Implementing Hybrid Index Method for Searching Dimension in Incomplete Database

Ms. Yogita M. Kapse
Computer Science And Engineering department
G.H. Raisoni Institute Of Engineering And Technology for womens
Nagpure, India
yogitakapse9@gmail.com

Ms. Antara Bhattacharya
Assistant professor
Computer Science And Engineering department
G.H. Raisoni Institute Of Engineering And Technology for womens
Nagpure, Indi
antarabhattacharya@raisoni.net

Abstract:— Incompleteness of dimension information is a common problem in many databases including web heterogonous databases, multi-relational databases, spatial and temporal databases and data integration. In recent times, querying incomplete data has represented extensive attention that poses new challenges to traditional querying techniques. The incompleteness of data introduces challenges in processing queries . More than few techniques have been proposed to process queries in incomplete database. Some of these techniques retrieve the query results based on the existing values . As providing accurate results that best meet the query conditions over incomplete database. To retrieve data from incomplete database is not a trivial task. Dimension incomplete problem causes due to collection of data from noisy network environment . The existing work addresses the problem where data values are uncertain and unknown on dimension incomplete database. Several time techniques are undesirable in many cases where as the dimensions with missing values might be the important dimensions of the user's query. Besides, the output is incomplete and might not satisfy the user preferences. In this work, we propose to investigate the problem of similarity search on dimension incomplete data. This Several techniques have been proposed to process queries in dimension incomplete database . A proposed 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. Mainly focus on index structure. Indexing schemes for improving the efficiency of data retrieval in high-dimensional databases that are incomplete. Which need to be examined when evaluating the similarity between the query and the data objects. The proposed work represent clustering , indexing, searching missing ratio . Each method try to model this dimension incomplete problem . Firstly , clustering is performing which form group of certain attributes using clustering based 'cihd' algorithm . After Index structure is also employed to further prune the search space and speed up the query process . Indexing scheme like BR-Tree, MOSAIC Tree , R* Tree which works on specific dataset . . This combination of three indexing scheme collectively called as hybrid index method. After this missing ratio will search out by filtering irrelevant data object. This process through the certain parameter i.e. precision and recall . This paper present proposed methodology for mitigating problem of searching dimension in incomplete database.

Keywords- *Dimension Incomplete, similarity query, hybrid index, query processing, index Structure .*

*****

I. Introduction

Missing Information regarding Dimension posses great computational challenges [2]. Incompleteness of data is a common problem in many databases including data-mining, information retrieval etc. In many database applications there are many reasons that lead into missing values which make database incomplete. Certain time when dimension information is missing we may not know whether it's data values is missing or not become dimension become un-known . Consider an example of real –life Application, not only data values is missing but also dimension information missing . when data collected from sensor network attribute value become missing. So that, it is not required which dimensions the values belongs to but we have to know the arrival order of data values . We have listed some of these problem 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. Inaccurate data from heterogeneous data sources: In many real life applications when data collected from wireless sensor network or in noisy environment, not only the data values but also the dimension information may be missing. That time bandwidth may be low.

C. Data type Missing: If certain data type is not properly mentioned at the time of data entry it responsible for dimension Incompleteness .

When we search regarding dimensionality of data in that case, it should verify the collected data is lower than its actual dimensionality, the correspondence relationship between dimensions and their associated values is lost . In this paper there are various methods and techniques are studied and also applied for implementation for searching dimension in incomplete database . we refer some of these method like clustering and indexing . Specific parameter are used for finding missing ratio. So that , here precision and recall these two parameter are included . CIHD algorithm is nothing but clustering incomplete high-dimensional database. Combination of indexing scheme implemented called as Hybrid indexing . which used to model BR-tree, MOSAIC and finally R+ Tree methods [2]. Implementation of Hybrid indexing is our contribution . . Hybrid indexing scheme are used to provide three scheme i.e. identical works individualy. In this way , It will prune the search space and increase the speed of user query .

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 . 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. Marking [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. 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 [8] 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. According to various approach given and we provide the comparative analysis according various methods and retrieval in multi-dimensional databases that are incomplete. 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. 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.

