

A Novel Approach for Searching Dimension in Incomplete Databases

Pagare Namrata M.

Prof.Mrs.J.R.Mankar

Abstract — In multidimensional database, retrieving similar data is an important task because of its applications in different areas. But it creates a problem when database is incomplete or data dimensions are missing in applications where data is collected from sensor network and some data values are lost due to faulty sensors or errors during transmissions. In such cases even the dimension information or on what position data loss occurred is not known. Hence, Similarity query becomes a problem since it has to check all possible combinations which increase the space and time complexity leading to performance degradation. To deal with this problem a probabilistic scheme is developed where lower and upper probability bounds are computed to reduce the search space and a probability triangle inequality method is used for filtering and increasing the speed of query process. The aim of proposed system is to improve the efficiency of query process. Indexing approach is used to speed up the process and different distance functions are used for similarity measurements.

Key Words — Dimension incomplete database; imputation; probability triangle inequality; similarity search

I. INTRODUCTION

Now-a-days, applications of multidimensional data are found in various areas. So the objective is to fetch similar data when query is given but databases occur with data missing, like in census database some attributes can have null values, or answering one question causes remaining questions to be skipped. Querying incomplete data is a difficult task since data values on some dimensions are not known or unsure which is also referred as missing value problem.

The incompleteness in data occurs when the sensors fail to perform properly or some errors appears when the data transferred leading to loss of data, even empty values are allowed for some attributes in census database, certain attributes not being available at the time the record was stored in a survey database, also in medical database users are unwilling to specify because of privacy issues. So, Imputation, statistical and regression based procedures are some methods used to estimate the missing entries. But, there are situations where we may not know the position of data loss or which dimension is missing in database.

In application like sensor network the database usually contains time series data objects, each of which is represented by a sequence of values x_1, x_2, \dots, x_n . The dimension information (like a time stamp) associated with

data elements can be implicitly inferred from the order of data arrival. This schema of data collection and storage is very common in resource-constrained applications since explicitly maintaining dimension information will cause additional costs. Hence, even a loss of single data element will destroy the dimension information of the entire data object. For instance, as shown in Fig. 1, the original data object is (3, 1, 2, and 5). When data element 1 is lost, the dimension information for the rest of data elements becomes uncertain. For instance, 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 preserved but the dimension indicator itself may be lost which also leads to the dimension incomplete problem like in time series data there can be temporal uncertainty because of surmise time stamps. The absence of clock synchronization in distributed system may also cause dimension incompleteness in time series data.

In dimension incomplete database, suppose that the original data dimensionality is M . Consider a query object $Q = (q_1, q_2, \dots, q_m)$ and a dimension incomplete data object $O = (o_1, o_2, \dots, o_n)$ where $(n < M)$, a solution to calculate the distance between these two objects by examining all possible missing dimension combinations for the data object O . The traditional similarity measurements which are used for any of any similarity query task may not be applied directly. So for each combination, the values are imputed at the positions with missing data, and then calculate the similarity between Q and the imputed data object O based on certain distance function such as the l_p -norms distance. But there are $\binom{m}{n}$ possible dimension combinations to be examined which is undesirable and inefficient.

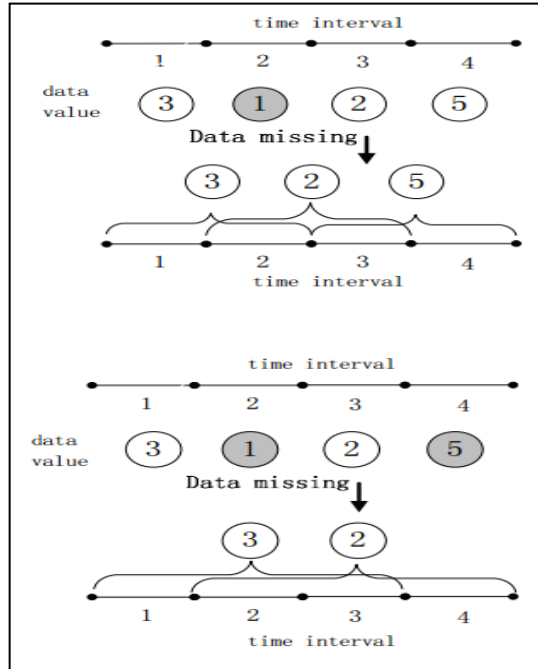


Fig.1. Dimension incomplete data due to dimension information not being explicitly maintained.

