International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

COPY RIGHT

IJIEMR Transactions, online available on 10th Febraury 2018. Link :
http://www.ijiemr.org/downloads.php?vol=Volume-7&issue=ISSUE-02

Title: Upper Bound Score Pruning Over Incomplete Data To Reduce Query Processing Time.

Volume 07, Issue 02, Page No: 165 – 174.

Paper Authors

***J.RADHAKRISHNA, MD.ISMAIL KHAN.**

* Eswar College of Engineering.

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

UPPER BOUND SCORE PRUNING OVER INCOMPLETE DATA TO REDUCE QUERY PROCESSING TIME

***J.RADHAKRISHNA, **MD.ISMAIL KHAN**
*PG Student, Eswar College of Engineering, Narasaraopet, Guntur, AP, India.
**Assistant Professor, Eswar College of Engineering, Narasaraopet, Guntur, AP, India.

ABSTRACT:

Incompleteness of data is a common problem in many databases including web heterogeneous databases, multirelational databases, spatial and temporal databases, and data integration. Merits of both Top-k Query and Skyline queries are put together to form a Top k dominating query. To solve this problem there are various algorithm for TKD queries on incomplete data upper bound score pruning, bitmap pruning, and binning strategy is used in order efficiently carry out TKD query processing. Weaims to find top elements from an incomplete dataset by providing priority values to each dimension in the data object. Skyline based algorithm is applied for that purpose. Since the priority value is used while determining the dominance this method return the most suitable and efficient result than other previous methods. The main column is enrollment vulnerability where each place with the database with a like LaHood, in the future called confidence. The methods are moreover examined, for example, upper bound score pruning, bitmappruning, and partial score pruning and so forth. The primary concentration is to support up the proficiency. Incomplete data exists in a wide range of genuine datasets, because of gadget disappointment, protection conservation, data misfortune, and etcetera. Thus, applying approximate query answering techniques, that are also typical for processing top-N queries in centralized database environments, seems to be the natural choice. We address this problem by presenting an approach that allows for reducing the number of queried peers as well as for giving probabilistic guarantees for the correctness of the answer.

Key words: Preference Queries, Incomplete Database, Query Processing, Estimating Missing Values. Top K, Extended Sky Band.

1. INTRODUCTION

Today, affiliations are managing gigantic and creating measures of data in different structure and particular databases [1]. Generally, data mining is the route toward separating data from exchange perspectives and abbreviating it into helpful data that can be used to addition salary, cuts costs, or both. Data mining is an able new procedure to recognize data inside the

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

enormous measure of the data. Additionally data mining is the path toward finding noteworthy new relationship, illustrations and examples by passing sweeping measures of data set away in corpus, utilizing outline affirmation developments and what's more true and numerical procedures [2]. There has been much interest on a new type of queries named skyline queries. Skyline queries prefer a data item p over the other data item q if and only if p is better than q in all dimensions and not worse than q in at least one dimension. The skyline queries are significant and mostly used in various application domains, like multi-criteria decision making applications decision support system and recommender system, where these systems combine various interests to help users to recommending a strategic decision [3]. The expected advantages of such a structure like robustness, scalability and self-organization come not for free: In a large-scale, highly dynamic P2P system it is nearly impossible to guarantee a complete and exact query answer. The reasons for this are among others possibly incomplete or incorrect mappings, data heterogeneities, incomplete information about data placement and distribution and the impracticality of an exhaustive flooding [4]. Top-K dominating query aims to find the top elements from a dataset. Movie recommendation system is a practical example for finding the top elements. That system will return the top movies from a set of movies based on ratings by different users. In real applications the datasets may be incomplete due to various reasons which will lead to uncertainty in data [5]. In along these lines, the deficient information show and the probabilistic thought are two techniques for dealing with missing information. It justifies saying that, differentiated and unverifiable information appear deficient information exhibit has one critical favorable position require any doubt on information relationship or, on the other hand prior learning we consider an insufficient dataset where a couple of items defy the missing of quality values in a couple of estimations, and study the issue of TKD address planning over insufficient information [6].

