A Parallel and Distributed Approach for Diversified Top-k Best Region Search*

Hamid Shahrivari
Eindhoven University of Technology
h.shahrivari.joghan@tue.nl

The amount of geospatial data generated from social networks, sensors, smart phone applications, tracking devices and so on is constantly increasing [1]. Analyzing big geospatial data at scale is of paramount importance for numerous applications in various areas such as geomarketing, mobile advertisement, urban planning, tourism and logistics. In many cases, the analysis involves identifying areas where the intensity of a studied phenomenon is maximized.

Given a set of points, the *Best Region Search* problem finds the optimal location of a rectangle of a specified size such that the value of a user-defined scoring function over its enclosed points is maximized. The *top-k Best Region Search* means detecting k regions that maximize the score function at the most, and in our case, these k regions are not overlapping each other. The k-BRS problem has been introduced in [2], presenting an algorithm that can compute top-k best regions progressively. Still, the state-of-the-art solution proposed in [2] operates in a centralized setting and cannot scale as the input dataset size grows; indicatively, it needs around 30 seconds to identify the top-10 non-overlapping regions of size 200×200 m^2 even on moderate-size datasets – around 100 thousand points.

We explore the solution space of the problem, and propose two Spark-based distributed algorithms for solving it. The first is a multi-round algorithm (*MR*), with the number of rounds bounded by the value of parameter k. The second is a single-round algorithm (*SR*), which guarantees that the top-k best regions can be identified within a single round of interactions. Also, we explore the space between the single-round and the multi-round algorithm, proposing a *hybrid* algorithm, and discussing additional optimizations.

As input of all algorithms, data points are spatially partitioned by applying a uniform grid over the data. Each grid cell (each partition) is then assigned to one of the N available nodes (in our case, the Spark workers).

Our first method is an incremental, multi-rounds algorithm (*MR*) that gradually builds the global top-k list of results by retrieving local top-k results for each partition at each round. While aggregating the local top-k results received at the end of each round, the coordinator resolves overlaps, and, if needed, contacts again the affected workers, informing them about the occurred overlaps and asking them for accordingly revised top-k results. The multi-round algorithm enables processing of datasets with sizes that could not be practically processed by the centralized algorithm. Still, the iterative nature of the algorithm (and overlapping regions from different partitions) may lead to a potentially high overhead and poor performance, and (MR) may take up to k rounds to complete.

To overcome this drawback, we introduce a single-round algorithm which maintains auxiliary information per region that is used during the reduction phase for handing overlapping regions from different partitions. The intuition for the single-round algorithm (*SR*) is the following. When processing each partition, a sufficient number of regions (typically larger than k) will be computed and sent to the coordinator, such that it is guaranteed that the coordinator has sufficient candidates to assemble the global top-k results (discarding the ones with overlap) without further communication to the workers. The single-round algorithm addresses this issue by forming a *dependency graph* of the identified candidate regions at each partition. These graphs allow each node to establish a lower bound on the number of regions contained in the local results that will be accepted by the coordinator, if their score is sufficiently high. However, creating and maintaining dependency graph for large value of k is a time consuming process, also at the aggregation step, (SR) takes small number of regions from each dependency graph, and rest of regions are ignored.

To address this issue, we propose a hybrid algorithm (*HY*), which covers the space between the single-round and multi-round algorithms, balancing the number of rounds and the number of results expected from each round. The intuition is that we can execute the single-round algorithm, but now requesting a smaller number of regions $k' << k$ per partition, aggregate the partial results, and then progressively ask for more results only from the partitions from which we already consumed at least one region as final results. The partial results collected per round are a sorted subset of the final results (at least the next k' answers, but typically much more), and can be presented progressively to the user.

We implement all proposed algorithms in Spark and we thoroughly evaluate them using real-world datasets. Our evaluation results demonstrate the scalability and benefits of the proposed algorithms, and analyze their tradeoffs on queries and datasets of different properties. All three presented algorithms rely on distribution of the dataset and the expensive computational part over cluster resources, thereby allowing the processing of large datasets. As shown with experiments on real data, our most advanced algorithm, Hybrid, drastically reduces the required time by one order of magnitude compared to the multi-round, and by a factor of two compared to the single-round algorithm. Through our experiments, we detected top-*300* regions all around the world among more than 25 million points of interest in less than 5 minutes on our Spark cluster including 10 workers.

REFERENCES

[1] A. Eldawy and M. F. Mokbel. The era of big spatial data: A survey. *Foundations and Trends in Databases*, 6(3-4):163–273, 2016.
[2] D. Skoutas, D. Sacharidis, and K. Patroumpas. Efficient progressive and diversified top-k best region search. In *SIGSPATIAL*, pages 299–308, 2018.

*Joint work with M. Olma and A. Ailamaki from EPFL (Switzerland), O. Papapetrou from TU/e (Netherlands), and D. Skoutas from Athena Research Center (Greece)