ranked queries based on uncertainty at the attribute level are studied. Finally, uncertain top-k query is studied under the setting of streaming databases where a compact data set is exploited to support efficient slide window top-k queries. Armed with sufficient set and necessary set, we develop a suite of algorithms for processing probabilistic top-k queries in two-tier hierarchical wireless sensor networks with PT- Topk as a case study, including 1) sufficient set-based (SSB) algorithm, 2) necessary set-based (NSB) algorithm, and 3) boundary-based (BB) algorithm. Moreover, we developed an adaptive algorithm that dynamically switches among the three proposed algorithms to minimize the communication and energy overhead, in responding to changing data distribution in the network. Furthermore, we discuss how to apply sufficient set and necessary set to devise a series of algorithms, namely SSB-T, NSB-T, and optimized NSB-T (NSB-T-Opt), for processing probabilistic top-k in a sensor network with tree topology. Finally, we evaluate the proposed algorithms both in two-tier hierarchy (i.e., SSB, NSB,BB) and tree topology(i.e.,SSB-T,NSB-T,NSB-T-Opt) in comparison with two baseline approaches.

The Existing paper is the full version of our preliminary work published as a short paper in Probabilistic top-k query processing in Distributed Sensor Network, where we introduce the targeted problem and propose the idea of employing sufficient set to develop the SSB algorithm for distributed processing of probabilistic top-k queries.

In this paper, we extend the idea of sufficient set with the notion of necessary set and sufficient/necessary boundaries; develop new distributed processing algorithms; and demonstrate the extend of our ideas to a sensor network with tree topology. The Experimental result validates our ideas and shows that the

proposed algorithms reduce data transmissions significantly without incurring excessive rounds of communications.

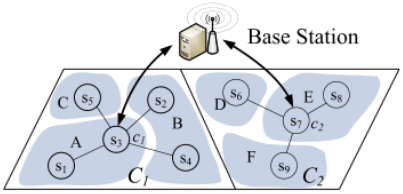


Fig- 1 : Wireless sensor network. These are 6 zones, denoted as A,B,..F which are organized into two clusters C1 & C2 with the corresponding cluster heads.

II. Distributed Data Management and Problem Definition

2.1 Distributed Database

In Distributed Data Management, The DDBMS synchronizes all the data periodically and, in cases where multiple users must access the same data, ensures that updates and deletes performed on the data at one location will be automatically reflected in the data stored elsewhere. A centralized distributed database management system (DDBMS) manages the database as if it were all stored on the same computer. In this paper, we assume that a wireless sensor network consisting of a base station and N sensor nodes is deployed in a monitoring field. The base station serves as the data collection and processing center to the external users and applications. We assume M clusters are formed and a cluster head is selected for each cluster. The whole sensor network can be logically treated as a distributed uncertain database. Sensor readings are collected by cluster heads and transformed into uncertain data, then set of uncertain tuples are distributed in cluster heads within the sensor network.

Definition 1 (Top-k Probability). Let N_a denote the top-k answer set in a possible world A. Given a tuple t € T, the top-k probability of t is the aggregate probability of t being in the top-k answers over all A €A , i.e.,

$$\text{Ptopk(t)} = \sum_{A \in a, t \in Na} P(A)$$

Definition 2 (Probabilistic Threshold Top-k Query (PT- Topk)). Given a probability threshold p(0<p≤1), PT- Topk finds the set of tuples whose top-k probabilities are at least p.

2.2 Centralized PT-Topk Query Processing

In this section, we present a general approach for processing PT-Topk queries in a centralized uncertain database, which provides a good background for the targeted distributed processing problem.

Given an uncertain table T, we first sort T in accordance with the ranking function f such that $t_1 <_f t_2 <_f \ldots <_f t_n$. The query answer can be obtained by examining the tuples in descending ranking order from the sorted table (which is still denoted as T for simplicity). We can easily determine that the highest ranked k tuples are definitely in the answer set as long as their confidences are greater than p since their qualifications as PT-Topk answers are not dependent on the existence of any other tuples. Nevertheless, the qualification of tuple tkþ1 as a PT-Topk answer is dependent on

1) Whether there exist some possible worlds where the tuples in front of t_{k+1} belong to less than k x-tuples; and
2) Whether the aggregated confidence of t_{k+1} over these possible worlds are greater than p. If the answer is positive, then tuple t_{k+1} is included in the answer set and tuple t_{K+2} is examined. This process continues until there are no more qualified tuples left to be examined.

III. Intracluster Data Pruning

In a cluster-based wireless sensor network, the cluster heads are responsible for generating uncertain data tuples from the collected raw sensor readings within their clusters.

