ISSN:0975-9646

Survey on Various Methods and Techniques for Searching Dimension in Incomplete Database

Ms. Yogita M. Kapse
G.H. Raisoni Institute Of Engineering And Technology for womens

Ms. Antara Bhattacharya
Assistant professor
G.H. Raisoni Institute Of Engineering And Technology for Womens

Abstract: Now a days, dimension incomplete problem is fundamental research problem in multidimensional database. Information regarding the missing dimension posses great computational challenges. In multidimensional database similarity query problem occur with numerous application in database area such as, data mining, information retrieval etc. Due to various practical issues like remote data accessing represent the dimension incomplete problem. This paper presents a framework for searching the dimension in incomplete database. The existing work proposed that, on certain dimension data values are unknown due to similarity query and query incomplete data. In this work we proposed to investigate the problem of information regarding incomplete dimension. In this paper we studied various methods and techniques such as, missing data recovery method, clustering algorithm as well as indexing method. This paper presents an approach for Indexing scheme like BR-Tree, MOSAIC Tree , R* Tree which works according to databases. Generating AFD based techniques define the approximation based approach. MDT's that is missing dimension technique proposed four techniques . Each method specifies their identical works according to field. So that, in proposed work we are using various methods and techniques. Probabilistic method is used to find no of possible sub queries using parameter. The probabilistic technique and scheme can be applied to whole as well as subsequence query.

Keywords:- Dimension Incomplete, similarity query, index Structure, subsequence query.

I. INTRODUCTION

Missing Dimension Information posses great computational challenges. Incompleteness of data is a common problem in many databases including web heterogonous databases, multi-relational databases, spatial and temporal databases and data integration. The incompleteness of data introduces challenges in processing queries, as providing accurate results that best meet the query conditions over incomplete database is not a trivial task. In the existing work dimension incomplete problem is studied which usually refers to the missing value problem that means, data values on certain dimension are unknown or uncertain. The existing work posses the common assumption regarding each dimension, whether it's data values is missing or not become known. In real life application, if data collected from noisy environment not only data values miss but also dimension information missing. So that, we have to know the arrival order of data values without knowing which dimensions the values belongs to. We have listed some of the causes that posses dimension incompleteness which are as follow.

A. Incomplete data entry: Users may intentionally or accidentally miss some values in one or more attributes (Dimensions) when entering data into the database.

B. Data type Missing: At the time of data entry if certain data type is not missing so that, it represents the problem of dimension Incomplete.

C. Low bandwidth or weak network: In many real life applications, such as data collected by a sensor network in a noise environment, not only the data values but also the dimension information may be missing.

When the dimensionality of the collected data is lower than its actual dimensionality, the correspondence relationship between dimensions and their associated values is lost. In this review there are various methods and techniques are studied. we refer some of these method from these references such as, high-dimensional databases, skyline missing values, Analysing large data sets using imputation method, similarity search in time sequence database, various scheme for indexing.

II. RELATED WORK

R. Agrawal , C. Faloutsos , and A.N. Swami [1] contributed the method for indexing in "Efficient Similarity Search in Sequence Databases" . This paper proposed an indexing method for time sequence for processing on similarity queries. R* trees method to index the sequence and efficiently work on answer similarity queries. similarity queries can be classified into two categories that are, whole sequence matching and subsequence matching .In whole sequence matching which represents two query . In which the first i.e Range query evaluate those sequence that are similar within distance 's' From given query sequence. Second is, All pair query which evaluate the pair of sequence which are within 't' of each other given a 'x' sequences. In subsequence matching it will consider large no of sequence . This paper present vital contribution on R* tree method R* Tree method applied for indexing. This method efficiently work for indexing. In this method where data value or dimension information missing it will place null or -1 value. so that, it will easy to search out missing dimension.

Beng Chin Ooi , Cheng Hian Goh , Kian-Lee Tan [2] has illustrated indexing scheme in "Fast high dimensional data search in incomplete database" . This paper propose two indexing schemes which are used for improving the

efficiency of data retrieval in high-dimensional databases that are incomplete. In this paper, we address the issues pertaining to the design of fast mechanisms that avoid the costly alternative of performing an exhaustive search. The sequence of the query becomes smaller. Subsequence can be search out from in large sequence and that are the best matches in query sequence. It represents two indexing scheme such as BR-Tree and MOSAIC index scheme. This first BR-Tree Scheme i.e multi-dimensional index structure called the Bit string-augmented R-tree (BR-tree).As we know in incomplete database missing information will replace as '?'. But when certain scheme applied in contribution of indexing, it will represent null value in place of missing data. Simultaneously, collected data entered at a time in a database through this scheme .In this proposed scheme it introduced the novel mapping function which randomly scattered in 'N' dimensional space that (ai…. an) be the search key which corresponding to tuple 'tp' and bit string is bi…. bm as follows:

$$bi = \begin{cases} 1 & \text{if } ai \text{ is known} \\ 0 & \text{otherwise} \end{cases} \quad \text{----------(2.1)}$$

The second scheme i.e Multiple one dimensional one attribute index called as MOSAIC .In this section index built on each attribute. The search keys may contain missing attribute values in that case these schemes are novel. Whereas, the second comprises a family of multiple one-dimensional one-attribute (MOSAIC) indexes. In this paper, we address the issues of pertaining to the design of fast mechanisms **.** It will create each data set for each attribute. so that storage cost will increase but data integrity will maintained.

