Journal of Multimedia Information System VOL. 6, NO. 1, March 2019 (pp. 37-42): ISSN 2383-7632(Online)
https://doi.org/10.33851/JMIS.2019.6.1.37

A Study on Representative Skyline Using Connected Component Clustering

Jong-Hyeok Choi, Aziz Nasridinov*

Abstract

Skyline queries are used in a variety of fields to make optimal decisions. However, as the volume of data and the dimension of the data increase, the number of skyline points increases with the amount of time it takes to discover them. Mainly, because the number of skylines is essential in many real-life applications, various studies have been proposed. However, previous researches have used the k-parameter methods such as top-k and k-means to discover representative skyline points (RSPs) from entire skyline point set, resulting in high query response time and reduced representativeness due to k dependency. To solve this problem, we propose a new Connected Component Clustering based Representative Skyline Query (3CRS) that can discover RSP quickly even in high-dimensional data through connected component clustering. 3CRS performs fast discovery and clustering of skylines through hash indexes and connected components and selects RSPs from each cluster. This paper proves the superiority of the proposed method by comparing it with representative skyline queries using k-means and DBSCAN with the real-world dataset.

Key Words: Skyline queries, Representative skyline queries, Density-based clustering, Hash index.

I. INTRODUCTION

A skyline query is a method of discovering points that are not dominated by any other points in a given data [1]. Here, a point dominates another point if it has the same value in all dimensions or better value in at least one dimension. Due to this feature, the skyline query is used in a variety of fields that require multiple-criteria decision making. However, as the volume and dimensions of data increase, discovering skyline points takes long computational time.

There have been many studies that mainly focus on two improvements: reducing the computational time taken to discover the skyline points and returning the most representative set of skyline points to the user [2–9]. The former case is essential because the number of discovered skyline points increases as the number of dimensions of the data increase. Thus, a representative skyline query can discover points that best represent the entire points resulting from the skyline query. Various techniques are used to discover representative skyline points (RSP), such as maximum domination, distance between data, dominant regions, and clustering techniques, e.g., k-means. However, it takes long computational time for existing studies to discover RSP as these studies perform a full scan of the entire data. Moreover, the representativeness of the RSP can be reduced in the process of selecting the k constraint of k-means, desired by the user.

In this paper, we propose a new Connected Component Clustering based Representative Skyline (3CRS) query that can quickly select RSP compared with existing studies. 3CSR does so by eliminating points that cannot become skyline points early in comparison process using hash index [10], and simultaneously creating clusters that can reflect the characteristics of RSP more efficiently. More precisely, the contributions of the paper are as follows:

- We use a method to create clusters quickly for finding RSP in high-dimensional data through the hash index and connected component clustering technique [11].
- We propose a dominating region-based representatives selection scheme for selecting RSPs from clusters obtained through connected component clustering.
- We show the superiority of the proposed method through a comparison experiment with existing methods using real-world dataset.

The remainder of this paper is organized as follows. In Section II, we discuss the related work. In Section III, we

Manuscript received March 26, 2019; Accepted March 27, 2019. (ID No. JMIS-19M-03-012)
Corresponding Author (*): Aziz Nasridinov, Chundaero 1, Seowon-gu, Cheongju, Chungbuk, Korea, aziz@chungbuk.ac.kr
Jong-Hyeok Choi, Dept. of Computer Science, Chungbuk National University, Cheongju, Chungbuk, Korea, leopard@chungbuk.ac.kr

describe the theoretical background of the proposed method. Section IV presents the experimental results. Section V summarizes and highlights the conclusions of the paper.

