ISSN (ONLINE) : 2395-695X
ISSN (PRINT) : 2395-695X
Available online at www.ijarbest.com
International Journal of Advanced Research in Biology, Ecology, Science and Technology (IJARBEST)
Vol. 1, Issue 9, December 2015

PROBABILISTIC SIMILARITY QUERY BASED SEARCHING IN INCOMPLETE DATABASES

P.Senthilraja[1], M.Kalidass2
Assistant Professors, Department of Computer Science and Engineering,
R.V.S Educational Trust's Group of Institution's, Dindigul, India
senthilrajamtech@gmail.com
kalidassme@gmail.com

***Abstract*-Databases with missing data occur in a wide range of research and industry domains. 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 missing of data is random. A probabilistic framework model is developed to investigate the problem of similarity search on dimension incomplete data. Using the framework users can get objects in the database that are similar to the query. Develop the lower and upper bounds of the probability that a data object is similar to the query. It enable efficient filtering of irrelevant data objects without explicitly examining all missing dimension combinations. we present a probability triangle inequality that can also be used to effectively prune the search space.**

Index terms-Dimension incomplete database, similarity search, whole sequence query.

I.INTRODUCTION

Similarity query in multidimensional 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. Recently, querying incomplete data has attracted extensive research efforts . 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. 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. 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 refer to such a problem as the dimension incomplete problem.

Data missing when dimension information is not explicitly maintained. Consider the sensor networks. The database usually contains time series data objects, each of which is represented by a sequence of values.. The dimension information associated with data values can be implicitly inferred from the data arrival order. This schema of data collection and storage is very common in resource-constrained applications because explicitly maintaining dimension information will cause additional costs. In this problem setting, missing a single data element will destroy the dimension information of the entire data object. In applications where dimension information is explicitly maintained, the dimension indicator itself may be lost. This will also cause the dimension incomplete problem.

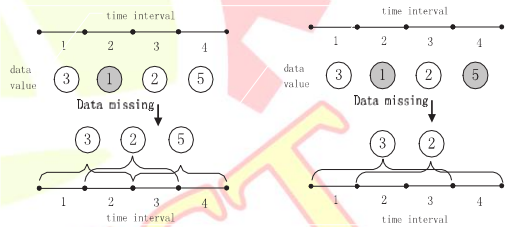

Fig.1.dimensionincomplete data due to dimension information not being explicitly maintained.

For example, in Fig. 1, the original data object is (3, 1, 2, 5). When data element 1 is missing, the dimension information for the rest of data elements becomes uncertain. For example, 3 can be the first or the second element, and 2 can be the second or the third element. When data elements 1 and 5 are missing, then both elements 3 and 2 may locate on three different dimensions. In applications where dimension information is explicitly maintained, the dimension indicator itself may be lost. This will also cause the dimension incomplete problem.

Time series data with temporal uncertainty due to imprecise time stamps. The imprecise time stamps due to granularity mismatch or data collection from distributed system that is lack of clock synchronization may also cause dimension incompleteness in time series data. For example, when time series data are collected from distributed environment, due to the lack of clock synchronization, each collected data value is

ISSN (ONLINE) : 2395-695X
ISSN (PRINT) : 2395-695X
Available online at www.ijarbest.com
International Journal of Advanced Research in Biology, Ecology, Science and Technology (IJARBEST)
Vol. 1, Issue 9, December 2015

assigned an uncertainty interval on the time line.

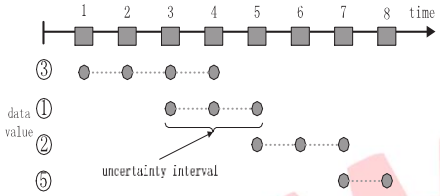

Fig. 2 shows a time series stream of four values 3, 1, 2, 5 and their un-certainty intervals on the time line. We may approximately infer the arrival order of the values as (3, 1, 2, 5). But the precise occurrence time of each element is uncertain. In real-world applications, we may want to automatically identify patterns(or say subsequences) satisfying given query condition on certain dimension incomplete time series data. This poses the subsequence matching problem on dimension incomplete data.

