International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 3, Issue 10, October 2014 **ISSN 2319 - 4847**

A Survey on Searching Dimension in Incomplete Databases using Indexing Approach

Miss Namrata M. Pagare [1], Prof.Mrs.J.R.Mankar [2]

[1] K. K. Wagh Institute of Engineering Education & Research, Department of Computer Engineering,
Savitribai Phule Pune University
Nashik, India

[2] K. K. Wagh Institute of Engineering Education & Research, Department of Computer Engineering,
Savitribai Phule Pune University
Nashik, India

ABSTRACT

Searching in multidimensional databases is tedious task because of its variety of applications in areas of databases, data mining and information retrieval. Querying incomplete databases creates various problems. Incomplete data refers to data with missing values or uncertain data. Also the database with missing dimension information is present. A probabilistic approach is one way to handle queries on uncertain data. Indexing techniques are used for multi-dimensional data when indexed attributes contains missing data like bitmaps and quantization. Query execution and interval evaluation are characterized for indexing structures based on whether missing data is considered to be a query match or not. Also to improve the efficiency of data retrieval in high dimensional database bit string augmented R-tree and multiple one dimensional attribute index structures is used. Dynamic time warping (DTW) and longest common subsequence are the widely used distance function in similarity search. The lower and upper probability bounds are used to reduce the search space and develop a probability triangle inequality for filtering and increasing the speed of query process and address the subsequence matching in dimension data. This provides effective way to deal with incomplete database and improve efficiency of similarity search

Keywords: incomplete databases, dynamic time warping, bitmaps, quantization, subsequence matching

1. INTRODUCTION

Databases with missing data occur in a wide range of research and industry domains. Some examples of these are: A census database that allows null values for some attributes, a survey database where answers to one question cause other questions to be skipped, a medical database that relates human body analyte measurements to a number of diseases, or patient risk factors to a specific disease. Similarity search in incomplete databases creates many problems. The aim is to access databases efficiently in the presence of missing data. In some cases, the missingness of data is random, i.e. the missingness of some value does not depend on the value of another variable. If the data are missing as a function of some other variable, a complete treatment of missing data would have to include a model that accounts for missing data. As the amount of uncertain data collected and stored is growing so analyzing these enormous collections of uncertain data has become an essential task.

2. INCOMPLETE DATABASES

There are a variety of reasons why databases may be incomplete or data may be missing. The data may not be available at the time the record was populated or it was not recorded because of equipment malfunction or adverse conditions. Data may have been unintentionally omitted or the data is not relevant to the record at hand. The allowance for and use of missing data may be intentionally designed into the database. Multi-dimensional indexing techniques work best when records are mapped to non-overlapping hypercube. When missing data are mapped to a single value, the overlaps associated with the index structure increase. Sequences constitute large portion of data stored in computers. It requires efforts to model time sequence data and query its different techniques are designed access such queries.

2.1 Searching Incomplete Databases

In incomplete database systems, query processing is challenging as in many cases a significant part of query answer may be neglected from the final answer due to the missing values in some dimensions (attributes). In addition, preference queries have not received much attention in incomplete database applications in which to evaluate the query, exhausted comparison needs to be performed in order to determine the best data items in the database that meet the query conditions. Preference queries in incomplete database are fundamentally different than the conventional preference queries in complete database because the transitivity property of preference techniques is no longer hold.

Solution to missing data problem includes the use of imputation, statistical and regression based procedures in order to estimate entries [9], [10]. The indexing of huge time series databases has created interest of the database community. The bulk of work in this area has concentrated on indexing under the Euclidean distance measure. The idea is to embed the sequences into Euclidean space such that the distances between them are approximately preserved, then classic multidimensional index structures can be utilized (Guttman 1984; Seidl and Kriegel 1988). Kim et al. introduced an exact algorithm for indexing of time series under DTW (Kim et al. 2001). The method extracts four features from the sequences and organizes them in a multidimensional index structure. In Park et al. (1999), the authors demonstrate a DTW indexing technique that is based on a piecewise linear representation of the data. They prove that this method can guarantee no false dismissals. Also, there are situations where we may not know the position of data loss or which dimension is missing in database. So it's necessary to address this issue to search in database. The dimension information may be missing due to faulty sensors in network, time series data with temporal uncertainty etc. The dimension information is crucial for the existing uncertain data querying methods hence the dimension incomplete data creates challenges to the similarity query task, [2], and [3]. Suppose that the original data dimensionality is m. Given a query object Q =(q1,q2,…….,q_m) and a dimension incomplete data object X =(x1,x2,……..,x_n) ($n < m$), a solution to calculate the distance between these two objects by examining all possible missing dimension combinations for the data object X. For each combination, we impute the values at the positions with missing data, and then calculate the similarity between Q and the imputed X based on certain distance function such as the 'p-norms distance.

