A model for processing skyline queries over a database with missing data

ABSTRACT

Skyline queries provide a flexible query operator that returns data items (skylines) which are not being dominated by other data items in all dimensions (attributes) of the database. Most of the existing skyline techniques determine the skylines by assuming that the values of dimensions for every data item are available (complete). However, this assumption is not always true particularly for multidimensional database as some values may be missing. The incompleteness of data leads to the loss of the transitivity property of skyline technique and results into failure in test dominance as some data items are incomparable to each other. Furthermore, incompleteness of data influences negatively on the process of finding skylines, leading to high overhead, due to exhaustive pairwise comparisons between the data items. This paper proposed a model to process skyline queries for incomplete data with the aim of avoiding the issue of cyclic dominance in deriving skylines. The proposed model for identifying skylines for incomplete data consists of four components, namely: Data Clustering Builder, Group Constructor and Local Skylines Identifier, k-dom Skyline Generator, and Incomplete Skylines Identifier. Including these processes in the proposed model has optimized the process of identifying skylines in incomplete database by reducing the necessary number of pairwise comparison through eliminating the dominated data items as early as possible before applying the skyline technique.