II. RELATED STUDY

The simple method to discover skyline points is to check dominance in the entire dataset. For example, Borzsonyi et al. [1] proposed Block-Nested Loop (BNL) and Divide-and-Conquer (D&C) skylines. However, these methods do not exploit the characteristics of the monotone order, and thus, cannot efficiently discover the skyline points. Chomicki et al. [2] proposed Sort-Filter-Skyline (SFS). SFS is a method first calculates an entropy score of points in the dataset, and then performs a procedure similar to BNL. As a result, SFS can discover the skyline points with a small number of comparisons. However, this method is still not suitable for quickly discovering the skyline from a large amount of data, as it must perform comparisons on whole data to discover the skyline points. To solve this problem, subsequent studies focused on eliminating the points that cannot become skyline points by using data partitioning technique such as angle-based space partitioning [3] or using indexing technique like BBS [4] and Z-SKY [5].

However, the increase in dimensions and data volume may increase the number of skyline points. Thus, even if the user discovers the skyline points quickly, it is difficult to utilize them. To solve this problem, representative skyline queries methods select the skyline points that are the most representative among the entire skyline point set. Lin et al. [6] proposed a top-k Representative Skyline Point (top-k RSP) that selects the top-k of RSPs that dominate the vast majority of the skyline point set. However, the top-k RSP does not adequately represent the entire skyline points because only a few skylines that are skewed and dominated by the data are selected as RSPs. To solve this problem, Tao et al. [7] proposed a Distance-based Representative Skyline (DRS) that selects k RSPs by applying a cluster distance optimization problem called a k-center. DRS can select a suitable type of RSP according to the distribution of the skyline. However, due to the problem of constant distance calculation to select the most optimal RSP, the proposed method takes long computational time to discovery RSPs. Bai et al. [8] proposed k-Largest dominance skyline (k-LDS) using the dominating region to select RSP from the data stream. The method selects k skyline points with the broadest range of dominating regions from the skyline as RSPs. It has the advantage of faster RSP selection than the previous methods. However, it is not suitable for higher dimensional data due to the scalability problem. In other words, it becomes difficult to detect redundancy of dominating region as dimension increases.

RSP can also be obtained using a clustering-based approach like SkyCluster proposed by Huang et al. [9]. SkyCluster divides skyline points into k clusters through k-means clustering and then selects the most RSPs from each cluster using a scoring function. However, this approach does not reflect the distribution of the actual skyline points when k, which differs from the actual number of clusters, does not reflect the distribution of the actual skyline. Also, the problem of reduction of representativeness caused by the use of the wrong k can be a problem for the SkyCluster. Thus, a new method to solve the disadvantages of k is required.

III. CONNECTED COMPONENT CLUSTERING BASED REPRESENTATIVE SKYLINE QUERY

In this section, we present the Connected Component Clustering based Representative Skyline Query (3CRS). The first step of the proposed method is to measure the dominating region through a hash index and remove the data that cannot become a skyline point early in the comparison stage. In the second step, we show how to perform connected component clustering through a hash index. In the final step, we describe the dominating region based representative selection technique for selecting RSPs from each cluster.

3.1 Hash Index for Dominating Region

Spatial hashing is a method of dividing a given data space into grids and managing each space through hash keys and hash buckets. It enables to manage the data of a multidimensional space in a one-dimensional hash table. Here, it is essential to identify the location information of the data by using the hash key. In our case, we generate a hash key combined with the location information of the data, such as the hash index proposed by Choi et al. [10]. Here, the

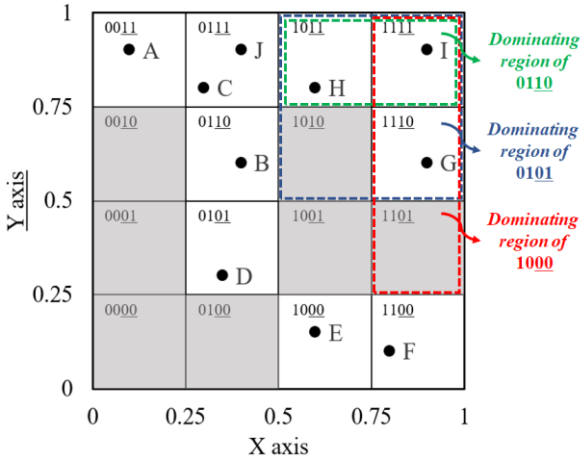

Fig.1. Example of dominating region.

generated hash key can represent the data stored in the hash bucket and can quickly detect the dominating region using the location information of the hash key.

