International Journal of Ethics in Engineering & Management Education
Website: www.ijeee.in (ISSN: 2348-4748, Volume 1, Issue 4, April 2014)

THE EFFICIENT PRUNING OF DISTRIBUTED UNCERTAIN DATA IN PROBABILISTIC TOP-K QUERIES PROCESSING.

MANJUNATH .K[1] K .S.PATIL[2]

[1] Dept. Computer Science and Engg, Basaveshware Engineering College, Bagalkot, Karnataka, India.
[2]Prof: Dept. Computer Science and Engg, Basaveshware Engineering College, Bagalkot, Karnataka, India.

Abstract: **Distributed database systems (DDS) are needed for applications where data and its access are inherently distributed and to increase availability during failures. In wireless sensor networks (WSN), the base station and the N sensor nodes are grouped into clusters and from each clusters one of the sensor node is selected as a cluster .the whole sensor network can be logically treated as a distributed uncertain data. The cluster head are responsible for generating uncertain data tuples from its cluster and performing the pruning on un-certain data based on the cost efficient performance and report pruned uncertain data and probabilistic values to base station (BS).The BS collects all types based from its cluster head and sorts all tuples based on the probabilities values and then process tuples in sorted order to reduce the communication and energy cost. The software constants implementation and testing are used to improve the performances of sensor network.**

Index terms: **Distributed Database system, Wireless sensor networks, Pruning uncertain data**

1. INTRODUCTION

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 or external noise and so on here we first use an environmental monitoring application of wireless sensor network to introduce some basics of probabilistic databases [1] thus 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 and to report [2]. The below figure1 shows that there are six zones denoted by A,B,C,D,E,F. which are grouped in to two clusters D1 and D2 with corresponding cluster head c2 in D1 and c7 in D2. The base station which collects all the aggregate data from its neighbours cluster head by inter communication (SSB and NSB) and sorted them in descending order the highly ranked k-tuples in the list.

2. SYSTEM DESIGN:

We briefly outline the concepts that are necessary for implementations [3].they include structure (architecture) of the as shown in the figure2, there are two system structure that can be distinguished: data storage based and data retrieval based.

2.1 Input Design:

The input design is the link between the information systems and user. It comprises the developing specification and procedures for data preparation and those step-s are necessary to put transaction data in to a usable form for processing can be achieve by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system[4].

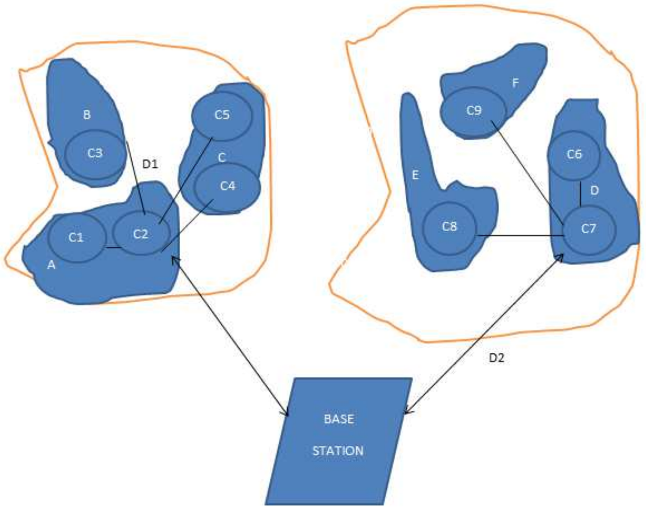

Fig-1: Distributed Network

The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple, the input is design in such a way so that it provides security and ease of use with retaining the privacy. Input design consider the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validation and steps to follow when error occur.

2.1.2 Objectives

- Input design is the process of converting a user-oriented description of the input into a computer-based systems. Input design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized systems [5].
- It is achieved by creating user friendly screens for the data entry to handle large volume of data. the goal of designing input is to make data entry easier and to be free from errors
- When the data is entered it will check for its validity. Data can be entered with the help of screens.

