Visualization Viewpoints

Editors: Theresa-Marie Rhyne and Lloyd Treinish

Data Signatures and Visualization of Scientific Data Sets

Pak Chung Wong, Harlan Foote, Ruby Leung, Dan Adams, and Jim Thomas

Pacific Northwest National Laboratory

Today, as data sets used in computations grow in size and complexity, the technologies developed over the years to deal with scientific data sets have become less efficient and effective. Many frequently used operations, such as Eigenvector computation, could quickly exhaust our desktop workstations once the data size reaches certain limits.

On the other hand, the high-dimensional data sets we collect every day don't relieve the problem. Many conventional metric designs that build on quantitative or categorical data sets cannot be applied directly to heterogeneous data sets with multiple data types.

While building new machines with more resources might conquer the data size problems, the complexity of today's computations requires a new breed of projection techniques to support analysis of the data and verification of the results. We introduce the concept of a data signature, which captures the essence of a scientific data set in a compact format, and use it to conduct analysis as if using the original. A time-dependent climate simulation data set demonstrates our approach and presents the results.

Background

In 1995, scientists at the Pacific Northwest National Laboratory (PNNL, on the Web at http://www.pnl.gov) had a challenging task: to analyze hundreds of thousands of unstructured text articles interactively on a desktop workstation. The solution—a system called Spire (Spatial Paradigm for Information Retrieval and Exploration)—has become one of the most powerful text analysis systems developed to date. (*R&D* magazine recognized it with an *R&D* 100 Award in 1996.) Among all the core technologies developed for this project, implementation of the document vectors, which represent individual topics of a corpus, plays a critical role in the system's success.

Because of the extremely compact design of the document vectors, many powerful—but potentially expensive—analysis techniques now can be applied to huge amounts of text data. Today we can interactively analyze more than half a million news articles, study their time trends, review topic correlation, and read the original text, all on a desktop workstation such as a Sun Ultra 10.

Figure 1 shows a visualization of a corpus with more than 60,000 medical research articles collected in 1997. We first project the corpus into individual document vectors before generating the terrain visualization using scaling and other analysis techniques. Refer to an earlier article[1] or to http://www.pnl.gov/infoviz on the Web for details of this visualization and the other interactive features the system provides.

Data signature

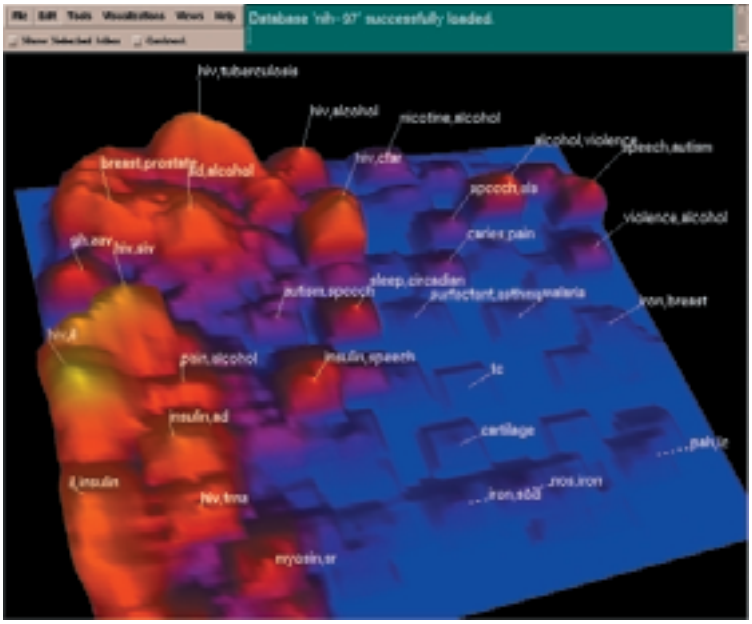
Our research on information abstraction is neither static nor complete. The idea of document vectors has since evolved into the powerful concept of a data signature that represents the content within the context of scientific data sets. In scientific computations such as climate and combustion simulations and modeling, we encounter large data sets with up to tens of gigabytes of data recorded per time step. Many conventional analysis techniques are hopelessly ineffective when faced with this much data, and the development of new tools seemingly lags behind. The data signature concept represents one promising approach to analyzing and understanding scientific data sets.

A data signature can be described as a mathematical data vector that captures the essence of a large data set in a small fraction of its original size. It's designed to characterize a portion of a data set, such as an individual timeframe of a scientific simulation or an article within a corpus. These signatures enable us to conduct analysis at a higher level of abstraction and yet still reflect the intended results as if using the original data. For example, we can now measure the dissimilarity between two text articles by computing the difference between the two corresponding signatures and return a quantitative answer.

