# Pivot Selection Methods Based on Covariance and Correlation for Metric-space Indexing

Kewei Ma, Yuanjun Liu, Honglong Xu, Pang Yue, Fuli Lei, Sheng Liu, Rui Mao*
Guangdong Key Laboratory of Popular High Performance Computers
Shenzhen Key Laboratory of Service Computing and Applications
Shenzhen University
Shenzhen, China, (+86)755 2653 4207-81
marknv1991@gmail.com

Jiaxin Han
Oracle Corporation
Redwood City, CA, USA, 1 650 918 8107
jiaxinjasonhan@gmail.com

***Abstract*—Metric-space indexing is a general method for similarity queries of complex data. The quality of the index tree is a critical factor of the query performance. Bulkloading a metric-space indexing tree can be represented by two recursive steps, pivot selection and data partition, while pivot selection dominants the quality of the index tree. Two heuristics, based on covariance and correlation, for pivot selection are proposed. Empirical results show that their performance is superior or comparable to existing methods.**

***Keywords-similarity query; metric-space indexing; pivot space model; pivot selection;***

## I. INTRODUCTION

With the advent of information age, the explosive growth of data causes the current research emphasis of internet and cloud computing has been shifted from computing to mass data processing. Indexing technique, which aims to improving search speed, is most fundamental and crucial due to search is the essential part of mass data processing. Content-based similarity search is an important information retrieval type that is widely used in databases and data mining applications. Accompany with the development of multimedia and bioinformatics technologies, complex data types spring up. Similarity search has become the primary need of content-based search in multimedia information systems (MIS), and its performance is the key indicator to evaluate the query function of MIS [1]. According to statistics, meantime, similarity search takes part in 35% of the whole research tasks in bioinformatics [2]. Currently, researches and applications of indexing technique are mainly focused on content-based similarity search methods.

The traditional way of similarity search is implemented by multi-dimensional indexing technique. Its basic idea is to extract the feature vectors from data objects and map them into a vector space, then use the coordinates to compute their Euclidean distances. However, there are two limitations while using vector space indexing: 1) Data object must be presented as feature vectors. 2) The similarity of each data objects must be measured only by Euclidean distance. Today, non-traditional data types is increasingly complex, more of them could not satisfy these two prerequisites. For instance, especially biological data, it is hard to consider them as spatial points and their similarity cannot be modeled efficiently by the Euclidean norm.

Meanwhile, many specified manage systems have been set up or are under construction, such as BLAST and Midomi [3] music search. Building independent specified systems which cost both money and manpower is not an effective method. As a result, a new choice called general purpose search system is badly needed.

Metric-space indexing also known as distance-based indexing is a general solution to the problem of searching based on similarity of complex data types. Metric-space indexing allow users to provide their own distance functions, and only distance information is maintained in metric space, therefore, same algorithm can be apply to different data types, which expands the scope of applications.

For a long time, the research emphasis of metric-space indexing lies in data partition, while the significance of pivot selection is not been fully recognized. In recent years, the proposal of Pivot space model [4] clarifies the importance of pivot selection again. This article aims to explore the problem of pivot selection based on Pivot space model, and proposes two brand new pivot selection methods using the properties of covariance and correlation in statistics.

## II. METRIC-SPACE INDEXING

A metric space [5-7] is a pair $(S, d)$, where $S$ is a nonempty set and $d: S \times S \rightarrow R$ is a real-valued function, called a metric on $S$, with the following properties:

1) For all x, $y \in S$, $d(x, y) >= 0$ and $d(x, y) = 0$ iff $x = y$. (Positivity)
2) For all x, $y \in S$, $d(x, y) = d(y, x)$. (Symmetry)
3) For all x, y, z $\in S$, $d(x, y) + d(y, z) >= d(x, z)$. (Triangle Inequality)

*Corresponding author

Only distance record is stored in general metric space, domain-specific and distance function information is completely segregated, then index tree could be built using Triangle Inequality. Metric-space indexing doesn't need coordinates of data objects; thus, it can solve many problems that multi-dimensional methods are not able to handle. The most common metric-space indexing methods are as following: CRT [8, 9], linear partition family has GHT [10-12], GNAT [13], and ball partition family has VPT [10, 14], MVPT [15].