III . Proposed Methodology

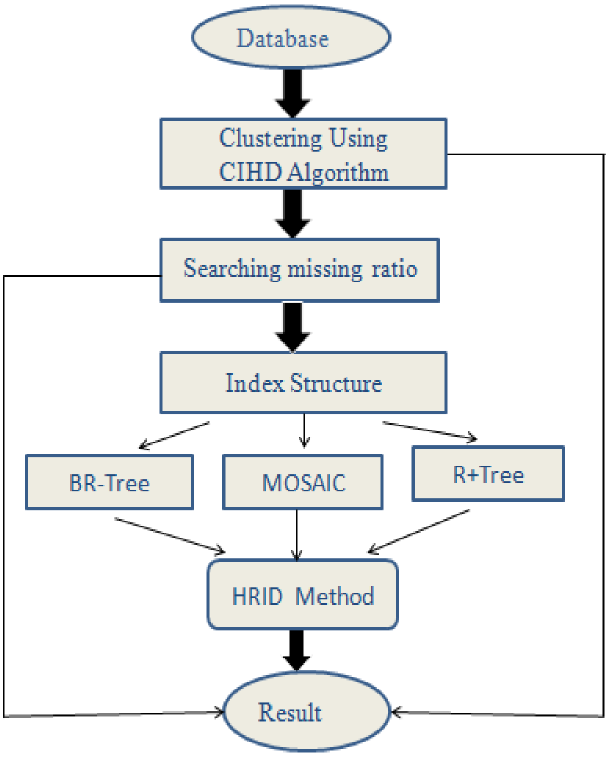

Fig.3.1.a Flow work of searching dimension in incomplete database using hybrid index method.

In this paper, we propose to investigate the problem of similarity search 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 nearly similar to the query . At the time of evaluating the similarity between the query and the data objects there are need to search out all possible combinations of missing dimensions [8]. The proposed framework will apply on both whole and subsequence queries. In this proposed Methodology introduce Index structure for improving the speed of query processing . Initially , clustering process are applied on database by finding missing dimension . After that Indexing will be applied . Due to indexing data will search in proper order and save time.

3.1. Clustering :

Clustering is nothing but a common technique in data mining to search out hidden patterns from massive datasets. With the development of privacy-maintaining data mining application, more usefull data regarding clustering incomplete high-dimensional data . CIHD is nothing but clustering Incomplete High Dimensional database.

ISSN: 2321-8169
106 – 111

The steps of algorithm CIHD are as follow :

CIHD Algorithm :

```
Input:    Dataset $Dt$, support threshold $minatr$;
                    Output:   A set of cluster's IDs;
Method:  CIHD($Dt$, $minatr$)
   1 :    $D \leftarrow$ determine and sort the order of $D$,dimensions;
     2 :    $UList \leftarrow$ FullDim($D$,$minatr$) //recognition on full
                                    dimensions;
  3 :    IncompleteDim($D$,$minatr$,$UList$, $DimID$) //recognition
                         On incomplete dimensions;
```

3. 2. Hybrid Indexing :

Indexing is the representation of summarized form of data. To Prune the search space and speed up the query process using Index structure. Hybrid Index scheme is used to applied for indexing, which are as follows :

1. BR-Tree Scheme (Bit String Augmented multidimensional index structure)
2. MOSAIC (Multiple one dimensional one attribute)
3. R+ Tree method

3. 2.1. BR-Tree Scheme :

BR-Tree is nothing but the Bit string augumented multidimensional index structure[2]. It provided identical id's to each entry in high-dimensional database. When simultaneously data enter in database there may possibility to miss out data . It may be responsible for missing dimension information. But due to BR-tree id's are provided to each tuple entry.

Steps are as follow :

BR-Tree scheme :

1. Start
2. Int j , Sr[ ]
3. Sr[ ] = j
4. j++
5. End

3. 2.2. MOSAIC :

MOSAIC is nothing but Multiple one dimensional one attribute . This scheme are used to provided specification for each attribute[2]. MOSAIC scheme introduced the process of forming single data set for single attribute. Due to this process data integrity will maintained. So thar user can seach data easily according to query.