3. RELATED WORK

In this section, an overview of missing data in large databases is provided. The objective of this survey is to clearly understand the limitations of existing missing values and problems in dimension incomplete databases. Ronald K. Pearson [5] defined the problem of disguised missing data where when missing data values are not explicitly represented as such, but are coded with values that can be misinterpreted as valid data. There are four different ways of dealing with explicitly coded missing data:

a. Deletion:

Deletion strategies simply omit some or all of the missing data records, depending on the details of the analysis considered. For example, Little and Rubin [11] distinguish between complete case analysis, based only on complete data records, and available case analysis, based on all records that are sufficiently complete for the analysis under consideration to be undertaken. The difference between these analysis strategies can be important in datasets with many fields per record since available case characterizations involving fewer variables (e.g., univariate characterizations like means and standard deviations) will generally be based on larger data subsets than those involving more variables (e.g., multiple regression analysis). For small fractions of missing data, these deletion strategies are used quite extensively.

b. Single imputation:

Single imputation strategies provide a single estimate for each missing data value. Examples of it are hot deck imputation where missing values are replaced by responses from other records that satisfy certain matching conditions (e.g., missing income values estimated by the recorded in- come value for another survey respondent from the same Zip code with similar age and educational background), and mean imputation where missing values are estimated by the mean of appropriately selected "similar" samples. A disadvantage of single imputation strategies is that they tend to artificially reduces the variability of characterizations of the imputed dataset. This observation provides the motivation for multiple imputation strategies where different imputed datasets are generated and subjected to the same analysis, giving a set of results from which typical (e.g., mean) characterizations and variability estimates can be computed. Deletion-based strategies and single imputation strategies may be regarded as filters because they yield modified datasets that are analyzed by standard methods without modification.

c. Multiple imputation:

Multiple imputation strategies do not require modification of the underlying analysis procedures and are non-iterative in nature. In contrast, iterative approaches analogous to the class of wrappers can also be developed for missing data.

d. Iterative procedures:

The Expectation-Maximization (EM) algorithm, formalizes ad hoc strategy [18] where it first, imputes the missing data values, next it estimate data model parameters using these imputed values, then, re-estimate the missing data values using these estimated model parameters and repeat, iterating until convergence. This approach is very general and has been applied to a wide range of missing data problems. Guadalupe Canahuate, Michael Gibas, and Hakan Ferhatosmanoglu [23] uses the indexing techniques for multidimensional data search when indexed attributes contain missing data. They apply the techniques of bitmap indexes and vector approximation (VA) files modified appropriately to account for missing data and to execute the query according to the query's semantics. The aim was to index each dimension independently and execute queries efficiently using bit operations for bitmaps and VA-Files for pruning multiple dimensions

a. Bitmap Indexes

In the bitmap index context, records are represented by a bit string. Each attribute A_i would be represented by at most C^i bits of the string where Ci is the cardinality of A_i, i.e. the number of distinct non-null values among all records for attribute A_i. A bitmap is a column wise representation of each position of the bit string. Each bitmap would have n bits where n is the number of records in the dataset. Given a dataset D = (A1, A2, ..., A_d) for each A_i attribute we build a certain number of bitmaps depending on Ci. To handle missing data using bitmaps, we map missing values to a distinct value, i.e. 0. By doing this we are increasing the number of bitmaps for each attribute with missing data by 1. While mapping missing data to a distinct value fails for multi-dimensional indexes, it is acceptable for bitmaps because the attributes are indexed independently and we are not creating an exponential number of subspaces that must be searched to answer a query.

b. Bitmap Equality Encoding (BEE)

