Science Forum (Journal of Pure and Applied Sciences)

journal homepage: https://atbu.edu.ng/science-forum/

Theoretical frameworks for improving the computation of skylines over uncertain data

Ma'aruf Mohammed Lawal[1,2*], Hamidah Ibrahim[2], Nor Fazlida Mohd Sani[2], Razali Yaakob[2]

[1]*Department of Computer Science, Faculty of Physical Sciences, Ahmadu Bello University, Zaria, Nigeria*

[2]*Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Seri Kembangan, Malaysia*

ABSTRACT

Skyline query is preeminent for realizing solutions that support multi-criteria decision-making, particularly for modern applications that often capture or generate a large volume of high-dimensional databases. The essence of the skyline query is to return a set of objects that are not dominated by any other objects. In this article, we proposed frameworks namely; *Skyline Query on Uncertain Database*, *Constrained Skyline Query on Uncertain Database*, and *Abating the computation of constrained Skyline Query on Uncertain Database* that deal with the computational problem associated with processing constrained skyline query over uncertain data with high dimensionality. Although solutions for processing such queries over uncertain data are available, they are inefficient as they do not utilize the result of earlier computations. Since multiple constrained skyline queries might span over the same part of the database, this research work attempt to avoid unnecessary skyline computations of these queries by identifying the conditions specified in these queries.

ARTICLE INFO

Artcle history:
Received 26 December 2022
Received in revised form
28 December 2023
Accepted 28 December 2022
Published 06 July 2023
Available online 06 July 2023

KEYWORDS

Constrained skyline query
Uncertain data
Indexing technique
Framework
Skyline query processing

1. Introduction

The essence of skyline query processing is to return a set of objects that are not dominated by any other objects. Skyline analysis is a popular way of deriving interesting and non-trivial answers from the underlying database based on user preferences. While conducting a literature review on preference evaluation techniques, a number of works have applied the concept of skyline evaluation technique to provide answers to a user's query. These include Branch and Bound Skyline (Nam and Sussman, 2004), Sort Filter Skyline (Afshani et al., 2011), Divide-and-Conquer, Block Nested Loop (Catania and Jain, 2013), Bitmap and Index (Wang et al., 2013), Linear Elimination Sort Skyline (Chomicki et al., 2005), and Nearest Neighbor (NN) (Kossmann et al., 2002). Common to all of these proposed algorithms, is the attempt to indirectly reduce the computational cost of computing the skylines, through the reduction in the number of pairwise comparisons among objects of the underlying database.

Several research works on skyline query processing over certain data exist, as most of the proposed algorithms for computing skylines assumed that the values along a dimension of the database are all exact values (Bartolini et al., 2006; Borzsonyi et al., 2001; Chomicki et al., 2005; Gothwal et al., 2018; Kemper and Moerkotte, 1990; Kossmann et al., 2002; Papadias et al., 2005; Wang et al., 2013). However, this assumption is not always true in the real world. Particularly, with

*** Corresponding author** Ma'aruf Mohammed Lawal ✉ mmlawal80@gmail.com ✉ Department of Computer Science, Faculty of Physical Sciences, Ahmadu Bello University, Zaria, Nigeria.

modern database applications that capture uncertain data with high dimensionality.

An uncertain database in our work is referred to as a collection of objects with continuous range values, whose exact values are not known at the point of computing the skylines of a given skyline query. With an uncertain database, performing dominance tests among objects without calculating the skyline probabilities is almost impossible. The conventional dominance test which derives skylines by directly comparing the values of objects to identify dominant objects no longer work in this setting, as continuous range value is neither comparable to an exact nor other continuous range values. Similarly, it becomes more complicated when a constrained skyline query is to be processed over uncertain database. Unlike conventional skyline queries, a constrained skyline query returns a set of skylines that strictly satisfies the constraints on specific dimensions that are attached to the query, that is, a constraint on a dimension is a range specifying the user's interest. Identifying whether an object satisfies these constraints is straightforward if the values of the object are all exact values; thus, those objects whose dimension values are within the range specified in the constraints are selected for skyline computation. However, objects whose dimension values are in a continuous range form that intersects or overlaps with the range specified in the constraint need further examination before the objects can be considered for skyline computation. Although there are works on skyline queries as well as constrained skyline queries over uncertain databases, these works which employ a probability-based approach require additional computational time, as they rely heavily on probability calculations. Moreover, processing every single constrained skyline query independently might incur unnecessary repetition of skyline computations as these constrained skyline queries might have similar constraints specified in the queries. Thus, avoiding these unnecessary skyline computations of multiple constrained skyline queries is crucial. The main contributions of this research work are presented as follows:

We proposed the Skyline Query on Uncertain Database (*SQUiD*) framework which is the baseline framework of the subsequent frameworks with the main aim of efficient process skyline query over uncertain data without employing either the arithmetic mean calculation or the probability methods.