The most advantage using metric-space indexing is its high general applicability, which means users only need to provide a distance function before proceeding similarity search. Like a double-edged sword, the advantage of metric-space indexing also leads to its disadvantage. Data objects have been abstracted to points without coordinates in metric space, so mathematical tools could not directly be used because the only information available is the values of distances. Therefore, a connection is needed between metric space and vector space in which mathematical tools are maturely used.

## III. PIVOT SPACE MODEL

Pivot space model maps data objects from metric space into vector space where data has coordinates without loss any distance information [4]. The existence of coordinates provides a platform on which mathematical tools could directly be performed. It builds a bridge between metric space and multi-dimensional mathematical methods.

### A. General Theory and Steps

Let $R^n$ denote a general real coordinate space of dimension n. $(M, d)$ is a metric space, $d$ is the distance oracle or distance function of $M$. Database $S$ is the finite subset of $M$, $S = \{x_i \mid x_i \in M, i = 1, 2, \ldots, n, n > 1\}$ . $P$ is the set of pivots, $P = \{p_j \mid j = 1, 2, \ldots, k\}$ , P⊆S. Fig. 1 shows the three steps which are needed when encountering a range query. [4]

**Step1.**

1) Map data into $R^k$.
2) Map query object into $R^k$.
3) Determine a region in $R^k$ that completely covers the range query ball.

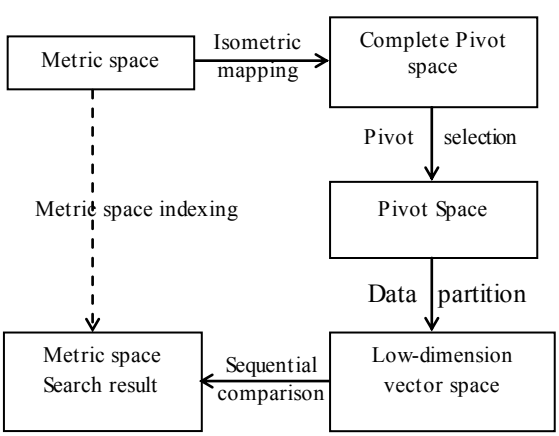


Figure 1. Pivot space model indexing process

**Step2.** Exploit multi-dimensional techniques to retrieve all the points in the region determined in Step1. The main task in this step is data partition, thanks to the result from Step1; many multi-dimensional methods can be used, such as $k$-d tree [16] and MVPT [15].

**Step3.** For each points retrieved in Step2, compute its distance to the query object to remove false positives.

These three steps show that in pivot space model, data objects are firstly mapped isometrically into complete pivot space, and then high-dimensional methods such as dimension reduction can be applied on it. After this step, the processed data is finally able to be partitioned by many data partition methods in low-dimension space in order to find the unfiltered query result. In the end, remove false positives and we can get the final query result.

### B. Pivot Space

Mapping data objects into $R^k$ by pivot selection also means picking specific points in database as pivots, and other points can be represented as the distances to these pivots. Pivot space $F_{P,d}(S)$ is a new metric space$(R^k, L^\infty)$, moreover, each of the k coordinate axis is corresponding to one of k the pivots. Therefore, every point in the original metric or data space can be represented as the distances to all the pivots in a pivot space:

$$F_{P,d}(S) = \{x^p \mid x^p = F_{P,d}(x) = (d(x, p_1), \ldots, d(x, p_k)), x \in S\} .$$

It is necessary to note that the distance in a pivot space is measured by $L^\infty$ rather than $L^2$ norm.

