From *SIAM News*, Volume 41, Number 10, December 2008

Geometry Analysis and Signal Processing on Digital Data, Emergent Structures, and Knowledge Building

By Ronald R. Coifman and Mauro Maggioni

This article (which is based on the invited talk of the first author at the 2008 SIAM Conference on Data Mining) discusses "diffusion geometry," which, by generalizing classic tools of harmonic analysis, provides a synthesis of different approaches to data analysis and processing.

In the last few years exciting developments in data mining and machine learning have been applied to the analysis of large data sets arising in a wide variety of disciplines. With millions of text documents being converted to digital format, for example, many users would benefit from automatic ways to organize and extract information from large collections of documents, automatic recommendations of interesting documents based on their reading history, and so on.

Many of the problems that arise in this area fall broadly into two classes. The first class encompasses problems related to the geometry of the data: low-dimensional, low-distortion embeddings of large data sets in high-dimensional space and graphs, permitting visualization, human interaction and information extraction, denoising of data, outlier detection, and other capabilities. The second class includes problems about the approximation/fitting/learning of functions on the data from a few samples, with the goal of predicting the values of the functions at new data points. Of particular importance have been methods based on the assumption that the intrinsic geometry of the data plays an important role, and that the smoothness of relevant functions on the data should be defined in a way that is adapted to the geometry.

Diffusion Geometry

Ideas from harmonic analysis and spectral graph theory have played a fundamental role in recent advances in this area. Diffusion geometry starts from the premise that a similarity measure $A(x,y)$ between any pair (x,y) of nearby data points can be meaningfully defined. A typical choice for data points lying in R^D is $A(x,y) = \exp(-\|x-y\|^2/t)$, where t is a fixed scale parameter. In general, the choice of A is both data- and goal-dependent. If N is the number of data points, A is an $N \times N$ matrix, which we think of as sparse because only nearby data points are connected by an edge with a weight above some threshold.

We can renormalize A to obtain a Markov matrix P, which represents a random walk on the data points—that is, $P(x,y)$ is the probability of jumping from x to y in one step. In diffusion geometry, P and its powers are used to gain insight into the geometry of the data, e.g., by finding coordinate systems, as well as to construct dictionaries of functions, à la Fourier or wavelet analysis, for learning functions on the data.

As an example (see Figure 1), we consider a body of 1000 articles from *Science News* (kindly provided by J. Solka). We can represent each document as a high-dimensional vector by fixing a vocabulary of d words and letting the kth coordinate of a document be the frequency of the kth word in the dictionary for that document. In our case we selected 10,000 common English words and then retained the top 1000 with respect to a score for significance in the data set, based on mutual information. We define similarity between documents as the correlation between their word vectors when larger than 0.95, and as 0 otherwise. The eigenvectors of the normalized similarity P can be used as coordinates on the data set, yielding a low-dimensional representation (Figure 1).

Notice how topics tend to cluster, suggesting that the subdivision of topics (not used in the construction of this embedding!) is related to the intrinsic geometry of the data set, and that automatic discovery of such topics and classification of the documents into categories are possible. This could be done with or without supervision. In the unsupervised framework, we look for clusters, defined as subsets of the data that have large amounts of inter-connectivity and low amounts of intra-connectivity. In a supervised setting, a few documents are labeled by hand as belonging to each corresponding scientific field; the given labels are then propagated on the graph to label the remaining documents automatically.

Another example is a hyperspectral image in which a spectrum of light absorption is associated with each pixel (Figure 2). The points (absorption spectra) are indexed by pixels, and we define affinity between two pixels as a correlation between their spectra that exceeds 0.8, and 0 otherwise.

If we label some regions as representing different tissue types, we can propagate these labels on the graph to label all the points in the image automatically and obtain a segmentation. We can then "propagate" the green, magenta, and blue labeled points, assuming that pixels with nonzero affinity are likely to be in the same class. The images shown in Figure 2 contrast the result of such "label diffusion" with the massive failure of the nearest-neighbor approach, in which a pixel is classified by the class of the label set with the highest correlation.