Amgun Myrtveit, Erik Stensrud, Member, IEEE, and Ulf H.Olsson [3] have illustrated four missing data technique in "Analyzing Data Sets with Missing Data An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods".. This paper, present four missing data techniques and comparision of mdt's techniques will contribute that Ld will give data set is to small that generate meaningful prediction model. It will indicate four missing data technique (MDTs). A first technique i.e Listwise deletion (LD) which define missing data technique sequential process perform. In this technique according to list deletion will perform. Secondly the Mean imputation (MI) technique. This method contributes the process of imputation in which no of possible combinations find out. On the basis of that mean value calculated and perform mean imputation method. Third MDT's technique i.e Smilar response pattern Imputation (SRPI) in this pattern of imputation sequence will find out in large sequence. Pattern will represent in the form of rows and column in database. If no of sequences will match according to query it called as similar response pattern imputation. Finally fourth technique is Full information maximum like hood (FIML)This missing data technique defines whole subsequence matching technique. It evaluate possible no of sequences on the basis of certain parameter such as, permutation and combination.

I. Waist and B. Mirkin [4] has been given a nearest neighbour approach in "Nearest Neighbour Approach in the Least-Squares Data Imputation Algorithms". This paper contribute the "global" method for least-square data imputation are reviewed and extension to them are proposed based on the nearest neighbors (NN) approach. Pattern of missing data are define in terms of rows and columns according to three different mechanisms that are denoted as Random missing, Restricted random missing, Merged Database. The first mechanism Random missing specify approach randomly data element missing, so that data uncertainty will increase. So that it is difficult to find out the no of possible neighboring places. It work on approximation basis model. The second mechanism i.e Restricted random missing approach no of data element may be missing in given sequence. So that nearest neighboring approach will work according by considering neighbor place of other data element. In this arrival order of data element can be known. In third mechanism , Merged Database give an approach incomplete and complete database become merged. If database will not merged properly it responsible for missing information. It will work on the basis of Prediction model according to arrival order of data in database.

Ali A. Alwan, Hamidah Ibrahim, Nur Izura Udzir, Fatimah Sidi [5] have given an approach for skyline missing values in this paper i.e "Estimating missing values of skyline in incomplete database". This paper, given approach for Approximate Functional Dependencies (AFDs) applied to generate, that captured the relationships between the dimensions for that utilizes the concept of mining attribute correlations. In addition to , identifying the strength of probability correlations for estimating missing values. Then, the skylines with estimated values are ranked. It will ensure that estimated value become evaluated on the basis of Precision and Recall. It contributes various methods and techniques are as follows :

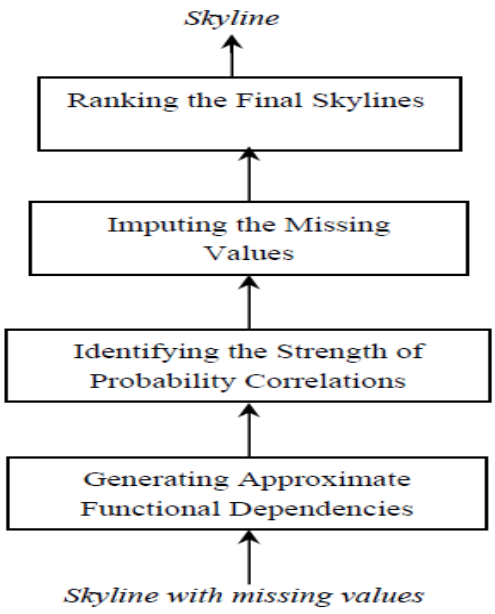

Fig.2.1. Flow work of Estimating skyline missing value

In first phase, Generating Approximate Functional Dependencies in this method, missing value estimated on the basis of approximation by capturing the relation between dimension. It represents the relation by arrow. for example if there are no of rooms related to rent (no of room rent of room). In second phase i.e. Identifying the Strength of Probability Correlations .It specify the strength of correlations between two dimensions is identified. It has evaluated the strength of probability correlations between the dimensions. In third phase, Imputing the Missing Values it define to impute the missing values of the dimensions in the skylines with the estimated values. In this by referring to the dimensions it has simply achieved. Dimension which have missing values it has replaced them with the estimated values. In this process there might be many estimated values that need to be considered. In fourth phase i,e Ranking the Final Skylines this section represent the last phase of ranking in which, skylines with the estimated values that have the highest confidence value of AFD and strength of probability correlations are place at the top of the skyline set.