The second framework that we have proposed named Constrained Skyline Query on Uncertain Database (*CSQUiD*) aims at computing constrained skyline queries over the uncertain database with high dimensionality

Finally, Abating the computation of constrained Skyline Query on Uncertain Database (*ASQUiD*) is introduced mainly to streamline the process of processing multiple constrained skyline queries over uncertain data

The rest of this article is organized as follows: Related works are presented in Section 2, while the proposed theoretical frameworks are presented in Section 3, and finally, this research work is concluded in Section 4.

2. Related Works

The quest to derive solutions for multi-criteria problems has resulted in the introduction of a number of preference evaluation techniques, as a traditional query is too rigid as it does not give users the privilege to specify their interests. Not until recently, preference-based queries have received considerable attention in databases (Arslan et al., 2010; Chen and Lian, 2008; Guttman, 1984; Lee et al., 2010; Pei et al., 2007; Saad et al., 2019). This is due to the effectiveness in returning interesting results that provide optimum solution. With databases emerging from processing environments, an extension of the standard structured query language with preference evaluation technique is required in order to provide answers to personalized queries. Among the preference evaluation techniques are the skyline, top-k, top-k dominating, *k*-dominance, and k-frequency. Each of these techniques is unique in the way preference evaluation has been performed (Chan et al., 2005; Kontaki et al., 2011; Li et al., 2012; Liu et al., 2013; Pei et al., 2007; Sajeev and Noorjahan, 2016; Tan et al., 2001; Wang et al., 2013). These preference evaluation techniques are discussed in detail with more emphasis on skyline queries. The most popular among the preference evaluation techniques used in a database is the top-k and skyline queries.

Several works on skyline computation have been inspired due to the benefits that can be achieved if applications for retrieving non-trivial information are developed (Afshani et al., 2011; Aggarwal and Philip, 2009; Alwan et al., 2014; Atallah and Qi, 2009; Bartolini et al., 2006, 2008; Borzsonyi et al., 2001; Chan et al., 2005; Cui et al., 2008; Dellis and Seeger, 2007; Gulzar et al., 2017, 2019; Han et al., 2013; Khalefa et al., 2010; Liu et al., 2013; Saad et al., 2016). Quite a number of works adopted the probabilistic approach in processing skyline queries in finding the dominant objects (Jiang et al., 2012; Khalefa et al., 2010; Liu et al., 2013; Pei et al., 2007; Saad et al., 2014, 2016, 2018, 2019). In respective of the underlying database to be processed,

basically, there are two main approaches employed by existing works for processing skyline queries over high dimensional databases, namely: the index and non-index-based approaches. The index-based approach requires a progressive preconstructed data structure that allows and supports the computation of skylines in an efficient manner.

The index-based techniques employ indexing techniques to organize data into Minimum Bounded Rectangles (MBRs) in order to streamline the process of computing skylines (Saad et al., 2018; Tan et al., 2001). Index-based techniques have been seen to be useful in query optimization, by playing a major role in reducing the computational time required for processing a query. The preconstructed data structure is meant to significantly speed up skyline computation by pruning off objects which do not satisfy the query when computing the skylines. Most of the works that adopt the index-based approach typically use *R*-based trees and spatial indexing techniques for supporting the process of computing skylines over a database (Jiang et al., 2012; Kossmann et al., 2002; Li et al., 2012; Papadias et al., 2005). The non-index-based approach requires preprocessing operations such as sorting and grouping. Unlike the index-based approach which required using an indexing technique to organize data, with a non-index approach, the operations to be executed are performed directly on the database (Li et al., 2012).