To answer a query, it's natural for the cluster heads to prune redundant uncertain data tuples before delivery to the base station in order to reduce communication and energy cost. The key issue here is how to derive a compact set of tuples essential for the base station to answer the probabilistic top-k queries. This is a very challenging issue for the following reasons: 1) the interplay of probability and ranking due to the semantic of probabilistic top-k queries; and 2) the lack of global knowledge to determine the probability and ranking of candidate tuples locally at cluster heads. In this section, we propose the notion of sufficient set and necessary set, and describe how to identify them from local data sets at cluster heads. Next, we use the PT-Topk query as a test case to derive sufficient set and necessary set and show that the top-k probability of a tuple t obtained locally is an upper bound of its true top-k probability. Thus, data tuples excluded from the sufficient sets and necessary sets in local clusters will never appear in the final answer set.

3.1 Definition of Sufficient and Necessary Sets

It would be beneficial if cluster heads are able to find the minimum sets of their local data tuples that are sufficient for the base station to answer a given query. Ideally, sufficient set is a subset of the local data set. Data excluded from the sufficient set, no matter which clusters they reside, will never be included in the final answer set nor involved in the computation of the final answer set. Here, we define the sufficient set more formally as follows:

Definition 3 (Sufficient Set). Given an uncertain data set T_i in cluster Ci, if there exists a tuple t_{sb}€Ti (called sufficient boundary) such that the tuples ranked lower than t_{sb} are useless for query processing at the base station, then the sufficient set of Ti, denoted as S(T), is a subset of T_i as specified below:

$$S(T_i)=\{t|t=_f t_{sb} \text{ or } t<_f t_{sb}\}$$

where f is a given scoring function for ranking. Note that a sufficient boundary may not exist for a given data set; then we consider the whole data set as a sufficient set and will discuss it in more detail later.

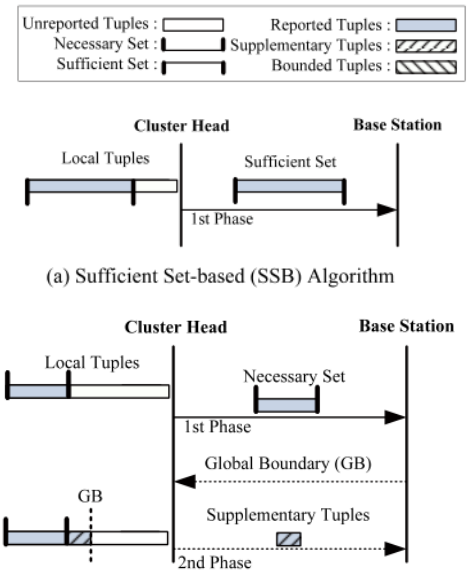
Definition 4 (Necessary Set). Given a local data set T_i in cluster C_i, assume that A_i is the set of locally known candidate tuples for the final answer and t_{nb} (called necessary boundary) is the lowest ranked tuple in Ai. The necessary set of Ti, denoted as $N(T_i)$, is

$$N(T_i)= \{t|t€ T_{i,} t <_f <t_{nb}\}$$

Note that, at the base station, for a given cluster, the probability computation of some candidate tuples from other clusters may still require data tuples outside its necessary set. In other words, while the data tuples in the necessary sets include the final answer set, they may not be sufficient to determine the final answer set. In the following theorem, we depict a relationship between the sufficient set and necessary set.

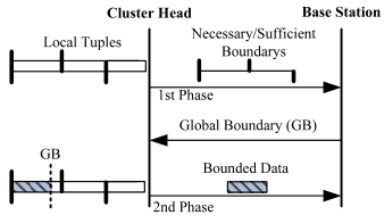
IV. InterCluster Query Processing

Response time is another important metrics to evaluate query processing algorithms in wireless sensor networks. All of those three algorithms, i.e., SSB, NSB, and BB, perform at most two rounds of message exchange there is not much difference among SSB, NSB, and BB in terms of query response time, thus we focus on the data transmission cost in the evaluation. Finally, we also conduct experiments to evaluate algorithms, SSB-T, NSB-T, and NSB-T-Opt under the tree-structured network topology.



(a) Sufficient Set-based (SSB) Algorithm

(b) Necessary Set-based (NSB) Algorithm



(c) Boundary-based (BB) Algorithm

Fig–2:Algorthim for intercluster query processing.

V. Adapative Query Processing :

In this section, we first perform a cost analysis on data transmission of the three proposed methods. Since their performance is affected by factors such as the skewness of data distribution among clusters which may change continuously over time, we propose a cost-based adaptive algorithm that keeps track of the estimated cost for all methods in order to switch as appropriate.