Given a range query $R(q, r)$ in general metric space, it's hard to determine the shape of the image of the query ball in a pivot space. Whereas, it can be proved that the image of the query ball is completely covered by a hypercube of edge length 2r in the pivot space using the triangle inequality, as Fig. 2 [4].

Complete pivot space refers to a specific pivot space that all points in the data space are selected as pivots. It is a known fact that any finite metric space (size n) is isometric to a metric space formed by a subset of $R^n$ with the $L^\infty$ distance [4]. As a result, if we consider distance only, data have been mapped from metric space that has no coordinates to complete pivot space which has coordinates. Thus, problems in metric-space indexing can be solved in complete pivot space (High-dimension vector space). It's also proved that evaluation of

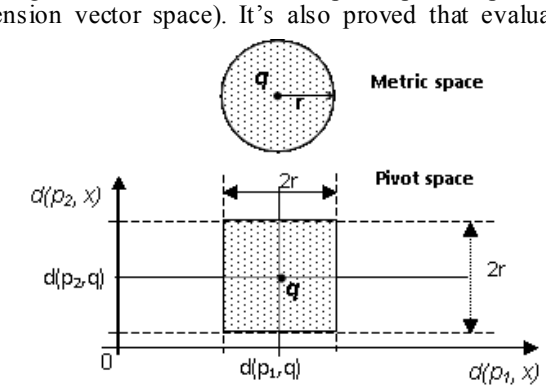


Figure 2. The query ball of a range query in a general metric space is covered by a square in the pivot space

similarity queries in the complete pivot space degrades the query performance to linear scan [4]. Therefore, dimension reduction or pivot selection is inevitable.

## IV. PIVOT SELECTION

The quality of index structure is a determinative factor of search performance. For a long time, the research emphasis of metric-space indexing lies in data partition, while the significance of pivot selection is not been completely recognized.

### A. Importance of Pivot Selection

Take Fig. 3 as example, three points A, B, C located on the axis. If B in the middle is selected as pivot, then the distances to A and C are both 1, and it's impossible to separate A, C after mapped into a pivot space. If we pick A or C as pivot, however, all three points are distinguishable in a pivot space. It follows that different pivots have a significant effect on query performance.

### B. Common Pivot Selection Methods

The M-tree[8] algorithm randomly selects points as pivots, while SA-tree[17] begin with a random picked point and selects the centers of neighbouring cells of a Voronoi diagram[18] as pivots.

Yianilos proves that in a unit square in which points are uniformly distributed, the best pivots are corners, so he applies this method in his VPT [19]. In addition, MVPT selects multiple corners as pivots, and the algorithm it uses is called Farthest-first-traversal(FFT) [20].

FFT is a k-center clustering algorithm often used to choose pivot. FFT minimize the maximum cluster diameter and it's proved that the result is at most twice the optimal result. It's a very fast algorithm to find corner points, the algorithm shows as Fig. 4.

The first cluster center is randomly picked. Each time selecting the next cluster center, find points that have minimum distance to selected points (pivots) first as the lower limit of distance, and then choose the point which has the maximum lower limit as pivot.

In pivot space model, the popular dimension reduction method PCA [21, 22] is applied for pivot selection. It selects the existing coordinate with the minimum angle with the new dimensions created by PCA as pivots. Experimental results demonstrate PCA is one of the best pivot selection methods at present [4].

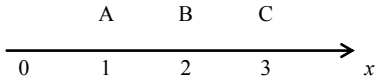


Figure 3. Importance of pivot selection: 1 dimension data case

## V. PIVOT SELECTION METHODS BASED ON COVARIANCE AND CORRELATION

Pivot space model shows that performing search in complete pivot space is equivalent to a linear scan, so selecting pivots, for example dimension reduction is inevitable. Pivot space is presented as a distance matrix, so we consider rows as points and column as dimension.