Using equality encoded bitmaps, bit $B_{i,\ j}[x]$ is 1 if record x has value j for attribute A_i and 0 otherwise. Using this encoding, if $B_{i,j}[x] = 1$ then $B_{i,k}[x]=0$ for all k = j. If attribute A_i has missing values, we add the bitmap $B_{i,0}$ that behaves in the same manner explained above. In this alternative, when missing is a match we make $B_{i,j}[x] = 1$ for all j if record x has missing data in attribute A_i; and when missing is not a match, we make $B_{i,j}[x] = 0$ for all j if record x has missing data in attribute A_i.Query Execution With equality encoded bitmaps a point query is executed by ANDing together the bit vectors corresponding to the values specified in the search key. Bitmap Equality Encoded is optimal for point queries. When missing data means a query match we need to use two bitmaps instead of one to answer the query, i.e. the bitmap corresponding to the value queried and the one for missing values. Range queries are initially executed by ORing together all bit vectors stated by each range in the search key and then ANDing the given answers together. If the query range for an attribute queried includes more than half of the cardinality then we execute the query by taking the complement of the ORed bitmaps that are not included in the range query.

c. Bitmap Range Encoding (BRE)

For range encoded bitmaps, bit $B_{i,j}[x]$ is 1 if record x has a value that is less than or equal to j for attribute A_i and 0 otherwise. Using this encoding if $B_{i,j}[x]=1$ then $B_{i,k}[x] = 1$ for all k>j. In this case the last bitmap B_i, C_i for each attribute A_i is all 1s. Thus, we drop this bitmap and only keep C_i-1 bitmaps to represent each attribute. If attribute A_i has missing values we add the bitmap $B_{i,0}$ which has $B_{i,0}[x] = 1$ if record x has a missing value for attribute A_i. Also in this case $B_{i,j}[x] = 1$ for all j. We are treating missing data as the next smallest possible value outside the lower bound of the domain, in our case, the value 0. In total the set of bitmaps required to represent attribute A_i with missing values is C_i.

d. Bitmap Compression

One of the biggest disadvantages of bitmap indices is the amount of space they require. The two most important techniques are byte aligned bitmap code and Word aligned Hybrid code. The compressed data in Bytes is preserved by BBC while WAH preserves it in words. WAH is simpler because it only has two types of words: literal words and fill words.

e. VA-Files

For traditional VA-files, data values are approximated by one of 2^b strings of length b bits. A lookup table provides value ranges for each of the 2^bpossible representations. For each attribute A_i in the database we use b_i bits to represent $2^b{}_i$ bins that enclose the entire attribute domain. In general bi << $\lg C_i$ when the cardinality is high. We made bi = [lg (C_i + 1)]. For our purposes, we use $2^{\ b}-1$ possible representations for data values and we use a string of b 0's to represent missing data values. A VA-file lookup table relates attribute values to the appropriate bin number. For VA-files we make a modification to the query based on the query semantics. For a range query where missing data is not a query match, we look for matches over the range of bins returned by the lookup table. In the case where missing data means a query match, we also include those records in the all 0's bin as a query match. Daqian Gu and Yang Gao[4] uses incremental Gradient descent imputation model where relationship among variables is used to estimate the missing value. Learning Classifier Systems (LCS) is a kind of self- adaptive, online learning systems. In LCS research, missing data has been one of the focuses. The first type of missing data is referred to as missing completely at random (MCAR). The second type of missing data is referred to as missing at random (MAR). In this case, missing data depends on known values and thus is described fully by variables observed in the data set. The last type of missing data is referred to not missing at random (NMAR).

a. InGrImputation model

The LCS adapts themselves to environments while evolving their classifiers which represent the relationship between input variables and classification. In the learning process, relationship among the input variables has not been utilized directly so to impute missing data based on the relationship an InGrImputation Model is proposed to impute missing data based on this relationship to handle LCS missing data. InGrImputation Model creates a universal model for the variable with missing data based on the relationship between the variable and other known variables. It may be inaccurate without any prior knowledge at the beginning. In each episode of training or testing in LCS, InGrImputation Model checks whether there is any missing data in the input, if the answer is 'no', InGrImputation Model adapts itself by gradient descent method according to the current input; if the answer is 'yes', InGrImputation Model simply creates a new value for the missing data based on other input variables and the current model. Wasito, B. Mirkin[6] proposed least-squares data imputation algorithms which adopts the Nearest neighbour approach. The problem of imputation of missing data is found in areas like data editing survey [12], preserving medical documentation [13] and DNA microarray data modeling [14]. This has given rise to expectation–maximization (EM) method for handling incomplete data. Different approaches for imputation of missing data are as follows

a. Prediction rules

The imputation method is substitution of a missing data entry by the analogous variable's mean, which will be referred as Mean algorithm. The flaw of above imputed method is that the variance of the imputed data underestimates the real variance. The other prediction models are used with Mean algorithm like hot deck imputation where the nearest neighbor's value is imputed, cold deck imputation where the modal value is imputed, and regression imputation where the regression-predicted value is imputed. For handling missing categorical data Decision trees are used. The typical aspect of the prediction rule based approach is that it depends on a limited number of variables.