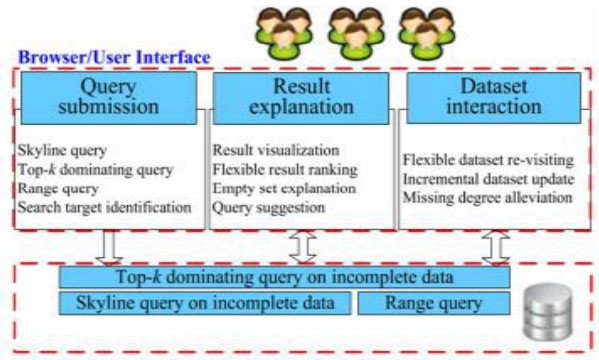

Fig1: System Flow

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

2. RELATED WORK

Information missing is a universal issue, and the investigation of inadequate information has pulled in much consideration [7]. There are numerous endeavors on displaying fragmented information table , the traditional rationale and modular rationale instruments for displaying and preparing fragmented information, display examinations for deficient information , I-SQL and world set variable based math dialect for deficient information and so on. Also, there are four normal list structures to list deficient information, in particular, bit string increased R-tree (BR-tree). Developing a suitable technique to manipulate an incomplete database is crucial as many real databases have missing attribute values and estimating them requires preprocessing before any operation on the database can be performed [8]. There are three different approaches that can be utilized in processing queries in incomplete database. The first approach retrieves exclusively the most relevant data items with no missing values. Index based algorithms as proposed require specialized index structures for processing skyline queries. These pre-computed indexes are designed for indexing data that is available locally and describe every queried dimension of each single data item. Due to the characteristics of P2P systems, dynamic behavior in particular, such index structures are not feasible [9]. The reasons for this are among others the need for global knowledge and high maintenance costs.There are various forms of top k dominating queries subspace dominating query that handles subset of dimensions in progressive manner, continuous top –k dominating queries over data streams, metric based top k dominating queries that processes top k dominating over distance-based dynamic attribute vectors, defined over metric space, top k dominating queries over massive data [10]. In the skyline approach basically steps as bucketing, local skyline is implemented. Bucketing means sort the data into different buckets based on the bit number of its dimension. Local skyline will be the dominating items from each bucket. A model for processing skyline queries on incomplete data is proposed. The proposed model components data clustering builder group constructor and local skylines identifier, k-dom skyline generator and incomplete skyline identifier [11]. This method divides the database into different clusters grouping the data items in the clusters based on local skyline.

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

3. SYSTEM ARCHITECTURE

We may need to highlight that TKD queries on insufficient information have an appealing great position, i.e., its yield is controllable by methods for a parameter k, and in this way, it is steady to the span of the divided dataset in different estimations [12]. We have to push the quality relationship definition on lacking information is truly vital. Take motion pictures m1 and m2 in the recommender framework. To the best of our knowledge, this is the central attempt to explore the TKD address on divided data. Regardless of the way that the TKD address over whole data or uncertain data has been particularly viewed as, TKD question get ready on lacking data still remains a noteworthy test [13]. To entirety things up, the key duties of this framework are consolidated as takes after. This formalizes the issue of TKD question in the particular situation of lacking data. To the extent anybody is worried, there is no prior work on this issue. This propose capable figurings for planning TKD queries on deficient data, using a couple of novel heuristics [14]. Framework demonstrates a flexible binning framework with a capable procedure for picking the appropriate number of repositories to limit the space of bitmap record for IBIG. This immediate expansive examinations using both honest to goodness what's more, made datasets to demonstrate the suitability of our made pruning heuristics and the execution of our proposed estimations [15].

