ISSN: 2321-7782 (Online)

Volume 3, Issue 6, June 2015

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Searching Dimension in Incomplete Database by Using Hybrid Indexing Method

Yogita M. Kapse[1]
M-tech CSE Student
Computer Science And Engineering department
G.H. Raisoni Institute Of Engineering And
Technology for women
Nagpure - India

Antara Bhattacharya[2]
Assistant professor
Computer Science And Engineering department
G.H. Raisoni Institute Of Engineering And
Technology for women
Nagpure - India

Abstract:** Now a days, 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. As providing accurate results that best meet the query conditions over incomplete database. To retrieve data from incomplete database is not a trivial task. This problem introduces challenges in processing query. Dimension incomplete problem causes due to collection of data from noisy network environment. Some of these techniques retrieve the query results based on the existing values rather than estimating the missing values. Such techniques are undesirable in many cases as the dimensions with missing values might be the important dimensions of the user's query. The existing work addresses the problem where data values are uncertain and unknown on dimension incomplete database. These Several techniques have been proposed to process queries in dimension incomplete database. In our proposed approach, we investigate the problem of similarity search on dimension incomplete data. A proposed framework is developed to model this problem so that the users can find data objects in the database that are similar to the query with probability guarantee . The proposed work introduces clustering, indexing, segmentation and searching and finally probabilistic approach. Each method tries to model this incomplete dimension problem. Firstly, clustering forms group of certain attributes using 'CLIHD' algorithm. After that, index structure is also employed to further prune the search space and speed up the query process. The combination of three schemes such as, BR-Tree, MOSAIC Tree, R* Tree collectively is called as hybrid index method. Segmentation is performed on random data set to filter out irrelevant data object. Searching represents result according to user query. The probabilistic approach represent missing dimension. Simulteneously missing ratio will be evaluated on the basis of standard measure such as, precision and recall. The proposed framework and technique can be applied to both whole and subsequence queries for providing effectiveness and efficiency to our approach.

***Keywords:** Dimension Incomplete, similarity query, hybrid index, query processing, index Structure, Heterogeneous database.*

I. INTRODUCTION

Similarity query in high dimensional database is a fundamental research problem with numerous applications in the areas of database, data mining, and information retrieval. Given a query object, the goal is to find similar objects in the database [1] [3]. Recently, querying incomplete data has attracted extensive research efforts [5] [6].In this problem, the data values may be missing due to various practical issues, For example, in sensor networks, the received data may become incomplete when sensors do not work properly or when errors occur during the data transfer process.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 [11]. 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 [3]. 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 list some of the problem as follows:

1. Incomplete Data Entry:

When user enter data in database, certain fields are not mandatory. It may responsible for data incompleteness problem. Users may intentionally or accidentally missed some values of different attributes (Dimensions) .

2. Data Type Missing:

When data enter by user that time if certain data type is not properly mentioned, it may responsible for incomplete dimension problem.

3. Weak Network or Low Bandwidth:

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 loss. Due to low bandwidth occur incomplete dimension problem.

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 become lost. The research methodology applied for sorting problem of searching dimension in incomplete database.

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 attributevalues 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 [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. 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. 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.

III. PROPOSED METHODOLOGY

In this paper, we propose to investigate the problem of similarity search on dimension incomplete data. The basic purpose of the proposed system is to model the problem of similarity search and search out dimension information of missing data. In proposed Methodology certain techniques have been introduced such as, clustering, indexing, segmentation and searching and finally probabilistic approach.

Initially, complete and partial database collected form uci repositorty machine. In research work, required to process on incomplete database. So that, clustering performs on partial data set. 'CLIHD 'algorithm is nothing but clustering incomplete high-dimensional database. It has been applied for clustering. Hybrid Index technique has been applied for improving traditional queries processing approach. Due to indexing data will search out within minimum span of time. Segmentation and searching remove irrelevant data object and easy to search out data according to attribute or keyword. Mainly, in probabilistic approach, missing dimension and missing data ratio become evaluated.

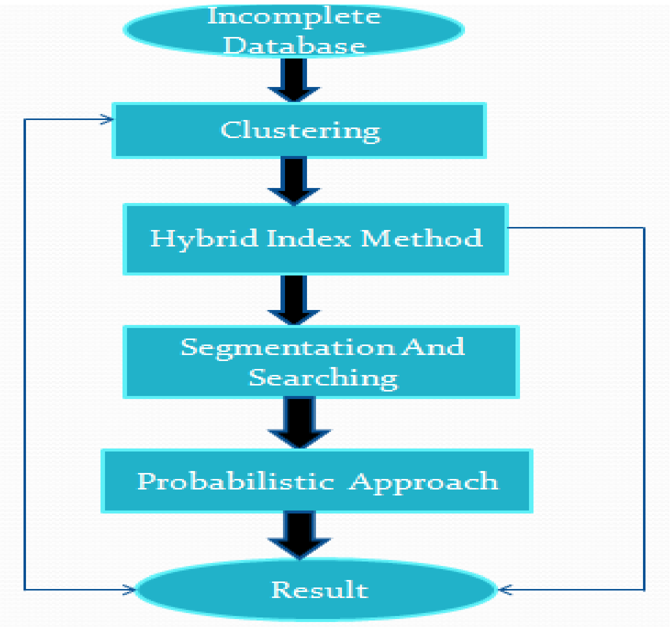


Fig.3 Flow work of searching dimension in incomplete database

The following are the proposed plan of work:-

1. Clustering

2.Indexing

3.Segmentation and searching

4. Probabilistics Approach

5. Analysis on the basis standard measure

3.1 Clustering :

Clustering is nothing but common technique in data mining to discover hidden patterns from massive datasets . With the development of privacy-maintaining data mining application. CLIHD is nothing but clustering incomplete high dimensional database. CLIHD algorithm applied for clustering. Clustering is nothing but the process of forming group of dataset according to attribute or keyword. Common applications of cluster analysis involve scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, marketing, computation biology, and many others.The steps of algorithm CLIHD are as follow :

