A SURVEY OF SKYLINE PROCESSING IN VARIOUS ENVIRONMENT

Angel C Bency
Department of computer Science and Engineering, Karunya University Coimbatore, Tamil Nadu
angelcbency@karunya.edu.in

S Deepa Kanmani
Department of computer Science and Engineering, Karunya University Coimbatore, Tamil Nadu
deepa_cse@karunya.edu

Abstract

Skyline queries which received an interesting attention to the database and data mining field and its main advantage is it is used for multi-criteria decision making. Advanced query operators, such as skyline queries are necessary in order to help users to handle huge amount of data from a large database by identifying set of non-dominated points. Skyline queries means it retrieve set of non-dominated points or better points from the given data points. It provides an interactive environment for information retrieval that help user to get answer for the given preference based query. This survey paper discusses about various skyline query processing algorithms and its different application in various fields includes spatial database, subspace, network, distributed environment etc. This study gives an overview about the existing algorithms for skyline query processing technique and different applications.

Keywords: skyline query, spatial skyline, dynamic skyline, static skyline, pareto-optimum skyline, Range based skyline query

1. Introduction

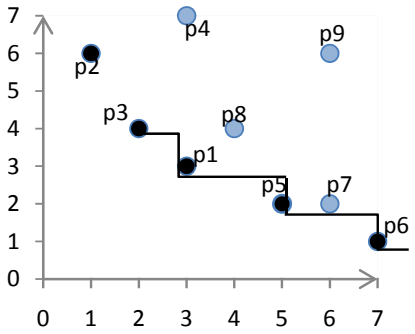
Before the entry of skyline into database management there is a problem called maximum vector problem or Pareto optimum [1].In recent years skyline queries receive an interesting attention in various applications such as multi-preference analysis and decision making. In such application skyline set which contain the most interesting points or best points i.e., it retrieve the points that are not dominated by any other points. In database systems, queries specialized to search for the non-dominated data points are called skyline queries and their corresponding result set is known skyline set. Individual data points in a skyline result set are known as skyline points. The skyline queries which first addressed problem in static environment and it later progressively find out for dynamic set. If the user is start moving or query is issued from a dynamic environment in that case it addresses the problem in dynamic environment. It also used in spatial network and all our data's are highly distributed so it addresses the problem in distributed environment.

For example, consider a database that contains information about hotels. Each tuples of the database is represented as a point in a data space consists of numerous dimensions. Assume a user is looking for a hotel that is cheap as possible and as close to the beach. To illustrate the idea of dominance relationships, Fig.1 gives hotel finding example in this example user is looking for a hotel based on two criteria, minimum price and minimum distance to the user standing location. Fig.1a lists 9 hotel records and their values and Fig.1b depict the representation of the hotel in a 2D space. Hotels p4, p7, p8 and p9 are all dominated by other points so skyline which return points that are not dominated by any other points. Consider the point, p7 which is dominated by p5 as it is more expensive than p5 but both have the same distance value. The skyline query retrieves all hotels for which no other hotel exists that is cheaper and closer to beach. So the skyline result set which consist of {p2, p3, p1, p5, p6}.

In the above example it returns the most interesting point in the database to the user. The skyline query processing over moving objects has received an interesting attention to numerous applications, such as object tracking and monitoring, location-aware computing, uncertain data stream, virtual environments, computer games, and visualization etc. Location based service using skyline is the attractive application. For example, in a taxi dispatching scenario it notify their locations within frequent time interval to the dispatcher. So it help to identify the last location and how much near to the user location etc. Skyline queries touch with many emerging applications and research fields. Besides all this advantage the main problem is skyline query processing and skyline result update are expensive application in database. The main cost is accessing data from storage and CPU cost spends for executing the user given query for dominance check. Search efficiency and update criteria are the two most important performance criteria to skyline query processing and skyline result maintenance.