b. Maximum likelihood

The maximum likelihood approach depends on a parametric model of data generation. A maximum likelihood method is used for proper model and imputation of the missing values. The expectation maximization (EM) algorithm [15], [16]. is implemented with multiple generation of successors for missing entries according to the proper probabilistic distribution, so that a missing data entry is imputed as the candidates' mean which is known as multiple imputation (MI) method [16],[17]. The maximum likelihood approach is based on a precise statistical model.

c. Least-squares approximation

Its build on approximation of the usable data with a below rank bilinear model similar to the singular value decomposition of a data matrix. The methods work linearly by creating one factor at a time to decrease the sum of squared differences between the usable data entries and those regenerated through bilinear modeling. To implement this approach:

(1) Searching an approximate data model using non missing data and then adding the missing data with data searched with the model.

(2) Begin with inserting all the missing data, then continuously approximate the completed data and update the imputed data which are hidden by the approximation. Jian Pei Ming Hua Yufei Tao Xuemin Lin [7] proposed the Technique of Query Answering on Uncertain
and Probable Data.

a. Query Types

In [19], Soliman et al. proposed 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. Thus, the tuples returned by a U-ranks query may not be a valid top- k tuple list in any possible world, and a tuple may appear more than once in the answer set.

b. Query Answering Methods

Yi et al. [20] proposed efficient 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. One of the fundamental ideas in those methods is to enumerate and prune possible answers systematically. For promising candidates, those methods (implicitly) search the possible worlds by estimating the probabilities of those promising candidates through considering the relationship between the candidates and other tuples. Eamonn Keogh, Chotirat Ann Ratanamahatana [8] proposed a technique for the exact indexing of DTW, where dynamic time warping (DTW) is a distance measure for time series. DTW does not follow the triangular inequality and thus has prevented experiments on exact indexing. So, researchers have introduced approximate indexing techniques and concentrated on speeding up sequential searches. The approaches to indexing time series under the Euclidean distance that guarantee no false dismissals use the GEMINI framework of Faloutsos et al. Using the GEMINI framework, all one has to do is to choose a high level representation of the data and define a lower bounding measure on it (Faloutsos et al. 1994). They introduces a indexing technique called piecewise aggregate approximation (PAA) (Keogh et al. 2000; Yi and Faloutsos2000). This technique is attractive because it is simple, intuitive, and competitive with the other more complex approaches. A K-NN algorithm is used to compute the exact K nearest neighbors of a query time series Q using a multidimensional index structure. Christos Faloutsos, M.ranganathan, Yannis Manolopoulos[1] used an indexing method to locate 1-dimensional subsequences within a collection of sequences such that subsequences match a given query within a specified tolerance. The aim is to map each data sequence into small set multidimensional rectangles in feature space. Mohamed E. K., et al., [21] tackled the issue 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 technique to retrieve the local skyline in every cluster. Iskyline method conducts two optimization techniques that reduce the number of local skyline in every cluster. However, Iskyline is time consuming as in each node there are many pairwise comparison need to be performed to find the local skyline. Most importantly, large amount of missing data in the skyline results does not give any insight to help user in selecting the most appropriate data item. Wei Cheng, Xiaoming Jin, Jian-Tao Sun, Xuemin Lin, Xiang Zhang, Wei Wang[22] proposed to find the problem of similarity search on dimensional incomplete data. A probability framework is designed to model the above problem for the users to search objects in the database which are similar to the query with probable guarantee. They designed two probabilistic strategies to reduce the search space by using the lower bound and upper bounds. These bounds enable efficient filtering of non-relevant data values without absolutely examining all possible combinations of missing dimension. A probability triangle inequality method is also used to reduce the search area and speed up the query process. This framework can be with whole queries and subsequence queries. The overall query process is shown in Fig. 1. The probability triangle inequality is first applied to evaluate the data objects. In this step, some data objects are judged as true results and some are filtered out. The bounds of the probability are then applied to evaluate the remaining data objects, from which some are determined as true results and some as dismissals. Only those data objects that cannot be determined in the former two steps are evaluated by the naive method.

