Spectral LPM: An Optimal Locality-Preserving Mapping using the Spectral (not Fractal) Order*

Mohamed F. Mokbel Walid G. Aref Ananth Grama

Department of Computer Sciences, Purdue University, West Lafayette, IN 47907-1398

{mokbel,aref,ayg}@cs.purdue.edu

Abstract

For the past two decades, fractals (e.g., the Hilbert and Peano space-filling curves) have been considered the natural method for providing a locality-preserving mapping. The idea behind a locality-preserving mapping is to map points that are nearby in the multi-dimensional space into points that are nearby in the one-dimensional space. In this paper, we argue against the use of fractals in locality-preserving mapping algorithms, and present examples with experimental evidence to show why fractals produce poor locality-preserving mappings. In addition, we propose an optimal locality-preserving mapping algorithm, termed the Spectral Locality-Preserving Mapping algorithm (Spectral LPM, for short), that makes use of the spectrum of the multi-dimensional space. We give a mathematical proof for the optimality of Spectral LPM, and also demonstrate its practical use.

1. Introduction

An important factor for multi-dimensional databases is how to place the multi-dimensional data into a one-dimensional storage media (e.g., the disk) such that the spatial properties of the multi-dimensional data are preserved. A mapping function f is required to map the multi-dimensional space into the one-dimensional space. *Locality-Preservation* is a desirable property for the mapping function f. Mapping data from the multi-dimensional space into the one-dimensional space is considered *locality-preserving* if the points that are nearby in the multi-dimensional space are nearby in the one-dimensional space.

Fractal space-filling curves (e.g., the Hilbert and Peano) have long been used as a locality-preserving mapping [4] for multi-dimensional similarity search queries, spatial join, R-tree packing, declustering, spatial access methods, and GIS applications. In this paper, we go beyond the idea of fractals for locality-preserving mappings and propose a novel and optimal algorithm that does not depend on fractals. Instead, the proposed algorithm, termed the *Spectral Locality-Preserving Mapping* algorithm (Spectral LPM, for short), depends on the spectral properties of the points in the multi-dimensional space.

*This work was supported in part by the National Science Foundation under Grants IIS-0093116 and IIS-0209120, the NAVSEA/Naval Surface Warfare Center Crane, and the Purdue Research Foundation.

2. The Fractal Mapping

Fractals divide the space into a number of fragments, visiting the fragments in a specific order. Once a fractal starts to visit points from a certain fragment, no other fragment is visited until the current one is completely exhausted. By dealing with one fragment at a time, fractals perform a local optimization based on the current fragment. Thus, fractals suffer from the *boundary effect* problem where points far from the fragment borders are favored. Points that lie near to the fragment borders fare the worst. Figure 1 gives an example of this boundary effect on three different fractal locality-preserving mapping algorithms, the Peano, Gray, and Hilbert space-filling curves. In each curve, the space is divided into four quadrants. P_1 and P_2 lie on two different quadrants. Although $|P_1 - P_2| = 1$ in the two-dimensional space, they are very far from each other in the one-dimensional space. The distance between P_1 and P_2 in the one-dimensional space will be $22, 47, 43$ if we use the mapping algorithms based on the Peano, Gray, and Hilbert space-filling curves, respectively. The boundary effect problem in fractals is unavoidable, and it results in non-deterministic results.

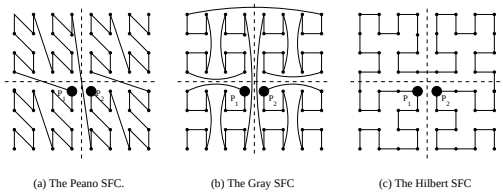


Figure 1. Fractal Mapping.

3. The Spectral LPM Algorithm

In this paper, we present the Spectral Locality-Preserving Mapping (LPM) algorithm, an optimal mapping with respect to all data points. Spectral LPM avoids the drawbacks of fractals by using a global optimization instead of a local one. By global optimization, we mean that all multi-dimensional data points are taken into account when performing the mapping. Unlike fractals, Spectral LPM does not favor any set of points over the others. In general, spectral algorithms use the eigenvalues and eigenvectors of the matrix representation of a graph. Spectral algorithms [5] have been widely used in graph partitioning, data clustering, linear labeling of a graph, and load balancing. The optimality of the spectral order in many applications is discussed in [1, 3]. Figure 2 gives the pseudo code of the Spectral LPM. An example of applying the Spectral LPM to a set of two-dimensional points in a 3×3 grid is given in Figure 3. Notice that Figures 3b and 3c correspond to the first and second steps of Spectral LPM, while Figure 3d corresponds to steps 3, 4, and 5 in Spectral LPM.

