Top-K Dominating Queries On Incomplete Data : A Survey

Jilu Sajeev, Noorjahan V. A.

Department of Computer Science and Engineering, Ilahia College of Engineering & Technology, Muvattupuzha, India

ABSTRACT

Top-k dominating queries output the k objects that are dominating all other objects in a dataset. In most of the existing systems the dataset is assumed as complete. But in practical examples the dataset may be incomplete due to various reasons. In this paper a survey on various methods used to find the dominating objects from an incomplete dataset.

Keywords : Top-K Query, Dominance Relation, Skyline, Bucketing

I. INTRODUCTION

Top-k dominating queries combine the advantages of top-k queries and skyline queries. There are many works based on top-k dominating queries on complete data. But in real-time applications it is not necessary that the datasets are complete. The incompleteness means that some dimensions in the dataset are missing.

The reasons for incomplete dataset may be dataloss, privacy preservation and so on. For example, consider the object A from a dataset. The dimensions of A is (1, 7, -, 4).There is 4 dimensions for the object given and the dimension „–„ indicates a missing value.

When using this type of dataset it is difficult to find the top-k objects because some dimensions are missing so that they are incomparable with others. So it is important that how to find dominating elements from the incomplete dataset.

To output the dominating objects from a dataset first of all we need to define the dominance relationship in an incomplete dataset.

Definition :(dominance relationship on incomplete data [1]). Given two objects o and o' in a dataset S. o dominates o' (i.e., o < o') if the following conditions hold: I) for every dimension i,either o. [i] is less than o'.[i] or at least one of them is missing. II) there is at least one dimension j in which both o. [j] and o'. [j] are observed and o.[j] is less than o'.[j]. Consider an incomplete dataset given in fig 1, in which 4 objects are given with 5 dimensions for each object. In object A1 third dimension value is missing and also in all other objects we can see that some dimensions values are not available. While checking the dominance relationship between objects by the above definition first we need to compare A1 with A2.For each dimensions available in both A1 and A2, A2 dominates A1 so score of A2 becomes 1.

In this way comparing each objects with others we can find the score of the entire dataset elements.

A1 (4,5,-,6,8)
A2 (3,2,1,-,5)
A3 (5,-,8,9,1)
A4 (6,-,3,2,9)

Figure 1. A Sample Dataset

But in case of large dataset it is not possible to compare each elements become complex and time consuming. So there may be simple and speedy methods to find the dominant elements. This paper explains some previous works done on this subject.

II. METHODS AND MATERIAL

Related Works

This section includes some details about previous works related to Top-K dominating queries on incomplete data.

Gosta Grahne[11] gives detail description about incomplete information in relational databases and how to represent them. This paper describes how to use the incomplete dataset in various relations. The applications of the incomplete databases can be view updates, data integrations and data exchange etc.In some cases on incomplete databases some replacing or filling technologies are used.

That is the missing data may be filled using some assumptions or probabilities. Kalbhor swati and Gupta shyam in[12] gives an ARIMA based substitution for data collected by sensors. An ARIMA model is constructed and that model is used to refill the missing information. A differentiation between incomplete dataset and uncertain dataset needed here. In incomplete dataset some dimensional values will be missing, but in case of uncertain dataset the uncertainty is described based on probability values. A ranking query based approach is proposed in [10]. It deals with query results from a database which has uncertain values. In that proposes a baseline algorithm which explains linear extension tree concept. A SkyQUD framework is proposed in [9] uncertain and autonomous database. The methodology is explained through two phases harvesting and strict selection. Both phases have steps as distinctive partitioning, range reduction, probability dependency and probability breakdown.Luyi Mo and Reynold Cheng [13] study how to quantify the ambiguity of results from probability top-k query. It also address the cleaning of probabilistic database.

Two indexing schemes for fast high dimensional data search in incomplete database is described in [8].The first one is Bit stringaugmented R-tree (BR tree) multi-dimensional indexing structure, in which a query is decomposed into 2k sub queries .Second indexing is MOSAIC in which B+ tree is used for indexing. Another work on k dominant skylines on fast high dimensional data [14] is proposing various algorithm finding k dominant skylines and its variants. It includes One Scan algorithm, Two Scan algorithm and Sorted Retrieval algorithm.

For evaluating top-k queries on incomplete data the common technique used in various papers are skyline based approach. In the skyline approach basically steps as bucketing, local skyline etc are implemented. Bucketing means sort the data into different buckets based on the bit number of its dimension. That is if the item A(1,-, 5,6) is in the dataset its bit number will be 1011.Like this dataitems with same bit numbers will be added to the same bucket. Local skyline will be the dominating items from each bucket.A model for processing skyline queries on incomplete data is proposed in [5]. The proposed model have 4 components, data clustering builder, group constructor and local skylines identifier, k-dom skyline generator and incomplete skyline identifier. This method divides the database into different clusters grouping the data items in the clusters based on local skyline.

In evaluating top-k queries on incomplete data stream [6], two algorithms are proposed. Sorted List Algorithm (SLA) and Early Aggregation Algorithm (EAA) describe tracking top-k items over multiple data streams in a sliding window. Sort-based Incomplete Data Skyline algorithm (SIDS) [7] also uses skyline algorithm. In SIDS first the dataset is presorted in non-increasing order of each dimension, and then each dimension is choosed in round robin fashion for comparison. On each iteration the dominated items are removed from the set, at the end an item which is not removed and processed k times are returned.

