Compressing High Dimensional Datasets by Fractals

Xintao Wu1  Daniel Barbará2
1UNC Charlotte, xwu@uncc.edu
2 George Mason University, dbarbara@gmu.edu

Fractal techniques based on iterated function systems have been successfully applied to the compression of one-dimensional signals, two-dimensional images, and three-dimensional volumes by finding a fractal representation that models the original data as closely as possible, and storing the model instead of the original data. In this paper, we extend fractal compression techniques to general datasets. The intuition behind our idea is some real world datasets (e.g., traffic in a network, stock prices, or datasets with hierarchical domains) may exhibit part of self-similarity which can be captured by affine transformations.

Just as an image is a collection of pixels which represent some measurable property of an object sampled on an integer 2-D grid, the high-dimensional dataset can be viewed as a collection of cells which represent some measure on an integer n-D grid. However, the pixels of an image can measure the property of an object at any resolution level by sampling on an integer 2-D grid (e.g., 512×512, 1024×1024 etc.) while the cells of general datasets do not contain infinite resolution levels and are characterized by fixed hierarchies along dimensions. For instance, the dimension stores can be grouped by city, state, region, and country. The measure over different scales along dimensions’ hierarchies may exist self-similarities, for instance, the product’s sale value (as measure) of one store may be similar to the aggregates of all stores of that city. We have applied a two-phase searching strategy to overcome the increased searching time caused by additional dimensions. The search scheme is to first check a small number of spatially close local domain chunks. This step is very fast as the statistics needed to compute the parameters of affine transformations are already available in the range chunk’s description. The data structure used here is 2^n-tree which is an natural extension of quadtree (for image) and octree (for volume). Each node, corresponding to a range chunk or a domain chunk, contains the summary information used for local matching. If a small localized search cannot provide sufficiently accurate domain maps, a larger global domain pool is interrogated during a second pass. Our experiments over stock dataset (2 dimensions with domain size 553, 224 for stock and date respectively) show the performance of fractal compression is comparable with rivals such as loglinear model. The experiments over synthetic datasets (4 dimensions with domain size 80, 40, 20, 10 respectively) show the scalability of fractal compression techniques when the underlying datasets display the self-similar characteristics.

There are some aspects of this work that merit further research. We are investigating some complex affine transformations between the range and domain chunks. To overcome high time complexity caused by additional dimensions, we are investigating approximate multi-dimensional nearest neighbors searching techniques (e.g., kd-trees) that run in expected logarithmic time. We are also investigating the techniques of fast characterizing the underlying datasets before choosing one data reduction technique to compress data.

1This work was supported, in part, by funds provided by the University of North Carolina at Charlotte.