Advanced Scientific Computing Research
Computer Science
FY 2007 Accomplishment

Variable Interactions in Query-Driven Visualization

E. Wes Bethel[1], Luke Gosink[2], John C. Anderson[2], Kenneth I. Joy[2]
[1]Lawrence Berkeley National Laboratory; [2]University of California, Davis

Summary

One fundamental element of scientific inquiry is discovering relationships, particularly the interactions between different variables in observed or simulated phenomena. Building upon our prior work in the field of Query-Driven Visualization, where visual data analysis processing is focused on subsets of large data deemed to be "scientifically interesting," this new work focuses on a novel knowledge discovery capability suitable for use with petascale class datasets. It enables visual presentation of the presence or absence of relationships (correlations) between variables in data subsets produced by Query-Driven methodologies. This technique holds great potential for enabling knowledge discovery from large and complex datasets currently emerging from SciDAC and INCITE projects. It is sufficiently generally to be applicable to any time of complex, time-varying, multivariate data from structured, unstructured or adaptive grids.

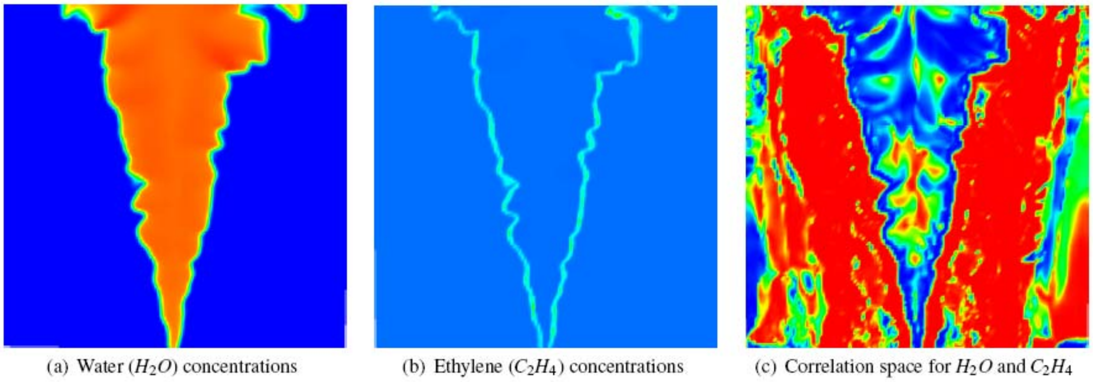
(a) Water (H_2O) concentrations
(b) Ethylene (C_2H_4) concentrations
(c) Correlation space for H_2O and C_2H_4

Figure 1. Water (H_2O) and ethylene (C_2H_4) concentrations from the methane combustion dataset are shown in (a) and (b), respectively (Sample data courtesy J. Bell and M. Day, Center for Computational Sciences and Engineering, LBNL). The derived correlation field for these two compounds is shown in (c). The switch from strong positive correlation to strong negative correlation in the reaction region corresponds to the area in which C_2H_4 is both produced and consumed, and H_2O is produced, in the process of combustion. The strong correlation (both positive and negative) in the center of the flame, as well as the atmospheric region, demonstrates the correlation field's ability to show fine-scale interactions.

Obstacles hindering scientific knowledge discovery from large and complex data may be broadly categorized into two separate but overlapping groups. The first category, concerned mainly with issues of throughput, includes the challenges inherent to efficiently managing and visualizing large-scale datasets. The second category includes the difficulties associated with attaining insight from datasets of high-complexity.