Hotel	Price	Distance
P1	3	3
P2	1	6
P3	2	4
P4	3	7
P5	5	2
P6	7	1
P7	6	2
P8	4	4
P9	6	6

(a)



(b)

Fig.1. An example of skyline query

In the above example it returns the most interesting point in the database to the user. The skyline query processing over moving objects has received an interesting attention to numerous applications, such as object tracking and monitoring, location-aware computing, uncertain data stream, virtual environments, computer games, and visualization etc. Location based service using skyline is the attractive application. For example, in a taxi dispatching scenario it notify their locations within frequent time interval to the dispatcher. So it help to identify the last location and how much near to the user location etc. Skyline queries touch with many emerging applications and research fields. Besides all this advantage the main problem is skyline query processing and skyline result update are expensive application in database. The main cost is accessing data from storage and CPU cost spends for executing the user given query for dominance check. Search efficiency and update criteria are the two most important performance criteria to skyline query processing and skyline result maintenance.

2. Skyline Query Processing

Skyline query processing has been extensively studied in recent years. Kung et al., [1] which proposed the first skyline algorithm which is maximum vector problem. The existing skyline query processing algorithms include Block-Nested Loop (BNL) [2], Divide and Conquer (D&C) [2], Sort Filter Skyline (SFS) [3], Bitmap [6], Nearest Neighbor (NN) [5], Branch and Bound Search (BBS) [4] etc. The existing skyline query algorithms can be classified mainly into three types

2.1 Static skyline query algorithms

The primitive skyline query processing algorithms are D&C [2] and BNL [2]. In D&C algorithm it considers the whole data space and makes m-way partition. For each partition it identifies the points that are not dominated by any other points and compute the skyline result by merging the result together. In BNL algorithm it iteratively checks each record with the main memory to find out the dominating point and return the skyline result to a temporary file.BNL and D&C are non-index based algorithms that scan the whole data iteratively. Sorting based algorithm such as SFS which sort the data based on the input monotonic function. The algorithms can be categorized mainly into two groups index-based and non-index based algorithm. The index based algorithms which include bitmap, R-tree, B-tree, NN etc and non-index based algorithm which include D&C, BNL etc. Index based algorithms which progressively return the result without scanning all data set. BBS which is based on iterative call of NN algorithm and it overcome the disadvantage of NN in terms of I/O cost and CPU.FAST-SKY [10] algorithm that improve skyline query processing with high dimensional data. A Stratification Technique which is used to index data on PODs and use two index structures (i) Stratified R-tree(SR-tree) for low dimensional data (ii) Stratified MinMax treaps(SM-treaps) for high dimensional data. FAST-SKY algorithm which helps to achieve IO optimality and progressives in skyline computation with both TODs and PODs, it is faster than all other existing algorithms.

2.2 Dynamic skyline query

Dynamic query means the query location is continuously changing so that skyline results are change frequently. The efficiency of skyline is computed in terms of accessing the data points and organizing the skyline result. Range based skyline query [14] in mobile environment which uses two algorithm I-SKY [14] and N-SKY [14]. I-SKY is mainly used static object and N-SKY is focus for dynamic object. Skyline query used with Location Based Service (LBS) [14] which receive an interesting attention in research field.LBS which provides timely information to the user. Skyline for Z-order space filling is done by using Z-SKY [7] framework. It support skyline query over a subset of dimension and ZBtree, it is the index structure used to organize the data. ZSearch, ZInsert, ZDelete, ZUpdate, ZRank, ZSubspace etc are used to support subset of dimensions. Keyword matched skyline search (KMS) [8] which provide an efficient textual search by using NN algorithm. It uses IR^2 tree index

for storing both spatial and textual information. Each keyword is represented by using signature and for query processing it mainly checks two conditions dominance check and signature check.