5.1 Cost Analysis

We develop a cost model on communication cost of the three proposed algorithms. Accordingly, we propose a cost-based adaptive algorithm that dyna-mically switches among the three algorithms based on their estimated costs. Let M denote the no. of clusters in the network, and C_q, C_b and C_d be the sizes of query messages, boundary messages, and data messages, respectively. Also let $|C(T_i)|$ and $|N(T_i)|$ denote the cardinalities of the sufficient set and necessary set of the data set T_i in a cluster $C_i(1 \leq i \leq M)$, respectively.

VI. Performance Evaluation

In this section, we first conduct a simulation-based performance evaluation on the distributed algorithms for processing PT-topk queries in two-tier hierarchical cluster- based wireless sensor monitoring system. As discussed, limited energy budget is a critical issue for wireless sensor network and radio transmission is the most dominate source of energy consumption. Thus, we measure the total amount of data transmission as the performance metrics. Notice that, response time is another important metrics to evaluate query processing algorithms in wireless sensor networks. All of those three algorithms, i.e., SSB, NSB, and BB, perform at most two rounds of message exchange, thus clearly outperform an iterative approach which usually needs hundreds of iterations. Note that, there is not much difference among SSB, NSB, and BB in terms of query response time, thus we focus on the data transmission cost in the evaluation. Finally, we also conduct experiments to evaluate algorithms, SSB-T, NSB-T, and NSB-T-Opt under the tree-structured network topology.

6.1 Simulation Model

Here we describe our simulation model. We assume a wireless sensor field consisting of I zones. Each zone is deployed with an average of λ sensor nodes. Here, I can also be seen as the number of x-tuples in the global database and λ is the average size of an x-tuple. The confidence values for sensor readings in an area are assigned randomly. The clusters in the network is realized by a simple grid partition. There are M clusters, which is varied in our experiments. The simulator models sensor mote behavior at a coarse level, similar to the TAG simulator in which time is divided into units of rounds. At the beginning of each round, users may issue PT-Topk queries at the base station and the query messages are passed to cluster heads and sensor nodes without delay since transmission latency is not our main concern in this evaluation.

6.2 Experimental Model

A series of experiments is conducted to evaluate the proposed algorithms for a two-tier network in the following aspects: 1) overall performance, 2) sensitivity tests, and 3) adaptiveness. Additionally, overall performance under the tree topology is evaluated.

6.2.1 Overall Performance

We first validate the effectiveness of our proposed methods in reducing the transmission cost against two baseline approaches, including 1) a naive approach, which simply transmits the entire data set to the base station for query processing; 2) an iterative approach devised based on the processing strategy explored. The iterative approach runs as follows: in each round, each cluster head delivers a single data tuple with the highest score in its local data set and the information of current local highest score (after removing the delivered data) to the base station. In response, the base station derives necessary set upon the data collected so far. One important reason

that our approaches outperform the Iterative approach is due to the small and constant query processing rounds in our approaches. In our experiment, our algorithms complete within two rounds; while the iterative approach incurs about 60-200 rounds. Note that the experiments on adaptive algorithm are conducted on a setting that exhibits dynamic changes with certain temporal locality. Since the algorithm dynamically adapts to the changes by switching to appropriate methods, it provides an additional saving over the other algorithms.

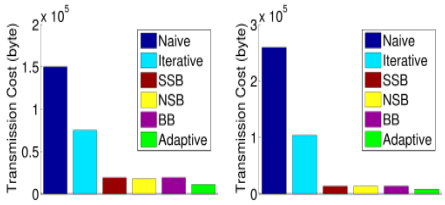


Fig -3 : Performance Evaluation

6.2.2 Sensitivity Tests

Next, we examine the impact of a variety of query and system parameters on the performance of the proposed algorithms. In the plots, we do not show there result of baseline approaches for clarity of presentation. We also omit the plots of experiments on real traces due to space limitation.

Here we first show the impact of query parameters, i.e., k and p, on performance. It shows the trend of transmission cost by varying k from 2 to 10. At the transmission cost increases for all algorithms because the number of tuples needed for query processing is increased. Among the SSB, NSB, and BB algorithms, BB does not perform as well as others when k is small but it becomes a good choice when k becomes larger.

6.2.3 Adaptiveness

The adaptive algorithm reaches our expectation by achieving the least transmission cost under all circumstances. In this section, we further test its adaptiveness to dynamic sensor network environments. One important factor that has an impact on the adaptive algorithm is the size of tuning window. To figure out the optimal setting of tuning window size, Additionally, we conduct an experiment to show the behaviors of algorithms under a dynamic environment. We simulate shows that the adaptive algorithm switches timely to match the best algorithms. The adaptive algorithm switches from the NSB algorithm to the BB algorithm at about time 10, then returns back to NSB at about time 20, and finally switches to the BB algorithm again at about time 30. While the different algorithms may outperform each other at different time, the adaptive algorithm adapts to the dynamic changes to achieve the least transmission cost.