2.2 Output Design:

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any systems results of processing are communicated to the users and to other systems through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the systems relationship to help user decision-making [6].

1) Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so the people will find the system can use easily and effectively. When analysis design computer output, they should identify the specific output that is needed to meet the requirements.
2) Select methods for presenting information.
3) Create document, report, or other formats that contain information produced by the system.

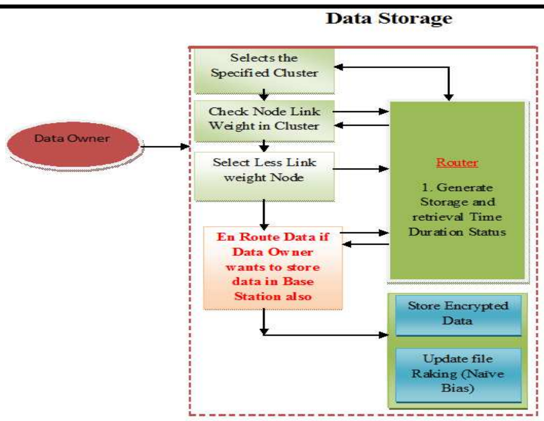
Data Storage
Data Owner
Selects the Specified Cluster
Check Node Link Weight in Cluster
Select Less Link weight Node
En Route Data if Data Owner wants to store data in Base Station also
Router
1. Generate Storage and retrieval Time Duration Status
Store Encrypted Data
Update file Raking (Naïve Bias)

Fig-2: Data Storage

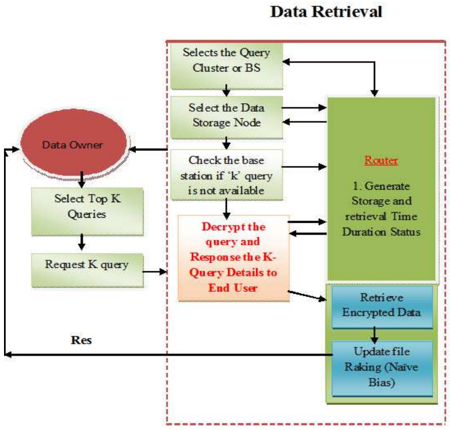
Data Retrieval
Data Owner
Select Top K Queries
Request K query
Selects the Query Cluster or BS
Select the Data Storage Node
Check the base station if 'k' query is not available
Decrypt the query and Response the K-Query Details to End User
Router
1. Generate Storage and retrieval Time Duration Status
Retrieve Encrypted Data
Update file Raking (Naïve Bias)
Res

Fig-3: Data Retrieval

2.2.1 Objectives:

- Convey information about past activities, current status or projections of the future.
- Signal important events, opportunities, problems or warnings.
- Trigger an action and confirm an action.

3. SYSTEM ANALYSIS

3.1 *Problem Statement:*

The efficient pruning of processing probabilistic topk queries over uncertain data in distributed sensor network and perform to minimize the transmission cost in wireless sensor network.

3.2 *IssuesIdentified:*

- We explore the problems of processing probabilistic topkqueries in distributed wireless sensor network [15].
- Data pruning is not done in accurate
- One station to another station delay the communication rate.

3.3 Overcoming the Issues:

- Necessary set based (NSB)and boundary based (bb) are effective for inter-cluster pruning.
- Transmission cost increases for all algorithms because the number of tuples need for queries processing is increased.

4. IMPLEMENTATION

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective.The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementations, designing of methods to achieve changeover and evaluation of changeover methods [7].

4.1 Module Description

- **PT-Topk Query Processing:** The PT-Topk queries in a centralized uncertain database, which provides a good background for the targeted distributed processing problem. 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 [8].

- **Sensor Networks:** The extensive number of research work in this area has appeared in the literature. Due to the limited energy budget available at sensor nodes, the primary issue is how to develop energy-efficient techniques to reduce communication and energy costs in the networks. Approximate-based data aggregation techniques have also been proposed. The idea is to trade off some data quality for improved energy efficiency [14]. Silberstein et al. develop a sampling-based approach to evaluate approximate top-k queries in wireless sensor networks. Based on statistical modelling techniques, a model-driven approach was proposed into balance the confidence of the query answer against the communication cost in the network. Moreover, continuous top-k queries for sensor networks have been studied in and. In addition, a distributed threshold join algorithm has been developed for top-k queries. These studies, considering no uncertain data, have a different focus from our study [9].