The dimension incomplete problem brings new challenges to the similarity query task, since the dimension information is essential for the existing uncertain data querying methods. In this paper, we formulate the problem of similarity query on dimension incomplete data within a probabilistic framework. Using the framework, a user can specify two thresholds: a threshold of the distance between the query object and the data object, and a threshold of the probability that the retrieved data objects are similar to the query object. An efficient method is developed to compute the lower and upper bounds of the probability that a data object is similar to the query. These bounds can be utilized to effectively prune the search space. Moreover, we develop a probability triangle inequality that can further speed up the query process. Our contributions are summarized as follows:

1. To the best of our knowledge, this is the first work to address the dimension incomplete similarity query problem. This problem has a wide range of applications and poses new technical challenges to traditional query methods. We propose a probabilistic framework to model and manipulate the uncertainty of the data. We also provide theoretical analysis of its computational properties.

2. We develop efficient algorithms to address the challenges in querying dimension incomplete data. The complexity of a naive approach to compute the probability that a data object is similar to the query object is O (m.($^{m}_{n}$)), An efficient method with time complexity O(n.(m-n)^2) is proposed to compute the lower bound and upper bound of the probability. These bounds can be utilized to prune the search space. We further develop a probability triangle inequality, which can be evaluated in O(m) time and used as a filtering tool to further speed up the query process.
3. Our method can be applied to both whole sequence matching and subsequence matching problems on dimension incomplete data. Moreover, the data of interest can be either static data or dynamical data streams.

4. We provide theoretical analysis of the relationship between the probability threshold and the quality of query results. This provides guidance for the users to effectively determine the probability threshold according to their search quality preference.
5. We conduct extensive experiments on real-life data sets. The results demonstrate the effectiveness and efficiency of our method.

II.Related Work

2.1Missing Data.

Although databases commonly deal with or contain missing data, relatively little work has been performed for this topic. Formal definitions for imperfect databases, of which databases with missing data is a subset, and database operations are provided in . Two techniques for indexing databases with missing data are introduced and evaluated in . This is the only paper we are aware of that focuses on indexing missing data. These are the bitstring augmented method and the multiple one-dimensional one-attribute indexes technique, called MOSAIC.

For the bit string augmented index, the average of the non-missing values is used as a mapping function for the missing values. The goal is to avoid skewing the data by assigning missing values to several distinct values. However, by applying this method it becomes necessary to transform the initial query involving k attributes into 2*k* sub queries, making the technique infeasible for large k.

MOSAIC is a set of B+-Trees were missing data is mapped to a distinguished value. Similarly to the previous method, it becomes necessary to transform the initial query involving k attributes into 2*k* sub queries, one for each attribute. What makes MOSAIC perform better than the Bit string Augmented index for point queries is that it uses independent indices for each dimension. However, by using several B+-Trees the query has to be decomposed and intersection and union operations need to be performed to obtain the final result. Queries that could gain a greater performance benefit by utilizing multiple-dimension indexes would not achieve it using this technique. Therefore, this method may not be useful for multiple-dimension range queries, or other queries where the number of matches associated with a single dimension is high.

Our work differs from in that we introduce and evaluate techniques that do not suffer the same weaknesses as their techniques. In our approach the query need not be transformed into exponential number of queries and no extra expensive computation, such as set operations, needs to be performed in order to obtain the final result set. Moreover, even though VA-File is not a hierarchical index it benefits from pruning multiple dimensions in one pass through the structure. In addition, our solution using bitmaps and VA-Files is also scalable

ISSN (ONLINE) : 2395-695X
ISSN (PRINT) : 2395-695X
Available online at www.ijarbest.com
International Journal of Advanced Research in Biology, Ecology, Science and Technology (IJARBEST)
Vol. 1, Issue 9, December 2015

with respect to the data dimensionality.

2.2 VA-Files.

The motivation for VA-files is introduced in theoretical limitations for the classes of data and space partitioning indexing techniques with respect to dimensionality. Since reading all database pages becomes unavoidable when the number of indexed dimensions is high, the authors suggest reading a much smaller approximate version, or vector approximation (VA), of each record in the database. An initial read approximately answers queries, and actual database pages are read to determine the exact query answer. To the best of our knowledge this is the first paper that compares and contrasts bitmaps and VA-files and discusses them together and the first paper in which these techniques are used to index incomplete databases.

