

PANORAMIC VIEW FOR SKYLINE QUERY PROCESSING IN MULTIDIMENSIONAL DATABASE ENVIRONMENT

¹S. Deepa Kanmani, ²E. Kirubakaran, ³Jasper Kathrine

^{1,3}Assistant Professor, Department of Computer Sciences Technology,
Karunya University, Coimbatore, India.

²Professor, Department of Computer Sciences Technology,
Karunya University, Coimbatore, India.

¹deepa_cse@karunya.edu, ²ekirubakaran@karunya.edu

³katherine@karunya.edu

Abstract: User preference queries are playing very significant role in the database community. Different preference based queries are available which yields very good results comparatively with results of traditional queries. Especially skyline query is playing a major role in multi-criteria decision-making applications. This query basically derived from the concept of Pareto optimality. It is involved in many different applications like wireless sensor networks, recommender system, e-commerce and so on. Many different approaches were proposed and developed for this skyline query but mainly for the centralized and static environment. This paper gives summary of existing traditional skyline query algorithms and also made a comparison with recent challenges of query in uncertain environment with high dimensions.

Keywords: Uncertain data, preference queries, skyline query.

1. Introduction

Data can be basically classified into two types such as exact data and inexact data. The traditional query is able to execute with complete and certain data. When it comes to the scenario of uncertain environment advanced query concepts are required. There are several types of query available such as range query, aggregate query, top-k query, skyline query, join query. This paper discusses the importance of skyline and its research directions. In recent years Skyline queries are executed in different domains such as distributed sensor networks, distributed clouds and multi- source data integration. Most of the existing approaches for skyline query deal with the completeness of the data sets i.e., no attribute values are missing and rely on transitivity property. Due to incomplete data, this property is not satisfied. So special attention is to be carried out to handle the above said scenario. Figure 1 gives the overall view of database with query concept.

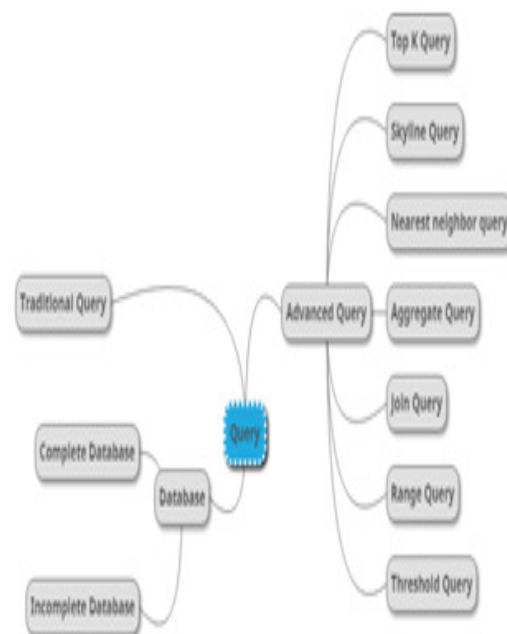


Figure 1. An overview of database query

Skyline query is used to retrieve relevant results based on the user's preferences from the database. Skyline query is used to deal the pareto optimum problem which means retrieving the objects which are not dominated by any other objects in the database. The following scenario gives the concept of skyline query. Scenario: Consider a used car dataset user who prefers a car with less price and more mileage but both are contradictory. if mileage is more then definitely price of the car will be high. In order to deal with such kind of scenario, skyline retrieves an object it should not be worse in all dimensions and should be better than at least one dimension. Figure 2 depicts the information

about used cars which contains two attributes such as price and mileage of the used car.

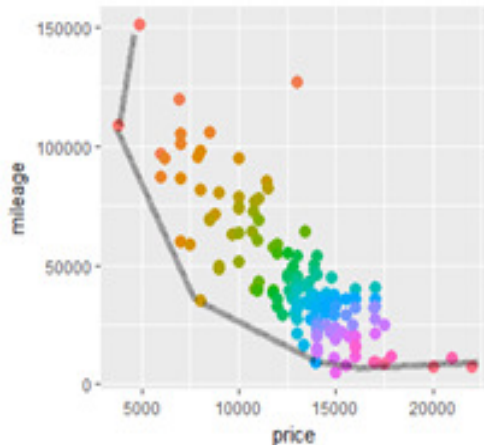


Figure 2. Two dimensional used car dataset

In the current trend, the world is operating on dynamic database. Such dynamicity in database leads to more uncertainty and huge volume of data. Executing the query in such kind of environments leads to inaccurate results and takes a long duration to respond. So preference queries are playing a significant role in the decision-making environment. Specifically, Skyline query and top-k queries are under the concept of preference based queries which gives very good impact in the field of database research. These queries are used to make good decisions in decision-making environment like purchasing a product with multi-criteria specifications.