Fig. 1 shows the dominating region through the hash index and hash key generated in 2-D space. From Fig. 1, we can observe that a hash key containing spatial information makes it possible to remove data that cannot become a skyline point. In 3CRS, a hash index is generated first for given data, and then the hash key is used to discover and remove the hash buckets belonging to the dominating region, thereby quickly removing data that cannot become a skyline point. Here, skyline points are discovered only from the hash buckets that have not been removed through the above process.

3.2 Connected Component Clustering using hash key

Clustering is a method of classifying similar data into the same group. Using this clustering feature, methods such as DRS and SkyCluster classify similar skyline points into identical clusters using k-center or k-means and then perform the RSP selection process. However, since the number of clusters varies according to k, there is a problem in that if the optimal k is not selected, the representativeness of RSPs can be reduced.

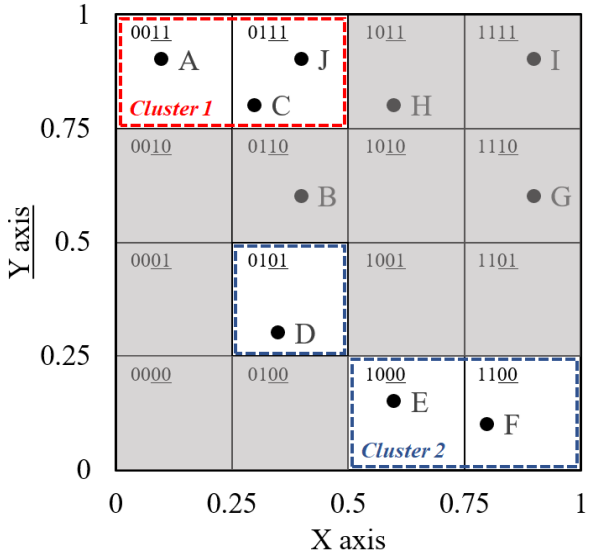

Fig. 2. Example of connected component clustering.

To solve this problem, 3CRS clusters the hash buckets spatially adjacent to each other using hash key location information and connected component labeling, and then discover the actual clusters necessary for RSP selection. At this time, in performing connected component clustering in the d- dimension, because data belonging to an arbitrary hash bucket can be classified into one cluster adjacent to data of a hash bucket located on a diagonal line. Thus, in 3CRS, the hash buckets corresponding to the (3d-1)-connected neighborhood for the d-dimension are classified into the same cluster.

Fig. 2 shows the connected component clustering of the skyline points and their hash buckets obtained from the Fig. 1. The figure shows the clusters obtained from the skyline {A, C, D, E, F} in Fig. 1. Assuming that B is not dominated and selected as a skyline, only one cluster is created in this case as Cluster 1. Cluster 2 can be connected via hash bucket 0101. The proposed clustering procedure enables the creation of clusters quickly and effectively reduces the computational time required for RSP discovery. Also, it does not have k dependency compared with similar methods.

3.3 Representatives selection using dominating regions

In the last stage of 3CRS, RSPs are selected from each cluster generated through connected component clustering. To do this, 3CRS uses the size of the dominating region as a method to identify representativeness. This notion occurs because the skyline explores the minimal set of data that can be represented through the concept of domination. However, the operation of performing the dominating count of the actual skyline point as in the top-k RSP has a problem of causing unnecessary operations to be generated in the process of discovering the skyline points. To solve this problem, we use the dominating region to measure representativeness.

We can easily calculate the range of the dominating region through the hash bucket using the hash key. Here, we select a hash bucket with the dominating region from each cluster as the representative bucket of the cluster. Then, the representative skyline point is selected as the data having the smallest entropy score among the data belonging to the corresponding bucket.