Steps of CLIHD Algorithm :

Input: Dataset Ds, support threshold minsup;

Output: A set of cluster’s IDs;

Method: CLIHD(Ds, minsup)

Where as ,

```
    DS ← determine and sort the order of Ds dimensions;
    UList ← fulldim(Ds, minsup) //recognition on full
                dimensions;
IncompleteDim(Ds,minsup,UList,DimID) //recognition on incomplete dimensions;
```

Minsup ⟵ attribute

CLIHD Algorithm :

```
1. Select Dataset Ds
2. Select Attribute (minsup)
3. Start Clustering process
4. Searching Input From Dataset
       If (minsup = = true)
       then
       Merge Record
5. Clustering Completed
6. Display Record
```

3.2 Indexing :

Indexing is the process which representation of data in sequence format. Due to indexing result will show in minimum span of time . It will Prune the search space and speed up the query process using hybrid Index method . Hybrid indexing method are as follows .

3.2.1 Hybrid Indexing :

Hybrid indexing is the combination of three method. So that it collectively called as hybrid indexing. Hybrid Index scheme is used to apply for indexing, which are as follows .

3.2.1 1 BR-Tree Scheme :

BR-Tree is nothing but the Bit string augumented multidimensional index structure. It provides identical id's to each entry in high-dimensional database. When simultaneously enter data in database, there may be possibility to lose data and their dimensions . It may be responsible for missing dimension information. But due to BR-tree, id's are provided to each tuple entry. The BR-tree is shown to be more efficient in supporting range queries and have lower insertion and storage costs . Steps of BR-tree are given as follows .

Where as,

dr ⟶ array list where value of 'i' will store

i ⟶ id's will increment

Steps of BR-tree

BR-Tree scheme :

```
1. Start
2. Int i , dr[]
3. dr[] = i
4. i++
5. End
```

3. 2.1.2 MOSAIC :

MOSAIC is nothing but Multiple one dimensional one attribute . This scheme are used to provides specification for each attribute. MOSAIC scheme introduced the process of forming single data set for single attribute. Due to this process data integrity will maintained. Storage cost will increase. So that user can seach data easily according to query. In this category, a one-dimensional index is built on each attribute (dimension) of the search key. This is essentially an inverted index that allows for rapid identification of the set of tuples having a given value in the dimension being indexed. Thus, there will be as many inverted indexes as there are dimensions in the search key.

.3. 2.1.3 R+ Tree :

R+ tree method called as conventional multidimensional index. 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
5. End
```

3. 3. Searching And Segmentation :

In segmentation and searching, Dataset may be complete or incomplete format According to that how much percent of dimensions are unknown or miss . In segmentation irrelevant data object removed by stop word removal process. In searching , data search out according to attribute or keyword form selected dataset . It provide form and additional attribute will add. In searching Process it introduced linear search method . Which are as follows .

Linear search process perform as follow :

Where as,

Number → it consider veriable value and increment position value

Loc → it store the position of value and return result

Linear Search

```
1. Start
 2.  Int  [] number  , searchvalue
 3. Int  loc = 0
 4. While ( loc <  number . length  &&  number [ loc] != searchvalue)
 5. loc++
 6. If  ( loc = = number . Length )
```