Earlier decades static centralized database were used but the modern world always keeps on updating the information in the database so this leads to dynamic and distributed database environment. So ordinary static database and normal query concept does not deal with the following conditions such as:

- Not able to manage huge volumes of data. Example: cloud, sensor databases
- Not able to provide feasible solutions.

A new query concept is required now called advanced query processing. It is able to address the following issues

- Heterogeneity of data
- incompleteness, and uncertainty
- resource limitations,

- huge data volumes and volatility, typical of the new applications and processing environments,

Because of above things the user will get empty set, or too many results or unsatisfied result. So user preferences are taken to account to get better results. Specifically, skyline query processing makes very useful to retrieve the results comparatively with Top-k processing. Skyline query can be involved in many different environments such as partially ordered domains, high dimensional data, mobile ad networks, web information systems and data mining.

2. Motivation

Uncertain data management is very crucial in database research field. There are many different practical applications of uncertain management includes sensor network, RFID network, data cleaning and extraction, location based services, market surveillance, and social data collection. So analyzing huge collections of uncertain data has become a very big challenging task now. To provide the interesting results to the user in multi criteria decision making applications such as e-commerce, recommender system this advanced query called skyline query is playing major role in database management field.

3. Comprehensive overview of traditional skyline query algorithms

There are several algorithms were developed for handling skyline queries in efficient manner. The existing algorithms basically classified into two different groups. They are Non Index based algorithms and Index based algorithms. The following algorithms are under non indexed based approach, They are Block Nested Loop, Divide and Conquer(D&C), Sort Filter Skyline(SFS), Linear Elimination Sort for Skyline (LESS), Sort and Limit Skyline Algorithm(SaLSa), Nearest Neighbor(NN), Branch and Bound Skyline (BBS). Under Index concept B-Tree, B+ Tree, Bitmap algorithms were implemented. The following Table 1 gives the summary of the above algorithms

Table 1. Summary of Existing Traditional Skyline Query Algorithms

Algorithms	Merits	Demerits	Additional Remarks	Complexity
BNL[18]	works Good for small data set	Most expensive for computation.	1) all focuses on centralized and static environment	• Best case $O(n)$ • Worst case $O(n^2)$
DC [18]	m-way partitions.	• More partitions • Not good for small memory systems.	2) No subset Considerations 3) Not able to handle with incomplete data	• Best case $O(n)$ • Worst case $O(n^2)$
NN[19]	Uses R-tree index concept due to this more dominating checks are easily avoided	More dimensions leads to Very high I/O Cost	4) Not able to support progressive skyline computation	• Average complexity $O(\log N)$ for constant dimensions.
BBS[20]		Unnecessary dominance checks due to more dimensions in the data set.		• Complexity depends on number of dimensions $O(d)$
SFS[21]	Presorting based on BNL algorithm	More time is spent on scanning on the data set	• No early termination • Cannot adapt with more preferences from the user	• Best case is $O(dn+n\log n)$ • Worst case is $O(dn^2)$
LESS[22]	Optimized version of SFS	More scanning needed for presorting based		• $O(nd)$ • $O(n^2d)$
SaLSa [17]	• Good comparatively with SFS and LESS • Sorts the dataset according to monotone scoring function	Not able to accommodate inconsistent preferences	• Able to handle many dimensions	• Depends on scoring function and dimensions
Index[23] (B-Tree, B+ Tree, Bitmap)	Good one for large data set Good performance comparatively with other algorithms	Dimensions increases affects the performance Not good for anti correlated dataset	• Progressive skyline computation supported	• Depends on the dimensions

The older algorithms for skyline which all considers only two to three dimensions and their skyline query complexity is $O(n \log n)$. For higher dimensions, the existing algorithms leads skyline query complexity is $O(n(\log n)^{d-2})$.

