International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

To Investigate the Problem of Similarity Search on Dimension Incomplete Data

Amol Patil, Ashwini Sagade

[1, 2]M.E. Student, Computer Engineering Department, IOKCOE, SP Pune University, India

Abstract: ***Similarity query in multidimensional database is a fundamental research problem with numerous applications in the areas of database, data mining, and information retrieval. The existing work on querying incomplete data addresses the problem where the data values on certain dimensions are unknown. Missing dimension information poses great computational challenge, since all possible combinations of missing dimensions need to be examined when evaluating the similarity between the query and the data objects. We develop the lower and upper bounds of the probability that a data object is similar to the query. These bounds enable efficient filtering of irrelevant data objects without explicitly examining all missing dimension combinations. A probability triangle inequality is also employed to further prune the search space and speed up the query process. The proposed probabilistic framework and techniques can be applied to both whole and subsequence queries. Extensive experimental results on real-life data sets demonstrate the effectiveness and efficiency of our approach.***

Keywords: Missing Dimensions, Similarity search, Whole sequence query, Probability triangle inequality, Temporal data.

1. Introduction

Recently, querying incomplete data has attracted extensive research efforts [1], [2], [3]. In this problem, the data values may be missing due to various practical issues. 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 [4], [5]. 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. Now day amount of data are retrieved in data mining is large and information retrieval. Data which is retrieved is mostly not complete the querying incomplete data has attracted greater attention as it poses new challenges to traditional querying techniques [6].

The existing work on querying incomplete data addresses the problem where only the data values on certain dimensions are unknown. In many applications, such as data collected by a sensor network there a noisy environment, so not only the data values but also the dimension information may be missing. The existing work is based on assumption that is for each dimension whether its data value is missing or not is know. We want to investigate the problem of similarity search on dimension incomplete data. We have to reduce the number of compares among similar data. Missing dimension information poses great computational challenge, since all possible combinations of missing dimensions need to be examined when evaluating the similarity between the query and the data objects. We have developed the lower and upper bounds of the probability that a data object is similar to the query. These bounds enable efficient filtering of irrelevant data objects without explicitly examining all missing dimension combinations. Due to this search space is reduce and it speed up the query process.

This paper is organized as follows: Section II consists of literature survey of existing work with comparisons. Some background details and formulation is given in section III. Section IV describes proposed system with mathematical models. Section V describes conclusion and finally future work.

1.1 Data incomplete due to dimension information is not explicitly maintained:

Consider the data send over sensor networks. The database usually contains time series data objects each data represented by a sequence of values (d1, d2, d3…………,dn). The dimension information associated with data values can be implicitly inferred from the data arrival order. Data arrival order store using time stamp. This schema of data collection and storage is very common in resource-constrained applications since 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.

1.2 Data dimension missing due to lack of clock synchronization

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.

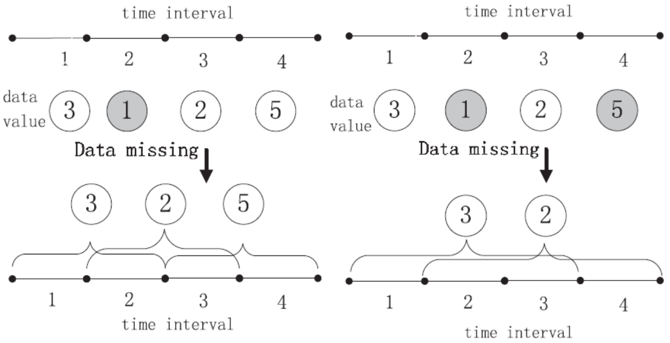

Figure 1: Missing Dimensions

2. Literature Survey

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 missing value of data is random. of some value does not depend on the value of another variable. Analysing uncertain data is an area that is also related to our work [7], [8]. The goal of these methods is to estimate a probability density function to model the uncertainty in the data. They do not address the problem of dimension incompleteness.

The authors [9] address the problem where there are missing elements in symbolic sequences. Our problem is more general in the sense that we consider real-value data and address the probabilistic query task. In [10], a temporal model is proposed to discover patterns in streams with imprecise time stamps. This work deals with pattern evaluations in event streams where event ids should be exactly matched. Moreover, the data arrival time intervals are needed to construct the temporal uncertainty model. In our work, such information is not available and only the data arrival order is known. To find common structure of two sequences, dynamic time warping [11], [12] and longest common subsequence [13], [14] algorithms are proposed. In these problems, the exact dimension information is not critical. These methods cannot be directly applied to the similarity query problem on dimension incomplete data.