III.PROBLEM DEFINITION

Incomplete databases, that is, databases that are missing data, are present in many research domains. It is important to derive techniques to access these databases efficiently. We first show that known
indexing techniques for multi-dimensional data search break down in terms of performance when indexed attributes contain missing data. This paper utilizes two popularly employed indexing techniques, bitmaps and quantization, to correctly and efficiently answer queries in the presence of missing data. Query execution and interval evaluation are formalized for the indexing structures based on whether missing data is considered to be a query match or not. The performance of Bitmap indexes and quantization based indexes is evaluated and compared over a variety of analysis parameters for real and synthetic data sets. Real world applications using databases with missing data are common. Databases with missing data occur in a wide range of research and industry domains. Some
examples of these are:
1. A census database that allow null values for some attributes
2. A survey database where answers to one question cause other questions to
be skipped 3. A medical database that relates human body analyze (a substance that can be measured in the blood or urine) measurements to a number of diseases, or patient risk factors to a specific disease
The goal of this paper is to provide techniques that access databases efficiently in the presence of missing data. There are a variety of reasons why databases may be missing data. 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. 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. In that case, the missingness is ignorable and the way of dealing with it is to "complete" the value using regression or other statistical model and treat the data as if it was never missing. However, 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.

IV.METHODOLOGY

Data mining

Data mining is also known as knowledge discovery in database. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. This process of extracts or mines large amount of data. It is the natural evaluation of information technology. In data mining concern is about the development of methods and techniques to identify sense of data. The data mining tools appeared with the intension of facilitating data analysis and visualization as well as the discovery of useful information for decision making. Data mining task is classified into two categories: descriptive and predictive. Descriptive mining describe the general properties of data in database.
Predictive mining perform inference on the current data to make predictions.

V.IMPLEMENTATION

The project Searching Dimension Incomplete Database, implementation of project is to find out the missing values by certain method. In Dimension the collection of data values can be stored and viewed. The admin is the one who is the master of the current project. Log on is the procedure used to get access to an operating system or application, usually in a remote computer. In the Admin set will be update the sensor values session by session. And the sensor data filled related to the database. If any values are missed, the database get matched with the previous database, It check the database and filled the value from the database. The number of data sets are calculated as dimensions.

A subset S of a partially ordered set P may fail to have any bounds or may have many different upper and lower bounds. By transitivity, any element greater than or equal to an upper bound of S is again an upper bound of S, and any element less than or equal to any lower bound of S is again a lower bound of S. This leads to the consideration of least upper bounds and greatest lower bounds. The bounds of a subset S of a partially ordered set K may or may not be elements of S itself. If S contains an upper bound then that upper bound is unique and is called the greatest element of S. The greatest element of S (if it exists) is also the least upper bound of S.

A special situation does occur when a subset is equal to the set of lower bounds of its own set of upper bounds. This observation leads to the definition of Dedekind cuts. 5 is a lower bound for the set { 5, 10, 34, 13 , 9 , 42 }, but 8 is not. 42 is both an upper and a lower bound for the set { 42 }, all other numbers are either an upper bound or a lower bound for that set. Every subset of the natural numbers has a lower bound, since the natural numbers have a least element (0, or 1

ISSN (ONLINE) : 2395-695X
ISSN (PRINT) : 2395-695X
Available online at www.ijarbest.com
International Journal of Advanced Research in Biology, Ecology, Science and Technology (IJARBEST)
Vol. 1, Issue 9, December 2015

depending on the exact definition of natural numbers). An infinite subset of the natural numbers cannot be bounded from above. An infinite subset of the integers may be bounded from below or bounded from above, but not both. An infinite subset of the rational numbers may or may not be bounded from below and may or may not be bounded from above. Every finite subset of a non-empty totally ordered set has both upper and lower bounds.

VI.SYSTEM ARCHITECTURE

The overall query process is the probability triangle inequality is first applied to evaluate the data object. In this step some data objects are judged as true results and some are filtered out. The lower and upper bounds of the probability are then applied to evaluate the remaining objects from which some are determined as true results and some as dismissals.