3. 2.3. R+ Tree :

R+ tree method called as conventional multidimensional index[2]. Subset form to operate data by 'n' number of ways. In scheme In this method incomplete database value may replace by 0 or 1. So that it is easy to verify the value of is available or not.

Steps are as follow :

R+ Tree method

1. select dataset
2. select attribute
3. comease (0, attribute)
4. comease (1, attribute)
5. End

3. 3. Searching missing Ratio :

In searching and segmentation missing ratio will search out from selected dataset ..Dataset may be image or text dataset. According to that how much percent of dimensions are unknown or miss , it will verify through this method.. we use two standard measures,precision and recall . This Parameter such as , Precision and Recall search out missing ratio versus threshold value .

3. 3. 1. Precision :

Precision represent that at what percent of relevant data collect from retrieved data.

Where,

$$\textbf{Precision} = |\textbf{Tp}| / |\textbf{Sresult}|$$

Tp -stands for true positives,

.**Strue** - stands for the "ground truth" results,

3. 3. 2. Recall :

Recall represent that at what percent of retrieved data collect from relavent field.

Where,

$$\textbf{Recall} = |\textbf{Tp}| / |\textbf{Sresult}|$$

Tp -stands for true positives,

Strue - stands for the "ground truth" results .

IV . Experimental Setup

Following are the screen shots of clustering and Indexing process :