A probability framework is used where the user can give a distance threshold to specify the allowed distance between the query and dimension incomplete data objects, and a probability threshold to specify that the fetched data objects are similar to the query. An efficient method is proposed to find lower bounds and upper bounds of the probability that a data object satisfies the query. These bounds can be used to eliminate data objects that are judged as dismissals, and keep qualified ones in $O(n(m/n)^2)$ time. Furthermore, based on the proposed probability triangle inequality, an approach with time complexity $O(m)$ is introduced to further speed up the similarity query process.

The remainder of paper is organized as follows: Section 2 discuss the related work done and its shortcoming. An overview of the proposed scheme is given in section 3. Section 4 discusses the expected results of the system. Finally, we conclude the paper.

II. LITERATURE SURVEY

In this section, an overview of existing methods for dealing with missing values in incomplete databases is provided.

R. Agrawal, C. Faloutsos, and A.N. Swami [1] proposed an indexing method for time sequence for processing

similarity queries. R* trees method is used to index the sequence and it efficiently works to explain similarity queries. In this method where data value missing it will place null or -1 value. So that, it is easy to search out missing values.

Beng Chin Ooi, Cheng Hian Goh, Kian-Lee Tan [2] has proposed two different indexing schemes for high-dimensional databases that are incomplete for improving the efficiency of data retrieval. In this paper, it addresses the issues pertaining to the design of fast mechanisms that avoid the costly alternative of performing an exhaustive search. It represents two indexing schemes such as BR-Tree i.e. multi-dimensional index structure called the Bit string-augmented R-tree (BRtree) and MOSAIC index scheme i.e. Multiple one dimensional one attribute index called as MOSAIC.

B. Mirkin and I. Wasito [3] has given a nearest neighbor way and the global method for least-square data imputations is reviewed and extensions to them are proposed. Patterns of missing data are in form of rows and columns according to three different mechanisms that are denoted as Random missing, restricted random missing and Merged Database.

Daqian Gu and Yang Gao [4] proposed an incremental Gradient descent imputation model where missing values are estimated using relationship among variables. Learning Classifier Systems (LCS) is a form of self-adaptive and online learning systems. In LCS research first type of missing data is referred to as missing completely at random and second as missing at random and last type of missing data is attributed to not missing at random.

K. Pearson [5] defined the problem of disguised missing data where when missing data values are not specifically represented but are coded with values that can be misinterpreted as valid data. There are different ways of handling explicitly coded missing data like Deletion, Single Imputation, Multiple Imputation, Iterative procedure.

Guadalupe Canahuate, Michael Gibas, and Hakan Ferhatosmanoglu [6] proposed the indexing techniques for multidimensional data search when indexed attributes contain missing data. Bitmap indexing technique is applied and vector approximation (VA) files are modified appropriately to report for missing data and to execute the query according to the query semantics. The aim was to index each dimension independently and execute queries efficiently by performing operations on bit for bitmaps and VA-Files for pruning multiple dimensions.

E. K. Mohamed [7] deal with the problem of skyline queries in incomplete database. He proposed Iskyline

algorithm that handles the skyline queries in incomplete relational database by dividing the initial database into distinct nodes depending on the missing dimensions and then applying the conventional skyline method to retrieve the local skyline in every cluster.

Jian Pei Ming Hua Yufei Tao Xuemin Lin [8] designed the method for Query Answering on Uncertain and Probable Data. Soliman[9] designed U-Topk queries and U-ranks queries. A U-Top query returns a k-tuple sorted list which has the highest probability to be the top-k list in possible worlds. A U-ranks query finds the tuple of the highest probability at each ranking position. K.Yi[10] proposed efficient Query Answering algorithms to answer U-Topk queries and U-ranks queries. Their algorithm for U-ranks uses the Poisson binomial recurrence. Lian and Chen developed the spatial and probabilistic pruning techniques for U-kRanks queries.