Among the proposed solutions for processing skyline queries over uncertain data are that of Saad et al. (2014, 2016, 2018, 2019) which employ a non-index-based approach that was inspired by the work of Pei et al. (2007). In their proposed solution, a probability calculation method was utilized just like in the work of Jiang et al. (2012) Pei et al. (2007). Even though the works of Khalefa et al. (2010); Li et al. (2012); Pei et al. (2007); and Saad et al. (2014, 2016, 2018, 2019), employ a method that captures the uncertainty associated with continuous range values, the solution seems inefficient especially when processing skyline queries over uncertain data with high dimensionality as many probability calculations need to be performed before the skylines are derived. Meanwhile, Li et al. (2012) employs the median value calculations method to compute the exact value of each continuous range value. However, the method is not realistic as there is no certainty that the exact value for a continuous range value is most likely to be at the central point of the continuous range value. Therefore, it has become imperative to support the design and development of an efficient sky-line computation framework for processing skyline queries over uncertain data by avoiding extensive probability calculations and not relying on the assumption that the exact value of each continuous range value is at the central point of the continuous range value.

3. Proposed Theoretical Frameworks

In dealing with the computational problem associated with processing skyline queries over an uncertain database, in this work, we proposed three distinct non-probability index-based frameworks, namely: Skyline Query Processing on Uncertain Database (*SQUiD*), Constrained Skyline Query Processing on Uncertain Database (*CSQUiD*), and Abating Skyline Query Recomputation over Uncertain Database (*ASQUiD*). Common to these indexed-based frameworks is the use of an *X*-tree structure in organizing the underlying data set. The choice of the *X*-tree is justified based on the analysis conducted in Lawal et al. (2020).

3.1. The SQUiD framework

This framework consists of two main modules, namely: Data Pre-processing module and real-time module as presented in Figure 1. Essentially, the *SQUiD* framework is meant for reducing the number of pairwise comparisons that need to be performed among objects and for minimizing the searching space which in-directly relates to an improved data access performance. The Data Pre-processing module aims at streamlining the process of skyline computation by employing the *X*-tree indexing technique to organize the uncertain database into MBR in such a way that objects that are being dominated by other objects within the same MBR can be filtered. Nevertheless, our choice of *X*-tree indexing technique is based on in-depth performance analyses conducted among the existing *R*-based indexing techniques, namely: *R*-tree, *R**-tree, and *X*-tree. Meanwhile, the real-time module aims at transforming the uncertain database (continuous range value) into certain data (exact value) so that the conventional dominance test can be directly employed. To realize the *SQUiD* framework, we have identified the following procedures: Grouping method that groups the objects within each MBR based on the values of each dimension; dominance relation that performs the conventional dominance test among objects with comparable values to realize the candidate skylines of each MBR; kNN algorithm that identifies the k nearest objects to an object—these objects have similar features that can be used to predict the continuous range values of an object, and fuzzy algorithm that consists of *trapezoid fuzzification*, *sum aggregation*,