IV. EXPERIMENT RESULTS

In this section, we compare the performance of discovering RSP using existing methods to show the superiority of proposed 3CRS. To do this, we compare the performance of 3CRS with the RSP discovery method of two types of k-means and DBSCAN based on SFS and SFS results, which are the most representative skyline discovery methods. For the objective analysis, a real-world dataset, Gas sensors for home activity monitoring dataset (below HT) [12] was used. The HT data consists of million data collected from 8 sensors related to the temperature and humidity.

All representative skyline query methods were implemented using VC ++ 12.0, and the experiments were carried out on an Intel i7-6700 3.4 GHz processor with 64-bit Windows 10 pro and 16 GB of main memory.

Finally, in the case of k-means, the number of clusters obtained as a result of 3CRS is used as the k parameter. In the case of DBSCAN, the epsilon distance of 0.158, which is calculated based on the dimension separation interval of 0.05 for the hash index generation of 3CRS, and 2 for minimum points are used.

4.1 Skyline performance test

We perform comparisons on skyline discovery using SFS and 3CRS. SFS is the most widely used in the skyline. In this experiment, we compare the computation time taken to discovery skyline points and number of comparisons used in skyline discovery. The result of the experiments is shown in Fig. 3.

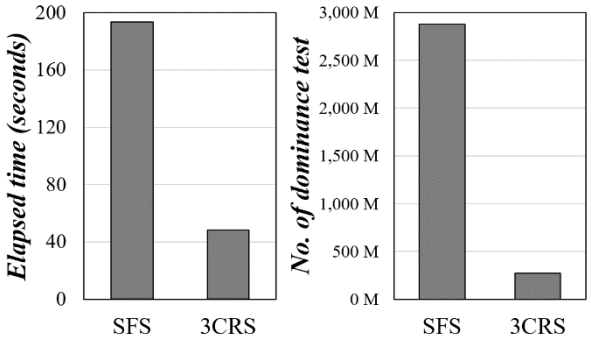

Fig. 3. Skyline performance.

Experimental results show that 3CRS has more than 4 times faster skyline discovery time compared to SFS, and the number of comparisons between data is also reduced 10.5 times. This improvement occurs because the data that cannot be skyline through the hash index is removed early. This result also indicates the reason why the next step of clustering can be performed quickly.

4.2 Clustering and RSP selection performance test

In the comparison of the clustering and RSP selection performance, we measured two parameters: cluster formation time from skyline generated from the previous step and RSP selection from each formed cluster. Fig. 4 shows the performance of three clustering techniques, such as 3CRS, k-means, and DBSCAN. The result of the experiment demonstrates that 3CRS performs clustering 46.3 times and 123.5 times faster than k- means and DBSCAN, respectively.

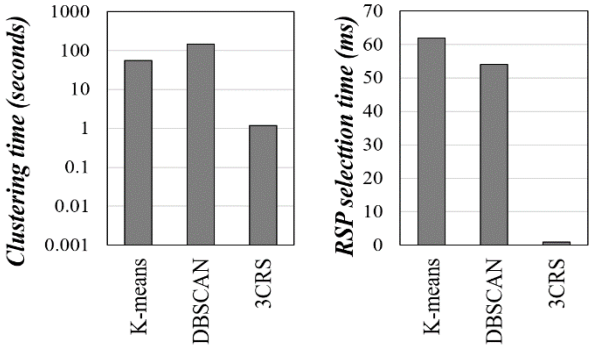

Fig. 4. Clustering and RSP selection performance.

This improvement is achieved due to the connected component clustering through the hash index. At the same time, the dominating region computation also greatly reduces the time required to find the data with the largest dominating region from the cluster to select the RSP. These results show that 3CRS can solve performance problems when selecting RSP in high-dimensional big data.

