IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 27-29
www.iosrjournals.org

A Review on Top-K Dominating Queries on Incomplete Data

Sreelekshmi B, Anoop S

[1](*Department of Computer Science, College of Engineering Perumon, Kollam, India)*
[2](*Department of Information Technology, College of Engineering Perumon, Kollam, India)*

Abstract: *Data mining is a powerful way to discover knowledge within the large amount of the data. Incomplete data is general, finding out and querying these type of data is important recently. The top-k dominating (TKD) queries return k objects that overrides maximum number of objects in a given dataset. It merges the advantages of skyline and top-k queries. This plays an important role in many decision support applications. Incomplete data holds in real datasets, due to device failure, privacy preservation, data loss. Here, we carry out a systematic study of TKD queries on incomplete data, which includes the data having missing dimensional value(s). We solve this problem, and introduce an algorithm for answering TKD queries over incomplete data. Our methods employ some methods, such as upper bound score pruning, bitmap pruning, and partial score pruning, to hike up the efficiency of queries. Extended experimental evaluation using both real and synthetic datasets shows the effectiveness of the developed pruning rules and confirms performance of algorithms.*

Keywords: *Algorithm , Dominance relationship, Incomplete data, Query processing, Top-k dominating query.*

I. Introduction

Data mining is a powerful new method to detect knowledge within the large amount of the data. Also data mining is the process of discovering meaningful new relationship, patterns and trends by passing large amounts of data stored in corpus, using pattern recognition technologies as well as statistical and mathematical techniques. Data mining sometimes called data or knowledge mining. Data are any facts, numbers, or sequence of characters that can be processed by a computer. Today, organizations are handling large and growing amounts of data in different structure and different databases.

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding reciprocity or patterns among lots of fields in large relational databases. Given a set S with d dimensional objects top k dominating queries ranks these objects base on the number of objects in S dominated by o, and returns k objects that dominates maximum number of objects. The TKD query identifies the most significant objects, and is a powerful decision making tool used to rank objects in real life applications. . In this paper, we take an incomplete dataset where some objects face the missing of attribute values in some dimensions, and study the problem of TKD query and processing over incomplete data. A TKD query on incomplete data returns k objects that dominates the maximum number of objects from a given incomplete data set.

TKD queries on incomplete data share some similarities with the skyline operator over incomplete data [1], because they both are based on the same dominance definitions. However, we would like to highlight that TKD queries on incomplete data have some advantages, i.e., its output is controllable by a parameter k, and hence, it is invariable to the scale of incomplete dataset in different dimensions. In addition to emphasize the dominance relationship definition on incomplete data ,is actually meaningful. We are developing the improved BIG (IBIG) algorithm by employing the bitmap compression techniques and the binning strategies for improving the efficiency for space in the TKD query over incomplete data. An efficient algorithm for processing

TKD queries on incomplete data, using several novel heuristics. We use an adaptive binning strategy with an efficient method for choosing the appropriate number of bins to minimize the space of bitmap index for IBIG. we propose the improved BIG (termed as IBIG) algorithm to efficiently address the storage issue by using the bitmap compression technique and the binning strategy. Specifically, the compression techniques are applied on the "vertical" bitsets, while the binning strategy compresses the bitmap index on the "horizontal" bitsets, i.e., for the bit-string of every object in the dataset. we introduce two most efficient and popular compression techniques.

II. Existing System

Data mining is the process of finding out knowledge from large amount of data stored in the database. Due to the availability of huge amount of data in electronic forms and turning such data into useful information and knowledge for broad application including business management and decision support in information industry in recent years. Data mining has a lot of benefits when using in a specific industry. Besides those sakes, data mining also have its own disadvantages. Data mining do good in business, society, governments as well as the individual. However privacy, security and misuse of information are the big problems, if they donot addressed and resolved properly. Data mining predicts future trends and customer purchase habits , it also helps in decision making and market basket analysis. The cons of datamining are privacy and security, great cost at implementation stage and possible misuse of information.

Xuemin Lin, extended the well-known skyline analysis to uncertain data, and developed two algorithms to tackle the problem of calculating probabilistic skylines on uncertain data using real and synthetic data sets. The Author, W.T.Balke, proposed Distributed Web Information services like [5] or [2] are premium examples benefiting from our contributions, they presented a first algorithm that allows to retrieve the skyline over distributed data sources with basic middleware access techniques and have well tried that it features an optimal complexity in terms of object accesses. To overcome the deterioration for higher numbers of lists he also proposed an efficient sampling technique to estimate the size of a skyline by assessing degree of data correlation.

The Author, M E Khalifa [4] (ET. Al) Skyline queries aim to prune search space of large numbers of multi-dimensional data to a small set of interesting items by eliminating items that are dominated by others. Existing skyline algorithms assume that all dimensions are available for all the data items. This paper goes beyond the restrictive assumption as we address the more practical case involving incomplete data items.

