ISSN 2350-1022

International Journal of Recent Research in Mathematics Computer Science and Information Technology
Vol. 2, Issue 1, pp: (247-252), Month: April 2015 – September 2015, Available at: www.paperpublications.org

Enhance The Technique For Searching Dimension Incomplete Databases

[1]Mr. Amol Patil, [2]Prof. Saba Siraj, [3]Miss. Ashwini Sagade

[1](M.E. Student), [2](Asst. Professor), [3](M.E. Student)
[1, 2, 3] Department of Computer Engineering, IOKCOE Pune, India

Abstract: **Data ambiguity is major problem in the information retrieval ambiguity is due to the loss in the data dimension it causes lot of problem in various real life application. Database may incomplete due to missing dimension and value. In previous work is totally based on the missing value. We focus on the problem is to find the missing dimension in our work. Missing dimension leads towards the problem in the traditional query approach. Missing dimension information create computational problem, so large number of possible combinations of missing dimensions need to be examined to check similarity between the query object and the data objects . Our aim is to reduce the all recovery version to increase the system performance as number of possible recovery data is reduces the time to estimate the true result is also reduces.**

Keywords: **Missing Dimensions, Similarity search, Whole sequence query, Probability triangle inequality, Temporal data.**

I. INTRODUCTION

Recently, querying incomplete data has attracted extensive research efforts [1], [2], [3]. In this problem, the data values may be missing due to various practical issues. 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 [4], [5]. 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. We refer to such a problem as the dimension incomplete problem. Now day amount of data are retrieved in data mining is large and information retrieval. Data which is retrieved is mostly not complete the querying incomplete data has attracted greater attention as it poses new challenges to traditional querying techniques [6]. The existing work on querying incomplete data addresses the problem where only the data values on certain dimensions are unknown. In many applications, such as data collected by a sensor network there a noisy environment, so not only the data values but also the dimension information may be missing. The existing work is based on assumption that is for each dimension whether its data value is missing or not is know. We want to investigate the problem of similarity search on dimension incomplete data. We have to reduce the number of compares among similar data. 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. We have developed 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. Due to this search space is reduce and it speed up the query process.

Query similar data is problem faced by the many application like data mining and information retrieval. we are introduce various probability methods to overcome the problem of querying incomplete data. To improve probability we are going to implement *Apriori algorithm* with probability methods.

A. *Time Series Data Sub Sequence Matching:*

In the financial and scientific application time series is and important format for data. Time series data is data of sequence of number with specified temporal instance. In the time series data the problem is to match the first sub sequence. The time series data is masseur by using the impression measuring device and clocking strategies Subsequence matching include the matching the sequence of any data like pattern matching in data. Let two sale data pattern may similar even sale volume is different we can differentiate terms by sequence.

B. *Data Incomplete Due To Dimension Information Is Not Explicitly Maintained:*

Consider the data send over sensor networks. The database usually contains time series data objects each data represented by a sequence of values *(d1, d2, d3…………,dn).* The dimension information associated with data values can be implicitly inferred from the data arrival order. Data arrival order store using time stamp. This schema of data collection and storage is very common in resource-constrained applications since explicitly maintaining dimension information will cause additional costs. In this problem setting, missing a single data element will destroy the dimension information of the entire data object.

C. *Data Dimension Missing Due To Lack Of Clock Synchronization:*

For example, in Fig. 1, the original data object is (3,1,2,5). When data element 1 is missing, the dimension information for the rest of data elements becomes uncertain. For example, 3 can be the first or the second element, and 2 can be the second or the third element. When data elements 1 and 5 are missing, then both elements 3 and 2 may locate on three different dimensions. In applications where dimension information is explicitly maintained, the dimension indicator itself may be lost. This will also cause the dimension incomplete problem.