Cheng, Xiaoming Jin, Jian-Tao Sun, Xuemin Lin, Xiang Zhang and Wei Wang [6] has been given an approach for searching Dimension incomplete database. It is used to sour a problem of similarity query. In this paper probabilistic framework and technique is applied to whole as well as subsequence query.When the dimensionality of the collected data is lower than its actual dimensionality, the correspondence relationship between dimensions and their associated values become lost. We refer to such a problem as the dimension incomplete problem. The first is Dimension information is not explicitly maintained and second is Time series data with temporal uncertainty Due to imprecise time stamps.

Suppose that, the original data dimensionality is 'D' Given a query object 'R' is (r1,r2,r3.. rx) and a dimension Incomplete data object i (i1, i2, i3…iy) (y < x) , a naïve Solution to calculate the distance between these two Objects. However, this approach is intractable in practice; since there is m (x/y) possible dimension combinations need to be examined. Efficient algorithms are highly desirable. This paper deal with the problem regarding similarity query on dimension incomplete data within a probabilistic framework. Using the framework, a user can identify two thresholds. There are two threshold consider that are the query object 'R' and the data object 'O'. So that, various method and techniques are applied to overcome this problem. Summarize process as follows:

1. To the best of our knowledge, this is the first work to denote the similarity query on dimension incomplete problem.
2. We develop efficient algorithms to specify the challenges in querying dimension incomplete data.
3. On dimension incomplete data , this method can be applied to both whole sequence matching as well as subsequence matching problem.
4. In this provide theoretical analysis of the relationship Between the probability threshold and the quality Query results.

Filter with Probability triangle inequality .The probability triangle inequality is first phase which applied to evaluate the data objects. In this phase , some data objects are verify as proper (true) results and algorithm work for filtering true result . At this phase result will show. The second phase i.e Filter with confidence lower and upper bounds, in this phase the remaining data objects filter out , from which some are determined as true results and some as dismissal. This phase also shows result. Third phase represents the Naive Probability verification. In which only those data objects can be filter out that cannot be determined in the former two steps are evaluated by the naive method. Small portion will filter out regarding data object in this phase. So that this phase will be considered as optional and finally result will have shown. Algorithm used to applied for Subsequence matching on dimension incomplete data.

PSM – DID δ,Ψ (Db,Qr,dr.cr) ---------- (2)

Where as, Db indicates database,
Qr for Query object
dr for distance threshold
Cr for threshold response

On the basis of comparative study approach BR-tree increase storage cost but data integrity will maintain. R* tree method which replace incomplete dimension by null or -1 value. AFD technique which specify the approximation relationship between dimension. It only work on the basis of prediction model . MDT's technique gives an approach regarding missing dimension technique such as, listwise deletion, similar response pattern etc which work on prediction model. So that in this way comparative analysis has done.

III. PROBLEM DEFINATION

According to various approach given and we provide the comparative analysis according various methods and technique for searching dimension incomplete database. So that certain technique are used to applied in proposed approach. So that, in proposed work we are using various methods and techniques. In this clustering technique perform on database by using 'CLINCH' i.e. clustering in incomplete high dimensional algorithm . Index structure which prune the search space and speed up the query process. For Indexing methods and scheme are applied i.e. BR-Tree, MOSAIC and R+ Tree method. Probabilistic method used to find no of possible sub queries using parameter. The probabilistic technique and scheme can be applied to whole as well as subsequence query.

IV. CONCLUSION

In this review various references considered for survey. So that we studied included methods and technique in the contribution of searching dimension incomplete database. In this each method and technique specifically works according to their fields. Efficiently search time sequence in database define the indexing scheme i.e R* tree method which place null value at the place of missing dimension information. In fast High dimensional in incomplete

database given approach of indexing for high dimensional database such as BR-Tree and MOSAIC. In estimating skyline missing values contributes the technique of AFD which work on approximation. On the basis of comparative analysis Proposed approach is better to overcome similarity search problem and provide solution to dimension incomplete problem. So that, certain indexing scheme will considered for implementing index structure which will represent the missing dimension information as well as maintain data integrity .

REFERENCES

1. R. Agawam , 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.
2. Beng Chin Ooi , Cheng Hian Goh , Kian-Lee Tan,"Fast High Dimensional Data Search In Incomplete Databases",Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD 94),1998.
3. Ingunn Myrtveit, Erik Stensrud, Member, IEEE, and Ulf H. Olsson "Analyzing Data Sets with Missing Data An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods" IEEE Transaction On Software Engineering, Vol. 27, No. 11, Nov 2001.
4. 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, 2005
5. Ali A. Alwan, Hamidah Ibrahim, Nur Izura Udzir, Fatimah Sidi,"Estimating Missing Values Of Skylines In Incomplete Database" Proc. 33rd Int'l Conf. Very Large Databases (VLDB '07), pp. 15-26, 2007.
6. Wei Cheng, Xiaoming Jin, Jian-Tao Sun, Xuemin Lin , Xiang Zhang, and Wei Wang, Member, IEEE, Searching Dimension Incomplete Databases ,vol.26, No.3, March 2014.