Eamonn Keogh, Chotirat Ann Ratanamahatana [11] proposed a method for the exact indexing of dynamic time warping (DTW) where DTW is a distance measure for time series. So, researchers have introduced approximate indexing techniques and concentrated on speeding up sequential searches uses the GEMINI framework and a indexing technique called piecewise aggregate approximation.

III. IMPLEMENTATION DETAILS

A. Proposed System:

The overall Block diagram of the system is shown in the Fig.2. And Table I lists the symbols used in this paper. Step by step architecture is explained in subsequent section.

1. Indexing

In proposed system, initially indexing is done to maintain the index count for number of missing dimension in dimension incomplete database. It is sorted in ascending order of number missing dimension and then the query process is evaluated in sorted order which will improve the efficiency of the system.

2. Probability triangle Inequality filter

The probability triangle inequality filter (PTI) is used to evaluate the data objects. With the help of assistant data objects, some data objects are judged as true results or true dismissals and the remaining are given to next block.

$$\Pr[\delta_U(R,X) + \delta(Q,R) < r] > c \quad (1)$$

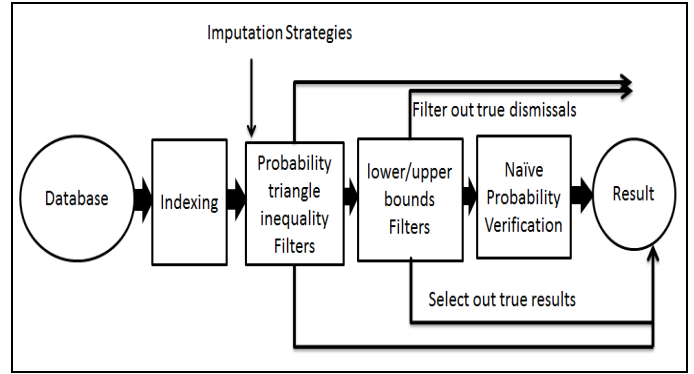


Fig 2: Proposed System Architectural Diagram

TABLE I . SUMMARY OF SYMSBOLS AND THEIR MEANINGS

Notation	Description
X	Complete multidimensional data
R	Assistant data objects
Q	Query
Pr	Probability to occur
δ	Distance function
R	Distance threshold
C	Probability threshold

Using equation (1) some data objects are determined as true results using PTI method.

$$\Pr[\delta_L(R,X) - \delta(Q,R) < r] \leq c \quad (2)$$

Similarly, some data objects are determined as true dismissals using equation (2).

3. Bounds of Probability

The lower and upper bounds are used to determine remaining data object where some belong to true results or true dismissals and some given to next block. The lower and upper bounds are computed for observed part and missing part to create recovery versions.

- i. Algorithm to calculate Lower and upper bound (δ_{Lob} and δ_{Umi})

Input: Query Q, |Q|=m and dimension incomplete data object

Output: δ_{Lob} and δ_{Umi}

- 1: Create two m x (2n + 1) matrices T and assistant array S

```

2: Initialize (i,j) element of T to  $(Q_i - X_j)^2$ 
3: Initialize S to (0,0)
4: for j=1 to 2n+1 do
5:   If j=1 then
6:     for i=1 to m-n do
7:        $S[i,j] \leftarrow ([i-1], 1)$ 
8:       if  $i > 1$  then
9:          $T[i][j] \leftarrow T[i,j] + T[i-1][j]$ 
10:      end if
11:    end for
12:   else if  $j > 2$  and  $j \bmod 2 = 1$  then
13:     for  $i=(j+1)/2+1$  to  $(j+1)/2+m-n-1$  do
14:        $p \leftarrow \operatorname{argmin}_{1 \leq k \leq (j+1)/2} T[i-k][j-2(k-1)]$ 
15:        $T[i][j] \leftarrow T[i][j] + T[i-p][j-2(p-1)]$ 
16:        $S[i][j] \leftarrow (i-p, j-2(p-1))$ 
17:     end for
18:   else if  $j > 2$  and  $j \bmod 2 \neq 0$  then
19:     for  $i=j/2$  to  $j/2+m-n$  do
20:        $T[i][j] \leftarrow T[i][j] + \min_{(j-2)/2 \leq k \leq i-1} T[k][j-2]$ 
21:     end for
22:   end if
23: end for
24: return  $(\min_{n \leq k \leq m} T[k][2n])^{1/2}$ 
    
```