X.Miao proposed an efficient probabilistic skyline query process on uncertain data streams. As data volume continuously grows its quality may not be high as in usual cases. The data can be defected ,cannot be precise or inaccurate due to the process called data acquiring. The skyline query is widely used for data analysis and to derive the results that meets more than in specific conditions simultaneously.

M.Kontaki The Continuous Top-k Dominating query(cTKDQ) method have some limitations. In order to overcome the existing cons, introduces a new indexing structure known as close dominance graph(CDG) to support and maintain the relationship between dynamic data records. However, CDG takes more time to search results. In this paper they introduce a dictionary based compression algorithm, which was efficient in answering cTKDQ with minimum time and memory. Papadias et al. [5] first introduce the top-k dominating query as a variation of skyline queries, and they present a skylinebased algorithm for processing TKD queries on the traditional complete dataset indexed by an R-tree.

Yiu andMamoulis [6], [7] propose two approaches based on the aR-tree to tackle the TKD query. More recently, some new variants of TKD queries are studied, including subspace dominating query continuous top-k dominating query metric-based top-k dominating query top-k dominating query on massive data.

Gao et al. [2] propose an efficient kISB algorithm for processing k-skyband queries over incomplete data. Lofi et al. present an approach to compute the skyline using crowd-enabled databases with the challenge of dealing with missing information in datasets.

III. Proposed System

Our proposed UBB algorithm limits the size of candidate set by utilizing upper bound score pruning technique for the TKD query on incomplete data. However, the upper bound score may be rather loose, thereby we have to derive the real scores for many objects (even the whole dataset) via exhaustive pair comparisons, which degrades search performance significantly. Thus, an efficient score computation method is in demand. As a solution, we introduce a newly proposed bitmap index on incomplete data and propose the bitmap index guided algorithm to solve the TKD query on incomplete data. we propose the improved BIG (termed as IBIG) algorithm to efficiently address the storage issue by using the bitmap compression technique and the binning strategy. Specifically, the compression techniques are applied on the "vertical" bitsets, while the binning strategy compresses the bitmap index on the "horizontal" bitsets, i.e., for the bit-string of every object in the dataset. we introduce two most efficient and popular compression techniques, i.e., Word Aligned Hybrid

(WAH) and Compressed 'n' Composable Integer Set (CONCISE) to compress the bitmap index vertically. In this paper, we choose CONCISE instead of WAH. This is because, as shown in CONCISE has better compression ratio than WAH, and its computational complexity is comparable to that of WAH. We also demonstrate that CONCISE does perform better than WAH via an empirical evaluation. IBIG consumes less storage is its advantage.

IV. Conclusion

In this paper, it study the problem of the TKD query on incomplete data where some dimensional values are missing in the dataset. To efficiently address this, we first propose ESB and UBB algorithms, which utilize novel to prune the search space. In order to further reduce the cost of score computation, we present BIG algorithm, which employs the upper bound score pruning, the bitmap pruning and fast bitwise operations based on the bitmap index to improve the score computation and boost query performance accordingly. In order to trade the efficiency for space, we propose IBIG algorithm by using the bitmap compression technique and the binning strategy over BIG, and develop a method to choose the appropriate number of bins. Experimental results on both real and synthetic datasets confirm the effectiveness and efficiency of our algorithms

References

[1]. W. Zhang, X. Lin, Y. Zhang, J. Pei, and W. Wang, "Thre shold based probabilistic top-k dominating queries," The Int. J. Very Large Data Bases, vol. 19, no. 2, pp. 283–305, 2010.
[2]. M. Kontaki, A. N. Papadopoulos, and Y. Manolopoulos, "Continuous top-k dominating queries," IEEE Trans. KnowlData Eng., vol. 24, no. 5, pp. 840–853, May 2012.
[3]. D. Papadias, Y. Tao, G. Fu, and B. Seeger, "Progressive skyline computation in database systems," ACM Trans. Database Syst., vol. 30, no. 1, pp. 41–82, 2005.
[4]. M. L. Yiu and N. Mamoulis, "Efficient processing of top-k dominating queries on multi- dimensional data," in Proc. 33rd Int. Conf.Very Large Data Bases, 2007, pp. 483–494.
[5]. M. L. Yiu and N. Mamoulis, "Multi-dimensional top-k dominating queries," The Int. J. Very Large Data Bases, vol. 18, no. 3, pp. 695–718, 2009.
[6]. Y. Gao, X. Miao, H. Cui, G. Chen, and Q. Li, "Processing k-skyband, constrained skyline, and group-by skyline queries on incomplete data," Expert Syst. Appl., vol. 41, no. 10, pp. 4959– 4974, 2014.
[7]. M. E. Khalefa, M. F. Mokbel, and J. J. Levandoski, "Skyline query processing for incomplete data," in Proc. IEEE 24th Int. Conf. Data Eng., 2008, pp. 556–565.
[8]. L. Antova, C. Koch, and D. Olteanu, "From complete to incomplete information and back," in Proc.
[9]. SIGMOD Int. Conf. Manage. Data, 2007, pp. 713–724.