```
7.  Return  Not Found
8.  End If
9.  Else  Return  loc    ( found  and return the position)
10.  End
```

3.4 Probabilistics Approach :

The probabilistic approach work on the basis of probabilistic method . It introduce novel method for searching missing ratio and dimension i.e probabilistic Searching Dimension method (PSDM) . It represent result by searching missing dimension and missing ratio . PSDM applied for Probabilistic result . This method classified in to two process. Which are as follows :

3.4.1 Searching Missing Data By comparing with Existing Dataset :

PSDM , firstly applied for searching missing data from dataset.

Steps are as follow :

1. Select dataset

2. Compare available data with missing data

3. Match the sequence of data

4. If row having missing data then placed whole data

3.4.2 Searching Dimension :

Searching dimension is the process of searching missing data position or which attribute having missing values.

steps are given as follows :

1. P(E) and P(M) is probability of existing data and

   missing data dimension respectively

2. P(E) and P(M) are the probabilities of E and M without

   regard to one other

3. Missing dimension is the ratio of probability of missing

   data dimension to the intersection of probability of

   existing data and missing data dimension.

4. Where as , P(E | M), a conditional probability, which is

   the probability of E given

   that M is true.

3.5 Standard Parameter :

3. 5. 1. Precision :

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

Where,

Select all record from data base

find how much percent of relevant data get

Precision (%)=(existing data/total data)*100

3. 5. 2. Recall :

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

Where,

Select all record from data base

find how much percent of data missing data

Recall(%)=(missing data/total data)*10

IV. Experimental Setup

Following are the screen shots of clustering and indexing process:

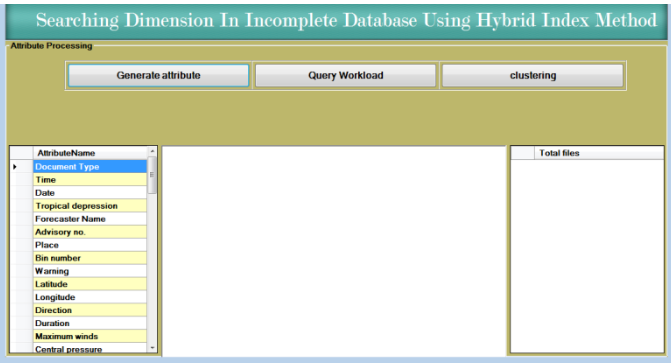
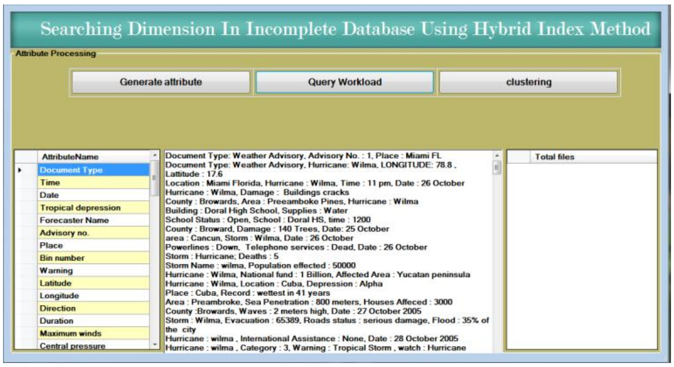
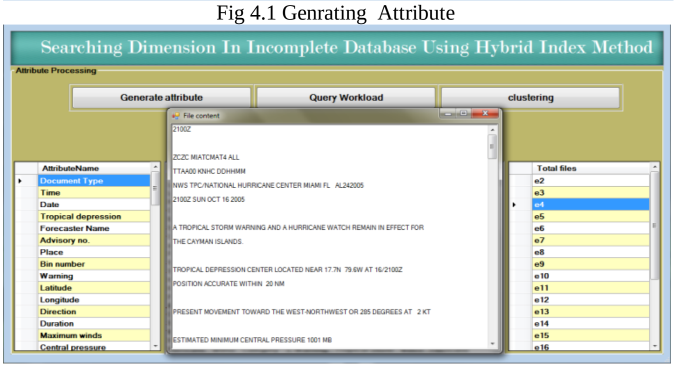


Fig 4.1 Genrating Attribute



Document Type: Weather Advisory, Advisory No. : 1, Place : Miami FL
Document Type: Weather Advisory, Hurricane: Wilma, LONGITUDE: 78.8 , Lattitude : 17.6
Location : Miami Florida, Hurricane : Wilma, Time : 11 pm, Date : 26 October
Hurricane : Wilma, Damage : Buildings cracks
County : Browards, Area : Preeamboke Pines, Hurricane : Wilma
Building : Doral High School, Supplies : Water
School Status : Open, School : Doral HS, time : 1200
County : Broward, Damage : 140 Trees, Date: 25 October
area : Cancun, Storm : Wilma, Date : 26 October
Powerlines : Down, Telephone services : Dead, Date : 26 October
Storm : Hurricane; Deaths : 5