### A. Pivot Selection Method Based on Covariance

In statistics covariance is a measure of how much two random variables change together. Assume X, Y are two random variables, the covariance between X and Y is $cov(X,Y)=E(X \bullet Y)-\mu\sigma$, $E(X)=\mu, E(Y)=\sigma$. If the variables show similar variation trend, the covariance is positive. Let $S=\{x_i \mid i=1,2,\ldots,n, n>1\}$ be a set of all points in a complete pivot space, $P=\{p_j \mid j=1,2,\ldots,k\}$ is the set of pivots, P⊆S. According to the properties of covariance we can give the pivot selection method, as shown in Fig. 5.

**Step1.** Select the row with maximum covariance as the first pivot. Large covariance indicates a high data fluctuation, and points are easier to be distinguished from each other.

**Step2.** If we still pick pivots the same way when selecting the second pivot, it may contain much redundant information compared to the first pivot. Take Fig. 3 as example; if A is the first pivot and C is the second one, the distances to other points are all the same, so picking two points like this is equivalent to picking only one point. As a result, when selecting the next pivot, for each rest points compute its covariance between pivots, and take maximum one as the upper limit of covariance.

**Step3.** Select the point with minimum upper limit value of covariance as the next pivot.

The essential methodology of this algorithm is making the covariance as the distance oracle through computing the covariance matrix of the original distance matrix, then using the theory of FFT to select pivots.

To find $k$ cluster centers:

(1) randomly choose one point $z \in S$ as the first center, set $D=\{z\}$

(2) for i = 2 up to k {

$z = \text{argmax}_{x \in S}(min(d(x,D)))$

$D = D \cup \{z\}$ }

Figure 4. FFT algorithm

(1) $p_1 = \text{argmax}_p(variance(S))$

(2) for each unselected point $x_i$ {

$maxcov(x) = max(cov(x_i, p_1), cov(x_i, p_2), \ldots, cov(x_i, p_{k-1}))$

(3) $p_k = \text{argmin}_{x \in S-P}(maxcov(x))$ }

Figure 5. Pivot selection method based on covariance

(1) $p_1 = \text{argmax}_p(\text{variance(S)})$

(2) for each selected point $p_j$ {

$$mincor(x) = min(abs(cor(x_1, p_j)), \ldots, abs(cor(x_n, p_j)))$$

(3) $p_k = \text{argmax}_{x \in S-P}(mincor(x))$ }

Figure 6. Pivot selection method based on correlation

### B. Pivot Selection Method Based on Correlation

The correlation indicates the strength of a linear relationship between two variables without considering their variance. Though correlation has some connection with covariance, in this article we propose a new and different algorithm for correlation. Assume X, Y are two random variables, the correlation between X and Y is $\rho_{X,Y} = Cov(X,Y)/(\sigma_X \sigma_Y)$ . Let $S = \{x_i \mid i = 1,2,\ldots,n, n > 1\}$ be a set of all points in a complete pivot space, $P = \{p_j \mid j = 1,2,\ldots,k\}$ is the set of pivots, P⊆S. According to the properties of correlation we can give the other pivot selection method, shown as in Fig. 6.

**Step1.** Select the row with maximum covariance as the first pivot. Same reason to the covariance method.

**Step2.** When picking the next pivot, for each pivot find points have minimum correlation (absolute value) with them. Small absolute value of correlation indicates two points are highly distinguishable, so it's helpful to optimize index tree.

**Step3.** Select the point with maximum correlation value among the points gathered from Step2 as the next pivot. In order to preventing a situation that x, y are pivots, z has small correlation with x while has large correlation with y, we use a heuristic method that select row with highest correlation value to avoid it.

## VI. EMPIRICAL RESULTS

### A. Test Suit

The empirical study involves in MoBIoS test suit [23], shown as Table I.

Table I. Test suit data set

| Data type | Size | Distance oracle | dimension |
|---|---|---|---|
| vector(uniform) | 100k | $L^2$ | 8 |
| DNA | 100k | Hamming distance | 9-18 |
| Protein | 100k | Weighted edit distance | 6-18 |
| image | 100k | L-norms | 66 |

The MoBIoS test suit consists of four data type and the size of databases are all 100,000. Two types of biological data are considered. (1) The amino-acid sequence fragments of the yeast proteome with weighted-edit distance based on the metric PAM substitution matrix, mPAM [24]. (2) The DNA sequence fragments of the Arabidopsis genomes with Hamming distance.