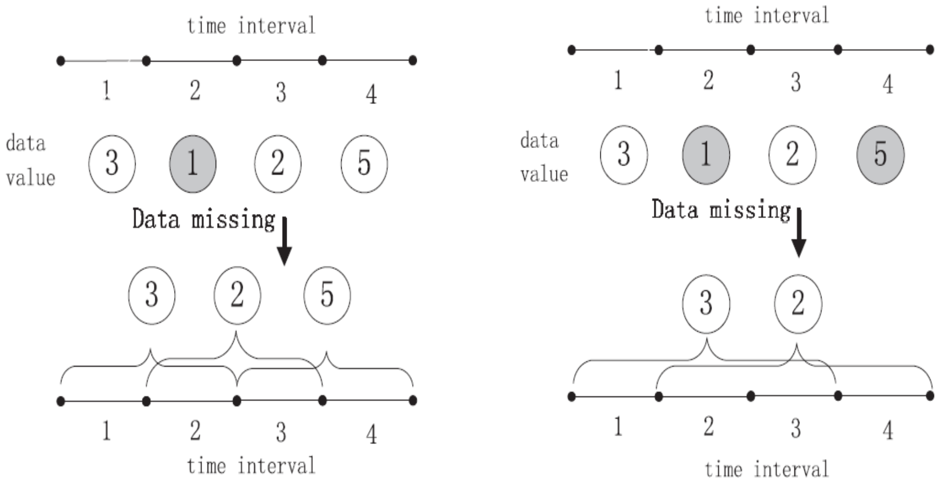

Fig.1. Missing Dimensions

II. LITERATURE SURVEY

There are a variety of reasons why databases may be missing data. The data may not be available at the time the record was populated or it was not recorded because of equipment malfunction or adverse conditions. Data may have been unintentionally omitted or the data is not relevant to the record at hand. The allowance for and use of missing data may be intentionally designed into the database. In some cases, the missing value of data is random. of some value does not depend on the value of another variable. Analysing uncertain data is an area that is also related to our work [7], [8]. The goal of these methods is to estimate a probability density function to model the uncertainty in the data. They do not address the problem of dimension incompleteness.

The authors [9] address the problem where there are missing elements in symbolic sequences. Our problem is more general in the sense that we consider real-value data and address the probabilistic query task. In [10], a temporal model is proposed to discover patterns in streams with imprecise time stamps. This work deals with pattern evaluations in event streams where event ids should be exactly matched. Moreover, the data arrival time intervals are needed to construct the temporal uncertainty model. In our work, such information is not available and only the data arrival order is known. To find common structure of two sequences, dynamic time warping [11], [12] and longest common subsequence [13], [14] algorithms are proposed. In these problems, the exact dimension information is not critical. These methods cannot be directly applied to the similarity query problem on dimension incomplete data.

III. BACKGROUND FORMULATIONS

The overall query process is shown in following Fig. 2. The triangular probability inequality is first which is applied to evaluate the data objects. In this some data objects are observed as true results and some are filtered out. The lower and upper bounds of the probability are then applied to estimate the remaining data objects, from which some are determined as true results and some as discarded. Only those data objects that cannot be determined in the above two steps are evaluated by the naïve probability method.

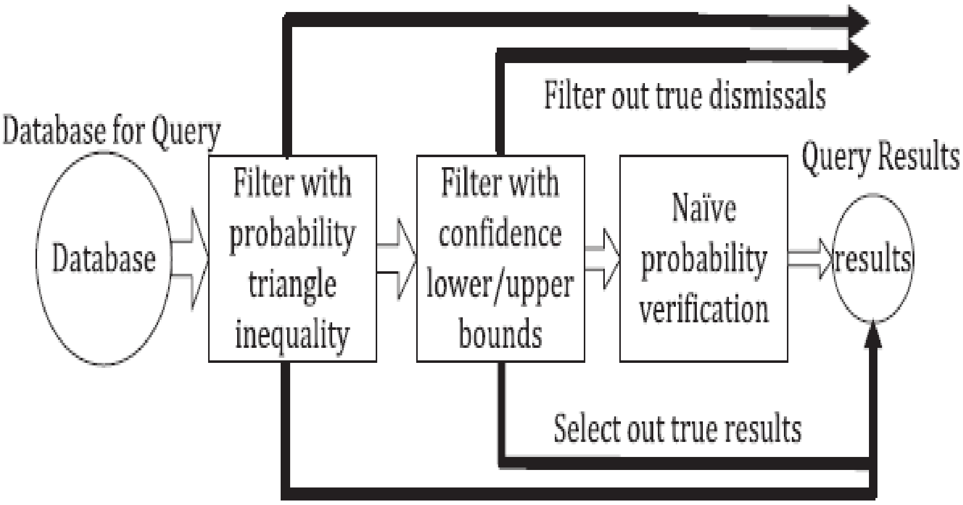
Filter out true dismissals
Database for Query
Query Results
Database
Filter with probability triangle inequality
Filter with confidence lower/upper bounds
Naïve probability verification
results
Select out true results