2.3 Spatial skyline query

Farthest spatial skyline queries which help to identify data points those are far away from query point. These queries are mainly used to identify spatial locations far away from undesirable locations. These can be achieved by two algorithms Threshold Farthest Spatial Skyline (TFSS) [11] and Branch and Bound Farthest Spatial Skyline (BBFS) [11].In TFSS it use standard set of accesses such as sorted access from distributed sources which uses R-tree for accessing node and it retrieves data objects in decreasing order of the attribute value. In BBFS it use minimum Bounding rectangle (MBR) of an R-tree for batch pruning. It also exploit spatial locality of spatial data points. Full space skyline can be supported incrementally by using naïve on-line maintenance module (NMA) [9]. Stream object in the data stream environment is maintained by using a naive approach called NMA and it uses mainly 3 types of pruning techniques they are Timestamp based Pruning, Attribute based pruning and Timestamp and attribute based pruning. Spatial skyline query which can be applied to many spatial applications. Branch and Bound Spatial Skyline (B^2S^2) [15] and Voronoi based Spatial Skyline (VS^2) [15] are the two algorithm which can be used for static query point. Voronoi based Continuous Spatial Skyline(VCS^2) [15] algorithm which is used for object that change their query location.B^2S^2 which uses minimum distance calculation and VS^2 which is based on Voronoi diagram.VCS^2 which avoid all the unnecessary dominance check.

2.4 Network skyline query

An efficient approach is proposed to handle distributed skyline query processing from the mobile unit (laptop, mobile phone etc) called efficient distributed skyline based on mobile computing (EDS-MC) [13].The EDS-MC algorithm which mainly consist of 5 phases they are Query shipping phase, Local processing phase, Reduction phase, Aggregation phase and Final Processing phase. Pruning strategy is used to reduce the intermediate skyline set. In the mobile network mobile unit is the initiator of the user request and base station which processes the request to generate intermediate skyline result set. Spatial network skyline query which considers the network distance between the data point and query point it is based on the nearest neighbor algorithm. It first finds all the skyline points and then calculates distance between query point and skyline point. The Network nearest neighbor skyline (N3S) [12] is mainly used in road network and its main advantage is it has few distance calculation.

Skyline computation

The evolution of the skyline query process is summarized in Table 1. Many number of papers are published in each year related to skyline query processing and research is till continue to apply skyline in various fields. The effectiveness of skyline query process is determined by two factors (i) accessing the data points and (ii) organizing the skyline result set. Skyline query which help to give preference based queries and it can be used for both centralized and distributed environment.

Table 1. Evolution of skyline query

1978	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**
Maximum vector Problem[1]	Skyline operator[2] Skyline query Processing with b-tree[6]	Skyline Processing with r-tree[5]	Skyline Variant[16] Skyline processing with Sorting [4]	web information systems[19]	Subspace Skyline queries [20]	Skyline in P2P systems [21]
2013	**2012**	**2011**	**2010**	**2009**	**2008**	**2007**
Textual database [8] Range based skyline [14]	Wireless sensor Network[18]	Spatial database [11]	Z-order curve[7]	Mobile Application [13]	Parallel skyline Computation [23]	Probabilistic Skyline [22]

The skyline variants [16] are introduced based on the user input search criteria and dominance relations they are Skyband query processing, Top Ranked skyline query processing, Dominant skyline query processing and constrained skyline query processing. In ranked skyline query it consider the d-dimensional space based on the mindist calculation .Constraints skyline query which means it define some constraints on the input ,commonly index method is used to modify the constraints. K-dominating query point which return k-points that dominate large number of points. Subspace skyline queries are mainly supported by distributed skyline approaches. The main distributed skyline applications are web information systems, Distributed data stream, Wireless sensor networks, uncertain data etc.

Conclusion

This paper made a detailed survey on skyline query processing for last two decades and made a overview of many algorithms and techniques. Recently skyline query processing which receives an interesting attention in data mining field. Skyline queries retrieve the non-dominated points from a large database system based on the user preference so it can be used in preference based applications. It successfully eliminates all the dominated points by using some efficient technique. This paper discusses different application of skyline query with their algorithms. Skyline query has been applied over both centralized and distributed system is discussed in recent years. Now research opens its area to discuss the skyline query in uncertain data, probabilistic skyline, peer-to-peer system, Location based service etc.

References

[1] J.L.Bentley, H.T. Kung, M. Schkolnick, C.D. Thompson, On the Average Number of Maxima in a Set of Vectors and Applications, J. ACM, vol. 25, no. 4, pp. 536-543 ,1978.

[2] S. Borzsonyi, D. Kossmann, K.Stocker, The Skyline Operator, in: Proc. Int'l Conf. Data Eng, pp. 421-430, 2001.