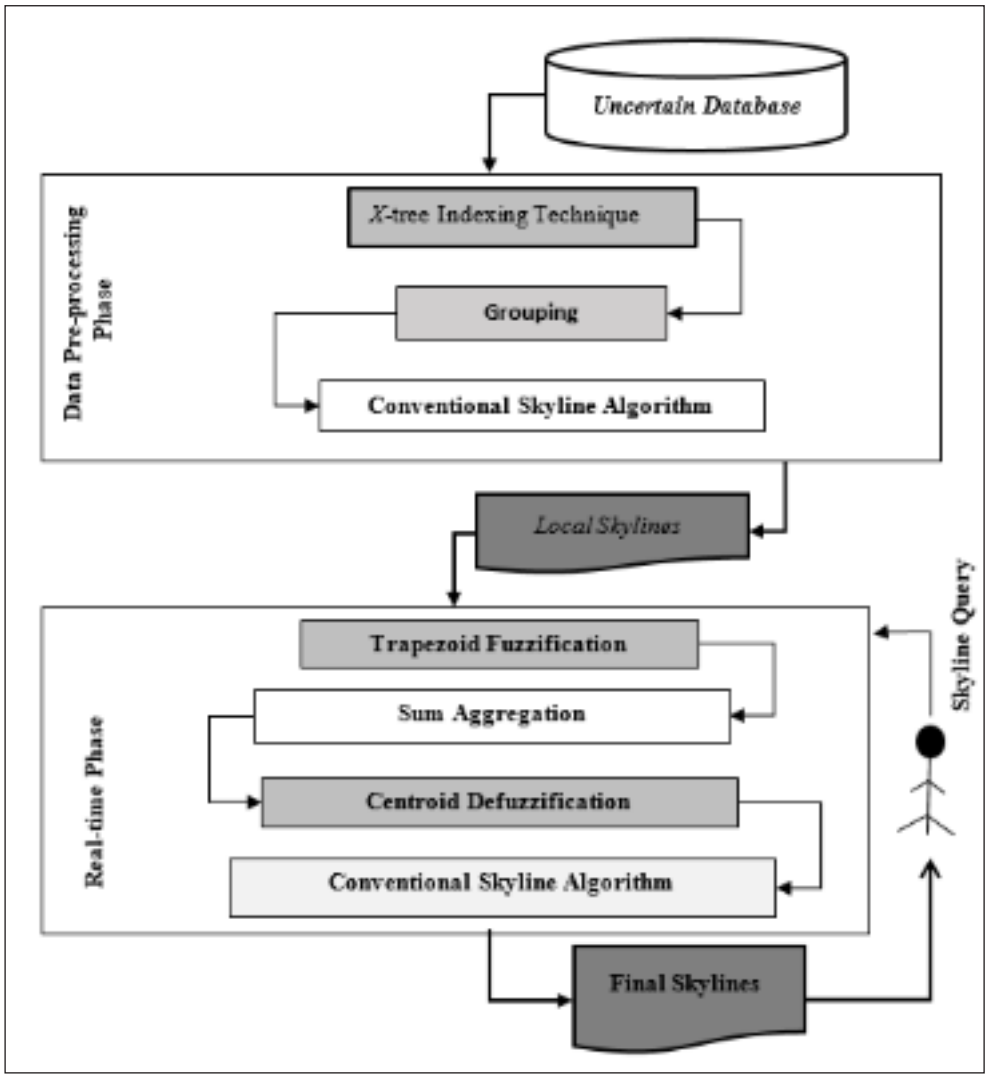

Figure 1. The proposed *SQUiD* framework for computing skylines over uncertain data.

and *centroid defuzzification* methods to predict the exact value of a continuous range value based on the features of *k* nearest objects. Eventually, as the values of these objects are now comparable the conventional dominance test can be employed to compute the final skylines.

The *SQUiD* framework is a two-phase framework. The first phase named the Data Pre-processing phase consists of two methods and utilizes the conventional skyline algorithm to derive the processed MBRs. These methods are (i) *X*-tree indexing technique which is employed for organizing both certain and uncertain data into MBRs, (ii) Grouping method which is introduced for eliminating unnecessary pairwise comparisons between an object having exact values and object represented with continuous range value, by grouping objects of each MBR is split into two distinct groups, with one group containing objects with certain data, while the other group containing objects with uncertain data, and (iii) conventional skyline algorithm which is utilized for computing the local skylines over the group with certain data.

The real-time phase incorporates the conventional skyline algorithm and three methods, namely: *Trapezoid Fuzzification*, *Sum Aggregation*, and *Centroid Defuzzification* for predicting the exact value of a continuous range value in order to seamlessly employ the conventional skyline algorithm to compute the final skylines.

3.2. The CSQUiD framework

The Proposed framework for processing constrained skyline queries over uncertain data (*CSQUiD*) as presented in Figure 2, is an extension of the *SQUiD* framework.