V. CONCLUSION

In this paper, we have proposed a Connected Component Clustering-based Representative Skyline Query (3CRS) which can select RSP quickly through connected component clustering. The proposed method effectively solves the problems caused by high-dimensionality and k dependency of existing representative skyline queries. 3CRS removes data that cannot be skyline early by using a hash index based on spatial hashing, classifies similar skyline points into the same cluster through connected component clustering using the hash index and then selects region-based RSPs. The 3CRS is superior in selecting the RSPs compared with SFS, k-means, and DBSCAN in terms of computational time required to discover skyline points and cluster formation. The experiment results highlight the academic contribution of the proposed method. In future studies, the efficiency of the RSP itself will be evaluated not only by comparing the performance with the k- based representative skyline queries such as DRS or SkyClust, but also by evaluating the representativeness objectively.

REFERENCES

[1] S. Borzsony, D. Kossmann, and K. Stocker, “The skyline operator,” in *Proceeding of ICDE.*, pp. 421–430, 2001.

[2] J. Chomicki, P. Godfrey, J. Gryz, and D. Liang, “Skyline with presorting,” in *Proceeding of ICDE*, pp. 717–719, 2003.

[3] A. Vlachou, C. Doulkeridis, and Y. Kotidis. "Angle-based space partitioning for efficient parallel skyline computation," in *Proceeding of SIGMOD*, pp. 227-238, 2008.

[4] D. Papadias, Y. Tao, G. Fu, and B. Seeger, "An optimal and progressive algorithm for skyline queries," in *Proceeding of SIGMOD*, pp. 467–478, 2003.

[5] K.C. Lee, W.-C. Lee, B. Zheng, H. Li, Y. Tian, "Z-Sky: An Efficient Skyline Query Processing Framework Based on Z-Order", *The VLDB J.*, Vol. 19, No. 3, 333-362, 2010.

[6] X. Lin, Y. Yuan, Q. Zhang, and Y. Zhang, "Selecting stars: The k most representative skyline operator," in *Proceeding of ICDE*, pp. 86–95, 2007.

[7] Y. Tao, L. Ding, X. Lin, and J. Pei, "Distance-based representative skyline," in *Proceeding of ICDE*, pp. 892–903, 2009.

[8] M. Bai, J. Xin, G. Wang, L. Zhang, R. Zimmermann, Y. Ye, X. Wu, "Discovering the k representative skyline over a sliding window," *IEEE Trans. Knowl Data Eng*, Vol. 28, No. 8, 2041–2056, 2016.

[9] Z. Huang, Y. Xiang, B. Zhang, X. Liu, "A clustering based approach for skyline diversity," *Expert Systems with Applications*, Vol. 38, No 7, 7984-7993, 2011.

[10] N. Feiping, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in *Proceeding of SIGKDD*., pp. 977-986, 2014.

[11] J. Choi, K. Yoo, A. Nasridinov, "An Index Structure for Efficiently Handling Dynamic User Preferences and Multidimensional Data," *Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, Vol.7, No.7, pp. 925-934, 2017.

[12] R. Huerta, T. Mosqueiro, J. Fonollosa, N. Rulkov, I. R. Lujan. *Online Decorrelation of Humidity and Temperature in Chemical Sensors for Continuous Monitoring*. Chemometrics and Intelligent Laboratory Systems, 2016

Authors

Jong-Hyeok Choi received BS degree in computer education from Chungbuk National University, South Korea in 2015, and MS degree in computer science from Chungbuk National University. He is currently a PhD candidate in Data Analytics laboratory led by Professor Aziz Nasridinov in the Department of Computer Science, Chungbuk National University. His research interests include database systems and data mining.

Aziz Nasridinov received BS degree in information technologies from Tashkent University of Information Technologies, Uzbekistan, in 2006, and MS and PhD degrees in computer engineering from Dongguk University, South Korea. He is currently an associate professor in Data Analytics laboratory, Department of Computer Science, Chungbuk National University. In the past, he has worked as a post-doctoral researcher at Sookmyung Women's University, and as a research professor at Dongguk University. His research interests include database systems, data mining and parallel and distributed computing. He has published more than 10 scientific papers in various international journals. He is also an editorial board member of several international journals.