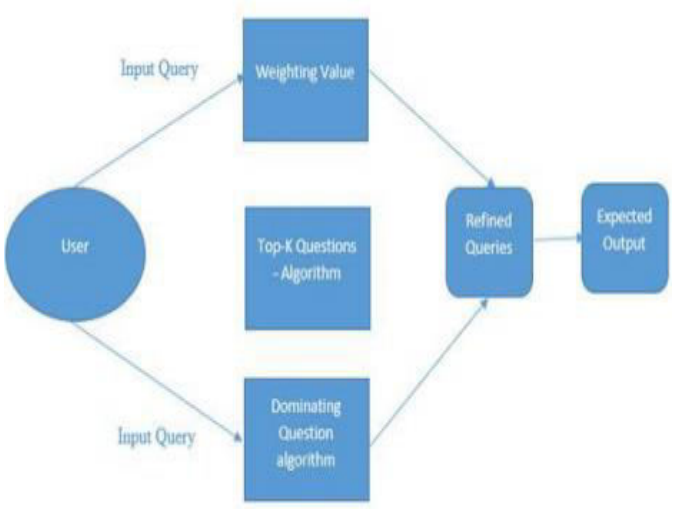

Fig2: System Architecture

4. PROPOSED SYSTEM

In this process there might be many estimated values for a dimension that need to be considered. Selecting the appropriate estimated value from several alternative values is based on the frequency of the value [16]. To gainfully address this, we initially propose ESB and UBB estimations, which utilize novel methodology to prune the interest space. With a particular true objective to also diminish the cost of score estimation, we appear Enormous estimation, which uses the upper bound score pruning, the bitmap pruning and fast bitwise

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

operations in perspective of the bitmap rundown to improve the score estimation besides, help request execution properly [17].

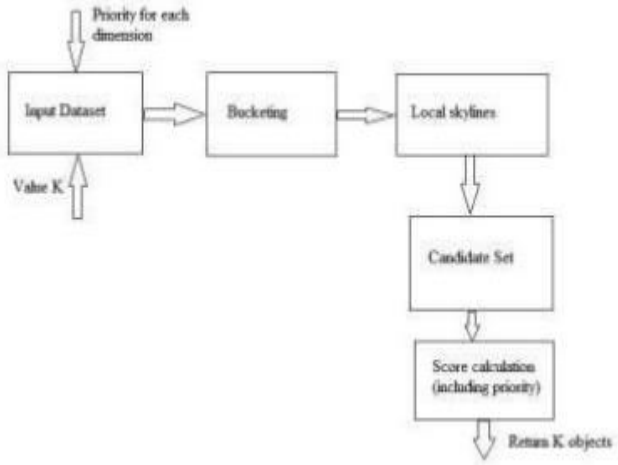

Fig 3: Architecture

A. Algorithm: Skyline Algorithm with Priority

Input: an incomplete data set S, a parameter K and Priorities for each dimension.

Output: The result set Sg.

1. Initialize Sc&Sg as 0.
2. for each object o ∈ S do
3. insert o into a bucket O based on βo (create if not exist)
4. for each bucket do
5. compare (,) for every , ∈ O
6. add K objects with highest score to the
7. add all o in
8. for each o ∈Sc
9. compare (,) for every , ∈Sc
10. add K objects objects in Sc having the highest score to the Sg
11. return Sg

The first step in the method is bucketing. In this step each object in the dataset is grouped into different buckets based on a bit number Compare function is based on the dominance relation. compare (,) returns the score for . Consider two objects A(1,-,2,1) and B(2,3,-,3) while considering the dominance between A and B , A dominates B since A have smaller value in all the available dimension than B. Then the score of A become one. In the score calculation stage if we have two objects O1(1,-,3,2) and O2(-,1,1,3). At the first dimension we have missing value in O2 and for second dimension in O1.So the third and fourth dimension determines the dominance. But at third dimension O2 dominates O1 and in fourth O1 dominates O2.According to the existing systems the dominance will not be valid in this case [18].

B. Ranking the Final Skylines

This section presents the last phase of the proposed approach for estimating missing values of skylines in incomplete database. This

phase emphasizes at ranking the skylines in a way such that the skylines with the estimated values that have the highest value of strength of probability correlations are placed at the top of the skyline set [19]. This is to ensure that the skylines are retrieved in the order of their precision to help users in selecting the most appropriate skyline. In addition, some skylines might have one dimension with estimated values, while other skylines might have estimated values in more than one dimension.

The final skylines each data item in the set of skylines, Sky, is analyzed (step 2). The highest value of the strength of probability correlations, p1 (step 3) and the number of dimensions with estimated values, m1 (step 4) of ai are determined. For each data item aj (step 5), the highest value of the strength of probability correlations, p2 (step 8), and the number of dimensions with estimated values, m2 (step 9) of aj are determined as well. If p1 is less than p2 and the number of dimensions with estimated values of ai is greater than aj (step 10), then swap ai with aj (step 11).

