LLE Based Pivot Selection for Similarity Search of Biological Data

Kewei Ma, 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*—Distance-based indexing is a widely used technique for general purpose search. Pivot selection is the most crucial step of bulkloading a metric-space indexing tree. Current pivot selection methods are mainly based on linear methods. A non-linear method based on Locally Linear Embedding is proposed. Empirical results demonstrate that the performance of new method is superior to existing methods.**

Keywords-similarity search; metric-space indexing; pivot selection; locally linear embedding; dimensional reduction

I. INTRODUCTION

In recent years, mass data processing has become the crucial part of internet and cloud computing. Indexing techniques which serve to enhance search performance are the fundamental component of mass data processing. For a decade, with the development of information technology, numerous new data types have emerged in many different areas. Traditional search methods like B^+-tree designed for numeric data and strings are no longer satisfied the current needs of those new unstructured data types which are multi-dimensional and cannot be sorted. Take biological data as example, the similarity of DNA sequence is measured by Hamming Distance and protein sequence is computing by Edit Distance. Moreover, the measurement of diverse data types differs from one to another. Building specific DBMSs for each data type costs money and time, so this is not an effective way. Consequently, new DBMSs for general purpose search are needed.

Metric-space indexing [1-3] is a general solution for similarity search of complex data types. Distance to each other data objects is the only information required to perform a similarity search. Domain-specific information is hidden and the distance function is provided by users which means it's transparent to search system. However, lacking of information apart from distance makes mathematical tools unable to be used in metric space and stunts the research progress in this area for a time.

Nevertheless, the appearance of pivot space model [4] makes it possible for mathematical tools directly to be used to solve metric-space indexing problems. Besides, bulkloading a metric-space indexing tree requires two steps: data partition and pivot selection. [4] proves that pivot selection has the determinative effect on the quality of indexing tree. [4] uses a popular linear dimension reduction method called Principal Component Analysis (PCA) [5] to select pivots.

This article focuses on studying the capability and query performance using a non-linear dimension reduction method called Locally Linear Embedding (LLE) [6] to select pivots under pivot space model.

II. PIVOT SPACE MODEL

This section is presented as background. Readers are referred to [4] for additional details.

Answering a similarity query in pivot space model, take range query as example, usually consists of two steps, as Fig. 1 shows:

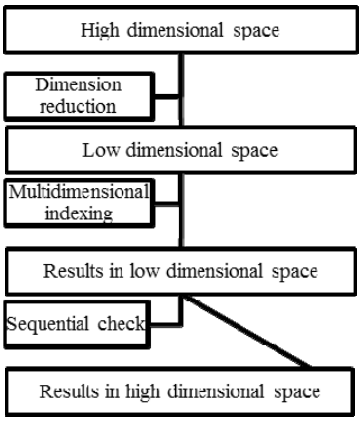


Figure 1 Process of answering a range query in pivot space model

Step1. Off-line initialization. In this step, in the first place, map data objects isometrically from metric space into a high-dimensional complete pivot space with L^∞ distance. Secondly, apply dimension reduction methods (equivalent to pivot selection in this case) on complete pivot space and the result is called low-dimensional pivot space. Next, multi-dimensional

* Corresponding author

indexing methods can be exploited to finish building the indexing tree.

Step2. Online search. In this step, the query object must be mapped into pivot space first. Though the shape of the image of the query ball in a pivot space is hard to determine, it has been proved that the image of the query ball is completely covered by a hypercube of edge length 2r in the pivot space. Under this circumstances, search the indexing tree uses the corresponding multi-dimensional indexing method. At last, in order to get final answer, check the result sequencely and remove false positivies.

III. EXISTING PIVOT SELECTION METHODS

A. FFT

Farthest-Fist-Traversal (FFT) [7] is a general pivot selection method. FFT is a fast, greedy algorithm that minimizes the maximum cluster radius. In FFT, k points are selected as cluster centers at first. Then the remaining point is added to the cluster whose center is the closest. Due to its fast speed and simplicity, indexing structures such as MVPT [8] choose FFT to select pivots.

B. PCA

Thanks to pivot space model, mathematical tools such as dimension reduction can be used to select pivots. [4] proves that performing search on a complete pivot space degrades to a linear scan, so that dimension reduction is inevitable.

Consequently, [4] uses PCA to select pivot. However, [4] also proves that any dimension reduction techniques which create new dimensions will still require a calculation of n dimensions. As a result, selecting the existing dimension is the only solution, so that [4] selects existing axis with the maximum correlation with the new dimensions created by PCA.