Storm Name : wilma, Population effected : 50000
Hurricane : Wilma, National fund : 1 Billion, Affected Area : Yucatan peninsula
Hurricane : Wilma, Location : Cuba, Depression : Alpha
Place : Cuba, Record : wettest in 41 years
Area : Preambroke, Sea Penetration : 800 meters, Houses Affeced : 3000
County :Browards, Waves : 2 meters high, Date : 27 October 2005
Storm : Wilma, Evacuation : 65389, Roads status : serious damage, Flood : 35% of the city
Hurricane : wilma , International Assistance : None, Date : 28 October 2005
Hurricane : wilma , Category : 3, Warning : Tropical Storm , watch : Hurricane

Fig.4.2 Query Workload Process



2100Z

ZCZC MIATCMAT4 ALL

TTAA00 KNHC DDHHMM

NWS TPC/NATIONAL HURRICANE CENTER MIAMI FL AL242005

2100Z SUN OCT 16 2005

A TROPICAL STORM WARNING AND A HURRICANE WATCH REMAIN IN EFFECT FOR

THE CAYMAN ISLANDS.

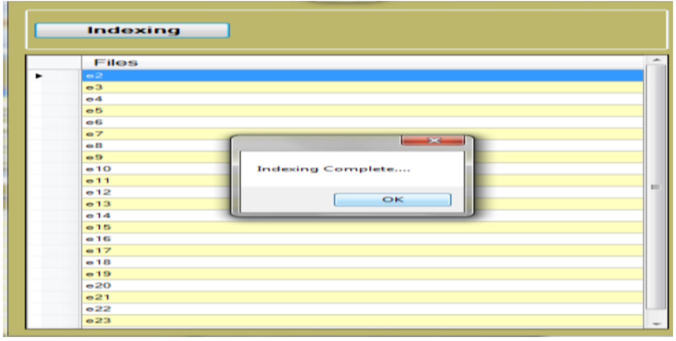
TROPICAL DEPRESSION CENTER LOCATED NEAR 17.7N 79.6W AT 16/2100Z

POSITION ACCURATE WITHIN 20 NM

PRESENT MOVEMENT TOWARD THE WEST-NORTHWEST OR 285 DEGREES AT 2 KT

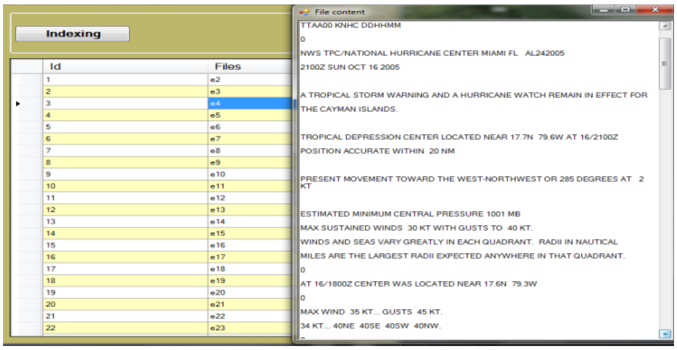
ESTIMATED MINIMUM CENTRAL PRESSURE 1001 MB

Fig.4.3 Clustering Process



Indexing Complete....

Fig.4.4 Indexing Complete



Id	Files
1	e2
2	e3
3	e4
4	e5
5	e6
6	e7
7	e8
8	e9
9	e10
10	e11
11	e12
12	e13
13	e14
14	e15
15	e16
16	e17
17	e18
18	e19
19	e20
20	e21
21	e22
22	e23

TTAA00 KNHC DDHHMM

0

NWS TPC/NATIONAL HURRICANE CENTER MIAMI FL AL242005

2100Z SUN OCT 16 2005

A TROPICAL STORM WARNING AND A HURRICANE WATCH REMAIN IN EFFECT FOR

THE CAYMAN ISLANDS.

TROPICAL DEPRESSION CENTER LOCATED NEAR 17.7N 79.6W AT 16/2100Z

POSITION ACCURATE WITHIN 20 NM

PRESENT MOVEMENT TOWARD THE WEST-NORTHWEST OR 285 DEGREES AT 2 KT

ESTIMATED MINIMUM CENTRAL PRESSURE 1001 MB

MAX SUSTAINED WINDS 30 KT WITH GUSTS TO 40 KT.

WINDS AND SEAS VARY GREATLY IN EACH QUADRANT. RADII IN NAUTICAL

MILES ARE THE LARGEST RADII EXPECTED ANYWHERE IN THAT QUADRANT.

0

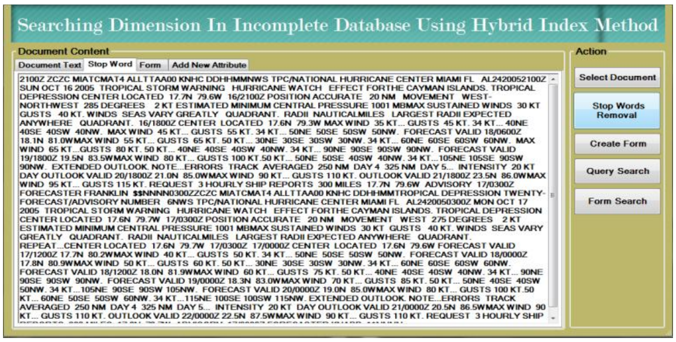
AT 16/1800Z CENTER WAS LOCATED NEAR 17.6N 79.3W

0

MAX WIND 35 KT... GUSTS 45 KT.

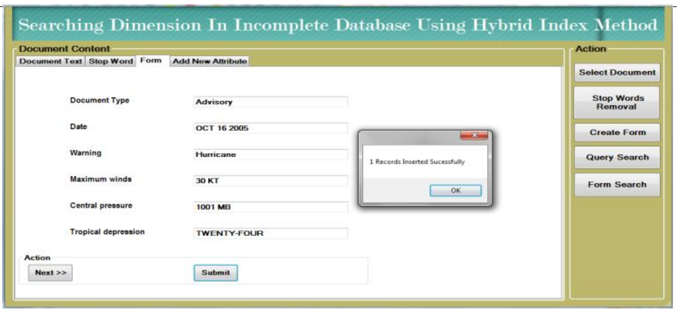
34 KT... 40NE 40SE 40SW 40NW.

Fig.4.5 Hybrid indexing provide id's to data set



2100Z ZCZC MIATCMAT4 ALLTTAA00 KNHC DDHHMMNWS TPC/NATIONAL HURRICANE CENTER MIAMI FL AL2420052100Z SUN OCT 16 2005 TROPICAL STORM WARNING HURRICANE WATCH EFFECT FORTHE CAYMAN ISLANDS. TROPICAL DEPRESSION CENTER LOCATED 17.7N 79.6W 16/2100Z POSITION ACCURATE 20 NM MOVEMENT WEST-NORTHWEST 285 DEGREES 2 KT ESTIMATED MINIMUM CENTRAL PRESSURE 1001 MBMAX SUSTAINED WINDS 30 KT GUSTS 40 KT. WINDS SEAS VARY GREATLY QUADRANT. RADII NAUTICALMILES LARGEST RADII EXPECTED ANYWHERE QUADRANT. 