Efficiently Producing the K Nearest Neighbors in the Skyline on Vertically Partitioned Tables

Marlene Goncalves, Departamento de Computación y Tecnología de Información, Universidad Simón Bolívar, Caracas, Venezuela

Maria-Esther Vidal, Departamento de Computación y Tecnología de Información, Universidad Simón Bolívar, Caracas, 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, making unfeasible for users to process this set of points. To identify the best points among the Skyline, the Top-k Skyline approach has been proposed. Top-k Skyline uses discriminatory criteria to induce a total order of the points that comprise the Skyline, and recognizes the best or top-k points based on these criteria. In this article the authors model queries as multi-dimensional points that represent bounds of VPT (Vertically Partitioned Table) property values, and datasets as sets of multi-dimensional points; the problem is to locate the k best tuples in the dataset whose distance to the query is minimized. A tuple is among the k best tuples whenever there is not another tuple that is better in all dimensions, and that is closer to the query point, i.e., the k best tuples correspond to the k nearest points to the query that are incomparable or belong to the skyline. The authors name these tuples the k nearest neighbors in the skyline. The authors propose a hybrid approach that combines Skyline and Top-k solutions and develop two algorithms: TKSI and k-NNSkyline. The proposed algorithms identify among the skyline tuples, the k ones with the lowest values of the distance metric, i.e., the k nearest neighbors to the multi-dimensional query that are incomparable. Empirically, we study the performance and quality of TKSI and k-NNSkyline. The authors’ experimental results show the TKSI is able to speed up the computation of the Top-k Skyline in at least 50% percent with respect to the state-of-the-art solutions, whenever k is smaller than the size of the Skyline. Additionally, the authors’ results suggest that k-NNSkyline outperforms existing solutions by up to three orders of magnitude.

Keywords: K Nearest Neighbors, Multi-Dimensional Points, Skyline XE “Skyline”, Top-k, Vertically Partitioned Table (VPT)

DOI: 10.4018/ijirr.2013040104
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 3.8 billion pages are indexed by the Web (Kunder, 2013) and the Linking Open Data cloud which by September 2011 it has at least 31 billion RDF triples, interlinked by around 504 million RDF links\(^1\).

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 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 & 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 points based on these criteria.

Several algorithms have been defined to compute the Top-k Skyline (Goncalves & Vidal, 2009; Chan et al., 2006b; Lin et al., 2007; Vlachou & Vazirgiannis, 2007), but they may be very costly. First, they require the computation of the whole Skyline; second, they execute probes of the multi-dimensional 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.

Table 1. Estimated skyline cardinality

<table>
<thead>
<tr>
<th>#Dimensions</th>
<th>Cardinality</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>191</td>
</tr>
<tr>
<td>3</td>
<td>2,637</td>
</tr>
<tr>
<td>4</td>
<td>36,431</td>
</tr>
<tr>
<td>5</td>
<td>503,309</td>
</tr>
<tr>
<td>6</td>
<td>6,953,471</td>
</tr>
<tr>
<td>7</td>
<td>96,065,749</td>
</tr>
<tr>
<td>8</td>
<td>1,327,197,371</td>
</tr>
<tr>
<td>9</td>
<td>18,335,909,288</td>
</tr>
<tr>
<td>10</td>
<td>253,319,948,365</td>
</tr>
</tbody>
</table>

\(^1\) RDF links: Resource Description Framework.
Efficient Materialized View Selection for Multi-Dimensional Data Cube Models
www.igi-global.com/article/efficient-materialized-view-selection-for-multi-dimensional-data-cube-models/161661?camid=4v1a