Chapter 4
Evaluating Top-k Skyline Queries Efficiently

Marlene Goncalves
*Universidad Simón Bolívar, Venezuela*

María Esther Vidal
*Universidad Simón Bolívar, Venezuela*

ABSTRACT

Criteria that induce a Skyline naturally represent user’s preference conditions useful to discard irrelevant data in large datasets. However, in the presence of high-dimensional Skyline spaces, the size of the Skyline can still be very large. To identify the best k points among the Skyline, the Top-k Skyline approach has been proposed. This chapter describes existing solutions and proposes to use the TKSI algorithm for the Top-k Skyline problem. TKSI reduces the search space by computing only a subset of the Skyline that is required to produce the top-k objects. In addition, the Skyline Frequency Metric is implemented to discriminate among the Skyline objects those that best meet the multidimensional criteria. This chapter’s authors have empirically studied the quality of TKSI, and their experimental results show the TKSI may be able to speed up the computation of the Top-k Skyline in at least 50% percent with regard to the state-of-the-art solutions.

INTRODUCTION

Emerging technologies such as Semantic Web, Grid, Semantic Search, Linked Data and Cloud and Peer-to-Peer computing have become available very large datasets. For example, by the time this paper has been written at least 21.59 billion pages are indexed by the Web (De Kunder, 2010) and the Cloud of Linked Data has at least 13,112,409,691 triples (W3C, 2010). The enormous growth in the size of data has a direct impact on the performance of tasks that are required to process on very large datasets and

DOI: 10.4018/978-1-60960-475-2.ch004
whose complexity depends on the size of the database. Particularly, the task of evaluating queries based on user preferences may be considerably affected by this situation.

Skyline (Börzsönyi et al., 2001) approaches have been successfully used to naturally express user preference conditions useful to characterize relevant data in large datasets. Even though, Skyline may be a good choice for huge data sets its cardinality may become very large as the number of criteria or dimensions increases. The estimated cardinality of the Skyline is $O(\ln^{d-1} n)$ when the dimensions are independent where $n$ is the size of the input data and $d$ the number of dimensions (Bentley et al., 1978).

Consider Table 1 that shows estimates of the skyline cardinality when the number of dimensions ranges from 2 to 10 in a database comprised of 1,000,000 tuples. We may observe in Table 1 that Skyline cardinality rapidly increases making unfeasible for users to process the whole skyline set. In consequence, users may have to discard useless data manually and consider just a small subset or a subset of the Skyline that best meet the multidimensional criteria. To identify these points, the Top-k Skyline has been proposed (Goncalves and Vidal, 2009; Chan et al., 2006b; Lin et al., 2007). Top-k Skyline uses a score function to induce a total order of the Skyline points, and recognizes the top-k objects based on these criteria.

Several algorithms have been defined to compute the Top-k Skyline, but they may be very costly (Goncalves and Vidal, 2009; Chan et al., 2006b; Lin et al., 2007; Vlachou and Vazirgiannis, 2007). First, they require the computation of the whole Skyline; second, they execute probes of the multidimensional function over the whole Skyline points. Thus, if $k$ is much smaller than the cardinality of the Skyline, these solutions may be very inefficient because a large number of non-necessary probes may be evaluated, i.e., at least Skyline size minus $k$ performed probes will be non-necessaries.

Top-k Skyline has become necessary in many real-world situations (Vlachou and Vazirgiannis, 2007), and a wide range of ranking metrics to measure the interestingness of each Skyline tuple has been proposed. Examples of these ranking metrics are skyline frequency (Chan et al., 2006b), k-dominant skyline (Chan et al., 2006a), and k representative skyline (Lin et al., 2007). Skyline frequency ranks Skyline in terms of the number of times in which a Skyline tuple belongs to a non-empty subset or subspace of the multi-dimensional criteria. k-dominant skyline metric identifies Skyline points in $k \leq d$ dimensions of multi-dimensional criteria. Finally, k representative skyline metric produces the $k$ Skyline points that have the maximal number of dominated object.
14 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the product's webpage:
www.igi-global.com/chapter/evaluating-top-skyline-queries-efficiently/52298?camid=4v1

This title is available in InfoSci-Database Technologies, InfoSci-Books, Business-Technology-Solution, Library Science, Information Studies, and Education, InfoSci-Library and Information Science, Advances in Data Mining and Database Management, Advances in Data Mining and Database Management, Advances in Data Mining and Database Management, Data Science, InfoSci-Computer Science and Information Technology, Science, Engineering, and Information Technology. Recommend this product to your librarian:
www.igi-global.com/e-resources/library-recommendation/?id=12

Related Content

A Methodology of Constructing Canonical Form Database Schemas in a Multiple Heterogeneous Database Environment
Jeongseok Lim and Dong-Guk Shin (1998). Journal of Database Management (pp. 4-11).
www.igi-global.com/article/methodology-constructing-canonical-form-database/51205?camid=4v1a

Object-Relational Modeling in the UML
www.igi-global.com/chapter/object-relational-modeling-uml/11183?camid=4v1a

MAMADAS: A Mobile Agent-Based Secure Mobile Data Access System Framework
www.igi-global.com/chapter/mamadas-mobile-agent-based-secure/4399?camid=4v1a

An Experimental Study of Object-Oriented Query Language and Relational Query Language for Novice Users
www.igi-global.com/article/experimental-study-object-oriented-query/51139?camid=4v1a