16/1800Z CENTER LOCATED 17.6N 79.3W MAX WIND 35 KT... GUSTS 45 KT. 34 KT... 40NE 40SE 40SW 40NW. MAX WIND 45 KT... GUSTS 55 KT. 34 KT... 50NE 50SE 50SW 50NW. FORECAST VALID 18/0600Z 18.1N 81.0WMAX WIND 55 KT... GUSTS 65 KT. 50 KT... 30NE 30SE 30SW 30NW. 34 KT... 60NE 60SE 60SW 60NW. MAX WIND 65 KT...GUSTS 80 KT. 50 KT... 40NE 40SE 40SW 40NW. 34 KT... 90NE 90SE 90SW 90NW. FORECAST VALID 19/1800Z 19.5N 83.5WMAX WIND 80 KT... GUSTS 100 KT.50 KT... 50NE 50SE 40SW 40NW. 34 KT...105NE 105SE 90SW 90NW. EXTENDED OUTLOOK. NOTE...ERRORS TRACK AVERAGED 250 NM DAY 4 325 NM DAY 5... INTENSITY 20 KT DAY OUTLOOK VALID 20/1800Z 21.0N 85.0WMAX WIND 90 KT... GUSTS 110 KT. OUTLOOK VALID 21/1800Z 23.5N 86.0WMAX WIND 95 KT... GUSTS 115 KT. REQUEST 3 HOURLY SHIP REPORTS 300 MILES 17.7N 79.6W ADVISORY 17/0300Z FORECASTER FRANKLIN $$NNNN0300ZZCZC MIATCMAT4 ALLTTAA00 KNHC DDHHMMTROPICAL DEPRESSION TWENTY-FORECAST/ADVISORY NUMBER 6NWS TPC/NATIONAL HURRICANE CENTER MIAMI FL AL2420050300Z MON OCT 17 2005 TROPICAL STORM WARNING HURRICANE WATCH EFFECT FORTHE CAYMAN ISLANDS. TROPICAL DEPRESSION CENTER LOCATED 17.6N 79.7W 17/0300Z POSITION ACCURATE 20 NM MOVEMENT WEST 275 DEGREES 2 KT ESTIMATED MINIMUM CENTRAL PRESSURE 1001 MBMAX SUSTAINED WINDS 30 KT GUSTS 40 KT. WINDS SEAS VARY GREATLY QUADRANT. RADII NAUTICALMILES LARGEST RADII EXPECTED ANYWHERE QUADRANT. REPEAT...CENTER LOCATED 17.6N 79.7W 17/0300Z 17/0000Z CENTER LOCATED 17.6N 79.6W FORECAST VALID 17/1200Z 17.7N 80.2WMAX WIND 40 KT... GUSTS 50 KT. 34 KT... 50NE 50SE 50SW 50NW. FORECAST VALID 18/0000Z 17.8N 80.9WMAX WIND 50 KT... GUSTS 60 KT. 50 KT... 30NE 30SE 30SW 30NW. 34 KT... 60NE 60SE 60SW 60NW. FORECAST VALID 18/1200Z 18.0N 81.9WMAX WIND 60 KT... GUSTS 75 KT. 50 KT... 40NE 40SE 40SW 40NW. 34 KT... 90NE 90SE 90SW 90NW. FORECAST VALID 19/0000Z 18.3N 83.0WMAX WIND 70 KT... GUSTS 85 KT. 50 KT... 50NE 40SE 40SW 50NW. 34 KT...105NE 90SE 90SW 105NW. FORECAST VALID 20/0000Z 19.0N 85.0WMAX WIND 80 KT... GUSTS 100 KT.50 KT... 60NE 50SE 50SW 60NW. 34 KT...115NE 100SE 100SW 115NW. EXTENDED OUTLOOK. NOTE...ERRORS TRACK AVERAGED 250 NM DAY 4 325 NM DAY 5... INTENSITY 20 KT DAY OUTLOOK VALID 21/0000Z 20.5N 86.5WMAX WIND 90 KT... GUSTS 110 KT. OUTLOOK VALID 22/0000Z 22.5N 87.5WMAX WIND 90 KT... GUSTS 110 KT. REQUEST 3 HOURLY SHIP

Fig.4.6 Stop word removal process



Searching Dimension In Incomplete Database Using Hybrid Index Method

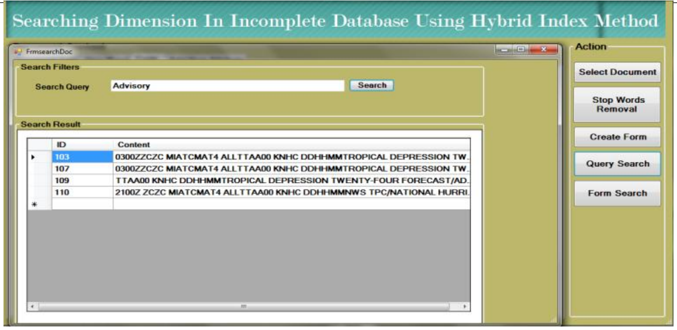
Document Content

Document Text | Stop Word | Form | Add New Attribute

Document Type	Advisory
Date	OCT 16 2005
Warning	Hurricane
Maximum winds	30 KT
Central pressure	1001 MB
Tropical depression	TWENTY-FOUR

1 Records Inserted Sucessfully

Fig.4.7 Create form and submit record



Searching Dimension In Incomplete Database Using Hybrid Index Method

FrmsearchDoc

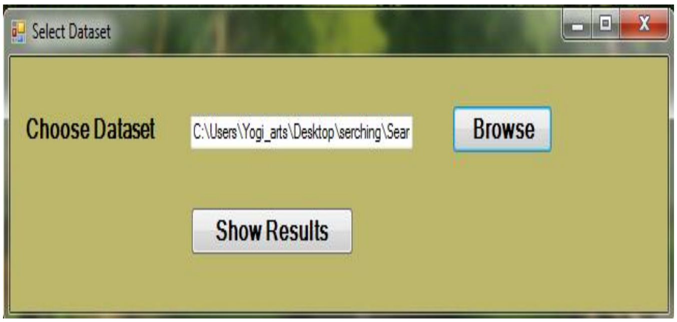