Fig.2. Query Processing

A. Probability Bounds Estimation:

We propose the probability bounds to estimate the incomplete data object by matching object of query to possible available recovery version (eg. Number of possible data object).there are lower bound and upper bound probability to find the correct data value to file the incomplete data. We proposed the distance threshold(r) and probability threshold or confidence(c). Following are the equation for upper and lower bound of distance function (Ə):

Lower bound:

$$\delta_{LB}(Q,X) = [\delta^2_{LB_1}(Q,X_1) + \delta^2_{LB_0}(Q,X_0)]^{\frac{1}{2}} \qquad (1)$$

Upper bound:

$$\delta_{UB}(Q,X) = [\delta^2_{UB_1}(Q,X_1) + \delta^2_{UB_0}(Q,X_0)]^{\frac{1}{2}} \qquad (2)$$

B. Triangular probability inequality:

In the probability inequality triangle the data objects in database can be determined to be true results which are less than the distance threshold and greater than the probability threshold.

$$Pr[\delta(Q,X) < r] \le Pr[\delta_{LB}(R,X) - \delta(Q,R) < r] \qquad (3)$$

The data objects in database can be determined to be dismissal results which are less than the distance threshold and less than the probability threshold

$$\Pr[\delta(Q,X) < r] \geq \Pr[\delta_{UB}(R,X) + \delta(Q,R) < r] \qquad (4)$$

It can be found that our method for subsequence matching can also tackle this problem, with only a little modification required. Specifically, when the sequence is updated, we only have to do the two steps on the new added part of the sequence. If we find the matching patterns that meet the query requirements, the system will output the matched sub-sequences.

IV. PROPOSED APPROACH

To increase the performance of the existing system we proposed our system which decrease response time and increase the overall performance of the system.

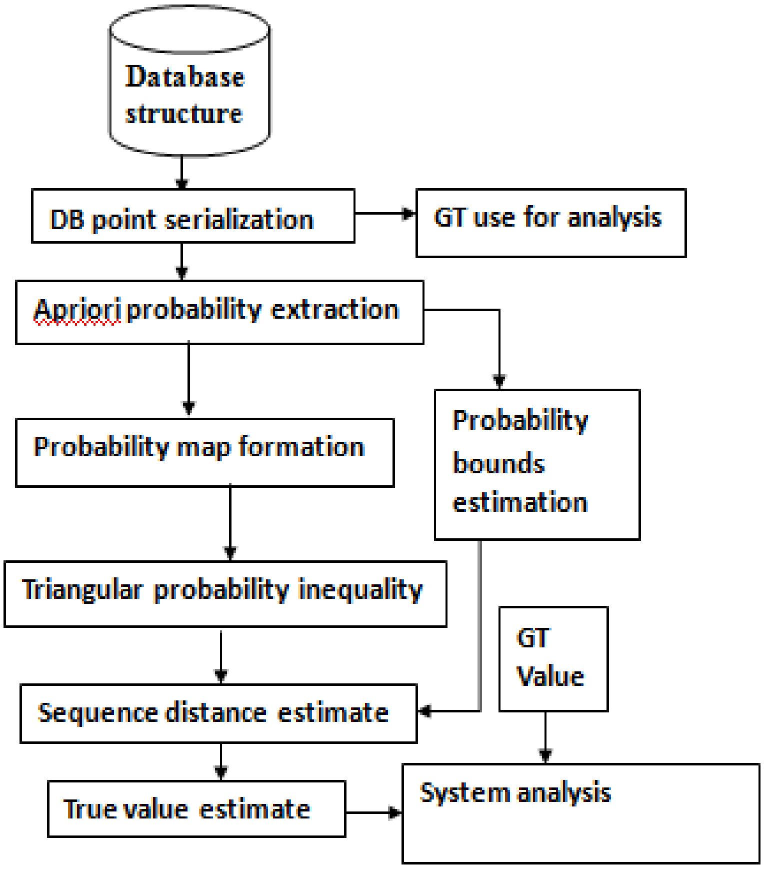
Database structure
DB point serialization
GT use for analysis
Apriori probability extraction
Probability map formation
Probability bounds estimation
Triangular probability inequality
GT Value
Sequence distance estimate
True value estimate
System analysis

Fig.3. System Design

We can apply various methods in proposed system to reduce the number of possible recovery version as follow.