Algorithm Spectral LPM
Input: *A set of multi-dimensional points P.*
Output: *A linear order S of the set P.*

1. *Model the set of multi-dimensional points P as a graph $G(V, E)$ such that each point $P_i \in P$ is represented by a vertex $v_i \in V$, and there is an edge $(v_i, v_j) \in E$ if and only if $|P_i - P_j| = 1$.*
2. *Compute the graph Laplacian matrix $L(G) = D(G) - A(G)$. $D(G)$ is the Diagonal matrix of G where $D(G)_{ii}$ = the degree of vertex v_i. $A(G)$ is the adjacency matrix of G where $A(G)_{ij}$ = 1 iff the edge $(i, j) \in E$.*
3. *Compute the second smallest eigenvalue λ_2 and its corresponding eigenvector X_2 of $L(G)$, known as the Fiedler Vector*
4. *For each $i = 1 \rightarrow n$, assign the value x_i to v_i and hence to P_i*
5. *The linear order S of P is the order of the assigned values of P_i's.*
6. **return** S.
7. **End.**

Figure 2. Pseudo code for the Spectral LPM.

The optimality of the Spectral LPM is proved with the following theorems. The proofs of these theorems are omitted for brevity.

Theorem 1 : *A vector $X = (x_1, x_2, \ldots, x_n)$ that represents the n one-dimensional values of n multi-dimensional points represented as a graph $G(V, E)$ is considered to provide the global optimal locality-preserving mapping if X satisfies: $Min.f = \sum_{(v_i,v_j)\in E}(x_i - x_j)^2 \quad S.t: \quad \sum_{i=1}^{i=n} x_i^2 = 1, \sum_{i=1}^{i=n} x_i = 0$*

Theorem 2 : *The optimization problem in Theorem 1 is equivalent to: $Min.f = X^T L X \quad S.t: \quad X^T X = 1, X^T e = 0$*

Theorem 3 *[2]***:** *The solution of the optimization problem in Theorem 2 is the second smallest eigenvalue λ_2 and its corresponding eigenvector X_2.*

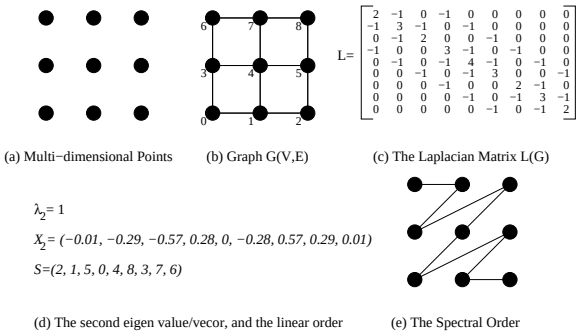


Figure 3. The Spectral LPM algorithm.

4. Extensibility of the Spectral Mapping

Assume that we need to map points in the multi-dimensional space into disk pages, and we know (from experience) that whenever point x in P_x is accessed, there is a very high probability that point y in page P_y will be accessed soon afterwards. To force mapping x and y into nearby locations in the one-dimensional space using Spectral LPM, we add an edge (x, y) to the graph G. By adding this edge, we inform Spectral LPM that x and y need to be treated as if they have Manhattan distance $M = 1$ in the multi-dimensional space.

Another extensibility feature in Spectral LPM is that we can change the way we construct the graph G. Figure 4 gives the modeling of two-dimensional points in a 4×4 grid with the resulting spectral order after applying Spectral LPM for four-connectivity (Figures 4a, 4b) and eight-connectivity graphs (Figures 4c, 4d). More generally, points in the multi-dimensional space can be modeled as a weighted graph, where the weight w of an edge $e(v_1, v_2)$ represents the priority of mapping v_1 and v_2 to nearby locations in the one-dimensional space[1]. Notice that the optimality proof of Spectral LPM is valid regardless of the graph type. The idea of Spectral LPM is that it is optimal for the chosen graph type.

[1]In this case, $L(G)_{ii} = \sum_{(i,j)\in E} w_{ij}$, and $L(G)_{ij} = -w_{ij}$ if $(i, j) \in E$, o.w., $L(G)_{ij} = 0$. The objective function of Theorem 1 will be $f = \sum_{(v_i,v_j)\in E} w_{ij}(x_i - x_j)^2$.

