ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

An Adaptive Algorithm for Distributed Processing in Multidimensional Data Sets

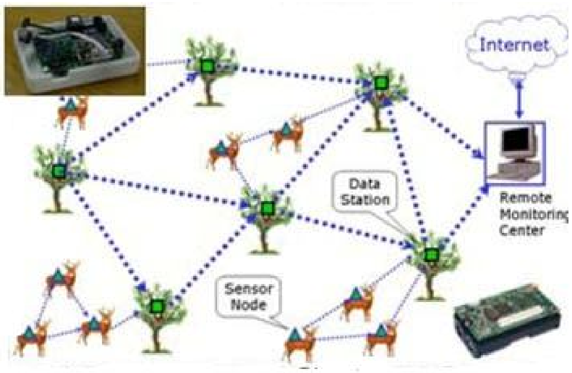
Sujithra.M[1], Gokulakrishnan.V[2]
Department of Computer Science And Engineering, Dhanalakshmi Srinivasan Engineering College, Perambalur, India [1, 2]

Abstract- The distributed processing of probabilistic top-k queries in cluster based wireless sensor networks is done by sufficient set and necessary set. System issues such as indexing techniques and query processing have been examined. Due to the centralized system setting, there is a failure occur in individual tuple means it will affects the neighboring tuples. The transmission cost is high in centralized system settings and more rounds of communication takes place. To overcome that, three algorithms are proposed for intercluster query processing with bounded rounds of communication. The proposed algorithms reduce data transmissions significantly. The algorithms namely Sufficient Set Based (SSB), Necessary Set Based (NSB), and Boundary Based (BB) are used. Therefore developing an Adaptive algorithm that switches among three algorithms to minimize the transmission cost. The concepts include data pruning and data aggregation are introduced. These two concepts have properties that can facilitate localized data pruning in clusters. These algorithms reduce data transmissions and incur only small constant rounds of communication. Hence to analyze the cost during data transmission a cost based adaptive algorithm is used. The least transmission cost is achieved by using the cost based adaptive algorithm.

Keywords— SSB, NSB, BB, Sensor Network, IntraCluster data Pruning, InterCluster Query Processing

I INTRODUCTION

In this network, sensor nodes are grouped into clusters, within each of which one of sensors is selected as the cluster head for performing localized data processing. By using statistic methods, a cluster head may generate a set of data tuples for each zone within its monitored region. In this example, we assume that each tuple is comprised of tuple id, zone, a derived possible attribute value, along with a confidence that serves as a measurement of data uncertainty.

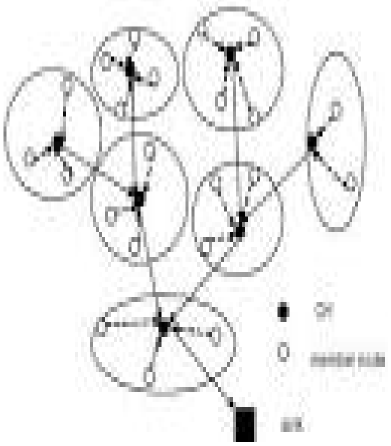


Thus, the data tuples corresponding to the same zone collectively represent the probabilistic distribution of derived possible values for the zone. Since the existence of possible values in the tuples is exclusive to each other, they naturally form a logical tuple, called x-tuple.

There are several top-k query semantics and solutions, including Topk, UkRank, and PT-Topk expected rank and so on. A common way to process probabilistic top-k queries is to first sort all tuples based on the scoring attribute, and then process tuples in the sorted order to compute the final answer set. Answer sets for the mentioned queries typically consist of highly ranked tuples in the sorted list because tuples positioned lowly usually do not have the required ranks and confidence to be included in the answer sets. The techniques require numerous iterations of computation and communication, introducing tremendous communication overhead and resulting in long latency. This is not desirable for many distributed applications, e.g., network monitoring that require the queries to be answered in a good response time, with minimized energy consumption. Hence, aiming at developing energy efficient algorithms optimized for fixed rounds of communications. It is general approach for efficient processing of the probabilistic top-k queries in distributed wireless sensor network.

1.1 Applications of Sensor Data Mining

Wireless sensor networks can be used for facilitating the collection of data for spatial data mining for a variety of applications such as air pollution monitoring.



A characteristic of such networks is that nearby sensor nodes monitoring an environmental feature typically register similar values. This kind of data redundancy due to the spatial correlation between sensor observations inspires the techniques for in-network data aggregation and mining. A sensor node might vary in size from that of a shoebox down to the size of a grain of dust, although functioning "motes" of genuine microscopic dimensions have yet to be created. Each such sensor network node has typically several parts, a radio transceiver with an internal antenna or connection to an external antenna, a micro controller, an electronic circuit for interfacing with the sensors and an energy source, usually a battery or an embedded form of energy harvesting.