Search Filters

Search Query: Advisory

Search Result

ID	Content
103	0300ZZCZC MIATCMAT4 ALLTTAA00 KNHC DDHHMMTROPICAL DEPRESSION TW..
107	0300ZZCZC MIATCMAT4 ALLTTAA00 KNHC DDHHMMTROPICAL DEPRESSION TW..
109	TTAA00 KNHC DDHHMMTROPICAL DEPRESSION TWENTY-FOUR FORECAST/AD..
110	2100Z ZCZC MIATCMAT4 ALLTTAA00 KNHC DDHHMMNWS TPC/NATIONAL HURRI..

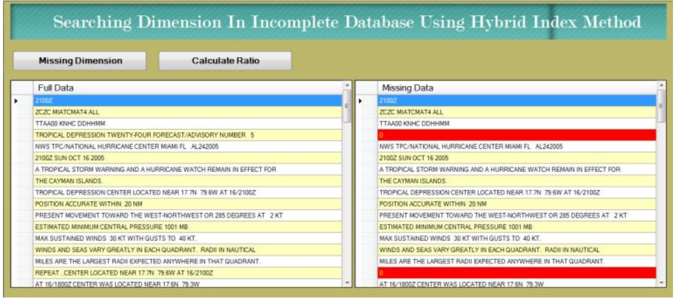
Fig.4.8 Query Search



Select Dataset

Choose Dataset: C:\Users\Yogi_arts\Desktop\serching\Sear

Fig.4.9 Select document

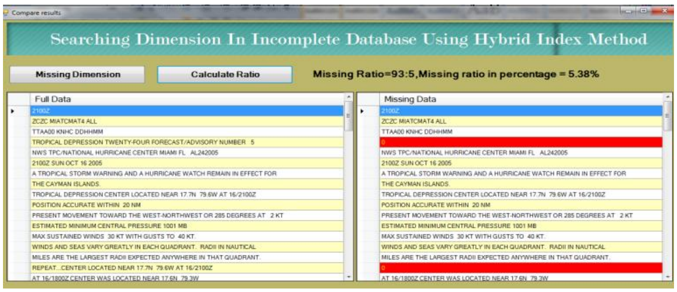


Searching Dimension In Incomplete Database Using Hybrid Index Method

Full Data
2100Z
ZCZC MIATCMAT4 ALL
TTAA00 KNHC DDHHMM
TROPICAL DEPRESSION TWENTY-FOUR FORECAST/ADVISORY NUMBER 5
NWS TPC/NATIONAL HURRICANE CENTER MIAMI FL AL242005
2100Z SUN OCT 16 2005
A TROPICAL STORM WARNING AND A HURRICANE WATCH REMAIN IN EFFECT FOR
THE CAYMAN ISLANDS.
TROPICAL DEPRESSION CENTER LOCATED NEAR 17.7N 79.6W AT 16/2100Z
POSITION ACCURATE WITHIN 20 NM
PRESENT MOVEMENT TOWARD THE WEST-NORTHWEST OR 285 DEGREES AT 2 KT
ESTIMATED MINIMUM CENTRAL PRESSURE 1001 MB
MAX SUSTAINED WINDS 30 KT WITH GUSTS TO 40 KT.
WINDS AND SEAS VARY GREATLY IN EACH QUADRANT. RADII IN NAUTICAL
MILES ARE THE LARGEST RADII EXPECTED ANYWHERE IN THAT QUADRANT.
REPEAT...CENTER LOCATED NEAR 17.7N 79.6W AT 16/2100Z
AT 16/1800Z CENTER WAS LOCATED NEAR 17.6N 79.3W

Missing Data
2100Z
ZCZC MIATCMAT4 ALL
TTAA00 KNHC DDHHMM
0
NWS TPC/NATIONAL HURRICANE CENTER MIAMI FL AL242005
2100Z SUN OCT 16 2005
A TROPICAL STORM WARNING AND A HURRICANE WATCH REMAIN IN EFFECT FOR
THE CAYMAN ISLANDS.
TROPICAL DEPRESSION CENTER LOCATED NEAR 17.7N 79.6W AT 16/2100Z
POSITION ACCURATE WITHIN 20 NM
PRESENT MOVEMENT TOWARD THE WEST-NORTHWEST OR 285 DEGREES AT 2 KT
ESTIMATED MINIMUM CENTRAL PRESSURE 1001 MB
MAX SUSTAINED WINDS 30 KT WITH GUSTS TO 40 KT.
WINDS AND SEAS VARY GREATLY IN EACH QUADRANT. RADII IN NAUTICAL
MILES ARE THE LARGEST RADII EXPECTED ANYWHERE IN THAT QUADRANT.
0
AT 16/1800Z CENTER WAS LOCATED NEAR 17.6N 79.3W

Fig.4.10 Comparison of available and missing data



Compare results

Searching Dimension In Incomplete Database Using Hybrid Index Method

Missing Ratio=93:5,Missing ratio in percentage = 5.38%

Full Data
2100Z
ZCZC MIATCMAT4 ALL
TTAA00 KNHC DDHHMM
TROPICAL DEPRESSION TWENTY-FOUR FORECAST/ADVISORY NUMBER 5
NWS TPC/NATIONAL HURRICANE CENTER MIAMI FL AL242005
2100Z SUN OCT 16 2005
A TROPICAL STORM WARNING AND A HURRICANE WATCH REMAIN IN EFFECT FOR
THE CAYMAN ISLANDS.
TROPICAL DEPRESSION CENTER LOCATED NEAR 17.7N 79.6W AT 16/2100Z
POSITION ACCURATE WITHIN 20 NM
PRESENT MOVEMENT TOWARD THE WEST-NORTHWEST OR 285 DEGREES AT 2 KT
ESTIMATED MINIMUM CENTRAL PRESSURE 1001 MB
MAX SUSTAINED WINDS 30 KT WITH GUSTS TO 40 KT.