VII. ImplementationResults

Implementation is the most crucial stage in achieving a successful system and giving the user's confidence that the new system is workable and effective. This type of conversation is relatively easy to handle, provide there are no major changes in the system. Each program is tested individually at the time of development using the data and has verified that this program linked together in the way specified in the programs specification, the computer system and its environment is tested to the satisfaction of the user.

The system that has been developed is accepted and proved to be satisfactory for the user. The final stage is to document the entire system which provides components and the operating procedures of the system. Systems implementation is the construction of the new system and the delivery of that system into production (that is, the day-to-day business or organization operation).

Implementation is the carrying out, execution, or practice of a plan, a method, or any design for doing something. As such, implementation is the action that must follow any preliminary thinking in order for something to actually happen. In an information technology context, implementation encompasses all the processes involved in getting new software or hardware operating properly in its environment, including installation, configuration, running, testing, and making necessary changes. The word deployment is sometimes used to mean the same thing. Generally implementation of the software is considered as the actual creation of the software.

VIII. Conclusion

In this paper, we propose the notion of sufficient set and necessary set for efficient in-network pruning of distributed uncertain data in probabilistic top-k query processing. Accordingly, we systematically derive sufficient and necessary boundaries and propose a suite of algorithms, namely SSB, NSB, and BB algorithms, for in-network processing of PT-Topk queries. Additionally, we derive a cost model on communication cost of the three proposed algorithms and propose a cost-based adaptive algorithm that adapts to the application

dynamics. Although our work in this paper is based mainly under the setting of two-tier hierarchical network, the concepts of sufficient set and necessary set are universal and can be easily extend to a network with tree topology. The performance evaluation validates our ideas and shows that the proposed algorithms reduce data transmissions significantly. While focusing on PT-Topk query in this paper, the developed concepts can be applied to other top-k query variants.

IX. Future Enhancements

Every web application has its own merits and demerits. The project has covered almost all the requirements. Further requirements and improvements can easily be done since the coding is mainly structured or modular in nature. Changing the existing modules or adding new modules can append improvements. Further enhancements can be made to the application, so that the web site functions very attractive and useful manner than the present one. We plan to develop algorithms to support other probabilistic top-k queries in the future.

References

[1] V. Bychkovskiy, S. Megerian, D. Estrin, and M. Potkonjak, "A Collaborative Approach to in-Place Sensor Calibration," Proc.Second Int'l Conf. Information Processing in Sensor Networks (IPSN), pp. 301-316, 2003.

[2] 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.

[3] 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 Sensor Networks (DMSN '05), pp. 61-66, 2005.

[4] X. Lian and L. Chen, "Probabilistic Ranked Queries in Uncertain Databases," Proc. 11th Int'l Conf. Extending Database Technology (EDBT '08), pp. 511-522, 2008.

[5] Y. Xu, W.-C.Lee, J. Xu, and G. Mitchell, "Processing Window Queries in Wireless Sensor Networks," Proc. IEEE 22nd Int'l Conf. Data Eng. (ICDE '06), 2006.

[6] S. Madden, M.J. Franklin, J. Hellerstein, and W. Hong, "TAG: A Tiny AGgregation Service for Ad-Hoc Sensor Networks," Proc. Fifth Symp. Operating Systems Design and Implementation (OSDI '02), 2002.

[7] 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.

[8] 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.

[9] K. Yi, F. Li, G. Kollios, and D. Srivastava, "Efficient Processing of Top-k Queries in Uncertain Databases with X-Relations," IEEE Trans. Knowledge and Data Eng., vol. 20, no. 12, pp. 1669-1682, Dec. 2008.

[10] J. Li, B. Saha, and A. Deshpande, "A Unified Approach to Ranking in Probabilistic Databases," Proc. Int'l Conf. Very Large Data Bases (VLDB), vol. 2, no. 1, pp. 502-513, 2009.

[11] Y. Diao, D. Ganesan, G. Mathur, and P.J. Shenoy, "Rethinking Data Management for Storage-Centric Sensor Networks," Proc. Conf. Innovative Data Systems Research (CIDR '07), pp. 22-31, 2007.

[12] M.A. Soliman, I.F. Ilyas, and K.C. Chang, "Top-k Query Processing in Uncertain Databases," Proc. Int'l Conf. Data Eng. (ICDE '07), 2007.