Applications

- Area monitoring
- Health care monitoring
- Environmental monitoring
- Industrial monitoring
- Water Quality monitoring

II LITERATURE SURVEY

"Top-k Query Processing in Uncertain Databases", Efficient processing of uncertain data is a crucial requirement in different domains including sensor networks, moving objects tracking and data cleaning. Several probabilistic data models have been proposed, to capture data uncertainty at different levels. According to most of these models, tuples have membership probability, e.g., based on data source reliability, or fuzzy query predicates. Tuple attributes could also contain multiple values drawn from discrete or continuous domains

"Semantics of Ranking Queries for Probabilistic Data and Expected Ranks", When dealing with massive quantities of data, top-k queries are a powerful technique for returning only the k most relevant tuples for inspection, based on a scoring function. The problem of efficiently answering such ranking queries has been studied and analyzed extensively within traditional database settings. The importance of the top-k is perhaps even greater in probabilistic databases, where a relation can encode exponentially many possible worlds.

"Working Models for Uncertain Data", uncertain data is the notion of data that contains specific uncertainty. Uncertain data is typically found in the area of sensor networks. When representing such data in a database, some indication of the probability of the various values. The fundamental difference between a traditional relational database and an uncertain relational database is that an uncertain relation represents a set of possible relation instances, rather than a single one.

"Ranking Distributed Probabilistic Data", Data are increasingly stored and processed distributive as a result of the wide deployment of computing infrastructures and the readily available network services. More and more applications collect data from distributed sites and derive results based on the collective view of the data from all sites. Due to the large amounts of data available nowadays and the network delay incurred, as well as the economic cost associated with such communication. Ranking queries are essential tools to process large amounts of probabilistic data that encode exponentially many possible deterministic instances.

III BACK GROUND

Ranking queries are a powerful concept in focusing attention on the most important answers to a query. To deal with massive quantities of data, such as multimedia search, streaming data, web data and distributed systems, tuples from the underlying database are ranked by a score, usually computed based on a user-defined scoring function. Only the top-k tuples with the highest scores are returned for further inspection. So, define several fundamental properties, including exact k, containment, unique rank, value invariance, and stability which are satisfied by ranking queries on certain data. By arguing these properties should also be carefully studied in defining ranking queries in probabilistic data, and fulfilled by definition for ranking uncertain data for most applications. Hence, an intuitive new ranking definition based on the observation that the ranks of a tuple across all possible worlds represent a well-founded rank distribution. The median and other statistics of this rank distribution for a tuple and derived the expected rank, median rank and quartile rank correspondingly. Hence, able to prove that the expected rank, median rank and quantile rank satisfy all these properties for a ranking query. A new approach based on the distribution of each tuple's ranks across all possible worlds. By leveraging statistical properties on such a rank distribution, such as the expectation, the median and the quartile, derive the expected rank, the median rank and quartile rank. The rest of the paper is organized as follows. Section IV shows the Architecture of the Proposed Approach, Section V reviews the related work on the proposed system, Section VI reviews the Conclusion.

IV SYSTEM ARCHITECTURE

In figure 4.1 represent the system architecture. Here, the Sensor Network uses three processes. They are Clusters, Zones and Records. The Clusters is mainly used for grouping similar set of data. Multiple sensors are deployed at certain zones in order to improve monitoring quality. In this network, sensor nodes are grouped into clusters, within each of which one of sensors is selected as the cluster head for performing localized data processing.

So, a cluster head may generate a set of data tuples for each zone within its monitored region and each tuple is comprised of tuple id, zone, a derived possible attribute value, along with a confidence that serves as a measurement of data uncertainty. Thus, the data tuples corresponding to the same zone collectively represent the probabilistic distribution of derived possible values for the zone. Record a data structure, Storage record is a basic input/output structure and Record (database), a set of fields in a database related to one entity, Record used to start an operating system. Here focus on addressing the communication overhead which is critical for wireless sensor networks and their applications.

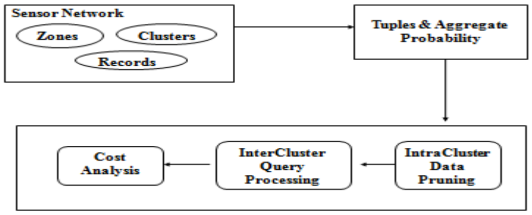

Sensor Network
Zones
Clusters
Records
Tuples & Aggregate Probability
Cost Analysis
InterCluster Query Processing
IntraCluster Data Pruning

Figure 4.1 *System Architecture*

For simplicity, logically assume single-hop transmission in both intra cluster and inter cluster communications. Nevertheless, algorithms are not restricted to this assumption and can be extended for the multi hop communications. As long as the base station receives all the candidate data tuples and supplementary tuples, So able to compute the final answer with a generic centralized algorithm.

A cost based adaptive algorithm is used, that switches dynamically among SSB, NSB and BB as the data distribution within the network changes. Here, approach is to keep track of the estimated cost for all three methods under different scenarios in order to trigger the switch as appropriate. The key issue is to estimate the cost of the running and alternative methods under different scenarios. In the following, consider three scenarios running SSB, NSB and BB respectively, and show the cost of alternative method is estimated.