WINDS AND SEAS VARY GREATLY IN EACH QUADRANT. RADII IN NAUTICAL
MILES ARE THE LARGEST RADII EXPECTED ANYWHERE IN THAT QUADRANT.
REPEAT...CENTER LOCATED NEAR 17.7N 79.6W AT 16/2100Z
AT 16/1800Z CENTER WAS LOCATED NEAR 17.6N 79.3W

Missing Data
2100Z
ZCZC MIATCMAT4 ALL
TTAA00 KNHC DDHHMM
0
NWS TPC/NATIONAL HURRICANE CENTER MIAMI FL AL242005
2100Z SUN OCT 16 2005
A TROPICAL STORM WARNING AND A HURRICANE WATCH REMAIN IN EFFECT FOR
THE CAYMAN ISLANDS.
TROPICAL DEPRESSION CENTER LOCATED NEAR 17.7N 79.6W AT 16/2100Z
POSITION ACCURATE WITHIN 20 NM
PRESENT MOVEMENT TOWARD THE WEST-NORTHWEST OR 285 DEGREES AT 2 KT
ESTIMATED MINIMUM CENTRAL PRESSURE 1001 MB
MAX SUSTAINED WINDS 30 KT WITH GUSTS TO 40 KT.
WINDS AND SEAS VARY GREATLY IN EACH QUADRANT. RADII IN NAUTICAL
MILES ARE THE LARGEST RADII EXPECTED ANYWHERE IN THAT QUADRANT.
0
AT 16/1800Z CENTER WAS LOCATED NEAR 17.6N 79.3W

Fig.4.11 Dimension of missing data

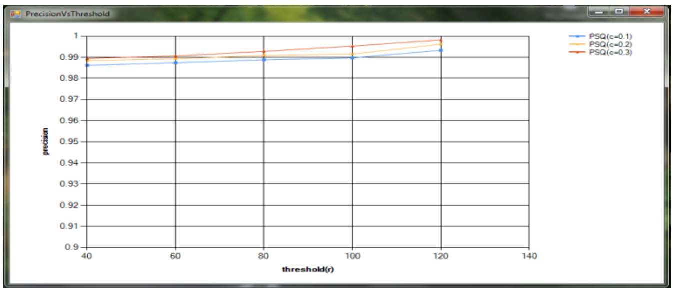


Fig.4.12   Precesion Vs Threshold Graph

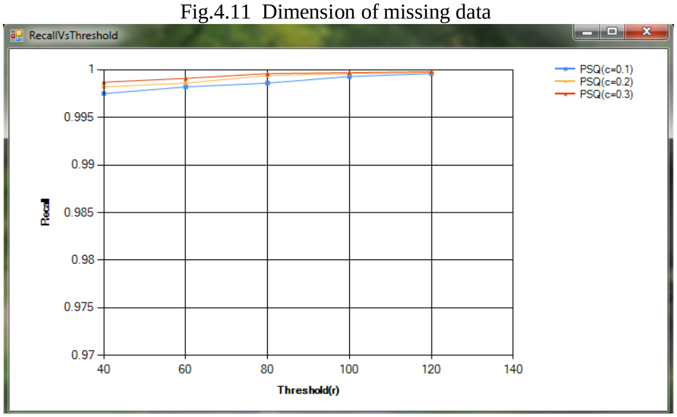


Fig.4.13 Recall Vs Threshold Graph

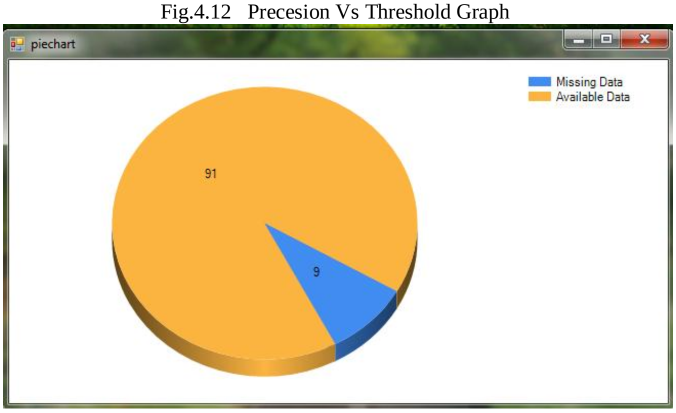


Fig.4.14 Missing Ratio Diagram

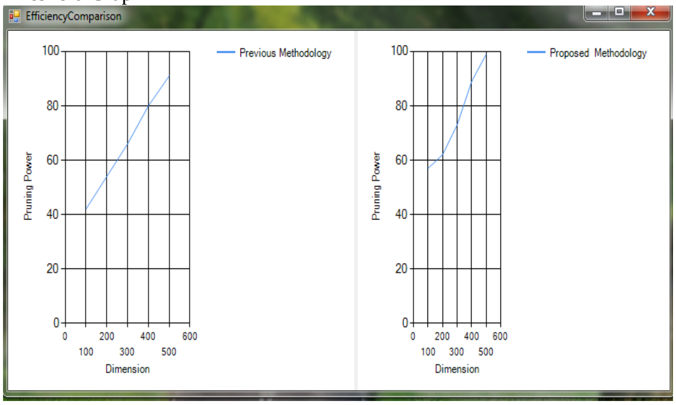


Fig.4.15 Efficiency Comparison Graph

V. CONCLUSION

In this paper we present an approach for mitigating the problem of dimension information missing. In our research work, certain method and techniques are applied. This proposed framework addresses the problem of dimension information which is of the both practical importance and technical challenge. To improve the performance of clusterization of partial or random data with the help of CLIHD (clustering incomplete high dimensional database) algorithm. Prominently, clustering has been posed as an optimization problem for minimum error or maximum attribute predictability. Clustering form group of data set according to attribute. In our research work novel indexing technique is introduced called as hybrid indexing technique. The hybrid index is the combination of three method i.e BR tree, MOSAIC and R+ tree method. Indexing represent data in arranged format and it become easy to search out. Hybrid index prune the search space and speedup the query process. Segmentation process removed irrelevant data object by stop word removal process. Searching represent result according to user query. The probabilistic searching dimension method represent missing dimension of partial database. Our approach will achieves satisfactory performance in querying incomplete

VI. FEATURE SCOPE

In future novel technique can also be implemented to the complex dataset by making some changes to the system and by changing some parameter of this methodology. The hybrid index technique can also be applied on multidimensional database in future. To investigate how to extend our query strategy on Heterogeneous service application.