Two types of vector data are considered. (1) Uniform vector of 8 dimensions with $L^2$ distance. (2) The image dataset contains 10221 images. Each image is represented as three feature vectors which stand for color, structure and texture. The length of each vector is 15, 3 and 48. For feature vectors of texture and structure, the distance functions are both $L^2$ norm, while color vector use $L^1$ norm. The final distance is a linear combination of the distances of each feature vector, and also has the metric properties.

MVPT is the index data structure we used in this article, and the partition algorithm is clustering partition [25]. The number of pivots is two, fan-out is three and the maximum number of data points in each index leaf node is 100.

Since distance evaluation in a metric space is usually costly, we use the average number of distance calculations to evaluate the performance. For each test 5000 range queries are picked sequentially from the beginning of the dataset files. The radii of range queries are chosen so that approximately 0.01% of the databases are returned as query results.

FFT is an easy, fast and generally used pivot selection method, and PCA is one of the best pivot selection methods known as so far. Thus, these methods and our two new methods are compared.

### B. Experimental Result

As shown in Table II and Fig.7, in biological data type, our two new algorithms perform better than PCA and much better than FFT. In vector data type, the result of two new algorithm is very close to PCA method, and they are both better than FFT.

Comparing out two new pivot selection algorithms, method based on correlation is better than the other one based on covariance in every data type.

Thus, we can draw the conclusion that in DNA and protein data types two new methods are both better than other methods and the correlation method is the best among all. Additionally, in vector and image data types, two new methods especially correlation method is close to PCA method and better than FFT method.

Table II. Comparison between four methods, evaluated by distance calculation times

| Data set | Radius | selectivity | FFT | PCA | Correlation Method | Covariance Method |
|---|---|---|---|---|---|---|
| | | | Distance calculation times | | | |
| DNA 18-mer fragments | 4 | 0.0036% | 53179 | 49711 | 48729 | 48763 |
| Vector(uniform) dimension 8 | 0.3 | 0.0147% | 7942 | 5841 | 6849 | 7461 |
| Protein 6-mer fragments | 4 | 0.01% | 35689 | 30689 | 29775 | 30571 |
| image | 0.08 | 0.118% | 1441 | 1157 | 1266 | 1303 |