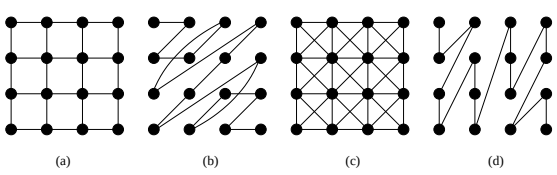
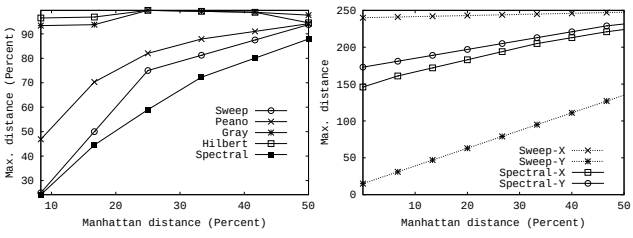


Figure 4. Variation of the Spectral LPM.

5. Experimental Results

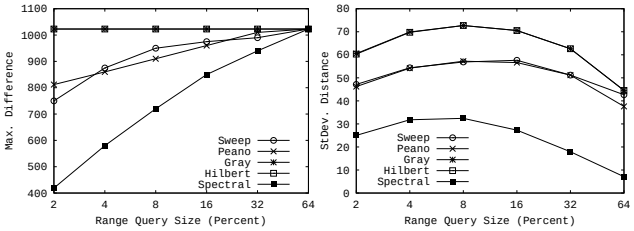
In the experiments, we concentrate on the effect of Spectral LPM on nearest-neighbor and range queries. Due to its optimality, Spectral LPM will give superior performance when applied to other applications. We compare Spectral LPM with three fractal algorithms based on the Peano, Gray, and Hilbert space-filling curves and a row-major (Sweep) mapping as a simple and straightforward non-fractal mapping algorithm. The experiments in Figure 5 answer the following question: If the Manhattan distance between any two points P_i, P_j in the multi-dimensional space is M_D, then what is the Manhattan distance M_1 between the same two points in the one-dimensional space? The lower M_1 the better the locality-preserving mapping for nearest neighbor queries. Figure 5a gives the maximum one-dimensional distance for any two five-dimensional points. In general, non-fractal algorithms have better performance than the fractals. In Figure 5b, we compute the Manhattan distance over only one dimension of the two-dimensional space. By the curves Sweep-X and Sweep-Y, we mean the Manhattan distance over the X and Y dimensions, respectively. The performance of the Sweep mapping have much variation when measuring the distance over the X and Y dimensions. However, for the Spectral mapping, the performance is very similar for the two dimensions. Thus, Spectral LPM provides fair mapping where it does not discriminate between the two dimensions.

Experiments in Figure 6 answers the following question: For any multi-dimensional range query, what is the difference between the minimum and the maximum one-dimensional values of the points that lie inside the range query? The smaller the difference the better the locality-preserving mapping. Keeping the difference as small as possible allows a sequential access from the minimum point to the maximum point while eliminating the records that lie outside the range query. Figure 6a gives the maximum difference between the maximum and minimum one-dimensional points for a certain range query in the four-dimensional space. Spectral LPM gives an outstanding performance compared to the other mappings. Figure 6b gives the standard deviation of the distance difference between the maximum and minimum one-dimensional values for any multi-dimensional range query in the four-dimensional spaces. For all possible partial range queries with a certain size and dimensionality, we plot the standard deviation of the difference between the maximum and the minimum values of the points that lie inside the query in the one-dimensional space. The lower the standard deviation the more fair the locality-preserving mapping.



(a) Worst case (b) Fairness

Figure 5. Nearest Neighbor queries



(a) Worst case (b) Fairness

Figure 6. Range queries.

6. Conclusion

In this paper, we argue against the use of fractals as a basis for locality-preserving mapping. Then, we propose the Spectral LPM; a provably optimal algorithm for mapping the multi-dimensional space into the one-dimensional space such that the points that are nearby in the multi-dimensional space would still nearby in the one-dimensional space. Experimental analysis shows the superior performance of Spectral LPM over the long used fractal algorithms for similarity search queries and range queries. We believe that Spectral LPM can effi ciently replace the fractal locality-preserving mapping algorithms in many other applications, e.g., multimedia databases, spatial join, declustering, multi-dimensional indexing, and GIS applications.

References

[1] T. F. Chan, P. Ciarlet, and W. K. Szeto. On the optimality of the median cut spectral bisection graph partitioning method. *SIAM Journal on Sci. Comp.*, 18(3):943–948, May 1997.

[2] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(98):298–305, 1973.

[3] M. Juvan and B. Mohar. Optimal linear labelings and eigenvalues of graphs. *Discrete Applied Math.*, 36:153–168, 1992.

[4] B. Moon, H. Jagadish, C. Faloutsos, and J. Salz. Analysis of the clustering properties of hilbert space-fi lling curve. *IEEE TKDE*, 13(1):124–141, 2001.

[5] L. A. Steen. Highlights in the history of spectral theory. *American Mathematical Monthly*, 80(4):359–381, Apr. 1973.