I have always been fascinated by the world of fractals and have been deeply enthusiastic exploring the maths behind them. This post is announcing the support of the fractal dimension computation in the MF - Matrix Factorization for Data Mining library. In the following paragraphs we shortly revise the most important concepts and definitions of the fractal dimension. The embedding dimensionality of a data set is the number of attributes in the data set. The intrinsic dimensionality is defined as the actual number of dimensions in which the {mathpublisher}n{/mathpublisher} m-dimensional original vectors can be embedded under the assumption that some distance in the reduced space is kept among them. Given a fractal as a self-similar set of points with {mathpublisher}r{/mathpublisher} self-similar pieces, where each is scaled down by a factor of {mathpublisher}s{/mathpublisher}, the fractal dimension&#65533;{mathpublisher}D{/mathpublisher}&#65533;of the object is defined as {mathpublisher}D = { log r} / { log s} {/mathpublisher}. Example: Sierpinski triangle (My seminar work for CG on L-systems - figure 1 in Appendix of the Report document) consists of three self-similar parts and each is scaled down by a factor of two, therefore its fractal dimension is {mathpublisher}D approx 1.58.{/mathpublisher} For the finite set of points in a vector space, we say the set is statistically self-similar on a range of scales{mathpublisher}(a, b){/mathpublisher}on which the self-similarity is true. However in the theory self-similar object should have infinitely many points because each self-similar part is a scaled-down version of the original object. As a measure of the intrinsic fractal dimension of a data set, the slope of the correlation integral is used. The correlation integral&#65533;{mathpublisher}C(r){/mathpublisher}&#65533;for the data set&#65533;{mathpublisher}S{/mathpublisher}is defined as {mathpublisher}C(r) = Count(dist(u, v) <= r; u in S, v in S, u != v). {/mathpublisher} Given a data set&#65533;{mathpublisher}S{/mathpublisher} which is statistically self-similar in the range{mathpublisher}(a,b){/mathpublisher}, its correlation fractal dimension {mathpublisher}D {/mathpublisher}is {mathpublisher}D = {partial log C(r)} / {partial log r}, r in [a, b]. {/mathpublisher} It has been shown that the correlation fractal dimension corresponds to the intrinsic dimension of a data set. Many properties hold for the correlation fractal dimension, see [1] and [2]. For us it is especially important, that the intrinsic dimensionality gives a lower bound on the number on attributes needed to keep the vital characteristics of the data set. A fast algorithm for the computation of the intrinsic dimension of a data set presented in [2] is implemented in the MF - Matrix Factorization for Data Mining library. Intuitive explanation of the correlation fractal dimension is that is measures how the number of neighbor points increases with the increase of the distance. It therefore measures the spread of the data and the fractal dimension equal to the embedding dimension means that the spread of the points in the data set is maximum. Of high importance is a Conjecture 1 in [1]: With all the parameters being equal, a dimensionality reduction method which achieves higher fractal dimension in the reduced space

is better than the rest for any data mining task. Therefore correlation fractal dimension of a data set can be used: <ul> for determining the optimal number of dimensions in the reduced space, as a performance comparison tool between dimensionality reduction methods, and all this can be done in way that is scalable to large data sets. &#65533; Recommended reading: <ol> Kumaraswamy, K., (2003). Fractal Dimension for Data Mining. Carnegie Mellon University. Jr, C. T., Traina, A., Wu, L., Faloutsos, C., (2010). Fast Feature Selection using Fractal Dimension. Science, 1(1), 3-16. In [1] a concept of intrinsic fractal dimension of a data set is introduced and it is shown how fractal dimension can be used to aid in several data mining tasks. In [2] a fast O(n) algorithm to compute fractal dimension of a data set is presented. On top of that a fast, scalable algorithm to quickly select the most important attributes of the given set of n-dimensional vectors is described.