We selected the threshold of 0.8 for correlations between spectra so that two spectra so tightly correlated would be for the most part in the same label class. In this case all that is required is a good "nearest-neighbors" model. This is analogous to a local linear differential equation model in calculus. In

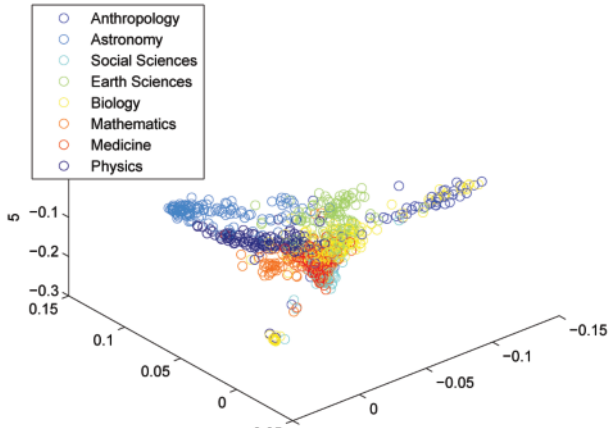

Figure 1. *Low-dimensional diffusion map of a body of documents from* Science News. *Data set courtesy of Jeff Solka.*

our case "integration from local to global" is achieved by diffusing the given labels on the graph induced by the local affinities.

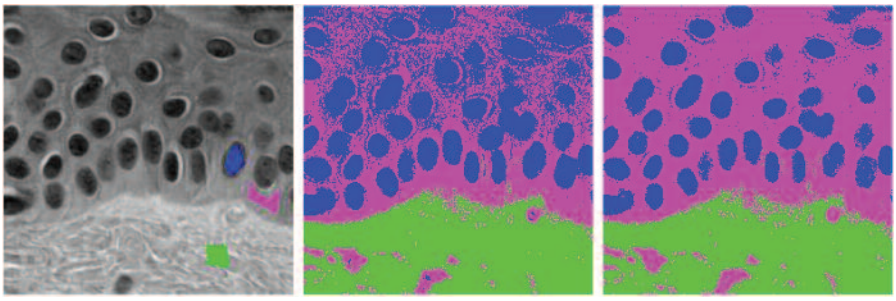

Figure 2. *Hyperspectral image of a dermatology sample. Left, regions with tissue of different types are marked by different colors. Center, prediction of tissue types by a nearest-neighbor approach. Right, prediction by diffusion of the training labels to all points.*

Organizing Data by Diffusion Geometry

The organization of digital data by diffusion geometry can be accomplished, roughly speaking, in two ways: (1) by a dimension-reduction approach that embeds the data in low-dimensional Euclidean space through the use of eigenvectors of the affinity matrix/kernel A (or a normalized related matrix), followed by processing and clustering in the lower dimension; (2) by hierarchical folder building and clustering, a bottom-up agglomeration approach that propagates or diffuses affinity between documents; this can be achieved through probabilistic mod-el building and statistical/combinatorial "bookkeeping" on the data.

For the first approach, which is based on the eigenfunctions of A, or on the random walk P, we let $P\varphi_i = \lambda_i\varphi_i$, assuming that $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_i \ldots$. We can use the eigenfunctions φ_i to map the data to m-dimensional Euclidean space by $\Phi_m^t(x) := (\lambda_1^{t/2}\varphi_1(x), \ldots, \lambda_m^{t/2}\varphi_m)$. This is closely related to the so-called spectral graph em-bedding long in extensive use for graph layouts; in fact, few properties of this embedding are known. Observe that the probability of a path of length t from x to y is $P^t(x,y) = \sum_i \lambda_i^t \varphi_i(x)\varphi_i(y)$. For large t, because all λ's are smaller than 1, λ_i^t is very small for i larger than, say, m. But the Euclidean distance between $\Phi_m^t(x)$ and $\Phi_m^t(y)$ is then equal to the Euclidean distance, in the N-dimensional space of data points, between the probability distributions $P^t(x,\cdot)$ and $P^t(y,\cdot)$, called the diffusion distance between x and y at time t. This embedding in Euclidean m-dimensional space thus reflects diffusion distances on the original data, which are intrinsic geometric properties of the data.

