International Journal of Thesis Projects and Dissertations (IJTPD)
Vol. 3, Issue 3, pp: (39-43), Month: July - September 2015, Available at: www.researchpublish.com

Searching Measurement for Imperfect Databases

G.Thilagavathi

Research scholar, Department of computer science, Vels University, Chennai, India

Abstract: **Uncertain data is inherent in a few important applications such as environmental surveillance and mobile object tracking. Top-k queries (also known as ranking queries) are often natural and useful in analyzing uncertain data in those applications. In this paper, we study the problem of answering probabilistic threshold top-k queries on uncertain data, which computes uncertain records taking a probability of at least p to be in the top-k list where p is a user specified probability threshold. I present an efficient exact algorithm, a fast sampling algorithm, and a Poisson approximation based algorithm. An empirical study using real and synthetic data sets verifies the effectiveness of probabilistic threshold top-k queries and the efficiency of our methods.**

Keywords: **Dimension incomplete database, similarity search, whole sequence query.**

1. INTRODUCTION

Similarity query in multidimensional database is a fundamental research problem with numerous applications in the areas of database, data mining, and information retrieval. Given a query object, the goal is to find similar objects in the database [1], [2], [3], [4]. Recently, querying incomplete data has attracted extensive research efforts [5], [6], [7]. In this problem, the data values may be missing due to various practical issues. For example, in sensor networks, the received data may become incomplete when sensors do not work properly or when errors occur during the data transfer process. The data incompleteness problem studied in the existing work usually refers to the missing value problem, i.e., the data values on some dimensions are unknown or uncertain. The common assumption of the existing work is that, for each dimension, whether its data value is missing or not is known. However, in real-life applications, we may not know which dimensions or positions have data loss [8], [9]. In these cases, we only have the arrival order of data values without knowing which dimensions the values belong to. When the dimensionality of the collected data is lower than its actual dimensionality, the correspondence relationship between dimensions and their associated values is lost. I refer to such a problem as the dimension incomplete problem.

- Data missing when dimension information is not explicitly maintained.
- Time series data with temporal uncertainty due to imprecise time stamps.

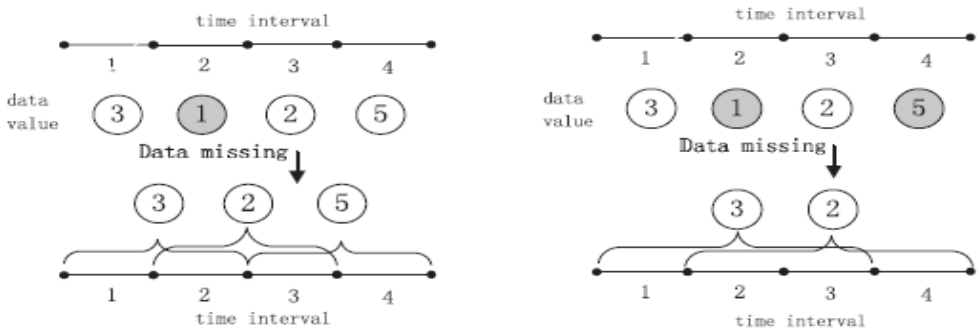


Fig.1.Dimension incomplete data due to dimension information not being explicitly maintained