Virtual Point based algorithm(VP) uses the bucketing concept of skyline technique and uses three concepts of virtual point, expired skyline and shadow skyline. It will help in reducing the processing of large dataset uses large buckets.K-skyband algorithm for incomplete dataset also uses the bucketing and skyline concepts[2].Skyline queries in [4] proposes two algorithms bucketing and replacement algorithm.In replacement algorithm int incomple value is replaced using infinity value.

In [1] four algorithms are proposed for finding top-k dominating elements from incomplete data. Extended Skyband based algorithm (ESB) uses the same skyline based approach used in [2].Another algorithm Upper Bound based(UBB) uses and a MaxScore value which is calculated for each dimension based on dominance. The third algorithm BIG(Bit map Index Guided) uses calculations based on bit map index and a MaxBitScore like MaxScore. Improved BIG algorithm uses a compression technique CONCISE to compress the bitmap index vertically and a binning strategy to cut down the bitmap storage consumption horizontally.

A restaurant recommendation system is implemented using preference query over incomplete information [3]. $SI2P$ Restaurant recommendation system have different interaction module like query submission, result explanation and dataset interaction. The user can submit query by specifying interest and constraints like region, price level etc for the restaurant. Query will be processed at the server and results will be returned. Users can write review about restaurant and rate them. Based on the rating the restaurant details will be updated.

At the server side the dataset is stored in PostgreSQL database. They integrate the PostgreSQL database by integrating two algorithms lksb[2] and UBB [1].Explaining the query will help the user to understand why an empty result set or mismatch occurred. This is an example of applying top-k queries on incomplete data.

III. CONCLUSION

This paper tries to go through different works related to Top-k Dominating queries on incomplete data. Top –k queries returns top elements from a dataset and it is very helpful in various realtime applications. Mainly skyline based approach is used in such cases. More methods have to be implemented to find top elements from incomplete dataset. This paper is not a complete reference but indenting to help students who are interested in researching on this topic and gives the brief idea of the same.

IV. ACKNOWLEDGEMENT

We would like to express our sincere thanks to all of the faculty members and students of Ilahia college of Engineering and technology for their immense support. Also we like to thank the friends and family and all others who give motivation and support for this work.

V. REFERENCES

[1] Xiaoye Miao, Yunjun Gaor "Top-k Dominating Queries on Incomplete Data", IEEE Transactions on Knowledge and Data Engineering,VOL. 28, NO. 1, January 2016.

[2] Yunjun Gao, Xiaoye Miao, Huiyong Cui Gang Chen, Qing Li, "Processing k-skyband, constrained skyline, and group by skyline queries on incomplete data", International Journal of Expert System with Applications, 2014.

[3] Xiaoye Miaoa,Yunjun Gao,"$SI2P$:A Restaurant Recommendation System Using Preference Queries over Incomplete Information", Proceedings of the VLDB Endowment, Vol. 9, No. 13,2016.

[4] Mohamed E. Khalefa, Mohamed F. Mokbel, Justin J. Levandoski,"Skyline Query Processing for Incomplete Data", DTC Digital Technology Initiative programme University of Minnesota,2006.

[5] Ali A. Alwan, Hamidah Ibrahim, Nur Izura Udzir, "A Model for Processing Skyline Queries over a Database with Missing Data", Journal of Advanced Computer Science and Technology Research, Vol.5 No.3, September 2015, 71-82.

[6] Parisa Haghani, Sebastian Michel, Karl Aberer," Evaluating Top-k Queries over Incomplete Data Streams ",2009 ACM 978-1-60558-512.

[7] Rahul Bharuka P, Sreenivasa Kumar," Finding Skylines for Incomplete Data ",Proceedings of the Twenty-Fourth Australasian Database Conference (ADC 2013), Adelaide, Australia.

[8] Beng Chin Ooi Cheng Hian Goh Kian-Lee Tan, "Fast High-Dimensional Data Search in Incomplete Databases ",Proceedings of the 24th VLDB Conference,USA.1998.

[9] Nurul Husna Mohd Saad, Hamidah Ibrahim, Ali Amer Alwan,Fatimah Sidi, Razali Yaakob, " A Framework for Evaluating Skyline Query over Uncertain Autonomous Databases", 14th International Conference on Computational Science, 2014.

[10] Mohamed A. Soliman, Ihab F. Ilyas, Shalev Ben-David," Supporting Ranking Queries on Uncertain and Incomplete Data".

[11] Gosta Grahne,"Incomplete Information",Department of Computer Science,Concordia university,Canada.

[12] Kalbhor swati, Gupta shyam,:" A Novel methodology for Searching Dimension Incomplete Database", International Journal of Computer Science and Information Technologies, Vol. 6 (1), 2015, 198-200.

[13] Luyi Mo, Reynold Cheng, Xiang Li, David W. Cheung, Xuan S. Yang:"Cleaning Uncertain Data for Top-k Queries",ICDE Conference 2013.

[14] Chee-Yong Chan,H.V Jagadish,Kian-Lee Tan:" "Finding kDominant Skylines in High Dimensional Space",SIGMOD 2006, June 27–29, 2006, Chicago.