Existing skyline query proposed and works only for completeness of data settings discussed in [6]. Skyline queries are very much useful in the stream, distributed and multi-core environments. Most of the existing algorithms work very well according to the assumption of completeness which means all data values are known in the dataset. But real-time scenario assumption needs to be handled with the incompleteness of data. This author proposed bucket algorithm based on bitmap representation. Every bucket has local skylines termed as skyline candidate sets. In every bucket, non-skyline points will be pruned. This algorithm is used to reduce the cost of comparing points within the bucket but still leads to unnecessary pair wise comparisons across the buckets. But this algorithm leads to unnecessary dominance test.

K-dominant skyline queries on incomplete data which is proposed in [7]. Existing algorithms very well suited for k-dominant skyline over complete data. They introduced different techniques to handle two interesting variants of skyline 1. Weighted dominant skyline query 2. And top- δ dominant skyline query over incomplete data. They introduced three different algorithms namely Baseline, DAG, BIB. Here they discussed search space complexity of skyline computation. Here they proposed baseline algorithm based on bucket algorithm. The performance of baseline algorithm highly depends on the size of the candidate set. When dimension gets increased the performance of these algorithms gets affected.

Skyline query playing a vital role in big data management because it is very useful in multi-criteria decision-making applications discussed in [5]. Skyline query also is known by Pareto optimum problem. They discussed distributed skyline query over the uncertain environment. Here they addressed about minimization of bandwidth consumption due to network delay and progressiveness. They presented grid summary approach called GDPS and presented about local pruning optimization and global pruning optimization. They maintain a fixed number of candidate tuples for reducing computation cost and storage.

An uncertain environment with time and space complexity discussed in [2]. They proposed Effective Probability Skyline Update (EPSU) is augmented with R-tree, SW tree with sliding window concept. This sliding window always getting updated because of uncertainty environment. They conclude with a future work suggested that this work has to be extended to distributed computing environment.

Uncertainty is the basic feature of uncertain data in [13]. Basically uncertain happens because of table-based, attribute-based or tuple based. There are many

uncertain queries are existing to handle inexact or uncertain data. They are Skyline query, Top-k query, Nearest Neighbor query, Aggregate query, Join query, range query, threshold query. Specifically, skyline query is a good solution for the multi-objective optimization problem, recommender system, and e-commerce. All skyline queries over uncertain data algorithms leads to worst-case computational complexity is $O(n^{2-1/(d+1)})$. The frequent change in data and uncertainty leads more computational complexity. This has to address properly. This is a very big challenge in the dynamic distributed database. They also discussed index techniques on uncertain data. The existing R-tree Biter indices are not providing an effective result with respect to uncertain data or fuzzy data environment.

Existing algorithms available for handling skyline queries over incomplete data which were discussed in [10], [6]. Replacement algorithm which replacing all the missing values with some preferred values. But it does not give the correctness of skyline results because it loses some skyline candidates. Bucket algorithm divides the entire set into buckets and every bucket will have known dimensions but this also leads to more pair wise comparisons. The ISkyline algorithm Presorted list concept on dimensions The SIDS algorithm Sorted list on dimensions so d sorted list will be created The ISkyline algorithm The SIDS algorithm the above algorithm also leads to more pair wise comparisons which give more computational complexity.

Most of the existing algorithms and techniques involved in skyline queries [12] focused only on the effectiveness of skyline query processing. Here they mainly discussed searching space is one of the important factors that influence the processing of the query. They made a comparison between both complete and incomplete databases. They introduced k-dom skylines with the help of clustering concept. So every cluster identified with local skylines and finally leads to k-dom skylines. Here this approach they deal with high dimensional data but not considered the uncertainty of the database.

As dimensions increase the number of skyline set also large [14] and it never offers good results. Unfortunately, skyline computation is very high. Consider a given dataset with n dimensions then skyline computation requires $2^n - 1$ skyline queries to be calculated. They proposed an algorithm based on counting the number of dominating subspaces. Here they introduced the concept of top-k frequent skyline queries. However computing skylines overall subspaces also leads high computational cost. Conventional skyline computation is inadequate to answer various queries since it needs to analyze the whole dataset and also their attribute combinations.