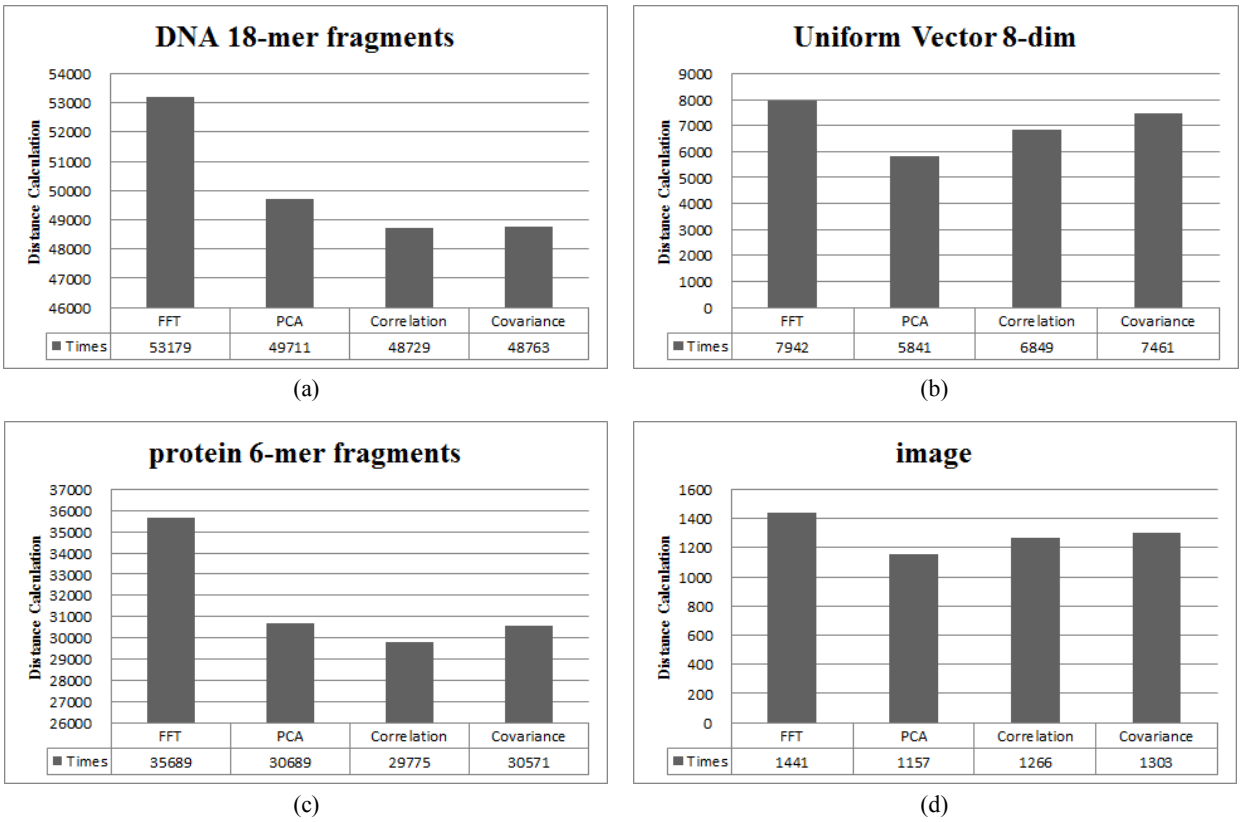


Figure 7. Comparison result: (a)DNA18, (b)vector 8-dim, (c)protein, (d)image

## VII. CONCLUSION

Pivot selection is the crucial factor that affects the performance of metric-space indexing. Under pivot space model, we propose two new pivot selection algorisms based on covariance and correlation. Empirical results shows that these two methods are both better than FFT, and is superior or comparable to the PCA method. The shortage of these algorithms is they take more time to bulkload an index tree. To this problem, we plan to use FFT to choose some candidates than perform out algorithm to a smaller matrix.

PCA method and our new methods only consider the relevancy of data objects, while ignoring the transforming magnitude of data. The basic idea of pivot selection is mapping points from high-dimension complete pivot space to low-dimension pivot space. After the mapping, the $L^{\infty}$ distance of data points decreased, which means information is lost. Thus, the target function of pivot selection is aim to minimize the loss of distance. This article attempts to selection pivot using covariance and correlation and get a good performance, future work will base on it.

Pivot space model makes it possible to apply mathematical tools to solve metric-space indexing problem. Current methods are all base on linear dimension reduction which only considers the linear correlation of data. Using non-linear methods is our next emphasis in future work.

## ACKNOWLEDGMENT

This research was supported by the following grants: China 863: 2012AA010239; NSF-China: 61033009, 61003272, 61170076; China NSF-GD grant: 10351806001000000; a grant from the Computer Architecture Key Lab of Chinese Academy of Sciences: ICT-ARCH201004; Shenzhen Foundational Research Project: JC201005280408A, JC200903120046A; a grant from the Shenzhen–Hong Kong Innovation Circle Project:

ZYB200907060012A. Shenzhen University Research Course Project: 0000132373.

## REFERENCES

[1] Y. Feng, K. Cao, Z. Cao, A Multidimensional Index Structure for Fast Similarity Retrieval, JOURNAL OF SOFTWARE, 13 (2002) 1678-1685.

