분산환경에서 스카이라인 질의를 처리하는 다차원 그리드 기법

이하, 장수민, 유재수
충북대학교 정보통신공학과
lihe@chungbuk.ac.kr, jsm@chungbuk.ac.kr, yjs@chungbuk.ac.kr

Multi-Layers Grid Method for Processing Skyline Queries in Distributed Environment

He Li, Sumin Jang, Jaesoo Yoo
Department of Information and Communication Engineering,
Chungbuk National University

1. Introduction

Most of the previous skyline literatures have primarily focused on providing efficient skyline algorithms on centralized data set. In practice, however, the vast numbers of independent data are often collected from multiple sources that stored in distributed servers. A naïve approach to process the distributed skyline queries is to send the skyline queries to all the connected servers, which in turn process the skyline queries locally and report the result to *QS*. The *QS* evaluates the global skyline. Joao B. et al. in [1] proposed a grid-based strategy for distributed skyline query processing (*AGiDS*), which use a grid-based data structure to capture the data of each server. However, if the cells of local servers that are transferred to the query server (*QS*) are overlapped, a lot of unnecessary data are transferred to the *QS*. Therefore, this paper proposes a multiple layers grid method for skyline queries (*MGSD*) in the widely distributed environment.

2. The proposed method

The proposed method assumes that each server shares a common grid structure. If cell *i* is dominated by cell *j*, which means that all the data points of cell *i* is dominated by any data point of cell *j*. We define the cells which contain skyline data as region-skyline, as shown in Figure 1, the shaded area corresponds to the region-skyline. If the region-skyline overlaps at different servers, we define these regions as overlap region-skyline, e.g. the cell *A* and *B*. If the overlap region-skyline contains more data points (*rCount*) than the predefined threshold value *k* (the value of *k* is defined according to the practical application), it is defined as hot overlap region-skyline, e.g. the cell *B*.

The proposed *MGSD* method comprises three basic stages: planning, analyzing and execution. At the beginning of planning stage, a skyline query can be initiated by a query server (*QS*). Each server S_i computes its local region-skyline by using an existing centralized skyline algorithm. The *QS* contacts all the connected servers and obtains the region-skyline information. In the analyzing phase, The *QS* analyze the received cells and evaluate global region-skyline. If the hot overlap region-skyline is occurred in the connected servers, it can be handled by creating an upper layer grid. As shown in Figure 1, cell *B* is hot overlap region-skyline both at server S_1 and S_2, which is converted to an upper layer grid. Then, the data points of cell *B* are managed by the upper layer grid and the local region-skyline of the upper layer grid is computed. Then, the *QS* requests the local region-skyline of the upper layer grid and computes the global region-skyline. If the hot overlap region-skyline still exists on the upper layer grid, more layers grid can be generated. Notice that, cell *A* is a overlap region-skyline both at S_1 and S_2, but it is not converted to a upper layer grid, this is because there are only a small amount of data in cell *A* both at S_1 and S_2, and

이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업의 결과임 (No. 2009-0089128).

processing these data is efficient than processing the upper layer grid. In the execution stage, only the data points existed in the global region-skyline of each server are requested for the final global skyline computation. Since most of unnecessary local skyline data points are filtered out, the *MGSD* method reduces both the communication cost and processing cost. In the execution stage, only the local skyline data within cells *A*, *D*, *B-1*, *B-2* of S_1's grid, and cells *A*, *C* of S_2's grid are transferred to the *QS* for final global skyline computation.

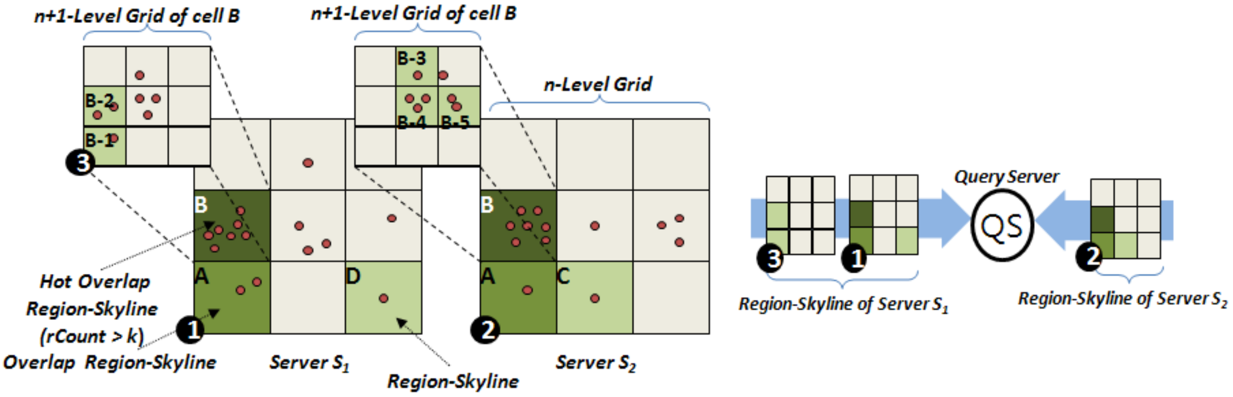
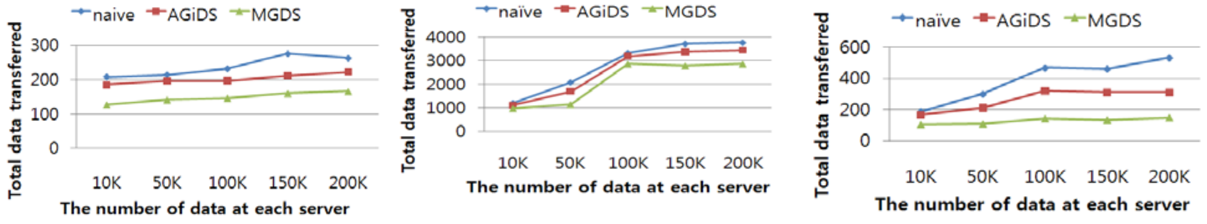


Figure 1. The proposed *MGSD* method.

3. Experiment evaluation

We present the experimental results comparing *AGiDS* method [1] and the naive method with our proposed method *MGSD*. We examine the performance of the methods by varying the number of data tuples from 10K to 200K at each server. The *MGSD* method outperforms the *AGiDS* method and the naive method in all of the three data sets since more non-promising data points are filtered out by the multiple layer grids mechanism of the *MGSD* method.



(a) Independent data set (b) Anti-correlated data set (c) Correlated data set

Figure 2. Evaluation of the total transferred data for various data distributions

4. Conclusions

This paper studies skyline queries over the horizontally partitioned data set. As the relevant data are scattered at several servers, the skyline query in distributed environment requires gathering a large number of data from the connected servers. The proposed *MGSD* method employs multiple layers grid mechanism to minimize the unnecessary data at each server before processing skyline queries.

References

[1] J. B. Rocha-Junior, A. Vlachou, C. Doulkeridis, K. Norvag, “AGiDS: A Grid-based Strategy for Distributed Skyline Query Processing,” *In Proc.* Globe, 2009, pp.12-23.