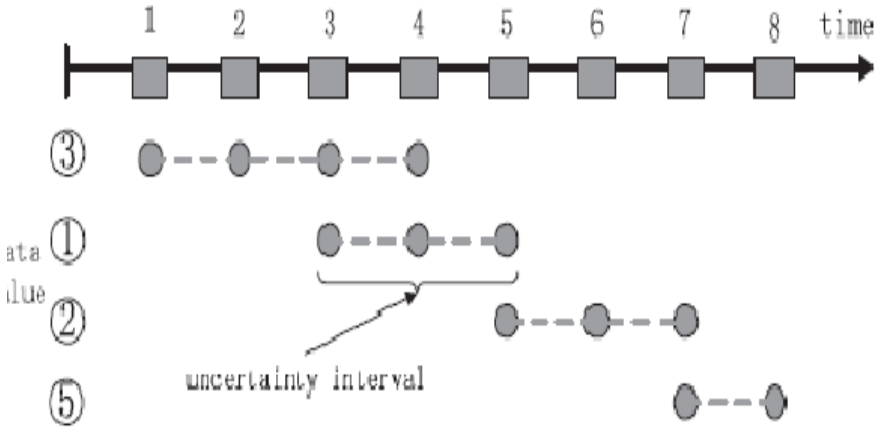


Fig.2.Dimension incomplete data due to imprecise time stamps

2. PROBLEM DEFINITION

Let D be the database. A data object X €D is a real valued vector (x1; x2; . . . ; xm), where $x_i(1 \leq i \leq m)$ is the data value for the ith dimension of X. |X| = m denotes the dimensionality of X. A data object X is dimension incomplete, if it satisfies

- at least one of its data elements is missing;
- The dimension of the missing data element is unknown.

The traditional range query on a multidimensional database is defined as follows: Given a database D containing N data objects of m dimensions, an m-dimensional query Q, a distance function δ, and a distance threshold r,

Traditional range queries retrieve all the data objects in D whose distances from Q are less than r. More formally,

$$RangeQuery_{\delta}(D,Q,r) = \{X \in D \mid \delta(Q,X) < r\}. \quad (1)$$

3. WHOLE SEQUENCE QUERY ON DIMENSION INCOMPLETE DATA

The probability triangle inequality is first applied to evaluate the data objects. In this step, some data objects are judged as true results and some are filtered out. The lower and upper bounds of the probability are then applied to evaluate the remaining data objects, from which some are determined as true results and some as dismissals. Only those data objects that cannot be determined in the former two steps are evaluated by the naive method.

- **Bounds of the Probability:**

I develop the lower and upper limit of the probability Pr [δ (Q,X) < r], the proof of their correctness, and an efficient algorithm for calculating them.

- **Efficient Bound Evaluation:**

To utilize this pruning strategy, need efficient algorithms for

- Calculating the probability P[δLB(Q,X) < r] and Pr[δUB(Q,X) < r], and
- Calculating the distance bounds.

- **The Overall PSQ-DID Algorithm:**

The triangle inequality and the bounds of probability can be evaluated efficiently and used to effectively prune the search space. Only a small portion of the data objects need to be evaluated by the naive verification algorithm. To further increase the efficiency, I can avoid the naïve verification step and simply treat the remaining candidates as query results (or dismissals, depending on the requirements Of query precision and recall).

This is reasonable for applications where the two probability bounds are effective for selecting candidates. This simplified strategy will dramatically increase the efficiency of the algorithm without causing significant change of the quality of the results.

4. SUBSEQUENCE MATCHING ON DIMENSION INCOMPLETE DATA

I discuss the problem of subsequence matching on dimension incomplete data.

Problem Description:

(Probabilistic Subsequence Matching on Dimension Incomplete Data).

Given a database D containing dimension incomplete sequences of real numbers Xo whose underlying complete version is of the length that is potentially different and unknown, a query sequence Q of length|Q|, a distance threshold r, a probability threshold c, an imputation method ' indicating the distribution of missing data values, and a distance function δ,

$$\begin{aligned}&\mathrm{PSM}-\mathrm{DID}_{\delta,\varphi}(D,Q,r,c)\\&\quad=\{X_o[i:j]|\Pr\{\delta(Q,X_{rv}[k:k+|Q|-1])<r\}\geq c,\\&\qquad X_o\in D,j\geq i-1\}.\end{aligned}$$

Algorithm for Subsequence Matching on Dimension Incomplete Data:

A probabilistic framework is developed to model this problem so that the users can find objects in the database that are similar to the query with probability guarantee. Missing dimension information poses great computational challenge, since all possible combinations of missing dimensions need to be examined when evaluating the similarity between the query and the data objects. I develop the lower and upper bounds of the probability that a data object is similar to the query. These bounds enable efficient filtering of irrelevant data objects without explicitly examining all missing dimension combinations. A probability triangle inequality is also employed to further prune the search space and speed up the query process. The proposed probabilistic framework and techniques can be applied to both whole and subsequence queries. Extensive experimental results on real-life data sets demonstrate the effectiveness and efficiency of our approach.

