M2 Research internship

Constrained Group-based Skyline

Supervisors: Tassadit Bouadi and Véronique Masson – IRISA Rennes

Keywords: Skyline queries, Constraints, Skyline optimization

When analyzing data, one often wants to optimize simultaneously several characteristics. For example, a customer may want a smartphone with a battery life as long as possible, but as cheap as possible. Or a football team coach may want a player which is good both at scoring goals and at making passes. It is difficult to optimize simultaneously multiple and potentially conflicting criteria (*i.e. dimensions*), and several compromises can be acceptable. Skyline queries are a tool to discover and present such compromises. In a multidimensional space where the dimension domains are ordered, skyline queries return the objects which are not dominated by any other object. An object dominates another object, *if it is as good or better in all dimensions and better in at least one dimension*. For example, with the smartphone example several skyline choices may be the cheapest smartphone having a very long battery life, or the most endurant of the budget smartphones, as well as the smartphone being average on both criteria.

Several studies were carried out on skyline analysis *[1]* as a retrieval tool in a decisional context. Skyline queries can formulate multicriteria queries associated with *preferences* and obtain the top answers, for example to find the best soccer players according to their *technical performance and behaviours*. In some applications, one may require the definition of *constraints* on some dimensions and/or objects to express *hard restrictions*. For example, find the best soccer players *that hold a permanent french nationality* to play for the representative french team. This issue was investigated in several works *[2,3]*.

However, conventional skyline queries are not adapted to answer queries that require to analyze not only individual objects but also groups of objects. For example, find the best soccer teams of eleven players. Recent works *[4,5,6,7]* have considered the issue of *group skyline computation*, and enable the user to perform skyline queries on object groups in order to select the most relevant one. But, to the best of our knowledge, none of these works have investigated group skyline queries, when dealing with *constraints*. It is not clear how to answer the following queries using only conventional group skyline computation when we wish to form a team of eleven players:

- How to find the best soccer teams with *six defenders, four attackers and one goalkeeper* ?
- We know that the *best players don't always make the best team*. So, how to take into account the teammates' cooperation to build the best team ?

An interesting problem arises when users are allowed to define constraints over group skyline analysis. This problem is challenging when there are many objects in the dataset. Indeed, constraints reduce the input size, yet, paradoxically, makes computing the skyline quite challenging.

The aims of this work are twofold :

- Propose an efficient *Group-based Skyline* algorithm that incorporate constraints into the search space and cope with dynamic constraints
- Formally define a novel Constrained Group-based Skyline by extending the definition of the *dominance relation*

This master internship takes place in the context of the MetaTNT2 project (2017-2019), funded by Region Bretagne. The main objectif of this project is to simplify the use of an agrohydrological model TNT2 (Topography-based Nitrogen Transfer and Transformations), that simulates transfer and transformation of nitrogen in agricultural catchments and predicts water and nitrate fluxes at a daily time step. In this context, a simulation is defined as: data fluxes (water and nitrate concentrations) at a daily time step, during a simulation period of 30 years, with a specific land-use distribution, and in a given geographic space. In this project we have a consequent agri-environnemental database of simulations (several Gbyte in size). Our goal is to identify the simulation groups (i.e. *group skyline*) in which the agronomic criteria variability is sufficient to cover all possible scenarios.

The experimental evaluation of the proposed method will be conducted on two real datasets : (1) European football players' technical statistics, and (2) simulations results from our agri-environnemental database. For the second dataset, we have experts in the field (*i.e. domain knowledge*) to validate the skyline results.

References:

1- Chomicki, Jan, Paolo Ciaccia, and Niccolo Meneghetti. "Skyline queries, front and back." *ACM SIGMOD Record* 42.3 (2013): 6-18.
2- Mortensen, Michael L., et al. "Efficient caching for constrained skyline queries." *EDBT*. 2015.
3- Zhang, Ming, and Reda Alhajj. "Skyline queries with constraints: Integrating skyline and traditional query operators." *Data & Knowledge Engineering* 69.1 (2010): 153-168.
4- Im, Hyeonseung, and Sungwoo Park. "Group skyline computation." *Information Sciences* 188 (2012): 151-169.
5- Zhu, Haoyang, et al. "Top-k Skyline Groups Queries." *EDBT*. 2017.
6- Liu, Jinfei, et al. "Finding pareto optimal groups: Group-based skyline." *Proceedings of the VLDB Endowment* 8.13 (2015): 2086-2097.
7- Zhang, Nan, et al. "On skyline groups." *IEEE Transactions on Knowledge and Data Engineering* 26.4 (2014): 942-956.

Ideally, candidates will have a strong academic background in mathematics, passion for data science, data mining and machine learning, and strong programming skills.

Applications should include results of Master, CV and a cover letter

Duration: ~6 months

Contact : tassadit.bouadi@irisa.fr, and veronique.masson@irisa.fr