Query-driven visualization (QDV) is well suited for performing analysis and visualization on datasets that are both large and highly complex. Tools like FastBit leverage highly efficient (in terms of speed and compression) data management techniques to rapidly identify and visualize "regions of interest" within a dataset. Specified as Boolean range queries, these regions of interest tend to be significantly smaller subsets of the original dataset; thus, these regions require

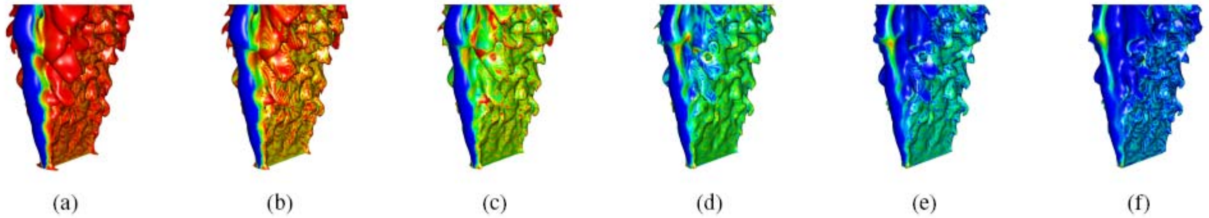

Figure 2. These images depict increase ((a) through (f)) isosurface values of temperature (isotherms) colored by values of the correlation field derived from water and ethylene. As temperature values increase, the predominant correlation between water and ethylene along the isotherms shifts away from strongly positive (red in (a)) to strongly negative (blue in (f)). This shift suggests that temperature is itself negatively correlated with the water-ethylene correlation. (Sample methane flame simulation data courtesy J. Bell and M. Day, Center for Computational Sciencs and Engineering, LBNL)

less time and effort to analyze, visualize, and interpret.

Well-characterized range queries are capable of identifying spatial regions where many domain-specific events occur: combustion flame fronts, vortices, chemical reaction fronts, etc. Beyond indicating these regions, however, queries reveal little about variable interactions or complex trends that lie in the domain of these characterizations. In such regions of interest, it is the behavioral trends *between* variables, or groups of variables, that are more important in providing insight than the traits or locations of individual variables alone. The challenge is to extend the strengths of QDV with methods that identify behavioral trends and provide insight into regions of interest through coherent and meaningful visualizations.

The novel contributions of this work are techniques that extend the capabilities of QDV by providing the basis for determining: how sets of variables in complex datasets interact throughout regions of interest, and the role other variables play in influencing these interactions.

We utilize the cumulative distribution functions (CDFs) of all variables in a query to reveal initial information about statistical regions of interest within the query's solution space. The CDF for each variable is computed by integrating over the query's solution space, then accumulating the variable's values as a histogram. Statistically, the solution set of a query is represented as an aggregate of histograms, one histogram for each variable expressed in the query.

We extend this analysis further by incorporating correlation fields, which provide insight into localized correlation between any two variables. By mapping a correlation field onto a third variable's isosurfaces (specifically, the statistically important isovalues suggested by the variable's CDF), statistically important interactions between any three variables in a dataset are readily visualized, allowing for trends between variables in a user's query to be identified.

In this method, CDFs and correlation fields are constrained to the query's solution space. By working exclusively in the query's solution space, this method takes full advantage of the performance benefits inherent to QDV strategies. Specifically, computational efforts are only focused on regions that have been rapidly identified (via a query engine) as "interesting" by the user's query. This method's integrated analysis extends current query solutions by revealing statistical trends of interactivity (i.e., dependency and independence) between any triad of variables in the solution space of the query.

Publications

L. Gosink, J. Anderson, E. W. Bethel, K. Joy. "Variable Interactions in Query-Driven Visualization." *Transactions on Visualization and Computer Graphics/IEEE Visualization 2007* (Accepted for publication.) LBNL-63254.

For further information on this subject contact:

Name: E. Wes Bethel.
Organization: Lawrence Berkeley National Laboratory.
Email: ewbethel@lbl.gov
Phone: (510) 486-7353

Name: Kenneth I. Joy
Organization: University of California, Davis
Email: joy@cs.ucdavis.edu
Phone: (530) 752-1077

Note: This work was funded in part by the SciDAC2 Visualization and Analytics Center for Enabling Technologies and ASCR's Visualization Base Program by the Director, Office of Science, Office of Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.