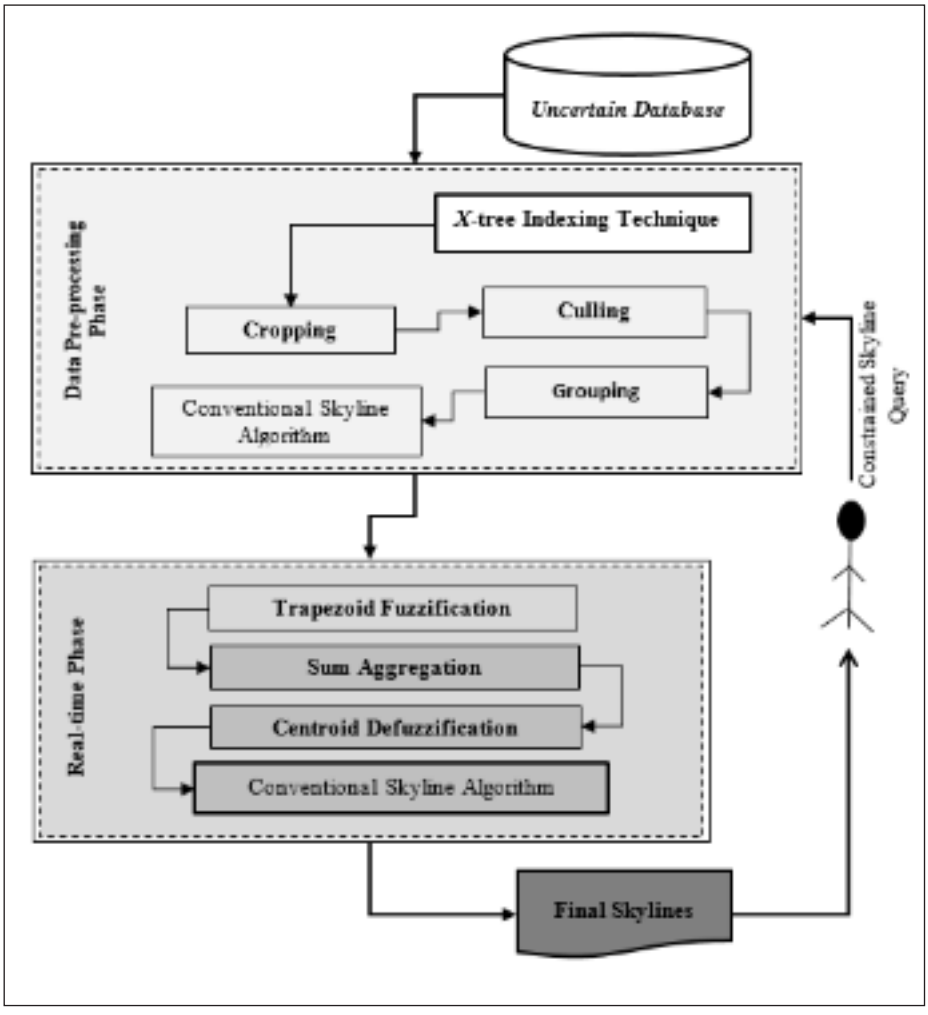

Figure 2. The Proposed *CSQUiD* framework for processing constrained skyline queries over uncertain data.

This has become imperative, as identifying whether an object satisfies a constrained skyline query is straightforward when dealing with certain data. Thus, objects whose exact values are within a constrained skyline query are selected for skyline computation. Likewise, objects with uncertain data need additional evaluation before computing the final skylines to the constrained skyline query. Invariably, these compounds and increases the process of evaluating constrained skyline queries over uncertain data. Thus, minimizing the computational cost is achieved by pruning off MBRs whose vertex is dominated by the vertex of other MBR. Instead of evaluating objects of each MBR, only objects within the MBR with the vertex are processed for the final skylines.

Consequently, the number of pairwise comparisons among objects is significantly reduced. Primarily, the CSQUiD framework is proposed for streamlining the process of computing constrained skyline query over uncertain data by evaluating only non-dominated objects within MBR, whose vertex lies within the constrained query. This was achieved by leveraging an indexing structure to minimize computational cost by pruning off dominated objects within each MBR in order to reduce the number of pairwise comparisons among objects. This framework consists of two phases, namely: Data Pre-processing and real-time. The Data Pre-processing phase employs an indexing technique to organize and store uncertain data into MBRs, while MBRs that either overlap or intersect or lie within a given constrained query are selected using the Crop-ping method. The Culling method is utilized for identifying the left coordinates and left-most coordinates in identifying the non-dominated MBRs whose objects will eventually be the local skylines. Meanwhile, the Grouping method group the objects of the selected MBRs into two distinct groups, and the conventional skyline algorithm computes the

local skylines of the selected MBRs. The real-time phase, the *Trapezoid Fuzzification*, *Sum Aggregation*, and *Centroid defuzzification* methods with conventional skyline algorithms are used to compute the final skylines.