C. Algorithm: Ranking the final skylines algorithm.

Input: A set of skylines, Sky, and the values of the strength of probability correlations, P

Output: A set of ranked skylines

1. BEGIN
2. FOR each ai in Sky DO
3. Let p1 = the highest value of P of ai
4. Let m1 = the number of dimensions with estimated values of ai
5. FOR each aj in Sky DO
6. BEGIN
7. IF i <> j THEN
8. Let p2 = the highest value of P of aj
9. Let m2 = the number of dimensions with estimated values of aj
10. IF p1 < p2 AND m1 >= m2 THEN
11. Swap (ai , aj)
12. END
13. END

This phase emphasizes at ranking the skylines in a way such that the skylines with the estimated values that have the highest value of strength of probability correlations are placed at the top of the skyline set [20].

D. Advanced Routing Filters

Routing filters mainly represent a combination of local index structures, XML synopsis, and histograms. Local index structures have been shown to be suitable for efficient routing in P2P environments. Histograms are successfully used in a wide variety of optimization and

International Journal for Innovative Engineering and Management Research
A Peer Revieved Open Access International Journal
www.ijiemr.org

estimation techniques. So we decided to combine them resulting in the concept of routing filters that allow for approximating the distribution of attribute values. At each peer one filter is maintained for each established connection to a neighbor [21]. The filter describes all XML data that can be accessed by forwarding a query to the neighbor. Routing filters are built and maintained using a query feedback approach. If they are not limited to any niter horizon and if we assume that no peers leave the network, they will converge to global knowledge as time passes. Proposed the work on skyline query processing on incomplete data, skyline queries aim to prune the search space of large number of multidimensional data items to small set of interesting items by eliminating items that are dominated by other items [22]

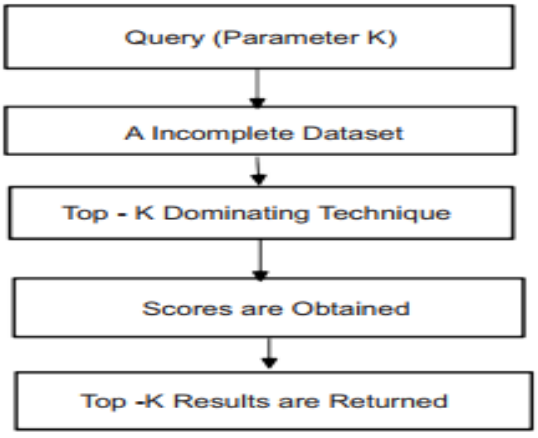

Figure 4: Flow of Execution

5. RESULT AND DISCUSSION

The result table shows the accuracy of TOP-K Dominant queries. Ranking of inquiry results is one of the crucial issues in information retrieval (IR), the logical order behind web indexes. Given an inquiry question and answer gathering of archives that match the question, the issue is to rank, that is, sort, the records in as indicated by some basis so that the best results seem ahead of schedule in the outcome list showed to the client. Traditionally ranking criteria are stated as far as significance of records as for information requires communicated in the question. First range is taken from the user(R) then calculate the count positive result generated as per the search (Pr). For each algorithm Precision=Pr/R and Recall=(R-Pr)/R.