We have so far investigated designing data signatures for text, scalar fields, tensor fields, and a combination of these for data sets with multiple parameters. Our design is flexible enough to process both scalar and tensor fields, and project them into one numerical signature. The construction of a data signature is based on one or more of the following features and approaches:

- Velocity gradient tensors (Jacobians)
- Critical points and their Eigenvalues
- Orthogonal and nonorthogonal edges
- Covariance matrices
- Intensity histograms
- Content segmentation
- Conditional probability

0272-1716/00/$10.00 © 2000 IEEE



1 The topics or themes within a corpus are shown as a relief map of natural terrain. The mountains indicate dominant themes. Both the height and the color of the mountains represent the amount of thematic content of documents. The peak labels use the terms that contribute most to the thematic content in that area.

The choice of feature selections often depends on the data type. In general, information such as velocity gradient and critical points proves most suitable for vector and tensor field data. Edge detection, covariance, and histograms can be applied to scalar data. Probability and segmentation can be used on scalars, vectors, and text documents.

An average article usually requires 200 to 400 numbers in its signature to fully describe the contents. For large scientific simulations, we sometimes rely on a multiresolution approach to determine the desirable size of the signatures. Determining the sufficient number of features in a signature also depends on the investigation's goal. For example, the goal of our demonstration example here is to select features sensitive to precipitation. Therefore, we emphasize features derived from moisture advection—strongly correlated with precipitation—instead of other parameters such as temperature.

Application

In theory, a data signature can never be as descriptive as its full-size original. In practice, however, a well-defined data signature can be as good or even better from many perspectives because it brings out specific details and eliminates less important information from consideration. The concept is particularly useful when we study the characteristics of scientific simulations such as global climate modeling.

Although atmospheric data sets are inherently multidimensional because of their 3D time varying characteristics and the complex relationships among their variables, conventional analysis in climate modeling often focuses on individual variables and their spatial or temporal means.[2] The selection of these variables is usually based on a priori knowledge of the atmospheric phenomena to be simulated.

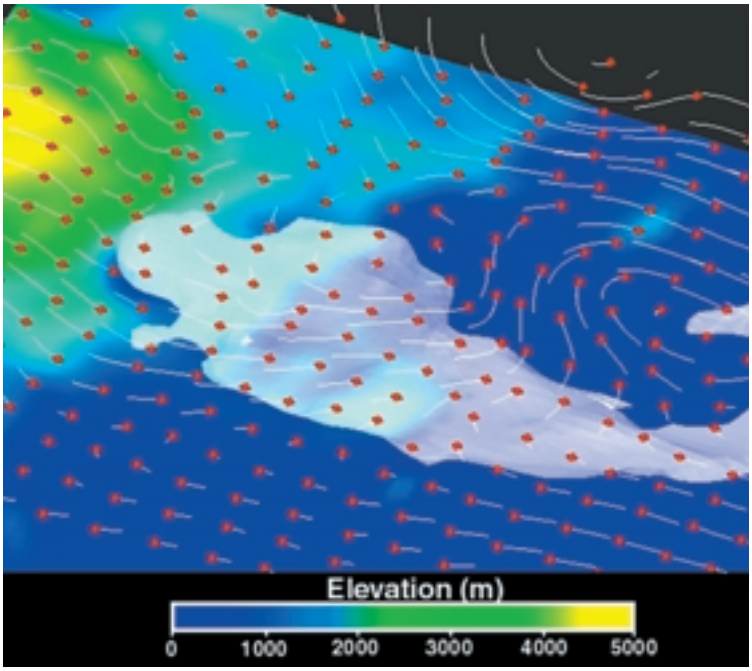
This approach provides a good synopsis of the simulation for us to understand the climate's general features and evaluate the simulation's accuracy. Based on this information, we can then perform fine analysis of certain time periods and variables. Our previous work[3] indicates that a low-dimensional overview, such as a multidimensional scatterplot generated by metric scaling, is an ideal tool to support these analyses. The scaling process, however, is intrinsically expensive, especially with large amounts of data, unless we use a smaller data signature to represent the original data in the calculation.

PNNL climate data modeling

We demonstrate our design using a climate modeling data set[3] provided by PNNL's Global Change program (http://www.pnl.gov/atmos_sciences/as_clim.html). The basic multivariate time-dependent data set, though smaller and relatively simple, has all the characteristics and features for our initial prototyping and experiments. The data set has five data variables (pressure, temperature, water-vapor mixing ratio, and two wind-velocity components) of different types (scalars and vectors) and dimensions (2D and 3D) recorded daily. Each of these variables contains more than 127,000 floating-point numbers. They add up to more than half a million numbers per time step and over 233 megabytes of data in the modeling data set.