Bucket algorithm and ISkyline algorithm for incomplete data discussed in [9]. Rational algorithms for incomplete data are bucket and replacement

algorithms. Both needs exhaustive pair wise comparison which needs more time for computing the skyline set. So Iskyline was introduced. The ISkyline algorithm employs two new concepts, namely, *virtual points* and *shadow skylines* to enable efficient execution of skyline queries over incomplete data. This algorithm contains three phase's namely local skyline, candidate skyline and global skyline.

Handling incomplete data discussed in [14]. The bitmap and Sort-Based Incomplete Data Skyline (SIDS) and Iskyline algorithms are existing algorithms to handle incomplete data. But all of the algorithms take more time consumption for pair wise comparisons. So here k dom skylines introduced with the help of clustering method. This also leads time complexity for the clustering is $O(n)$ where n represents number of data items in the database.

Parallel skyline queries over uncertain environment is discussed in [15]. Analyzing huge amount of uncertain streaming data is an important and challenging field in real time applications. Here three models have been proposed namely Simple Parallel Model, Alternate Parallel Model, and Distributed Parallel Model (DPM). All the above models are executed with the help of sliding window partitioning. Finally to optimize the result for sliding window improvised with granularity adaptive concept. For further optimization in load balancing is done for all the peers in distributed environment.

Different types of preference queries like top-k query, Skyline query, K-dominance, ranked skyline, and K-frequency are analyzed in [16]. Here the analysis is made according to size of database in which always fixed and number of dimensions varied between 10-60. Whenever the dimensions are getting increased the corresponding result set also getting affected.

4. Variants of advanced skyline query

1. Top-K dominant query
2. Skyline query
3. Multi- Objective skyline
4. Top- K dominant query
5. K-dominance query
6. Ranked Skyline
7. K-frequency

5. Research Challenges in skyline query processing

The feature of uncertainty in tuple, attribute and table based or group based data cannot be handled efficiently by existing index and sorting based algorithms. Curse of dimensionality major challenge in skyline query processing is discussed in [3], [6].

1. When dimensions increases grouping of data becomes tedious.

2. Very difficult to identify the proper subsets in order to provide better results.

In the theoretical study when dimensionality getting increased tends to increase the skyline points also in exponential manner i.e., $\Theta((\ln n)^d / d!)$

Another issue is decentralized data sets and massive data this leads to communication overhead. This also has to be addressed properly otherwise this affects the bandwidth consumption and more computational cost.

6. Conclusions

Big data and its management are compulsory for database research field because of its nature of uncertainty. Updating everyday in sequence is important for uncertain since the database management needs to be updated every change in the world. So this is a challenging task nowadays. Preference queries are playing a crucial role in the dynamic database environment. Even many different algorithms are available to handle these queries. But still, it is an ongoing research area in order to develop the optimized result to the users. This paper gives an overall view of existing algorithms of skyline query one of the types of preference based query and gives certain insights about pros and cons of different approaches.

References

- [1] Cheng Sheng, Yfei Tao.: Worst -Case I/O Skyline algorithms Journal ACM transactions on Database Systems, Vol 37, Issue 4(2012).
- [2] Chuan-Ming Liu , Syuan-Wei Tang .:An Effective Probabilistic Skyline Query Process on Uncertain Data Streams In: Proceedings of 6th International Conference on Emerging Ubiquitous Systems and Pervasive Networks, pp 40-47(2105).
- [3] Amardeep Kaur, Amitava Datta. : A novel algorithm for fast and scalable subspace clustering of high-dimensional data. Journal of Big Data(2015)
- [4] Stefan Mandal , Markus Endre .: Preference Analytics in EXASolution , In proceedings of 27th International on Database and Expert Systems Applications (2016)
- [5] Xiaoyong Li, Yijie Wang , Xiaoling Li, Xiaowei Wang, Jie Yu GDPS: An efficient approach for skyline queries over distributed uncertain data: Big Data Research, pp23-36(2014)
- [6] JongwukLee, HyeonseungIm , Gae-wonYou Optimizing Skyline Queries over Incomplete Data: Information Sciences, pp.14-18(2016)