3.3. The ASQUiD framework

The *ASQUiD* framework consists of two phases, namely: Data Pre-processing and real-time, as presented in Figure 3. An index-based materialized view framework to avoid the unnecessary skyline computations of multiple constrained skyline queries. Intuitively, the information of the MBRs as well as their candidate skylines that have been identified in the computations of earlier constrained skyline queries are saved to be utilized in processing subsequently constrained skyline queries in which constraints lie within, overlap, or intersect with the space of the saved MBRs. At the Data Pre-processing phase, the whole search space that represents the data is partitioned into a unit space, while the local skylines for each MBR are derived. Subsequently, the derived local skylines are organized into MBRs in order to seamlessly compute the final skylines to a constrained skyline query in the real-time phase.

The real-time phase consists of the Merging and Glean methods for identifying the candidate skylines, while the skyline algorithm is employed to compute the final skylines over the candidate skylines derived. Basically, this is meant to speed up the central processing unit processing time in processing a constrained skyline query as it eliminates the need to always

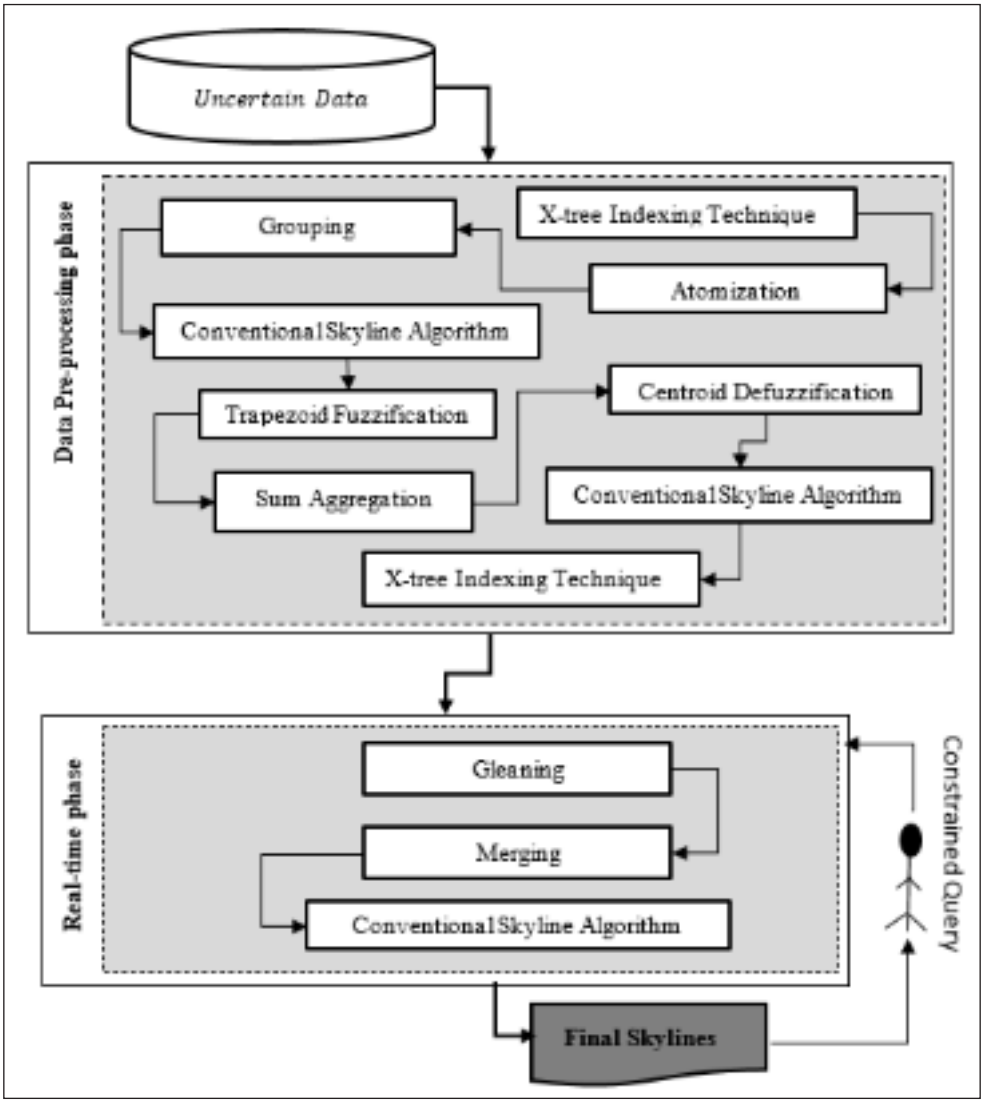

Figure 3. The proposed *ASQUiD* framework for Abating recomputation of constrained skyline queries over uncertain data.

process skylines over a similar set of objects each time a constrained skyline query is to be evaluated.

