ISSN:0975-9646

# A Novel methodology for Searching Dimension Incomplete Database

Kalbhor Swati[#1], Gupta Shyam[#2]

[#]*Department of Computer Engineering*
*Siddhanth college of engineering*
*Sudumbre, Pune.*

***Abstract*— This manuscript deals with the similarity querying problems for cases where data loss exists. Limitations in traditional methodologies for querying incomplete data in database, data mining and information retrieval research has urged to shift into development of different new innovative models. This Investigation is done based on a model developed based on ARIMA constructional model to check the performance related to dimensional incomplete data. In actual problems, the data collected by the sensor network in noisy environment results in loss in data values and also dimensional information in some cases. An attempt has been made to verify the effect of time series model on recovering the missing dimension information. The complications encountered while checking out all the possible combinations of missing dimensions has been studied while evaluating the query and the data objects. Implication of the ARIMA model will result in reduction in total time required for searching and recovering the data. It is predicted that time series approach will perform faster that probability triangle model. To reduce the search time, evaluation is done without generating all the recovery versions. Basic concept that sensors give values which have strong resemblance in different times is used in this model. In the time series values, samples at different time will be evaluated to establish consistent ARIMA coefficients.***

***Keywords*—Dimension incomplete database, similarity search, ARIMA model.**

## I. Introduction

Extensive research efforts have been taken in similarity querying in the field of database, data mining, and information retrieval. If a query object is given, then efforts are been put to find similar objects. In many cases, it is observed that the data is incomplete or it is a missing value problem i.e., the data values on some dimensions are unknown or uncertain. For example, in sensor networks, the received data might be incomplete owing to the inaccuracy in sensors or when errors occur during the data transmission process. This study is based upon the assumption that missing data information is known. Whereas the same information may be unknown to the engineer in actual practical environment. For these cases, order of arrival of data values would be known without any information regarding its respective dimensions. When the dimensionality of the collected data is minor than its actual dimensionality, the dependency of dimensions on their related values is lost. We refer to such a problem as the dimension incomplete problem.

1) Data missing when dimension information is not clearly maintained. In this case consider the sensor networks. The database usually contains time series data objects, each of which is represented by a sequence of values (x1; x2; . . . ; xm). The dimension information associated with data values can be indirectly inferred from the data arrival order. This schema of data gathering and storing is very common in resource-constrained applications because clearly maintaining dimension information will cause additional costs. In this problem setting, missing a single data element will terminate the dimension information of the whole data object.

For example, the original data object is (3, 1, 2, 5). When data element 1 is lost, then dimension information for the rest of data elements becomes unclear. 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 lost, then both elements 3 and 2 may locate on three altered dimensions. In applications where dimension information is clearly maintained, the dimension pointer itself may be missing. This will also cause the dimension incomplete problem.

Instead of using mean and variance, we will use ARIMA based reconstruction. ARIMA is a time series construction model. We use the concept that sensors give values which have strong resemblance in different times. In the time series values, we sample at different times and find out which particular of ARIMA coefficients are consistent. Once the most stable ARIMA model is constructed, we use this model to refill the incomplete data. By this way, we hope that accuracy of search improves.

## II. Literature survey

W. Cheng, X. Jin, J. Sun, X. Lin, X. Zhang, and W. Wang [1] developed probabilistic framework to model the problem of similarity search on dimension incomplete data so that the users can find objects in the database that are similar to the query with probability. They developed both upper and lower bounds of probability that a data object will become similar to the query. This bounds enabled efficient filtering of irrelevant data objects without explicitly examining all missing dimension combinations. A probability triangle inequality is employed to further prune the search space and speed up the query process.

R. Fagin, R. Kumar, and D. Sivakumar [2] introduced rank aggregation as approach towards doing similarity search and classification. In this method, query and candidates are classified as points in multidimensional space. Each coordinate has been individually pointed out as voter. It ranks the points based on closeness to the corresponding coordinate of query. The winners are those points with the highest aggregated ranks which when are combined with dimensionality reduction yields a simple, database-friendly algorithm and gives a very good approximate answer to the nearest neighbour problem. This algorithm is observed to be efficient. Median rank aggregation is an efficient and useful form of rank aggregation.

E. Keogh [3] Dynamic Time Warping (DTW) is a much more robust distance measure for time series. DTW allowing similar shapes to match even if they are out of phase in the time axis. First, consideration in the case where the two sequences are of the equal length. This is not really a limitation because the user can always re-interpolate the query to any desired length. Secondly, only index sequences if we assume the warping path is constrained. Approach it degenerates to Euclidean indexing using PAA.