Empirical results show that PCA is one of the best pivot selection methods currently.

IV. LLE BASED PIVOT SELECTION METHOD

PCA is a linear dimension reduction method designed for linear data set. It only focuses on the linear correlation between data objects, while the transforming magnitude and non-linear relation between data objects are not considered. Biological data are mostly not linear, therefore, we propose to use a non-linear dimension reduction technique to optimize the quality of pivot selection.

A. Locally Linear Embedding

Locally Linear Embedding [6] is a relatively fast non-linear dimension reduction method. It has several advantages including faster optimization due to take advantage of sparse matrix algorithm.

The LLE algorithm can be separated into three steps [6]:

Step1. For each point x_i in high-dimensional space, find its k nearest neighbors. The distance function is $d_{ij} = [\sum_{k=1}^{D} | x_{ik} - x_{jk} |^p]^{\frac{1}{p}}$, and we use p=2 i.e. Euclidian Distance for simple computing.

Step2. Solve for reconstruction weight matrix. The reconstruction error is given by the cost function:

$$\min\varepsilon(\mathrm{W}) = \sum_{i}^{N} | x_i - \sum_{j}^{k} w_{ij} x_{ij} |^2 \tag{1}$$

x_{ij} is k nearest neighbors of point x_i, w_{ij} is the weight between x_i and x_{ij}. Moreover, the sum of every row in the weight matrix equals 1:

$$\sum_{j=1}^{k} w_{ij} = 1 \tag{2}$$

Combine (1) and (2), the cost function can be rewritten to:

$$\min\varepsilon(\mathrm{W}) = \sum_{i}^{N} w_i^T Z_i w_i \tag{3}$$

The goal of this step is to minimize the cost funtion, and we could solve (3) using Lagrange multiplier. The result is $Z_i w_i = 1$.

Step3. Use the weight matrix W to map data points into lower dimension *d*, and the cost function is:

$$\min\phi(\mathrm{Y}) = \sum_{i}^{N} | y_i - \sum_{j}^{k} w_{ij} y_{ij} |^2 \tag{4}$$

The restrain of Y is $\sum_{i=1}^{k} y_i = 0$ and $\frac{1}{N}\sum_{i-1}^{N} y_i y_i^T = I$, I is a unit matrix of N dimension. Then use Lagrange multiplier again to solve (4) and we get $MY^T = \lambda Y^T$. Because of the smallest eigenvalue of M is very close to zero, so we pick from the feature vector whose eigenvalue is secondly smallest to *d*+1th feature vector as the output result.

B. Pivot Selection Method Base on LLE

```
PivotSelection ((S, d): dataset, p: pivot number, c: constant)
{  // run FFT to create a candidate set in size of p*c
1.    candidate set  = FFT(S, p*c);
   // generate pivot space with candidate set as the pivot set
2.    PS = G(S, candidate, d);
   // run LLE on PS
3.    LLE output = LLE(PS, p)
   // for each row, find the vector has smallest angle with it
4.    for each row ∈LLE output:
      Pivots = Pivots ∪  argmax_x(Angle(LLE output, x))
      Return Pivots; }
```

Figure 2 Pivot selection algorithm based on LLE

Table I. TEST SUIT DATA SET

Data type	Size	Distance oracle	dimension
DNA	100k-500k	Hamming distance	9-18
protein	100k-500k	Weighted edit distance	6-18

Our LLE based pivot selection method is shown in Fig. 2 However, the size of weight matrix in LLE algorithm is N*N, if we directly perform LLE in data which have 50k points, it takes too much time and memory, which make it unpractical in real application. In addition, [9] argues that good pivots are usually corners, but the reverse is false. In step 1, as a result, we use FFT algorithm to select some points in the original distance matrix as candidate set [4, 7].

In step 2, the distance between the corners and the point in database are computed to build the distance matrix of corner pivot space.

In step 3, the classical LLE algorithm is performed on the corner pivot space, and the number of pivot is the output dimension of LLE.

In step 4, although no new dimension is created, the exact shape of the query ball which is mapped into pivot space by LLE is hard to determine. Thus, we propose a heuristic method to solve the problem that is for each row, also means points here, of the LLE output, find the vector in corner pivot space which has smallest angle with it to be the pivot.

V. EMPIRICAL RESULTS

The empirical study involves in MoBIoS test suit [10], shown as Table I.

A. Test Suit

In order to qualify the query performance of biological data, two types of biological data in MoBIoS test suit are considered. (1) The amino-acid sequence fragments of the yeast proteome with weighted-edit distance based on the metric PAM substitution matrix, mPAM [11]. (2) The DNA sequence fragments of the Arabidopsis genomes with Hamming distance.