V PROPOSED SYSTEM

Based on the notion of sufficient and necessary sets for in-network processing of PT-Topk queries in a two-tier hierarchical sensor network the algorithms exploit individual and combined strengths of sufficient and necessary sets in query processing. These algorithms are used to minimize the communication and energy overhead in responding to changing data distribution in the network. The algorithms are Sufficient set-based (SSB) algorithm, Necessary set-based (NSB) algorithm, and Boundary-based (BB) algorithm. Therefore, developing a cost model based on communication cost of the three algorithms. Accordingly, a cost-based adaptive algorithm that dynamically switches among the three algorithms based on their estimated costs.

Advantages of Proposed System

- The least transmission cost is achieved using the cost based adaptive algorithm.
- These algorithms reduce data transmissions significantly with bounded rounds of communication
- Repeated set of data gets avoided.
- Include accurate probability method
- Provide expected ranking method
- Accurate results for both probability & top-k results

A Modules:

- Sensor Network
- Top-k Probability
- Intracluster Pruning
  - o Sufficient Boundary
  - o Necessary Boundary
- Intercluster Query Processing
  - o Sufficient Set Boundary
  - o Necessary Set Boundary
  - o Boundary-Based Algorithm
- Cost Analysis
  - o Cost for SSB
  - o Cost for NSB
  - o Cost for BB

a) Sensor Network

A wireless sensor network that consists of a large number of sensor nodes deployed in a geographical region. Feature readings (e.g., moisture levels or speed of wind gust) are collected from these distributed sensor nodes. In this network, sensor nodes are grouped into clusters, within each of which one of sensors is selected as the cluster head for performing localized data processing

b) Top-k Probability

Define the tuple structure for each Zone. Then calculate the aggregate probability for all zones. Let W denote a possible world which consists of a subset of tuples in T and W denote the set of all possible worlds. The probability that w϶W exists.

c) Intracluster 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. So, the notion of sufficient set and necessary set, and describe how to identify them from local data sets at cluster heads. Next, 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 obtained locally is an upper bound of its true top-k probability.

d) Intercluster Query Processing

The notion of sufficient and necessary sets as a basis, proposing three 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. Here, focus on addressing the communication overhead which is critical for wireless sensor networks and their applications. For simplicity, logically assume single hop transmission in both intracluster and intercluster communications.

e) Cost Analysis

A cost analysis is performed on data transmission of the three proposed methods. A cost based adaptive algorithm that switches dynamically among SSB, NSB and BB as the data distribution within the network changes. Hence, approach is to keep track of the estimated cost for all three methods under different scenarios in order to trigger the switch as appropriate. The key issue is how to estimate the cost of the running and alternative methods under different scenarios.

VI CONCLUSION

Here, the notion of sufficient set and necessary set for efficient in-network pruning of distributed uncertain data in probabilistic top-k query processing. Accordingly 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, derive a cost model on communication cost of the three proposed algorithms and find out a cost based adaptive algorithm that adapts to the application dynamics. Although work 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.

REFERENCES

[1] Benjelloun O., Halevy A., Sarma A.D. and Widom J. (2006) “Working Models for Uncertain Data,”Proc. 22nd Int’l Conf. Data Eng.

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

[3] Chang K.C., Ilyas I.F., and Soliman M.A. (2007) “Top-k Query Processing in Uncertain Databases,”Proc. Int’l Conf. Data Eng.

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

[5] Chen L., Jin C., Lin X., Yi K., and Yu J.X. (2008) “Sliding-Window Top-k Queries on Uncertain Streams,”Proc. Int’l Conf. Very Large Data Bases.

[6]Cheng R., Kao B., Ngai W.K., Tao Y., Prabhakar S., and Xiao X. (2005) “Indexing Multi-Dimensional Uncertain Data with Arbitrary Probability Density Functions,” Proc. 31st Int’l Conf. Very Large Data Bases , pp. 922-933, 2005.

[7] Dalvi N., Re C., and Suciu D. (2007) “Efficient Top-k Query Evaluation on Probabilistic Data,” Proc. Int’l Conf. Data Eng.

[7] Cormode G., Li F., and Yi K. (2009) “Semantics of Ranking Queries for Probabilistic Data and Expected Ranks,”Proc. IEEE Int’l Conf. Data Eng.

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

[10] Diao Y., Ganesan D., Mathur G., and Shenoy P.J. (2007) “Rethinking Data Management for Storage-Centric Sensor Networks,”Proc. Conf. Innovative Data Systems Research.

[11] Jestes J., Li F., and Yi K. (2009) “Ranking Distributed Probabilistic Data,” Proc. 35th SIGMOD Int’l Conf. Management of Data.

[12] Lee W.C., Liu X., and Xu J. (2010) “A Cross Pruning Framework for Top-k Data Collection in Wireless Sensor Networks,”Proc. 11th Int’l Conf. Mobile Data Management,pp. 157-166