- The traditional similarity measurements (or distance functions), which are the bases of any similarity query task, may not be applied when the dimension information is missing However, it cannot be calculated if the dimensions of the two objects do not match.

- The data incompleteness problem studied in the existing work usually refers to the missing value problem, i.e., the data values on some dimensions are unknown or uncertain. The common assumption of the existing work is that, for each dimension, whether its data value is missing or not is known.

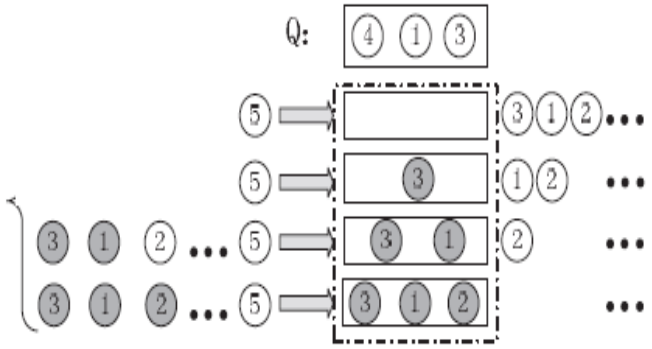
I will show that our framework on whole sequence queries can be extended to handle subsequence queries. Without taking into account the influence of boundary elements in determining the expected values of random variables, we can divide the process of the subsequence query problem as follows:

- Tackle the case matches the subsequence

From the first element of the target dimension Incomplete sequence.

- Remove the first element of target dimension

Incomplete sequence and repeat Step 1; terminate until there is no element left in target sequence. One straightforward method for Step 1 is to divide the problem into several cases and examine each case separately.For example, given query Q = (4, 1, 3) and dimension incomplete sequence Xo= (3, 1, 2, 5), The first case assumes that all elements to be matched with the query are missing, and there are three random variables need to be imputed. Next, we consider the case where only one element is chosen to match the query, and the remaining two elements are assumed to be missing. The third and fourth cases are similar to the second case. Obviously, the first and fourth cases can be evaluated easily, and the second and the third cases can be processed with the algorithms discussed in this figure.



Subsequence matching on dimension incomplete data.

5. CONCLUSION

I conduct extensive experimental evaluation using real data sets. The results indicate that 1) our approach achieves satisfactory performance in querying dimension incomplete data for both whole sequence matching and subsequence matching; 2) both the probability triangle inequality and the probability limits have a good pruning power and improve query efficiency significantly; our work will focus on the following directions. Since a probability triangle inequality holds, I plan to develop an index structure that can utilize the inequality to further improve the efficiency of the query process. Furthermore, we plan to investigate how to extend our query strategy to incorporate a wide range of distance functions.

Algorithm:

Exact matching algorithm:

- Our method can be applied to both whole sequence matching and subsequence matching problems on dimension incomplete data. Moreover, the data of interest can be either static data or dynamical data streams.
- I develop the lower and upper probability limits and the probability triangle inequality that can be used to dramatically prune the search space.

REFERENCES

[1] R.K. Pearson, "The Problem of Disguised Missing Data," ACM SIGKDD Explorations Newsletter, vol. 8, pp. 83-92.

[2] I. Wasito and B. Mirkin, "Nearest Neighbour Approach in the Least-Squares Data Imputation Algorithms," Information Sciences: An Int'l J., vol. 169, pp. 1-25

[3] J. Pei, M. Hua, Y. Tao, and X. Lin, "Query Answering Techniques on Uncertain and Probabilistic Data: Tutorial Summary," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '08), pp. 1357-1364, 2008.