Computing constrained skyline queries without utilizing the result of the previous computations of skyline queries is unwise as some of the queries might cover the same part of the database which can be easily identified based on the conditions specified in the queries.

4. Conclusion

Today, we are in the era of making multi-criteria decisions based on analysis of available data collected from autonomous databases that usually contain uncertain data. Without a doubt, streamlining the process of processing skylines in providing answers to user-specified queries is inevitable. In our work, we proposed *SQUiD*, *CSQUiD*, and *ASQUiD* frameworks that combine an index-based technique with a prediction method to reduce the computational time for computing skylines and processing skylines over uncertain data with high dimensionality.

Acknowledgment

This work was supported by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS/1/2016/ICT04/UPM/01/2) and the Universiti Putra Malaysia.

Referenc es

Afshani P, Agarwal PK, Arge L, Larsen KG, Phillips JM. (Approximate) Uncertain skylines. In Proceedings of the 14th International Conferenceon Database Theory, ACM, Uppsala, Sweden, 2011, pp 186–96.

Aggarwal CC, Philip SY. A survey of uncertain data algorithms and applications. IEEE Trans Knowl Data Eng 2009; 21(5):609–23.

Alwan AA, Ibrahim H, Udzir NI. A framework for identifyingskylines over incomplete data. In 2014 3rd International Conference on Advanced Computer Science Applications and Technologies, IEEE, Amman, Jordan, 2014, pp 79–84.

Arslan S, Sa¸can A, Acar E, Toroslu IH, Yazici, A. Comparisonof multidimensional data access methods for feature-based image retrieval. In 2010 20th International Conference on Pattern Recognition, IEEE, Istanbul, Turkey, 2010, pp 3260–3.

Atallah MJ, Qi Y. Computing all skyline probabilities for uncertain data. In Proceedings of the Twenty-Eighth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, ACM, Providence, RI, 2009, pp 279–87.

Bartolini I, Ciaccia P, Patella M. Salsa: computing the skyline without scanning the whole sky. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management, ACM, Arlington, VA, 2006, pp 405–14.

Bartolini I, Ciaccia P, Patella M. Efficient sort-based skyline evaluation. ACM Trans Database Syst 2008; 33(4):31.

Borzsonyi S, Kossmann D, Stocker K. The skyline operator. In Data Engineering, 2001. Proceedings. 17th International Conference, IEEE, Heidelberg, Germany, 2001, pp 421–30.

Catania B, Jain LC. Advanced query processing. Springer, Berlin, Heidelberg, Germany, 2013.

Chan CY, Eng PK, Tan KL. Efficient processing of skylinequeries with partially-ordered domains. In 21st International Conference on Data Engineering (ICDE'05), IEEE, 2005, vol. 189, pp. 190–191.

Chen L, Lian X. Efficient processing of metric skyline queries. IEEE Trans Knowl Data Eng 2008; 21(3):351–65.

Chomicki J, Godfrey P, Gryz J, Liang D. Skyline with presorting: theory and optimizations. In Kłopotek MA, Wierzchoń ST, Trojanowski K (Eds.). Intelligent information processing and web mining, Springer Berlin, Heidelberg, Germany, pp 595–604, 2005.

Cui B, Lu H, Xu Q, Chen L, Dai Y, Zhou, Y. Parallel distributed processing of constrained skyline queries by filtering. In 2008 IEEE 24th International Conference on Data Engineering, IEEE, Cancun, Mexico, 2008, pp 546–55.

Dellis E, Seeger B. Efficient computation of reverse skyline queries. In Proceedings of the 33rd International Conference on Very large data bases, VLDB Endowment, Sanya, China, 2007, pp 291–302.

Gothwal H, Choudhary J, Singh D. The survey on skyline queryprocessing for data-specific applications. In Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT), Jaipur, India, 2018, pp 26–7.

Gulzar Y, Alwan AA, Salleh N, Shaikhli I. Skyline query processing for incomplete data in cloud environment. In Proceedings of the 6th InternationalConference on Computing & Informatics, Kuala Lampur, Malaysia, 2017, pp 567–76.

Gulzar Y, Alwan AA, Abdullah RM, Xin Q, Swidan MB. Scsa: evaluating skyline queries in incomplete data. Appl Intell 2019; 49(5):1636–57.