We choose MVPT as the index data structure in this article, and the partition method used is clustering partition [12]. The number of pivots is two, fan-out is three and the maximum number of data points in each index leaf node is 100.

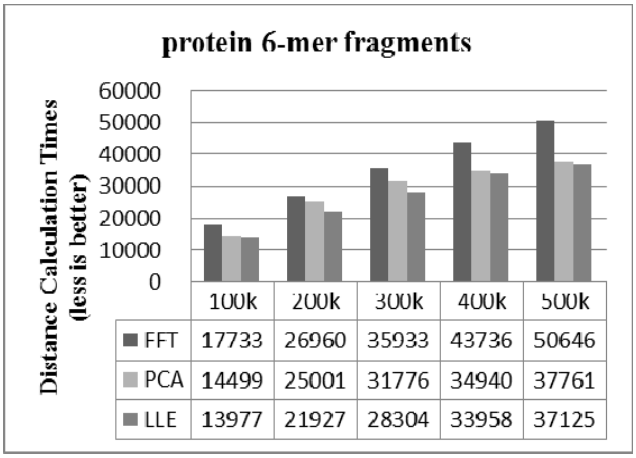
In the respect of LLE algorithm, the number of neighbors we used is 18 for DNA sequence and 17 for Protein data. Besides, the constant we used in FFT algorithm is 100.

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 data set files. The radii of range queries are chosen so that approximately 0.01% of the 100k databases are returned as query results. In order to prove the expandability of our algorithm we test databases with different size from 100k to 500k, and the radii of range queries are same to ensure the consistency of variables.

FFT is an easy, fast and most widely used pivot selection method, and PCA is one of the best pivot selection methods at present. Thus, LLE method is compared with these two methods.

B. Empirical Result

The experiment result is shown as Fig. 3. According to protein query results, in each database size the distance calculation times of FFT is the most among three. PCA method has less distance calculation times than FFT, while LLE has the least and best distance calculation times in any size of database. Additionally, the increase of number of calculation from100k to 500k database is 32913 for FFT, 23262 for PCA and 23148 for LLE, which also proves LLE a better pivot selection method for protein data type.



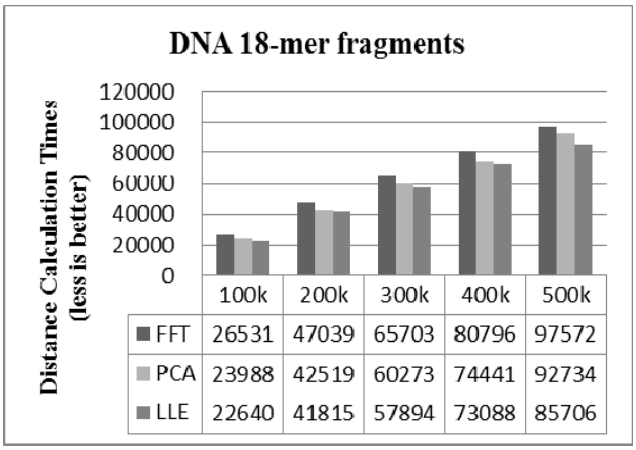


Figure 3 Experiment results of protein and DNA

In the DNA query result, the number of distance calculation for three pivot selection methods in each size of databases still shows that LLE is better than the other two methods. Besides, the increase of calculation times from100k to 500k database is 71041 for FFT, 68746 for PCA and 63066 for LLE, and LLE method still performs the best.

Based on experiment result, we can draw the conclusion that in biological data DNA and protein, LLE method performs the best pivot selection quality among all, and best fits these biological data types.

VI. CONCLUSION

The significance of pivot selection for metric-space indexing is clear without doubt. It's a crucial factor to build a better indexing tree. Current pivot selection methods are all linear. In this article, we propose a non-linear pivot selection method based on LLE under pivot space model in order to get better pivot selection quality of biological data types. Empirical results show that LLE method is the best method.

There is a slight pity that LLE method performs better than FFT but not better than PCA method when dealing with vector data. The next emphasis of our research will focus on how to improve the performance of LLE method on vector data.

Moreover, studying the exact shape of the image of query ball in a pivot space is a challenging task; future work will also include this problem.

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

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

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

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

[6] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, SCIENCE, 209 (2000) 2323--2326.

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

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

[9] B. Bustos, E. Chavez, G. Navarro, Pivot selection techniques for proximity searching in metric spaces, Pattern Recognition Letters, 24 (2003) 2357-2366.

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

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

[12] 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.