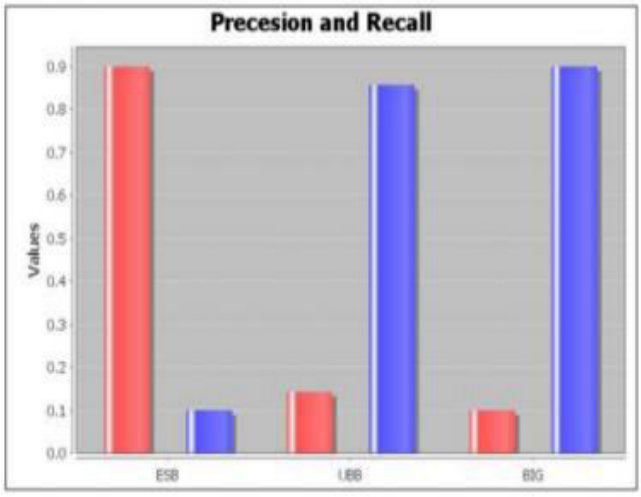

Fig .5.: Graphical Result

6. CONCLUSION

This system tries to experience diverse works identified with Top-k Dominating queries on incomplete information. Top –k queries returns top components from a dataset and it is extremely useful in different real time applications. The approach attempts to estimate values for the dimensions with missing values in the skylines. To the best of our knowledge, our approach is the first attempts in estimating missing values in the skylines. Skyline based algorithm with priority is used here for returning the top elements. This method can be used in systems like movie recommendation with user preferred priorities for more accurate outputs. wepropose IBIG computation by using the bitmap weight system and the binning philosophy over BIG, and develop a methodology to pick the fitting number of canisters. Critical exploratory happens on both bona fide and produced datasets certify the ampleness besides, viability of our presented heuristics and counts.

7. REFERENCES

[1] M. E. Khalefa, M. F. Mokbel, and J. J. Levandoski, “Skyline query processing for incomplete data,” in ICDE, pp. 556–565, 2008.

[2] YunjunGao, 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] XiaoyeMiaoa,Yunjun Gao,"□□2□: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, NurIzuraUdzir, "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] Bharuka R. and Kumar P., “Finding Superior Skyline Points from Incomplete Data,” Proceedings of the 19 thInternational Conference on Management of Data, Ahmadabad, pp. 35-44, 2013.

[7] Börzsönyi S., Kossmann D., and Stocker K., "The Skyline Operator," in Proceedings of the 17th International Conference on Data Engineering, Cancun, pp. 421-430, 2001.

[8] Bruyère V., Decan A., and Wijsen J., "On FirstOrder Query Rewriting for Incomplete Database Histories," in Proceedings of the 16th International Symposium on Temporal Representation and Reasoning, BressanoneBrixen, pp. 54-61, 2009.

[9] Canahuate G., Gibas M., and Ferhatosmanoglu H., "Indexing Incomplete Databases," in Proceedings of the 10th International Conference on Advances in Database Technology, Munich, pp. 884-901, 2006

[10] A. Lotem, M. Naor, and R. Fagin.Optimal aggregation algorithms for middleware.In PODS'01, May 2001.

[11] D. Papadias, Y. Tao, G. Fu, and B. Seeger. An optimal and progressive algorithm for skyline queries. In ACM SIGMOD 2003, pages 467–478, 2003.

[12] F. P. Preparata and M. I. Shamos.Computational Geometry - An Introduction.Springer, 1985.

[13] K.-L. Tan, P.-K.Eng, and B. Chin Ooi. Efficient progressive skyline computation. In VLDB 2001, pages 301–310, 2001.

[14] M. Theobald, G. Weikum, and R. Schenkel.Topk query evaluation with probabilistic guarantees. In VLDB 2004, pages 648–659, 2004.

[15] L. Antova, C. Koch, and D. Olteanu, "From complete to incomplete information and back," in Proc. [16] A. Colantonio and R. Di Pietro, "Concise: Compressed 'n' composable integer set," Inf. Process. Lett., vol. 110, no. 16, pp. 644–650,2010.

[17] 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

[18] W. Zhang, X. Lin, Y. Zhang, J. Pei, and W. Wang, "Threshold- based probabilistic top-k dominating queries," The Int. J. Very Large Data Bases, vol. 19, no. 2, pp. 283–305, 2010.

[19] E. Tiakas, G. Valkanas, A. N. Papadopoulos, Y. Manolopoulos, and D. Gunopulos, "Metric-based top-k dominating queries," in Proc. Int. Conf. Extending Database Technol., 2014, pp. 415–426.

[20] L. Antova, C. Koch, and D. Olteanu, “From complete to incom- plete information and back,” in Proc. SIGMOD Int. Conf. Manage. Data, 2007, ap. 713–724.

[21] GostaGrahne,"Incomplete Information",Department of Computer Science,Concordiauniversity,Canada.

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

Authors profile:

J RADHAKRISHNA is a student pursuing M.Tech (CSE) in Eswar College of Engineering, Narasaraopet, Guntur, India..

MD ISMAILKHAN is having 5

years of experience in the field of teaching in various Engineering Colleges. At present he is working as Asst. Prof. in Eswar College of Engineering, Narasaraopet, Guntur, India.