References

1. Xiang Zhang, and Wei Wang, Member, IEEE, Searching Dimension Incomplete Databases, vol.26, No.3, March 2014.
2. Y. Kapse , A. Bhattacharya, “Survey On Various Method and Technique For Searching Dimension In Incomplete Database “, Proc. International Journal Of Computer Science and Information Technology(IJCSIT ’15), pp. 7078- 7081,2015.
3. Y. Kapse , A. Bhattacharya, “Implementing Hybrid Index Method For Searching Dimension in Incomplete Database” , Proc. 2nd International Conference on Science & Technology for Society (ICSTS 2015). IJRITCC.pp.106-111,2015.
4. Ali A. Alwan, Hamidah Ibrahim, Nur Izura Udzir, Fatimah Sidi,“ Estimating Missing Values Of Skylines In Incomplete Database” Proc. 36th Int’l Conf. Very Large Databases (VLDB ‘13) , 2013.
5. Dimitrios Skoutas, Dimitris Sacharidis, Alkis Simitsis, and Timos Sellis, “Ranking and Clustering Web Services UsingMulticriteria Dominance Relationships”, Proc. IEEE Tran on Service computing, Sept 2010 .
6. J. Pei, B. Jiang, X. Lin, and Y. Yuan,” Probabilistic Skylines on Uncertain Data”, Proc. 33rd Int’l Conf. VeryLarge Databases (VLDB ’07), pp. 15-26, 2007.
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. 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.
9. 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.
10. 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.
11. Beng Chin Ooi, Cheng Hian Goh, Kian-lee Tan , “Fast High –Dimensional data Search In Incomplete Databases”, Proc.24th Into’l conf. very large Databases(VLDB) 2000.
12. 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.
13. S.Otsuka, N.Miyazaki ,” An incomplete database approach” , Proc.12th International conference,Information Networking (ICIN ‘ 98) , 1998 .
14. 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.
15. 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
16. Y. Kapse , A. Bhattacharya, “Survey On Various Method and Technique For Searching Dimension In Incomplete Database “, Proc. International Journal Of Computer Science and Information Technology(IJCSIT ’15),pp. 78-81,2015.

AUTHOR(S) PROFILE

Yogita .M. Kapse, received the BE degree in Information Technology from K.D.K College of Engineering in 2009and 2013 , Now perceiving M-Tech final year in Computer science and engineering from G.H.Raisoni Institute of Engineering and Technology for women During the session 2013 to 2015 respectively. Now, working as lecturer in Balaji polytechnique, 2015. This research work represents Novel Hybrid Indexing Technique. This research work completed in domain datamining by refering paper searching dimension incomplete database.