Generating the signatures takes about 17 CPU seconds (about 24 wall-clock seconds) on an SGI O2 workstation. Figure 2 (next page) shows the renderings of the water-vapor ratio and wind-velocity data set of the first time frame (1 May 1995) in a 3D volume. The terrain elevation maps to the accompanying legend. The particle-tracking icons show the wind-velocity fields at the 850-mbar level. The semi-opaque isosurface (isovalue = 0.4 kg/kg m/s) represents the magnitude of the moisture transport.

The data set is a climate simulation of the extreme flood conditions of Eastern Asia (China and Japan) from May to July 1991. During that period, the weather was characterized by three rainfall episodes, the first

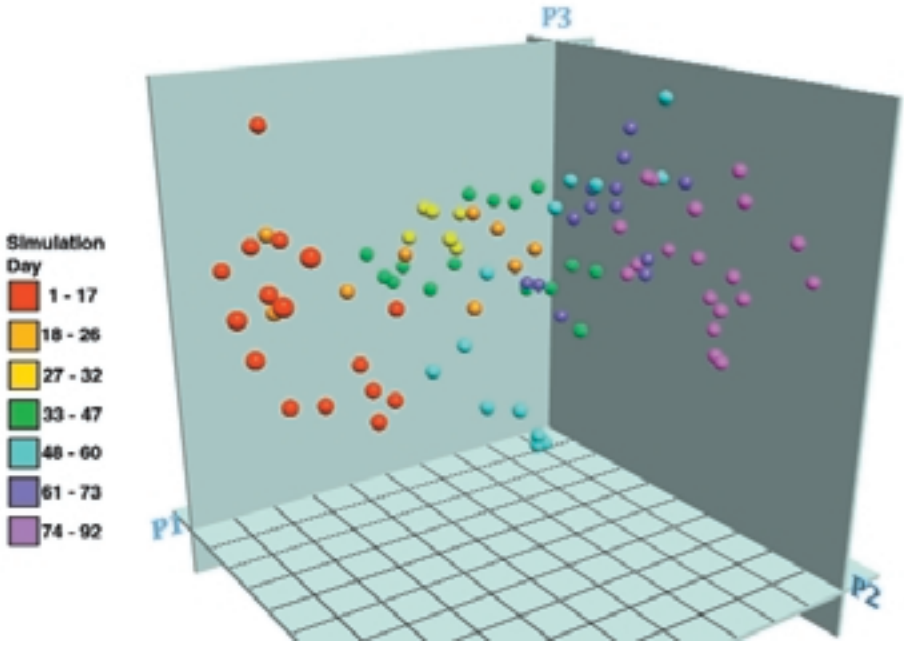


2 The color-coded terrain elevation overlaid with the wind-velocity field at the 850-mbar pressure level. The semi-opaque, centrally located isosurface represents the magnitude of the moisture advection. The particle-tracking icons depict the wind-velocity field.

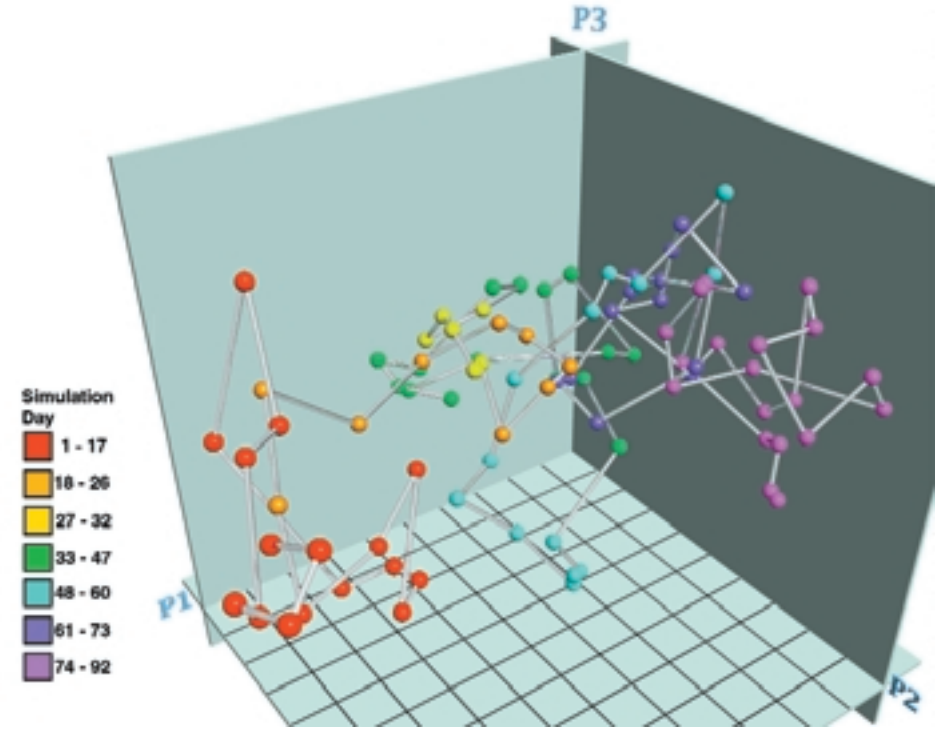
between 18 and 26 May, the second between 2 and 16 June, and the last between 30 June and 13 July. Together, the three episodes of the unusually early and long rainy season clearly separate the summer into seven periods. These characteristics, however, don't show in Figure 2 because the simulation starts with a relatively calm day in the region.