Fig.4.a selecting dataset

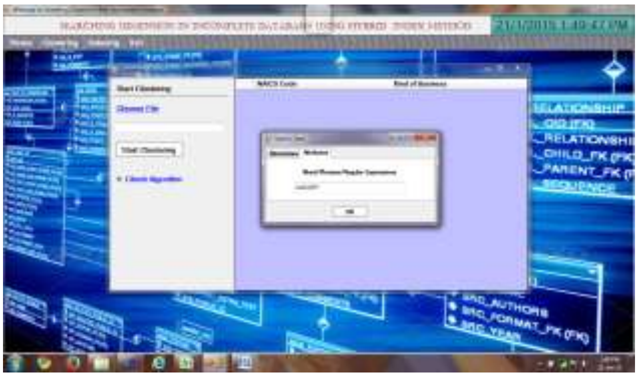

Fig.4.b selecting attribute

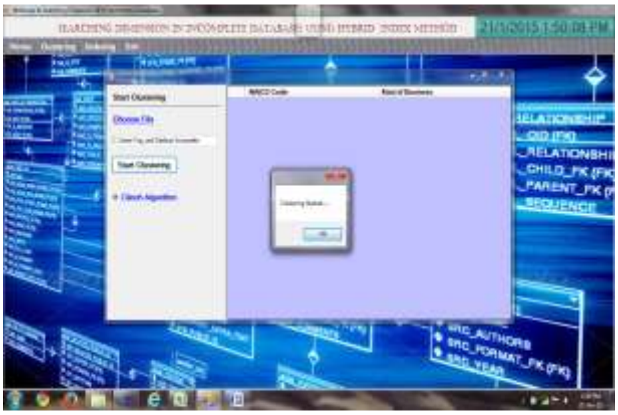

Fig.4.c clustering process completed

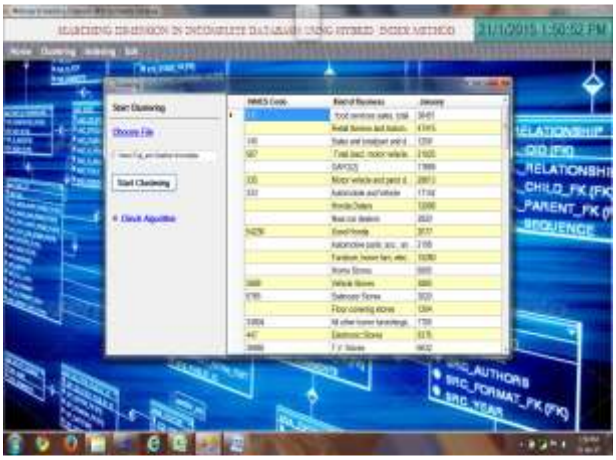

Fig.4.d selecting dataset

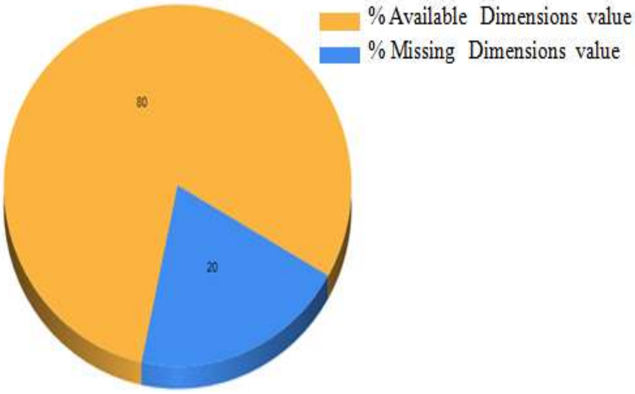

Fig.4.e Missing ratio

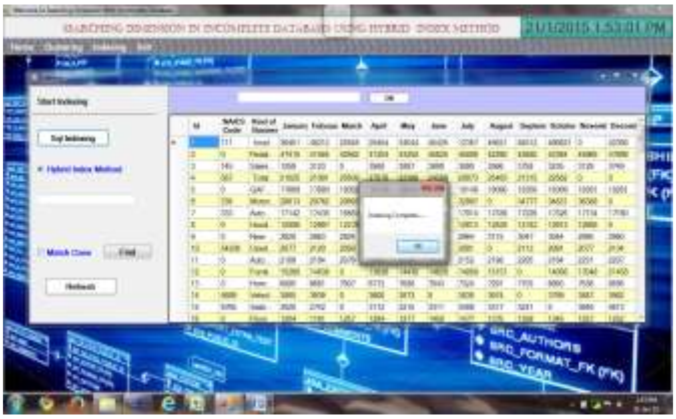

Fig.4.f Indexing completed and value place at the blank space

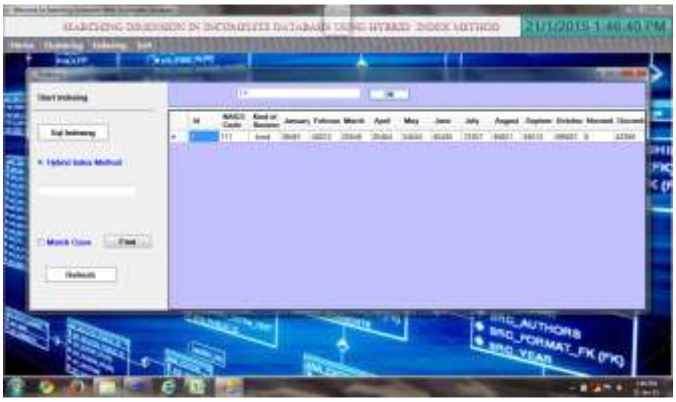

Fig.4.g Result shown

V . Conclusion

In this paper we present an approach for sorting the problem of dimension information missing. In this proposed approach certain method and techniques are applied . Our approach will achieves acceptable performance in querying incomplete. A probabilistic framework is developed to model this problem so that the users may get objects in the database that are similar to the query with probability guarantee. This framework is used to applied on for both whole sequence matching and subsequence matching. So that, we studied and included these methods and technique in our contribution. Clustering and indexing are implemented . clustering search out the hidden patterns from massive data set. It from groups of attributes. Indexing is used to provide for 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. BR-Tree and MOSAIC play vital role in indexing for providing id's and data integrity . so that clustering form cluster for attribute searching . Indexing prune the search space and increase the speed of user query process. Missing ratio represent by standard measure such as , precision and recall. So that , implementing this hybrid index method for precise result.

VI. Future work

To investigate how to extend our query strategy on Heterogeneous service application**.** Comparative study of data retrieving with the help of graph .

VII. 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] 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,1999

[7] G. Canahuate , M. Gibas , and H. Ferhatosmanoglu ,“ Indexing Incomplete Database,” Proc. 10th Int’l Conf. Advances in Database Technology (EDBT ’06), pp. 884-901, 2006.

[8] 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.