Reduction of Distance Computations in Selection of Pivot Elements for Balanced GHT Structure

László Kovács

University of Miskolc, Department of Information Technology, Miskolc, Hungary
kovacs@iit.uni-miskolc.hu

Abstract. In general metric spaces, one of the most widely used indexing techniques is the partitioning of the objects using pivot elements. The efficiency of partitioning depends on the selection of the appropriate set of pivot elements. In the paper, some methods are presented to improve the quality of the partitioning in GHT structure from the viewpoint of balancing factor. The main goal of the investigation is to determine the conditions when costs of distance computations can be reduced. We show with different tests that the proposed methods work better than the usual random and incremental pivot search methods.

Keywords: pivot based indexing, general metric space, interval based computation.

1 Introduction

In many application areas the objects can't be represented with appropriate feature vectors, only the distances between the objects are known. If the distance function $d()$ is a metric it fulfills the following conditions:

\[
\begin{align*}
  d(x,y) &\geq 0 \\
  d(x,y) = 0 &\iff x = y \\
  d(x,y) &\equiv d(y,x) \\
  d(x,z) + d(z,y) &\geq d(x,y)
\end{align*}
\]

The different application areas may have very different and complex distance functions. For example the detection and comparison of components for sound data objects, are relatively expensive operations. In the case of applications with huge collection of these objects, the objects are clustered and indexed to reduce the computational costs on the collection. The most widely used indexing methods in general metric spaces use pivot elements. The pivot element $p$ is a distinguished object from the object-set. The distance from an object $x$ to $p$ is used as the indexing key value of $x$ to locate the bucket containing $x$. Usually more than one single pivot element are used in the algorithms.

It is known that the efficiency of indexing methods depends significantly on the position of the pivot elements [4], thus the appropriate selection of the pivot elements is a crucial optimization component of object management. The usual measure to calculate the fitness of a pivot-set is the mean of distance distribution [1]:

© Springer-Verlag Berlin Heidelberg 2012
where $N$ denotes the number of objects in the set. This measure is tailored to reduction of the pruning operation in the search tree. Having a nearest neighbor query $Q(q,r)$ where $q$ is the query object and $r$ is the threshold distance, a branch of the search tree assigned to pivot element set $\{p_i\}$ can be excluded if for all elements $x$ of the subtree

$$|d(x, p_i) - d(q, p_i)| > r$$

is met for some index $i$. As the exact calculation of the measure $\mu$ is an $O(N)$ operation, only a sampling with $O(1)$ is used to estimate the fitness parameter value.

There are many variants of indexing trees in general metric spaces. The Generalized Hyperplane Tree (GHT) [5] and Bisector Tree [6] are widely used alternatives. These structures are binary trees where each node of the tree is assigned to a pair of pivot elements $\{p_1, p_2\}$. If the distance of the object to $p_1$ is smaller than the distance to $p_2$, then the object is assigned to the left subtree, otherwise it is sent to the right subtree. According to authors, the GHT provides a better indexing structure than the usual vantage point trees [5].

Based on the survey of [1], the following methods are usually used for pivot selection. The most simple solution is the random selection of the pivot elements. In this approach, more tests are run and the pivot set with best parameter is selected. The second method is the incremental selection method. In the first step of this algorithm, a $p_1$ with optimal fitness is selected. In the next step, the pivot set is extended with $p_2$, yielded by a new optimization process where $p_1$ is fixed already. On this way, the pivot set is extended incrementally to the required size. The third way is the local optimization method. In this case, an initial pivot set is generated on some arbitrary way. In the next step, the pivot element with worst contribution is removed from the set and a new pivot element is selected into the set.

The work [1] analyzed the pivot selection methods from the viewpoint of subtree pruning operation. Usually a heuristic approach is used in the applications. The core elements of the heuristics are the following rules: the pivot elements should be far from the other not pivot elements and they should be far from each others too. The paper concluded that the incremental selection method provides the optimal solution of this heuristics.

An improved pivot selection method called Sparse Spatial Selection (SSS) is presented in [2]. The SSS method generates the pivot elements dynamically when a new outlier elements is inserted into the object pool. A new incoming element is selected as a new pivot if it is far enough from the other pivot elements. A loss minimization method was proposed by [8] where the loss is measured as the real distance between the object and its nearest neighbor in the index tree.

A conceptually different approach for object indexing is the family of computation methods based only on the distance matrix. In the AESA [4] algorithm, the distances between every pairs of objects are known and thus every objects can be considered as