A Eigenvalue-based Pivot Selection Strategy for Improving Search Efficiency in Metric Spaces

***Abstract*—Utilizing pivot spaces is a popular method to perform similarity search queries in metric spaces when it is difficult to search the objects with their plain representation. In the pivot space, an object is transformed into a vector whose coordinates are its distances to pre-defined pivots. Based on this transform, any distance-based and multi-dimensional data structures can be used to perform various types of search queries. Although it has been observed that the search performance in terms of query throughput highly depends on which pivots are chosen, it still has been unclear how to choose good pivots despite of a number of work presented over decades. In this paper, we present a pivot selection strategy based on their correlation. By computing eigenvalues, independent pivots are chosen to improve the efficiency of the searching process in the pivot space. We also propose randomized sampling method to reduce the time complexity involved by exhaustive computation of correlation between pivot candidates. Experimental results show that selecting uncorrelated pivots improve the performance, and outperforms other previous pivot selection approaches.**

***Index Terms*—Similarity Search, Distance-based Index, Metric Space, Pivot Selection**

I. Motivation

Finding similar objects is an essential task for data processing. Range search is widely used in multimedia retrieval systems [1], spam [2] and swear filters [3]. Nearest neighbor search and similarity join are also utilized for spell correctors [4], data deduplication [5], clustering and classification [6]. A number of data structures to support such search queries have been developed, and they are usually designed for vector spaces [7].

Pivot space transform is one of the popular methods for indexing and searching in metric spaces [8]. In the pivot space, each object is represented as a vector whose coordinates are comprised of distances of the object to the pre-selected pivots. It transforms any object in any domain into the vector space once we have a well-defined metric distance function on that domain, so that we can index objects regardless of the complexity and the dimensionality of their own nature. As a result, we can use partition-based data structures or multi-dimensional data structures to execute various search queries on the data.

It has been observed that the search performance is significantly affected on how to select pivots [8], [9], [10]. Many methods based on the pivot space, however, use just randomly picked objects or first k objects visited by the farthest-first traversal as pivots [9]. There are two reasons. First, it is still unclear how to select the *best* pivots, but only partial observations have been made over decades. The other reason is that, even though we have any statistical clues to select good pivots, their time and space complexity is excessively high so the computation becomes intractable as the data amount increases. Several attempts to tackle this problem have been made, but it is still remaining challenging.

In order to address these problems, this paper presents a pivot selection strategy based on eigenvalues. The main contributions are summarized as follows:

- We propose an incremental pivot selection method based on the eigenvalue of the covariance matrix.
- We show that selecting uncorrelated pivots can improved the performance of the distance-based indexing methods.
- We also present experimental results to demonstrate that our approach outperforms the other pivot selection methods.

The organization of this paper is as follows. Section II gives some background knowledge for the pivot-based similarity search. Selection III introduces the existing approaches on pivot-based searching and pivot selection. The proposed method is presented in Section IV. After showing experimental results in Section V, we conclude the paper in Section VI.

II. Preliminaries

A. Similarity Search

Given an object set $X \subset \mathbb{U}$ and a distance function d, a range search query $(q, r) \in \mathbb{U} \times \mathbb{R}$ is to find all the object x in X such that $d(q, x) \leq r$. This paper is devoted to improve the performance on processing range search queries of pivot-based indexing methods for metric spaces by presenting a new approach to select good pivots.

B. Metric Space

A distance function d on a set X is *a metric* if the following properties hold:

1) $\forall x \in X,\ d(x, y) = 0 \Leftrightarrow x = y$.
2) $\forall x, y \in X,\ d(x, y) \geq 0$.
3) $\forall x, y \in X,\ d(x, y) = d(y, x)$.
4) $\forall x, y, z \in X,\ d(x, y) \leq d(x, z) + d(z, y)$.

Among the properties, the fourth one, which is so-called *triangle inequality*, is the key property for utilizing various similarity search queries in the metric space. Using the triangle