- [7] Xiaoye Miao, Yunjun Gao, Gang Chen, Tianyi Zhang.:K-Dominant skyline queries on incomplete data: Information Sciences (2016)
- [8] Skyline computation for frequent queries in update intensive environment.
- [9] M.E. Khalefa, M. F. Mokbel, J.J. Levandoski .: Skyline Query Processing over incomplete data in ICDE, pp.556-565 (2008)
- [10] Yonis Gulzara , Ali A. Alwan1 , Norsaremah Salleha , Imad Fakhri Al Shaikhliia , Syed Idrees Mairaj Alvib.: A Framework for evaluating skyline queries over incomplete data: In 13th International Conference on Mobile Systems and Pervasive Computing, pp. 191-198 (2016)
- [11] Y Wang, X Li, X Li, Y Wang.: A Survey of queries over uncertain data: Springer Knowledge and information systems, (2013)
- [12] CY Chan, HV Jagadish, KL Tan, AKH Tung.: On High Dimensional Skylines In: International Conference on Extending Database Technology, pp 478-495 (2006)
- [13] Hyeonsung Im, Sungwoo Park.: Group Skyline Computation: Information Sciences, pp 151-159 (2012)
- [14] Ali A. Alwan, Hamidah Ibrahim, Nur Izura Udzir, Fathima sidi.: An Efficient Approach for Processing Skyline Queries in Incomplete Multidimensional Database: Arabian Journal for Science and Engineering, pp 2927-2943 (2016)
- [15] X.Li, Y. Wang .: Parallel Skyline Queries Over Uncertain Data Streams in Cloud Computing Environments: Int. Journal web Grid Serv, pp 24-53 (2014)
- [16] Ali. A. Alwan, Hamidah Ibrahim, Tan C. Yip, Fatimah Sidi.: A Performance Evaluation of Preference Evaluation Techniques in Real High Dimensional Database: The International Symposium on Advances in Transaction Processing (ATP), pp. 894-901(2012)
- [17] Ilaria Bartolini, Paolo Ciaccia, Marco Pantella.: Salsa: Computing Skyline without scanning the whole sky : Proceedings of the 15th ACM international conference on Information e and Knowledge Management pp 405-414 (2006)
- [18] Stephan Börzsönyi, Donald Kossmann, and Konrad Stocker. The skyline operator. In Proceedings of the 17th International Conference on Data Engineering, pp 421-430 IEEE Computer Society (2001)
- [19] Donald Kossmann, Frank Ramsak, and Steffen Rost. : Shooting stars in the sky: an online algorithm for skyline queries. In Proceedings of the 28th international conference on Very Large Data Bases, VLDB '02, pp 275-286. (2002)
- [20] Dimitris Papadias, Yufei Tao, Greg Fu, and Bernhard Seeger. :An optimal and progressive algorithm for skyline queries. In Proceedings of ACM SIGMOD international conference on Management of data, pp. 467-478 (2003)
- [21] Jan Chomicki, Parke Godfrey, Jarek Gryz, and Dongming Liang. Skyline with presorting. In ICDE, pp. 717-719 (2003)
- [22] Parke Godfrey, Ryan Shipley, and Jarek Gryz. Maximal vector computation in large data sets. In VLDB, pp 229-240 (2005)
- [23] Kian-Lee Tan, Pin-Kwang Eng, and Beng Chin Ooi. Efficient progressive skyline computation. In Proceedings of the 27th International Conference on Very Large Data Bases, VLDB '01, pp. 301- 310, (2001)
- [24] T. Padmapriya and V. Saminadan, "Inter-cell Load Balancing Technique for Multi- class Traffic in MIMO - LTE - A Networks", International Conference on Advanced Computer Science and Information Technology , Singapore, vol.3, no.8, July 2015.
- [25] S.V.Manikanthan & D.Sugandhi "Interference Alignment Techniques For Mimo Multicell Based On Relay Interference Broadcast Channel" International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE) ISSN: 0976-1353 Volume- 7 ,Issue 1 -MARCH 2014.
- [26] R. Jeevidha, V. Sowmiya , K. Kiruthiga3, R. Priya, "Collabration Complexity Reducing Strategy In Cloud Computing", International Innovative Research Journal of Engineering and Technology, Vol 02, No 04, pp.124-128, 2017.