[3] J. Chomicki, Godfrey,J. Gryz , D. Liand ,Skyline with Presorting, in: Proceedings of ICDE, pp. 717–816,2003.

[4] D. Papadias, Y. Tao, G. Fu, B. Seeger,Progressive skyline computation in database systems, in: ACM TODS 30(1), 41–82 ,2005.

[5] D. Kossmann, F. Ramsak and S. Rost, Shooting stars in the sky: an online algorithm for skyline queries, in: Proceedings of VLDB Conference, pp. 275–286, 2002.

[6] K.L.Tan, P.K.Eng and B.C.Ooi, Efficient Progressive Skyline Computation, In: Proceedings of VLDB Conference, pp. 301–310, 2001.

[7] C.K. Lee, W.C. Lee, B. Zheng, H. Li, Y. Tian, Z-SKY: An Efficient Skyline Query Processing Framework Based on ZOrder,In:VLDB J., vol. 19, no. 3, pp. 333-362, 2010.

[8] H. Choi, H. Jung, Y.K Lee, and Y.D.Chung, Skyline queries on keyword-matched data" Information Science 449-463,2013.

[9] Z. Huang, S. Sun and W. Wang, Efficient mining of skyline objects in subspaces over data streams, in: Knowl Inf Syst 22:159–183,2010.

[10] H. jung, H. Han, H.Y. Yeom, S. Kang, A fast and progressive algorithm for skyline queries with totally- and partially-ordered domains, in: journal of system and software 429-445, 2010.

[11] G.W You, M.W Lee, H.Im and S.W Hwang, The Farthest Spatial Skyline Queries,in: Information System 286-301,2013.

[12] M. Safar, D. El-Amin,D.Taniar, Optimized skyline queries on road networks using nearest neighbor, Pers Ubiquit 15:845-856,2011.

[13] Y.Y Xiao,Y.G Chen ,Efficient distributed skyline queries for mobile applications,in: Journal Of Computer Science And Technology 25(3): 523-536 May 2010.

[14] X. Lin, J. Xu, H. Hu,Range Based Skyline Query in Mobile Environment,in: IEEE Transactions On Knowledge and Data Engineering, vol. 25, no. 4, APRIL 2013.

[15] M. Sarifzadeh ,C.Shahabi,The Spatial Skyline Queries,in: VLDB '06 Proceedings of the 32nd international conference on Very large data bases, Pages 751-762 ,2006

[16] D. Papadias, Y. Tao, G. Fu, B. Seeger, An Optimal and Progressive Algorithm for Skyline Queries,In: Proc. ACM SIGMOD Int'l Conf. Management of Data, 2003.

[17] X. Ding, X. Lian, L. Chen ,H, Continuous monitoring of skylines over uncertain data streams, Information Sciences 184 196–214, 2012.

[18] B. Chen, W. Liang, J.X Yu, Energy -efficient skyline query optimization in wireless sensor networks,In: Media, LLC 2012.

[19] W.T. Balke, U.Guntzer, J. X. Zhen, Efficient Distributed Skylining for Web Information Systems,in: In: Proceedings of International Conference on Extending Database Technology (EDBT) 2004,LNCS 2992,pp.256-273,2004.

[20] Y.Yuan, X.Lin, Q.Liu, W.Wang, J.X.Yu, Q.Zhang, Efficient computation of the skyline cube, in: Proceedings of International Conference on Very Large Data Bases (VLDB), pp. 241–252, 2005.

[21] P. Wu , C. Zhang , Y. Feng , B. Y. Zhao , D. Agrawal, A.E Abbadi, Parallelizing skyline queries for scalable distribution, In: Proceedings of International Conference on Extending Database Technology (EDBT), pp. 112–130, 2006.

[22] Pei, J., Jiang, B., Lin, X., Yuan, Y, Probabilistic skylines on uncertain data. In: Proceedings of International Conference on Very Large Data Bases (VLDB), pp. 15–26, 2007.

[23] A.Vlachou, C.Doulkeridis, Y.Kotidis, Angle-based space partitioning for efficient parallel skyline computation, In: Proceedings of International Conference on Management of Data (SIGMOD), pp. 227–238, 2008.