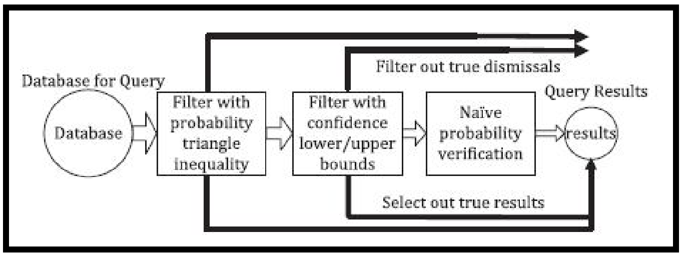

Fig. 1. The overall query process of the proposed approach.

4. CONCLUSION AND FUTURE WORK

This paper is based on missing data values. In this we summarized different methods for similarity query search in incomplete database. The techniques presented are easy to work and allow effective indexing of missing data and exhibit linear performance for query execution time with database and query dimensionality. The methods are used to estimate a probability density function to model the uncertainty in data. It also addresses the similarity query problem on dimension incomplete data, which is of both practical importance and technical challenge. A probability framework is proposed to where the probability bounds and the probability triangle are used to increase efficiency and decrease the search space. In future will study probability framework to address this problem and extend the work for subsequence matching in dimension incomplete database and use different indexing techniques to improve the efficiency of search and explore various distance functions

References

[1] 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.

[2] R. Cheng, D.V. Kalashnikov, and S. Prabhakar, “Evaluating Probabilistic Queries over Imprecise Data,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’03), pp. 551-562, 2003.

[3] M. Hua, J. Pei, W. Zhang, and X. Lin, “Ranking Queries on Uncertain Data: A Probabilistic Threshold Approach,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’08), pp. 673- 686, 2008.

[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] 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.

[7] J. Pei, B. Jiang, X. Lin, and Y. Yuan, “Probabilistic Skylines on Uncertain Data,” Proc. 33rd Int’l Conf. Very Large Databases (VLDB ’07), pp. 15-26, 2007.

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

[9] R.A. Little and D.B. Rubin, Statistical Analysis with Missing Data, Wiley Series in Probability and Statistics, first ed., pp. 2-278. John Wiley & Sons, 1987.

[10] T. Mathew and K. Nordstrom, “Inequalities for the Probability Content of a Rotated Ellipse and Related Stochastic Domination Results,” The Annals of Applied Probability, vol. 7, no. 4, pp. 1106- 1117, 1997.

[11]. R.little,D.Dubbin.Statistical analysis with Missing Data values.Wiley series in prob. & stat.,1987

[12]P. Davies, P. Smith, Model Quality Reports in Business Statistics, ONS, UK, 1999.

[13] N. Kenney, A. Macfarlane, Identifying problems with data collection at a local level: survey of NHS maternity units in England, British Medical Journal 319 (1999) 619–622.
[14] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Hastie, R. Tibshirani, D.Botsein, R.B. Altman,"Missing value estimation methods for DNA microarrays," Bioinformatics17 (2001) 520–525.
[15] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society 39 (1977) 1–38.
[16] J.L. Schafer, "Analysis of Incomplete Multivariate Data," Chapman and Hall, 1997.
[17] D.B. Rubin, Multiple Imputation for Nonresponse in Surveys, John Wiley & Sons, 1987.
[18] R. Pearson. Mining Imperfect Data: Dealing with Contamination and Incomplete Records. SIAM, 2005.
[19] M. A. Soliman, I. F. Ilyas, and K. C.-C. Chang. Top-k query processing in uncertain databases. In Proceedings of the 23nd International Conference on Data Engineering (ICDE'07), Istanbul, Turkey, April2007. IEEE.
[20] 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.
[21] 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.
[22] Wei Cheng, Xiaoming Jin, Jian-Tao Sun, Xuemin Lin, Xiang Zhang, and Wei Wang. Searching Dimension Incomplete Databases. IEEE Trans. Knowledge and Data Eng., vol. 26,no.3,March 2014
[23] 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.

AUTHOR

Namrata Pagare received the B.E. degree in Information Technology Engineering from SandipInstitute 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.

Prof.Jyoti Mankar, Assistant Professor, Department of Computer Engineering, K. K. Wagh Institute of Engineering Education & Research, Nashik, India.