- **Database structure:**

It is the data base which in which our experimental data set is store it may be incomplete.

- **DB point serialization:**

The serialization of data base is the process of translating data structures or object state into such a format that can be able to stored and reconstructed in the same storage. When the series of bits is in serializing format, it can be used to create a semantically identical clone of the original object.

- **Apriori probability extraction:**

In this step we use the apriori algorithm to estimate Dependency. Apriori is an algorithm in which frequent item set are extracted by using the set of the dependency rule from databases. It proceeds by identifying the frequent individual items in the database and include those item set in recovery version which are dependent on data object which we have to find.

- **Probability map formation:**

The probability map formation is the process in which the particular range of the probability is mapped so the data object which are able to complete the incomplete data.

- **True value estimation:**

Once the data is filter out from all the above method they are include in to the true value. True value are the values which are perfectly complete the data.

- **Probability bounds estimation:**

The data object which are extracted from the apriory estimation they are passed through the probability bound. There are the lower and upper bounds of the probability. In this only those object are filter which are fit in:

1) Data objects in database can be determined to be true results which are less than the distance threshold and greater than the probability threshold

2) The data objects in database can be determined to be dismissal results which are less than the distance threshold and less than the probability threshold

- **System analysis:**

Finally the object or data value included in the true result t are compare with the ground true values which are theoretically formed and analyze the database.

V. CONCLUSION

we focused the similarity query problem on dimension incomplete data and the technical challenge in the incomplete data to solve this problem we form various probability framework To solve this problem efficiently, we develop the lower and upper probability bounds and the probability triangle inequality that can be used to dramatically prune the search space . We propose the various methods to estimate the close dependant data object or value so it reduces the number of possible dada object or simply recovery version which are cheeked by system it increases the response time of our system than the existing one.

ACKNOWLEDGMENT

I am Amol Patil and would like to thank the publishers, researchers for making their resources material available. I am greatly thankful to Prof. Saba Siraj for their guidance. We also thank the college authorities Prof. Pratap Singh, PG coordinator and Principal Dr. R.S. Jahagirdar for providing the required infrastructure and support. Finally, we would like to extend a heartfelt gratitude to friends and family members.

REFERENCES

[1] C. Faloutsos, M. Ranganathan, and Y. Manolopoulos, “Fast Subsequence Matching in Time-Series Databases,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’94), pp. 419-429, 1994.

[2] Ankerst, B. Braunmller, H.-P. Kriegel, and T. Seidl, “Improving Adaptable Similarity Query Processing by Using Approximations,” Proc. 24th Int’l Conf. Very Large Data Bases (VLDB ’98),pp. 206-217, 1998.

[3] M R. Agrawal, C. Faloutsos, and A.N. Swami, “Efficient Similarity Search in Sequence Databases,” Proc. Fourth Int’l Conf. Foundations of Data Organization and Algorithms (FODO ’93), pp. 69-84, 1993.

[4] D. Gu and Y. Gao, "Incremental Gradient Descent Imputation Method for Missing Data in Learning Classifier Systems," Proc.Workshops Genetic and Evolutionary Computation (GECCO '05),pp. 72-73, 2005.

[5] R.K. Pearson, "The Problem of Disguised Missing Data," ACM SIGKDD Explorations Newsletter, vol. 8, pp. 83-92, 2006.

[6] 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, 2005.

[7] J. Pei, B. Jiang, X. Lin, and Y. Yuan, "Probabilistic Skylines on Uncertain Data," Proc. 33rd Int'l Conf. Very Large Databases (VLDB '07), pp. 15-26, 2007.

[8] 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.

[9] E. Keogh, "Exact Indexing of Dynamic Time Warping," Proc. 28th Int'l Conf. Very Large Data Bases (VLDB '02), pp. 406-417, 2002.

[10] G. Navarro, "A Guided Tour to Approximate String Matching," ACM Computing Surveys, vol. 33, pp. 31-88, 2001.

[11] R.A. Little and D.B. Rubin, Statistical Analysis with Missing Data, Wiley Series in Probability and Statistics, first ed., pp. 2-278. John Wiley & Sons, 1987.

[12] T. Mathew and K. Nordstrom, "Inequalities for the Probability Content of a Rotated Ellipse and Related Stochastic Domination Results," The Annals of Applied Probability, vol. 7, no. 4, pp. 1106-1117, 1997.