- **Data pruning:**The cluster heads are responsible for generation 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 [10].

- **Structured network topology:** To perform in-network query processing, a routing tee is often formed among sensor nodes and the base station. A query is issued at the root of the routing tree and propagated along the tree to all sensor nodes. Although the concepts of sufficient set and necessary set introduced earlier are based in two-tier hierarchical sensor networks, they are applicable to tree-structured sensor network [11].

- **Data transmission:**The total amount of the 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 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 experiment to evaluate algorithms, SSB-T, NSB-T, and NSB-T-Opt under the tree-structured network topology [12].

- **Performance evaluation:** The 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 [13], [15].

5. SYSTEM TESTING

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product .It provides a way to check the functionality of components, sub-assemblies,assemblies and a finished product. It is the process of exercising software with the intent of ensuring that the software system meets its requirements and user expectations and not fail in an unacceptable manner [8].

5.1 Unit Testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decisions branches and internal code flow should be validated. It is the testing of individual software units of the application. It is the done after the completion of an individual units before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected [8].

5.2 Integration Testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were invidiously satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

5.3 Functional Testing

Functional test provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation , and user manuals.Functional testing is centred on the following items:

- Valid input : Identified classes of valid input must be accepted.
- Invalid input : Identified classes of invalid input must be rejected.
- Functions : Identified functions must be exercised.
- Output : Identified classes of application must be exercised.
- System /procedures: Interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, data fields, predefined process, and successive processes must be considered for testing. Before testing is complete,additional test are identified and the effective valueof current testsis determined.

CONCLUSION

In this paper normally, the whole sensor networks can be logically treated as a distributed uncertain data, measure the pruning on uncertain data based on the cost efficient performances. Although our work in this paper is based mainly under the setting of two-tier hierarchical network, the algorithms help to understand the concepts of data owner (service provider), cluster router and base station which can be easily extended to a network with tree topology. While focusing on PT-Topk query in this paper, the query variants and sorts all tuples based on the probabilistic values and then process tuples in sorted order to reduce the communication and energy cost. The software constraints are implemented and tested to improve the performances of sensor networks.

REFERENCES

[1]. N. Dalvi and D. Suciu. "Efficient Query Evaluation on Probabilistic Databases". Proc. 30th Int'l Conf. Very Large Data Bases (VLDB'04), pages pp.864-875, 2004.

[2]. 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), pages pp. 301-316, 2003.

[3]. 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.

[4]. R. Cheng D.V. Kalashnikov and S. Prabhakar"Evaluating Probabilistic Queries over Imprecise Data". Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMO '03).

[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]. G. Cormode F. Li and K. Yi. "Semantics of Ranking Queries for Probabilistic Data and Expected Ranks". Proc. IEEE Int'l Conf. Data Eng. (ICDE '09), 2009.

[7]. M.A. Soliman and I.F. Ilyas. "Ranking with Uncertain Scores". Proc. IEEE Int'l Conf. Data Eng, ICDE-09:pp.11-15, 2009.

[8]. www.google.co. /testing methods. Knowledge and Data Eng, 2008.

[9]. www.google.com. /distributed system. Knowledge and Data Eng, 2010.

[10]. X. Liu J. Xu and W.-C. Lee. "A Cross Pruning Framework for Top-k Data Collection in Wireless Sensor Networks". Proc. 11th Int'l Conf. Mobile Data Management,:pp. 157-166, 2010.

[11]. D. yanlei D. ganesan G. mathur and P. shenoy."Rethinking Data Management for Storage centric Sensor Networks". Department of Computer Science (MA1003), 2003.

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

[13]. 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, pages pp. 1669-1682,2008.

[14]. 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.

[15]. 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.