Figure 3 shows a simple example of such an embedding. Each data point is a small (100 × 100-pixel) image of the symbol 3D; the correlation between them is just the dot product in dimension 10,000, and the affinity matrix A is defined as above. The first two eigenfunctions organize the small images, which were provided in random order in the assembly of the 3D puzzle. The eigenvectors integrate and piece together the information measured through the local similarities between spatially nearby patches.

The second approach leads to a generalized wavelet analysis, based on so-called diffusion wavelets, which is associated with and tuned to the random walk P described earlier. The approach shares some features with classic multiscale and wavelet analysis, and it allows generalization of classic signal processing techniques—for, say, compression, denoising, fitting, and regularization—to functions on data sets.

We can show that these two seemingly different approaches are mathematically related, much as Fourier and wavelet analysis are related in Euclidean spaces. The multiscale construction leads to basis functions that are hierarchically organized according to diffusion distances at different scales. The eigenvectors are global functions on the data that "integrate" precisely the local "infinitesimal" affinity geometry.

Techniques based on these ideas of diffusion on data sets have led to machine-learning algorithms that perform at state-of-the-art levels or better on standard community benchmarks. We refer interested readers to "Regularization on Graphs with Function-adapted Diffusion Processes," A.D. Szlam, M. Maggioni, and R.R. Coifman, *Journal of Machine Learning Research*, 9 (2008), 1711–1739, and the references therein.

In conclusion, we see emerging a "signal processing toolbox" for digital data as a first step in the development of methods for analyzing the geometry of large data sets in high-dimensional space and functions defined on such data sets. Among the numerous problems and applications are multi-dimensional document rankings, extension of the Google ranking algorithm, information navigation, heterogeneous material modeling, and multiscale complex structure organization.

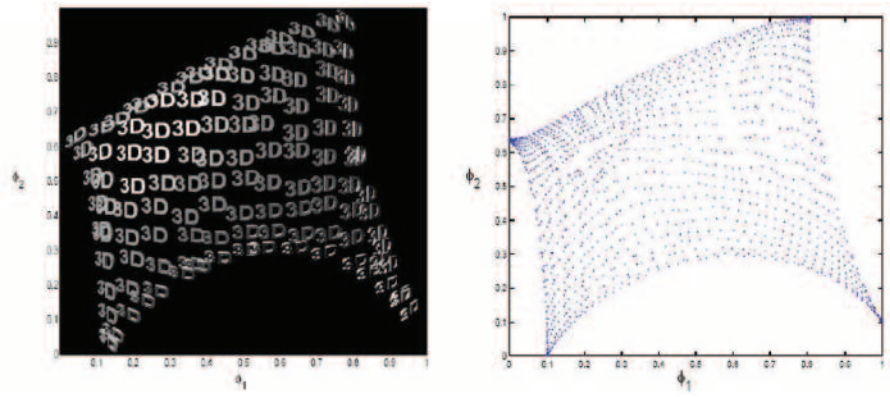

Figure 3. *Embedding of a set of images of the symbol 3D under different rotation angles and illumination. Courtesy of Stéphane Lafon.*

The ideas described in this short article are strongly related to nonlinear principal component analysis, kernel methods, spectral graph embedding, and many other techniques lying at the intersection of various branches of mathematics, computer science, and engineering. They have been documented in literally hundreds of papers by researchers from various communities. A simple description of these and other ideas from a "diffusion geometry perspective" can be found in the July 2006 issue of *Applied and Computational Harmonic Analysis*, and the references therein.

Ronald Coifman is the Phillips Professor of Mathematics and a professor of computer science at Yale University. Mauro Maggioni is an assistant professor of mathematics at Duke University.