D. Gu and Y. Gao [4] Incremental gradient descent imputation(InGrImputation) Model creates an universal model for the variable with missing data based on the relationship between the variable and other known variables. InGrImputation model uses a relationship among variables to estimate the missing value and therefore improves the performance of Learning Classifier Systems(LCS).

E. Keogh and M. Pazzani [5] Dynamic time warping (DTW) has been suggested as a technique to robust distance calculations for time series data, however it is computationally expensive. DTW is distance measure for time sequence, allowing related shapes to match even if they are out of phase in the time alignment. Related shapes to match even if they are out of phase in time aligment for time sequence is nothing but DTW distance measure. Modification of DTW that exploits a higher level representation of time series data. This produces one to three orders of magnitude speed-up without compromising in accuracy.

B. Bollobas, G. Das, D. Gunopulos, and H. Mannila [6] Gave a pair of unidentical complex objects, dening (and determining) however their similarrity to each other is a nontrivial problem. In data mining applications, one needs to determine the similarity between two time series. Analyze a model of time-series similarity that allows outliers, different scaling functions, present deterministic and randomized algorithms for computing this notion of similarity. Nontrival tools and methods from computational geometry on which algorithms are based. Use the properties of well-separated geometric sets. The randomized algorithm for computing similarity between two time series has provably good performance.

## III. PROCESS EXECUTION

The process which will be followed to establish the methodology is explained in Fig. 1. Key points are:

### A. Data uploader

This module processes the datasets from their native format and uploads them to databases.

### B. ARIMA based substitution

This module will generate full dataset by filling the missing dimension elements by using ARIMA model generation and prediction of missing elements.

### C. Probabilistic query processor:

This module will find solution to the users queries by Probabilistic matching on the ARIMA generated results.

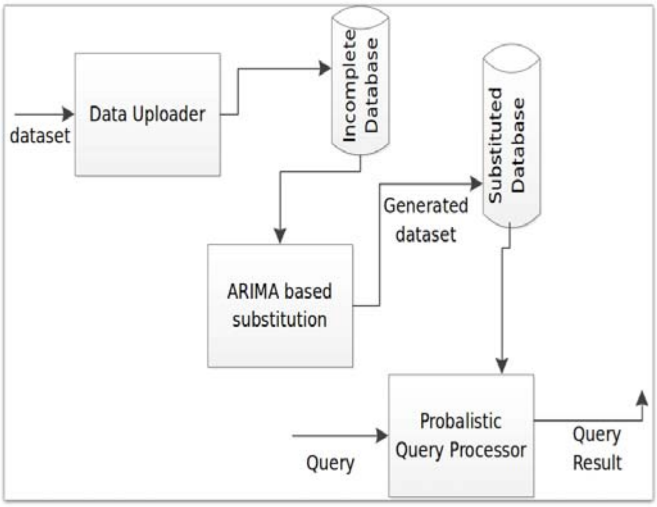

Fig. 1: Flow chart for process diagram.

## IV. CONCLUSION

This paper addresses the similarity query problem on dimension incomplete data. ARIMA model is proposed to solve this problem. The sensors give values which have strong resemblance in different times. In the time series values, sample at different times and find out which particular of ARIMA coefficients are consistent. Once the most stable ARIMA model is constructed, use this model to refill the incomplete data .With the use of these model accuracy of search improves.

## ACKNOWLEDGMENT

I would like to express my sincere thanks to my guide Prof. Shyam Gupta for his motivation and useful suggestions which truly helped me in improving the quality of this paper. I take this opportunity to express my thanks to my teacher, family and friends for their encouragement and support.

## REFERENCES

[1] W. Cheng, X. Jin, J. Sun, X. Lin, X. Zhang, and W. Wang., “Searching Dimension Incomplete Databases" IEEE Transactions on knowledge and data engineering, pp. 725-738, 2014.

[2] R. Fagin, R. Kumar, and D. Sivakumar, “Efficient Similarity Search and Classification via Rank Aggregation,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’03), pp. 301-312, 2003.

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

[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] E. Keogh and M. Pazzani, “Scaling up Dynamic Time Warping to Massive Data Sets,” Proc. Third European Conf. Principles of Data Mining and Knowledge Discovery (ECML/PKDD ’99), pp. 1-11, 1999.

[6] B. Bollobas, G. Das, D. Gunopulos, and H. Mannila, “Time-Series Similarity Problems and Well-Separated Geometric Sets,” Proc. 13th Ann.Symp. Computational Geometry (SCG ’97), pp. 454-456,