Processing Distance-Based Queries in Multidimensional Data Spaces Using R-trees

Antonio Corral[1⋆], Joaquin Cañadas[1], and Michael Vassilakopoulos[2]

[1] Department of Languages and Computation, University of Almeria
04120 Almeria, Spain
acorral,jjcanada@ual.es
[2] Department of Information Technology
Technological Educational Institute of Thessaloniki
P.O. BOX 14561, 541 01, Greece
vasilako@it.teithe.gr

Abstract. In modern database applications the similarity, or dissimilarity of data objects is examined by performing distance-based queries (DBQs) on multidimensional data. The R-tree and its variations are commonly cited multidimensional access methods. In this paper, we investigate the performance of the most representative distance-based queries in multidimensional data spaces, where the point datasets are indexed by tree-like structures belonging to the R-tree family. In order to perform the K-nearest neighbor query (K-NNQ) and the K-closest pair query (K-CPQ), non-incremental recursive branch-and-bound algorithms are employed. The K-CPQ is shown to be a very expensive query for datasets of high cardinalities that becomes even more costly as the dimensionality increases. We also give ϵ-approximate versions of DBQ algorithms that can be performed faster than the exact ones, at the expense of introducing a distance relative error of the result. Experimentation with synthetic multidimensional point datasets, following Uniform and Gaussian distributions, reveals that the best index structure for K-NNQ is the X-tree. However, for K-CPQ, the R*-tree outperforms the X-tree in respect to the response time and the number of disk accesses, when an LRU buffer is used. Moreover, the application of the ϵ-approximate technique on the recursive K-CPQ algorithm leads to acceptable approximations of the result quickly, although the tradeoff between cost and accuracy cannot be easily controlled by the users.

1 Introduction

Large sets of multidimensional data are used in modern applications such as multimedia databases [12], medical images databases [21], CAD [18], metric databases [2], etc. In such applications, complex objects are stored. To support distance-based queries (DBQ), multidimensional feature vectors are extracted

⋆ The author has been partially supported by the Spanish CICYT (project TIC 2002-03968).

Y. Manolopoulos et al. (Eds.): PCI 2001, LNCS 2563, pp. 1–18, 2003.

from the objects and organized in multidimensional indices. The most important property of this feature transformation is that these feature vectors correspond to points in the Euclidean multidimensional space. We need to calculate the distance between them for obtaining the similarity or dissimilarity of the original objects in the underlying application.

The distance between two points is measured using some metric function over the multidimensional data space. We can use the Euclidean distance for expressing the concepts of "neighborhood" and "closeness". The concept of "neighborhood" is related to the discovery of all the multidimensional points that are "near" to a given query point. The δ-distance range query and the K-nearest neighbors query are included in this category. The concept of "closeness" is related to the discovery of all pairs of multidimensional points that are "close" to each other. The δ-distance join query and the K-closest pairs query are included in this category.

Usually, distance-based queries are executed using some kind of multidimensional index structure [13] such as the R-trees, since the result can be found in logarithmic time, applying pruning techniques. The multidimensional access methods belonging to the R-tree family (the R-tree [14], the R*-tree [4] and particularly the X-tree [5]) are considered a good choice for indexing multidimensional point datasets in order to perform nearest neighbor queries. The branch-and-bound algorithms for DBQs employ distance metrics and pruning heuristics based on the MBR characteristics in the multidimensional Euclidean space, in order to reduce the searching space.

The main objective of this paper is to study the performance of the K-nearest neighbor and the K-closest pairs queries in multidimensional data spaces, where both point datasets are indexed by tree-like structures belonging to the R-tree family. To the authors knowledge, this paper and its successor [8] are the first research efforts in the literature that study performance of distance join queries for more than one tree structures and dimensionality of data larger than 2. We compare the results of exact recursive algorithms, applied in different kinds trees, in terms of the I/O activity and the response time.[1] We also test ϵ-approximate versions of the algorithms using the average relative distance error (ARDE) with respect to the exact solutions as a metric of the accuracy of the result. Experimental results are presented for several dimensionalities, numbers of elements in the result and fixed cardinalities of multidimensional point datasets following Uniform and Gaussian distributions in each dimension. Based on these experimental results, we draw conclusions about the behavior of these multidimensional access methods over these kinds of queries.

The paper is organized as follows. In Section 2, we review the literature (distance-based queries on multidimensional data spaces) and motivate the research reported here. In Section 3, a brief description of the R-tree family, and the definitions of the most representative distance-based queries are presented. In Section 4, recursive branch-and-bound algorithms based on metrics and pruning heuristics over MBRs for answering K-NNQs and K-CPQs are examined.

[1] A preliminary version of such a comparison appears in [10].