Data Transformation for Improved Query Performance

Alok Watve

April 27, 2012

Abstract

A database management system stores data in order to facilitate fast and efficient retrieval at a later point of time. The typical queries run on a database can be classified into three broad classes. They are range queries, k-nearest neighbor (k-NN) queries and box queries. Implementation of a box query typically involves simple comparisons among the query objects and index objects, however, implementing range queries and k-NN queries may be slightly involved.

In this dissertation, we study mapping of one type of query on to other. From performance perspective, an index structure may favor one type of query over other. Hence, such a mapping provides a way of improving query performance. It also highlights relationships among the various query classes.

Our first transformation maps a range query in $L_1$ space on to a box query. This mapping provides a similar interface between the query space and data space of each of the index pages for Bounding Box based indexes. In 2-dimensional space, the mapping is exact with no false positives or false negatives. However, it cannot be used directly for higher dimensional spaces. We propose a novel approach called disjoint planar rotation in order to alleviate the problems in higher dimensions. We also develop a new type box query (called pruning box query) which is equivalent to the range query in original space. Our theoretical analysis shows that this mapping can improve I/O performance of the queries. Further, performance improvement increases with increasing number of dimensions. Experiments with some of the well known indexing schemes verify these findings. Due to underlying similarity in implementation, k-nearest neighbor queries can also be optimized using a similar transformation. We successfully apply this transformation to improve I/O performance of k-nearest neighbor queries as well.

We next use a similar transformation to map box queries on to range queries $L_1$ space. But the inherent property of box queries to allow varying degree of selectivity along each dimension, poses some challenges for the transformation. We propose square tiling approach to map each box query on to a number of square box queries. Each of the square box queries can then be transformed in to a range query. We demonstrate practical
application of this mapping using M-Tree which is a well known index scheme for metric spaces.

Euclidean distance (or $L_2$ norm) is another popular distance measure. We discuss the challenges in mapping of range queries in Euclidean space to box queries. Although, it is challenging to improve performance of range queries in disk based databases, it is possible to reduce the CPU time required for query execution in main memory databases. This is because comparing data objects with a box query is generally computationally easier than computing $L_2$ distance of objects from the range query. Existing work on vantage point based indexing uses mapping from range queries to bounding box queries to reduce the number of costly distance computations. However, these schemes use data points as vantage points. As a result they are static and cannot allow dynamic insertion, deletion and updating of data. Further, their computational complexity depends on the size of the database which can be a problem for large databases. We analyze impact of vantage points on false positives and the number of duplicates. Based on our analysis, we present a heuristic algorithm for selecting vantage. The proposed algorithm is polynomial in number of dimensions and the number of vantage points and being independent of data allows dynamic insertions and deletions. Comprehensive experimental evaluation with several synthetic and real databases shows effectiveness of our vantage point selection scheme.