Guttman A. R-trees: A dynamic index structure for spatial searching. In Proceedings of ACM management of data (SIGMOD), ACM, Boston, MA, 1984, vol. 14.

Han X, Li J, Yang D, Wang J. Efficient skyline computation onbig data. IEEE Trans Knowl Data Eng 2013; 25(11):2521–35.

Jiang B, Pei J, Lin X, Yuan Y. Probabilistic skylines on uncertaindata: model and bounding-pruning-refining methods. J Intell Inf Syst 2012; 38(1):1–39.

Kemper A, Moerkotte G. Advanced query processing in object basesusing access support relations. In Proceedings of the 16th International Conferenceon Very Large Data Bases, Morgan Kaufmann Publishers, Burlington, MA, 1990, pp 290–301.

Khalefa ME, Mokbel MF, Levandoski JJ. Skyline query processing for uncertain data. In Proceedings of the 19th ACM International Conferenceon Information and knowledge management, ACM, 2010, pp 1293–6.

Kontaki M, Papadopoulos AN, Manolopoulos Y. Continuous top-k dominating queries. IEEE Trans Knowl Data Eng 2011; 24(5):840–53.

Kossmann D, Ramsak F, Rost S. Shooting stars in the sky: anonline algorithm for skyline queries. In Proceedings of the 28th International Conference on Very Large Data Bases, VLDB Endowment, 2002, pp 275–86.

Lawal MM, Ibrahim H, Sani NFM, Yaakob R. Analyses of indexing techniques on uncertain data with high dimensionality. IEEE Access, 2020; 8:74101–17.

Lee KC, Lee WC, Zheng B, Li H, Tian Y. Z-sky: an efficient skyline query processing framework based on z-order. VLDB J 2010; 19(3):333–62.

Li X, Wang Y, Li X, Wang G. Skyline query processingon interval uncertain data. In 2012 IEEE 15th International Symposium on Object/Component/Service-Oriented Real-Time Distributed Computing Workshops, IEEE, Shenzhen, China, 2012, pp 87–92.

Liu X, Yang DN, Ye M, Lee WC. U-skyline: a new skyline queryfor uncertain databases. IEEE Trans Knowl Data Eng 2013; 25(4):945–60.

Nam B, Sussman A. A comparative study of spatial indexing tech-niques for multidimensional scientific datasets. In Proceedings. 16th International Conference on Scientific and Statistical Database Management, IEEE, 2004, pp 171–80.

Papadias D, Tao Y, Fu G, Seeger B. Progressive skyline computation in database systems. ACM Trans Database Syst 2005; 30(1):41–82.

Pei J, Jiang B, Lin X, Yuan Y. Probabilistic skylines on uncertaindata. In Proceedings of the 33rd International Conference on Very large data bases, VLDB Endowment, 2007, pp 15–26.

Saad NHM, Ibrahim H, Alwan AA, Sidi F, Yaakob R. A framework for evaluating skyline query over uncertain autonomous databases. Proc Comput Sci 2014; 29:1546–56.

Saad NHM, Ibrahim H, Sidi F, Yaakob R. Non-index based sky-line analysis on high dimensional data with uncertain dimensions. In International Baltic conference on databases and information systems, Springer, Cham, Switzerland, pp 272–86, 2018.

Saad NHM, Ibrahim H, Sidi F, Yaakob R. Skyline probabilities with range query on uncertain dimensions. In Advances in computer communication and computational sciences, Springer, The Gateway, Singapore, pp 225–42, 2019.

Saad NHM, Ibrahim H, Sidi F, Yaakob R, Alwan AA. Computing range skyline query on uncertain dimension. In International Conference on Database and Expert Systems Applications, Springer, Cham, Switzerland, 2016, pp 377–88.

Sajeev J. Noorjahan V. Top-k dominating queries on incomplete data: a survey. IEEE Trans Knowl Data Eng 2016; 28:252–66.

Tan KL, Eng PK, Ooi BC. Efficient progressive skyline computation. In Proceedings of the 27th VLDB Conference, Rome, Italy, 2001, vol. 1, pp 301–10.

Wang Y, Li X, Li X, Wang Y. Asurvey of queries over uncertaindata. Knowl Inf Syst 2013b; 37(3):485–530.