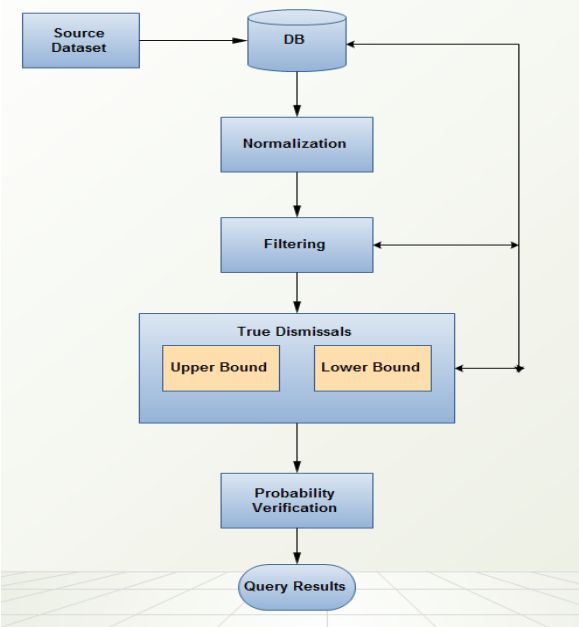

Fig 3. System Architecture

CONCLUSION

Querying incomplete data addresses the similarity query problem, which is of both practical importance and technical challenge. A probability framework is proposed to model this problem. To solve this problem efficiently, the lower and upper probability bounds and the probability triangle inequality that can be used to dramatically prune the search space. Furthermore, the similarity query framework is extended to tackle subsequence matching in dimension incomplete data. For a query Q and a dimension incomplete data object Xo, the brute force method is of complexity. Our method achieves a significant improvement: most data objects can be handled or even time. This approach achieves satisfactory performance in querying dimension incomplete data for both whole sequence matching and subsequence matching. Both the probability triangle inequality and the probability bounds have a good pruning power and improve query efficiency significantly.

FUTURE WORK

Future work will focus on the following directions. Since a probability triangle inequality holds, so plan to develop an index structure that can utilize the inequality to further improve the efficiency of the query process. Furthermore, plan to investigate how to extend our query strategy to incorporate a wide range of distance function.

REFERENCES

[1] B. Bollobas, G. Das, D. Gunopulos, and H. Mannila, "Time-Series Similarity Problems and Well-Separated Geometric Sets," Proc. 13th Ann. Symp. Computational Geometry (SCG '97), pp. 454-456, 1997.
[2] Beng Chin Ooi, Cheng Hian Goh , and Kian-Lee Tan , " Fast High-Dimensional Data Search in Incomplete Databases," Proc VLDB '98 Proceedings of the 24rd International Conference on Very Large Data Bases ,pp. 357-367 , 1998
[3] D. Burdick, P.M. Deshpande, T.S. Jayram, R. Ramakrishnan, and S. Vaithyanathan, "Olap over Uncertain and Imprecise Data," Proc. Int'l Conf. Very Large Databases (VLDB '05), pp. 970-981, 2005
[4] D. Gu and Y. Gao, "Incremental Gradient Descent Imputation Method for Missing Data in Learning Classifier Systems," Proc. Workshops Genetic and Evolutionary Computation (GECCO '05), pp. 72-73, 2005
[5] 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
[6] H. Zhang, Y. Diao, and N. Immerman, "Recognizing Patterns in Streams with Imprecise Timestamps," Proc. VLDB Endowment, vol. 3, pp. 244-255, 2010
[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] J. Pei, M. Hua, Y. Tao, and X. Lin, "Query Answering Techniques on Uncertain and Probabilistic Data: Tutorial Summary," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '08), pp. 1357-1364, 2008
[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] R.K. Pearson, "The Problem of Disguised Missing Data," ACM SIGKDD Explorations Newsletter, vol. 8, pp. 83-92, 2006.

ISSN (ONLINE) : 2395-695X
ISSN (PRINT) : 2395-695X
Available online at www.ijarbest.com
International Journal of Advanced Research in Biology, Ecology, Science and Technology (IJARBEST)
Vol. 1, Issue 9, December 2015

Mr. P.Senthilraja was born on 15.05.1985. He did his B.E – CSE Sethu institute of technology Anna University in 2006 and M.Tech (Computer & Information Technology) Manonmaniam sundaranar university in the year 2012. He is currently working as Assistant Professor in Department of Computer Science Engineering in RVS Educational Trust's Group of Institutions.

Mr. M.Kalidass was born on 30.07.1984 . He did his B.E – CSE in Maharaja prithivi Engineering College, Anna University in 2008 and M.E(Computer Science Engineering) in Nandha Engineering collegein the year 2010. He is currently working as Assistant Professor in Department of Computer Science Engineering in RVS Educational Trust's Group of Institutions.