論　文　の　要　旨

Information Filtering of Very Large Databases by Using Skyline Queries

（スカイライン問合わせを利用した大規模データベースの情報選別）

D105660 Mohammad Shumsul Arefin

Extended Abstract

Conventional SQL queries take exact input and produce complete result set. However, with massive increase in data volume in different applications, the large result sets returned by traditional SQL queries are not well suited for the users to take effective decisions. Therefore, there is an increasing interest in queries like top-k queries and skyline queries those produce a more concise result set.

Top-*k* queries rely on the scores of the objects to evaluate the usefulness of the objects. In this type of queries, users require to define their own scoring function by combining their interests. Based on the user defined scoring function, the system sorts the objects by their scores and outputs the top-k objects in the ranking list as the result. However, defining a scoring function by the users is a major draw of the top-k queries as in the large data sets where there are many conflicting criteria exist, it is very difficult for the users to define the scoring functions by themselves.

To overcome this disadvantage of top-k queries, skyline queries were proposed. Skyline queries do not rely on scoring functions to retrieve objects. Instead they use the concept of dominance relation. An object is said to dominate another object if it is not worse in any of the dimensions and is better in at least one of the dimensions. Given a set of objects with multiple dimensions, an object would not be retrieved if it is dominated by some other objects. From the result of skyline objects, the user can choose promising objects for them and make further inquiries. Therefore, such skyline query functions are important for several database applications, including customer information systems, decision support, data mining and visualization, and so forth.

In this thesis, we focus on information filtering of very large databases by using skyline queries. We proposed several novel techniques that can filter less important information from the databases. The main contributions of this thesis are on the following types of skyline queries:

(1) Privacy aware skyline sets queries from distributed databases
(2) Skyline queries for selecting spatial objects by utilizing surrounding environments
(3) Skyline queries for selecting spatial objects for groups

From the introduction of skyline queries in 2001, skyline queries are treated as an important approach for information filtering and there are many research works on skyline queries considering either a sole database or distributed databases with almost no consideration about the privacy of data. Although there are few considerations about the privacy of data while computing skyline queries from a sole database, there is no consideration about individual's privacy during the computation of skyline results from distributed databases. However, with the rapid growth of data volume and network infrastructure, in most cases data are stored at distributed databases nowadays.

Considering these facts, first part of this deals with preserving the privacy of data while computing skyline results from distributed databases. In this part, at first, we introduce a parallel computation framework for skyline sets queries from distributed databases. Let DB_1, DB_2, …,DB_m be the m databases with same schema. In addition, let s is the number of objects in each set and n is the total number of objects in m databases. We propose an agent-based technique to compute skyline s-sets from m such databases in such a way that the privacy of individual's is almost preserved. We used the concept of convex hull to retrieve skyline sets. In our approach, we only retrieve the skyline sets those are in the convex-hull. In this thesis, we mention such skyline sets as convex skyline sets. We do not consider other skyline sets as they cannot be the optimal point in any linear objective function and we can reduce the number of retrieved objects. In our framework, we use a coordinator who is responsible for the task of skyline sets queries.

The coordinator utilizes a divide-and-conquer strategy. It divides the distributed databases into several clusters and creates sub-coordinators for each cluster. For each cluster, the sub-coordinator computes the "local" top-s among the databases in the corresponding cluster.

After computing all "local" top-*s*, the coordinator merges all "local" top-*s* and finds "global" top-*s*. During the process, agents are used to preserve privacy of all "local" databases.
In addition of preserving individual's privacy, our approach is robust against the outliers and the frequently update situation.

However, in our above approach there are possibilities of disclosure of record's values from the return skyline sets in statistical compromisable situations. Considering this fact, in this part, we also introduce an efficient protection mechanism against statistical compromisable situations. In this part, we also consider the mechanism of dealing with missing values in the databases during skyline sets queries.

We can apply our above approach in several application domains such as in planning for stock markets investment, secure online business. We can also apply our above technique in the situations where records' values of the databases change frequently and in the databases with outliers.

The second part of this thesis focuses on selecting spatial objects from spatial databases considering environmental influences such as the presence of restaurants and supermarkets while selecting skyline hotels. Here, we utilize the concept of skyline queries. Conventional skyline queries select such spatial objects like hotels based on non-spatial attributes such as price and rating of hotels and there is no consideration of utilizing surrounding environments. In this thesis, we propose two methods for utilizing surrounding environments. Our first approach considers the best value in each attribute of each surrounding facility. In this approach, we used a grid-based data structure to keep the spatial information. For each grid, we pre-compute best value for each attribute of each surrounding facility. Then for each target facility in the user specified location, we compute the derived attributes' values for each object of the target facility based on the best values in the surrounding areas. Finally, we use a conventional skyline query algorithm to return the final skyline result. This approach is well suited for hotel recommendation systems.
Our second method of this part considers the number of objects of each type of facility in the surrounding environments. Here, we use the concept of *aR*-tree. This approach can help in real estate recommendations.

Besides theoretical guarantees, our comprehensive performance studies indicate that the techniques are very effective and efficient.

The last part of this thesis considers a problem of selecting spatial objects for a group of users located at different positions, since recent social network services can connect users and make such groups. Here, we also consider the concept of skyline queries. If a group wants to find a restaurant to hold a meeting, we have to select a convenient place for all users.

Although there are many research works on spatial skyline queries, none of them can efficiently compute skyline objects in such a scenario. Considering this fact, we propose an efficient skyline query algorithm to select spatial objects, considering both spatial and non-spatial information. In our approach, we compute the skyline results in two phases.

In the first phase, we compute skyline results in the spatial sub-space. Here, we utilize the concept of *Sum-Distance*, i.e. distances of an object from the query points, for spatial processing which can easily eliminate a large number of objects during the computation of skyline objects in the spatial sub-space.

Based on the skyline result of the spatial sub-space, the second phase efficiently computes whether some other objects can be in the skyline in the non-spatial sub-space. In this phase, we check the dominance of non-skyline objects of spatial sub-space against the skyline objects of spatial sub-space. Such an approach can easily eliminate many objects from domination check. Finally, we can retrieve all the objects considering both spatial and non-spatial features.

In addition of selecting groups' need in selecting restaurants or other recreation places, this approach is also applicable for disaster management and attack planning.

To summarize, this thesis addresses three different sophisticated information filtering methods based on the concept of skyline queries: skyline sets queries from distributed databases, skyline queries by utilizing surrounding environments, and spatial skyline queries for a group of users.