3. Problem Statement

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 lower and upper bound of probability and manipulate the uncertainty of the data. We also provide theoretical analysis of its computational properties.

Let D be the database. A data object $X \epsilon D$ is a revalued vector $(x1, x2,...,xm)$, where $xi(1 \leq i \leq m)$ is the data value for the ith dimension of X. $|X| = m$ denotes the dimensionality of X. A data object X is dimension incomplete, if it satisfies 1) at least one of its data elements is missing; 2) the dimension of the missing data element is unknown. For example, given a complete data object X, if its k data elements are missing, the resulting dimension incomplete data object is of the form $Xo\ (xo1, xo2,... ; xon\ )$, where $oj < oj+1$, $n = |X| - k$. The traditional range query on a multidimensional database is defined as follows: Given a database D containing N data objects of m dimensions, an m-dimensional query Q, a distance function f, and a distance threshold r, traditional range queries retrieve all the data objects in D whose distances from Q are less than r.

The above problem formulation cannot be directly applied to dimension incomplete data because the distance between the query object and the dimension incomplete data object is not well defined. Intuitively, the distance between the query and the data objects depends on both the dimension alignment and the values estimated for the missing dimensions. In the following, we develop a probabilistic framework to formulate the query problem on dimension incomplete data. The goal is to find all data objects that have high probability to be similar to the query.

4. Proposed System

The overall query process is shown in Fig. 2. 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 lower and upper 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.

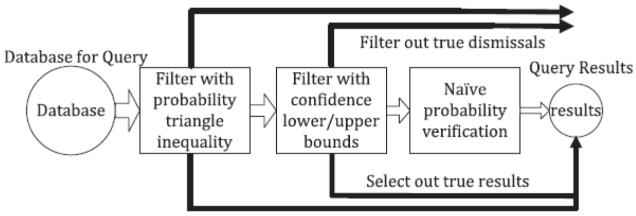
Database for Query
Database
Filter with probability triangle inequality
Filter with confidence lower/upper bounds
Naïve probability verification
results
Filter out true dismissals
Query Results
Select out true results

Figure 1: System Design

4.1 Bounds of the Probability

The triangle inequality and the bounds of probability can be evaluated efficiently and used to effectively prune the search space. Only a small portion of the data objects need to be evaluated by the naive verification algorithm. To further increase the efficiency, we can avoid the naïve verification step and simply treat the remaining candidates as query results (or dismissals, depending on the requirements of query precision and recall).

This is reasonable for applications where the two probability bounds are effective for selecting candidates. This simplified strategy will dramatically increase the efficiency of the algorithm without causing significant change of the quality of the results.

4.2 Probability Triangle Inequality

In addition, in our algorithm for subsequence matching on dimension incomplete data, the data of interest can be either static data or dynamical data streams. A special case of this problem would be monitoring the data streams, where the sequence in database is appended over time, and once the sequence is updated, we examine if the newest part of the sequence is similar to a user specified query.

It can be found that our method for subsequence matching can also tackle this problem, with only a little modification required. Specifically, when the sequence is updated, we only have to do the two steps on the new added part of the sequence. If we find the matching patterns that meet the query requirements, the system will output the matched sub-sequences.

5. Conclusion

The results indicate that, 1) Our approach achieves satisfactory performance in querying dimension incomplete data for both whole sequence matching and subsequence matching; 2) Both the probability triangle inequality and the probability bounds have a good pruning power and improve query efficiency significantly; Our future work will focus on the following directions. Since a probability triangle inequality holds, we plan to develop an index structure that can utilize the inequality to further improve the efficiency of the query process. Furthermore, we plan to investigate how to extend our query strategy to incorporate a wide range of distance functions.

6. Acknowledgment

I am Amol Patil and would like to thank the publishers, researchers for making their resources material available. I am greatly thankful to Prof. Saba Siraj for their guidance. We also thank the college authorities Prof. Pratap Singh, PG coordinator and Principal Dr. R.S. Jahagirdar for providing the required infrastructure and support. Finally, we would like to extend a heartfelt gratitude to friends and family members.

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] Ankerst, B. Braunmller, H.-P. Kriegel, and T. Seidl, "Improving Adaptable Similarity Query Processing by Using Approximations," Proc. 24th Int'l Conf. Very Large Data Bases (VLDB '98),pp. 206-217, 1998.

[3] M 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.

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

[10] G. Navarro, "A Guided Tour to Approximate String Matching," ACM Computing Surveys, vol. 33, pp. 31-88, 2001.

[11] 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.

[12] 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.