After calculating the bounds of probability the recovery versions are created.

$$\Pr[\delta_U(Q,X) < r] > c \quad (3)$$

The data objects are determined as true results using equation (3)

$$\Pr[\delta_L(Q,X) < r] \leq c \quad (4)$$

The data objects are determined as true dismissals using equation (4)

4. Naïve Probability Verification

The data objects which cannot be determined using above steps are judged using naïve verification steps where a pos and neg strategy is used where the pos strategy considers the remaining data objects as query answers and neg strategy eliminates the data objects.

5. Ranking

The obtained similar results are ranked in ascending order of probability to select top similar results.

IV. RESULTS AND DISCUSSION

All Experimentation is performed using Intel core i5 processor and 4 GB RAM. The operating system is windows 8(64 bit) with visual studio 10.

A. Dataset:

The data sets are used in the experiments are:

i. Standard and Poor 500 index historical stock data .It consists of stock prices of 541 companies. The opening stock prices are used as data objects. Initially the dataset is preprocessed by segmenting the data, resulting in $541 * 8 = 4328$ data objects with 30 dimensions

ii. Color histograms of Corel image features (denoted by IMAGE) [13]. It contains 32-dimensional image features extracted from 68,040 images of the Corel image collection.

For both data sets, the original data objects are complete. We construct the dimension incomplete data sets by randomly removing some dimensions of each complete data object.

B. Performance Measure:

The performances are measured by using precision and recall on the effectiveness of probabilistic similarity query on dimension incomplete data, probability threshold and different pruners. The query results are used on complete data as ground truth.

$$\text{Precision} = \frac{|\text{true positive}|}{|S_{\text{results}}|}$$

$$\text{Recall} = \frac{|\text{true positive}|}{|S_{\text{true}}|}$$

Where, true positive means retrieved data objects whose complete forms are in ground truth, S_{true} represents ground truth and S_{results} is retrieved dimension incomplete data objects.

C. Results:

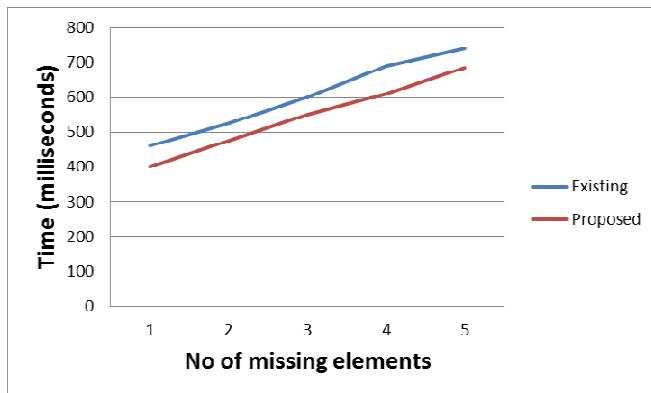
To test the quality of query results color histogram dataset and S & P500 dataset are given as input. The number of missing data elements are controlled by missing ratio which is varied. The distance threshold(r) and confidence threshold(c) are chosen to obtain good query results. Fig .3 shows the results where proposed system achieves better time efficiency as compared to existing system. Table II & III shows that the S & P500 dataset obtains good query results as compared to IMAGE dataset. As the missing ratio increases the precision and recall decreases but still the method achieves good query results.