We first project the water-vapor mixing ratio (scalar) and the wind-velocity (vector) variables of each time frame into a single data signature. We then create a symmetric matrix by measuring the similarities (or dissimilarities) of the signatures representing individual time frames. Figure 3 depicts a 3D scatterplot generated by the three most significant principal components of the similarity matrix, P1, P2, and P3.

Each data signature (or its corresponding time frame) of the climate simulation is represented by a rainbow color icon in the 3D scatterplot, with the red icons showing the first period and the purple the last period. We selected the rainbow approach partly because we know a priori that the data set includes only seven distinctive time periods. Otherwise, an isoluminant, segmented



3 The three axes represent the three most significant principal components of the similarity matrix. The seven time periods are shown in red (1 through 17), orange (18 through 26), yellow (27 through 32), green (33 through 47), cyan (48 through 60), blue (61 through 73), and purple (74 through 92).



4 **An enhanced version of Figure 3 with a different camera viewpoint and extra lines connecting the colored spheres to show the ordering of the simulation.**

color map would be ideal to depict the data. In the 3D scatterplot, as we can see, the simulation starts from the left (red and orange), across the middle (yellow, green, and cyan), and stops at the right (blue and purple).

Figure 4 shows an enhanced version of the scatterplot with a different camera viewpoint and extra lines connecting the colored spheres in the order they appeared in the simulation. Although the line icons create partial occlusions, they show the exact ordering of the simulation, which is missing in Figure 3.

The 3D scatterplot based on the water-vapor mixing ratio measurements and wind directions correctly reflects the trend, history, and similarity of characteristics from different time periods of the simulation. Interesting or unexpected characteristics of the simulation detected at this level will draw our attention to particular time intervals and physical interactions.

Conclusions

Data signatures let us combine a multivariate scientific data set into a single quantitative data vector for scientific computation and analysis. Our design is flexible enough to unify different data types, including those to which it is seemingly difficult to apply any scale measures. The much smaller data signatures make good substitutes for their original large data sets in scientific computations such as climate modeling.

Our preliminary results indicate that this approach shows promise in simplifying the analysis and understanding of large data sets. It's still under active development, including addressing the problem of task-oriented feature selection. Therefore, we welcome ideas on extensions and improvements to our current methodology. ■

Acknowledgments

The Pacific Northwest National Laboratory is managed for the US Department of Energy by Battelle Memorial Institute under contract DE-AC06-76R1-1830. Special thanks go to Lloyd Treinish and Theresa-Marie Rhyne, who provided valuable suggestions to improve our work. Thanks also to Ray Bair, Jackie Chen (Sandia National Laboratory [SNL]), Kris Cook, Dave Dixon, Thom Dunning, Sharon Eaton, George Fann, Habib Najm (SNL), Marty Peterson, Larry Rahn (SNL), and Paul Whitney, who provided assistance of many forms throughout this research.

References

1. J. Wise et al., "Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Documents," *Proc. IEEE Information Visualization 95*, IEEE Computer Society Press, Los Alamitos, Calif., 1995, pp. 51-58.
2. L.R. Leung et al., "Intercomparison of Regional Climate Simulations of the 1991 Summer Monsoon in Eastern Asia," *J. of Geophysical Research*, Vol. 104, No. 6, Mar. 1999, pp. 6425-6454.
3. P.C. Wong and R.D. Bergeron, "Multivariate Visualization Using Metric Scaling," *Proc. IEEE Visualization 97*, ACM Press, New York, NY, 1997, pp. 111-118.

Readers may contact the authors by e-mail: pak.wong@pnl.gov, harlan.foote@pnl.gov, ruby.leung@pnl.gov, dan@pnl.gov, and jim.thomas@pnl.gov.

Contact department editors Rhyne and Treinish by e-mail at theresa.rhyne@epa.gov and lloydt@us.ibm.com.