[2] R. Stevens, C. Goble, P. Baker, A. Brass, A classification of tasks in bioinformatics, Bioinformatics, 17 (2001) 180-188.

[3] Midomi, http://www.midomi.com/.

[4] R. Mao, W.L. Miranker, D.P. Miranker, Pivot selection: Dimension reduction for distance-based indexing, Journal of Discrete Algorithms, 13 (2012) 32-46.

[5] E. Chávez, G. Navarro, R. Baeza-Yates, J.L. Marroquín, Searching in metric spaces, ACM Comput. Surv., 33 (2001) 273-321.

[6] G.R. Hjaltason, H. Samet, Index-driven similarity search in metric spaces (Survey Article), ACM Trans. Database Syst., 28 (2003) 517-580.

[7] H. Samet, Foundations of Multidimensional and Metric Data Structures, Morgan Kaufmann, 2006.

[8] P. Ciaccia, M. Patella, P. Zezula, M-tree: An Efficient Access Method for Similarity Search in Metric Spaces, in: Proceedings of the 23rd International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc., 1997, pp. 426-435.

[9] F. Dehne, H. Noltemeier, Vorono trees and clustering problems, Inf. Syst., 12 (1987) 171-175.

[10] J.K. Uhlmann, Satisfying general proximity / similarity queries with metric trees, Information Processing Letters, 40 (1991) 175-179.

[11] Z. Zhang, J. Li, An Algorithm Based on RGH-Tree for Similarity Search Queries, JOURNAL OF SOFTWARE, 13 (2002) 1969-1976.

[12] J. Li, Z. Zhang, Haperplane Tree: A Structure of Indexing Metric Spaces for Similarity Search Queries, JOURNAL OF COMPUTER RESEARCH AND DEVELOPMENT, 40 (2003) 1209-1216.

[13] S. Brin, Near Neighbor Search in Large Metric Spaces, in: Proceedings of the 21th International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc., 1995, pp. 574-584.

[14] W.A. Burkhard, R.M. Keller, Some approaches to best-match file searching, Commun. ACM, 16 (1973) 230-236.

[15] T. Bozkaya, M. Ozsoyoglu, Indexing large metric spaces for similarity search queries, ACM Trans. Database Syst., 24 (1999) 361-404.

[16] J.L. Bentley, Multidimensional binary search trees used for associative searching, Commun. ACM, 18 (1975) 509-517.

[17] G. Navarro, Searching in metric spaces by spatial approximation, The VLDB Journal, 11 (2002) 28-46.

[18] G. Navarro, Searching in Metric Spaces by Spatial Approximation, in: Proceedings of the String Processing and Information Retrieval Symposium \& International Workshop on Groupware, IEEE Computer Society, 1999, pp. 141.

[19] P.N. Yianilos, Data structures and algorithms for nearest neighbor search in general metric spaces, in: Proceedings of the fourth annual ACM-SIAM Symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, Austin, Texas, United States, 1993, pp. 311-321.

[20] D. Hochbaum, D. Shmoys, A Best Possible Heuristic for the k-Center Problem, Mathematics of Operations Research, 10 (1985) 180-184.

[21] K. Pearson, On lines and planes of closest fit to systems of points in space, Philosophical Magazine, 2 (1901) 559-572.

[22] S. Roweis, EM algorithms for PCA and SPCA, in: Proceedings of the 1997 conference on Advances in neural information processing systems 10, MIT Press, Denver, Colorado, United States, 1998, pp. 626-632.

[23] MoBIoS Test Suit, http://aug.csres.utexas.edu/mobios-workload/.

[24] W. Xu, D.P. Miranker, A metric model of amino acid substitution, Bioinformatics, 20 (2004) 1214-1221.

[25] R. Mao, W. Xu, S. Ramakrishnan, G. Nuckolls, D.P. Miranker, On Optimizing Distance-Based Similarity Search for Biological Databases, in: Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference, IEEE Computer Society, 2005, pp. 351-361.