TABLE II. QUERY RESULTS ON S & P500 DATASET

r	c	Missing Ratio					
		5%		10%		15%	
		Precision	Recall	Precision	Recall	Precision	Recall
40	0.2	0.963033	0.999074	0.960920	0.996666	0.952108	0.999107
	0.3	0.990488	1	0.991953	0.997222	0.969213	0.995793
	0.4	0.987302	0.996756	0.986425	0.999122	0.983892	0.998528

TABLE III. QUERY RESULTS ON IMAGE DATASET

r	c	Missing Ratio					
		5%		10%		15%	
		Precision	Recall	Precision	Recall	Precision	Recall
0.4	0.2	0.959748	0.962011	0.773887	0.632111	0.769137	0.648959
	0.3	0.911507	0.868957	0.837384	0.599492	0.777432	0.596052
	0.4	0.915373	0.890726	0.930467	0.584410	0.798827	0.480565


Fig.2.Time required for execution

SUMMARY AND CONCLUSION

Similarity query on dimension incomplete data is a problem for traditional querying techniques. A probability framework is proposed to model this problem. The proposed system improves efficiency using indexing approach where different distance functions are explored for similarity measurements. The lower and upper probability bounds are used to prune

the search and the probability triangle inequality which increases the speed of query process. The method provides the solution for searching in different incomplete datasets with good query results.

ACKNOWLEDGMENT


I would like to express my special thanks to all those people who have helped me to complete this work. I am very grateful to my guide, Mrs.J.R.Mankar,Computer Engineering,K.K.W.I.E.E.R., Nashik for her guidance, encouragement and the interest shown in this project She has continuously helped and encouraged me in my work.

REFERENCES

- [1] R. Agarwal, C. Faloutsos , and A.N. Swami, Efficient Similarity Search in Sequence Databases, Proc. Fourth Intl Conf. Foundations of Data Organization and Algorithms (FODO 93), pp. 69-84, 1993.
- [2] Beng Chin Ooi , Cheng Hian Goh , Kian-Lee Tan, Fast High Dimensional Data Search In Incomplete Databases, Proc. ACM SIGMOD Intl Conf. Management of Data (SIGMOD 94), 1998.
- [3] I. Wasito and B. Mirkin, Nearest Neighbour Approach in the Least-Squares Data Imputation Algorithms, Information Sciences: An Intl J., vol. 169, pp. 1-25, 2005

- [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] G. Canahuate, M. Gibas, and H. Ferhatosmanoglu, Indexing Incomplete Database, Proc. 10th Intl Conf. Advances in Database Technology (EDBT 06), pp. 884-901, 2006.
- [7] Ali A. Alwan, Hamidah Ibrahim, Nur Izura Udzir and Fatimah Sidi. Skyline queries over incomplete multidimensional database. Proceedings of the 3rd International Conference on Computing and Informatics, ICOCI2011, 8-9 June, 2011 Bandung, Indonesia
- [8] J. Pei, M. Hua, Y. Tao, and X. Lin, Query Answering Techniques on Uncertain and Probabilistic Data: Tutorial Summary, Proc. ACM SIGMOD Intl Conf. Management of Data (SIGMOD 08), pp. 1357-1364, 2008
- [9] M. A. Soliman, I. F. Ilyas, and K. C.-C. Chang. Top-k query processing in uncertain databases. In Proceedings of the 23rd International Conference on Data Engineering (ICDE'07), Istanbul, Turkey, April2007. IEEE.
- [10] K. Yi, F. Li, D. Srivastava, and G. Kollios. Efficient processing of top-k queries in uncertain databases. In Proc. 2008 International Conference on Data Engineering (ICDE'08), April 2008.
- [11] E. Keogh, Exact Indexing of Dynamic Time Warping, Proc. 28th Intl Conf. Very Large Data Bases (VLDB 02), pp. 406-417.

AUTHOR'S PROFILE

	<p>Namrata Pagare received the B.E. degree in Information Technology Engineering from Sandip Institute of Technology and Research Center, Nashik, Savitribai Phule Pune University in 2012. Now pursuing M.E. from K. K. Wagh Institute of Engineering Education & Research, Nashik, India.</p>
	<p>Prof. Jyoti Mankar, Assistant Professor, Department of Computer Engineering, K. K. Wagh Institute of Engineering Education & Research, Nashik, India.</p>