Multivariate Temporal Features in Scientific Data

Venue

IEEE Visweek 2009 Tutorial

Duration

The tutorial is half-day, including 180 minutes presentation and discussion, and 30 minutes coffee break.

Organizers

Jian Huang	University of Tennessee, Knoxville huangj@eecs.utk.edu
Chaoli Wang	Michigan Technological University chaoliw@mtu.edu
Heike Jänicke	Swansea University csheike@swansea.ac.uk
Jonathan Woodring	Los Alamos National Laboratory woodring@lanl.gov

Abstract

A wave of large datasets amounting beyond terascale is now being produced by scientific applications on a daily basis. The ensuing challenge to manage and make sense of these data demands systematic breakthroughs in several areas of computer science. In this tutorial, we survey recent progresses made in addressing a particular difficulty of pressing user need. That is, the gap between users' conceptual domain knowledge *vs.* the way features have to be specified in traditional scientific visualizations. This gap is particularly acute at terascale and beyond, where a large amount of parallel automation is necessary to study the data at full-scale. Interactive techniques alone cannot solve the whole problem. Algorithmic methods must be studied.

Recent research to visualize time-varying multivariate data has led to the study of several key technical hurdles. For instance, when visualization is used for exploratory purposes, it could be the case that even application scientists themselves do not fully understand the phenomena and are using visualization to start and refine their research. There are also general cases where a user has clear concepts of a pattern but the knowledge lacks specifics, as exemplified by the concept of "the start of a growing season" that is commonly studied in climate modeling. When a feature can be defined with rigor, there is still a common situation that the data cannot be viewed altogether, and that straightforward ways to generate hierarchical representations of the features do not lead to satisfactory results.

In an effort to survey recent progress in addressing the above set of diverse challenges, our tutorial covers the following topics: (i) Programming language interfaces for time-varying multivariate visualization; (ii) Purely mathematical ways to specify features for visualization; (iii) Importance-driven data analysis and visualization; (iv) Chronovolumes, comparative visualization, and time-varying transfer functions. The tutorial also demonstrates the applications of these techniques in highly visible recent application areas, such as modeling and simulation of climate, combustion, astrophysics, earthquake and hurricane. The goal of this tutorial is to inform visualization researchers and practitioners the state-of-the-art technologies that have greatly enriched the toolset for visualizing large-scale time-varying multivariate scientific data.

Level

Beginner/Intermediate.

Prerequisite

An intermediate knowledge level of volume visualization is required. Some prior experience with a time-varying multivariate simulation would be very helpful.

Audience

The intended audience includes visualization researchers and practitioners who are interested in learning about an enriched set of feature specification in time-varying volumetric data of multiple variables.

Importance

Time-varying multivariate data visualization has received increasing attention over the years. Most scientific datasets are in the form of time-varying, multivariate. As the size and complexity of data increase, there is a pressing need to design new algorithms and techniques that extract, organize and visualize multivariate temporal features in scientific data. In the recommendations from the DOE/ASCR 2007 Workshop on Visual Analysis and Data Exploration at Extreme Scale, visualization experts around this country listed *Multifield and Multimodel Data Understanding* and *Time-Varying Datasets* among the key research areas for visual analysis and knowledge discovery. As such, new visualization techniques and user interfaces must enable the users to understand extreme-scale, time-varying, multivariate datasets by interactively browsing through different spatial and temporal scales, identifying scientific phenomena of different temporal length, and tracking salient features in both time and space.

Since 2006, the IEEE Visualization Conference also has dedicated sections on multivariate data visualization. However, for the past ten years (1999-2008), there was no tutorials at this conference that focused on time-varying multivariate data visualization. We believe that our proposed tutorial answers this timely call, informing the audience of recent advances in time-varying multivariate data visualization towards effective analysis, viewing and understanding of data.

Schedule

Introduction	All	10 minutes
Talk1	Jian Huang	40 minutes
Talk2	Heike Jänicke	40 minutes
Break	—	30 minutes
Talk3	Chaoli Wang	40 minutes
Talk4	Jonathan Woodring	40 minutes
Discussion	All	10 minutes

Description

The description and outline of each topic is in the following:

Programming Language Interfaces for Time-Varying Multivariate Visualization

Jian Huang

Abstract

Recent advances in the field of query-based visualization have enabled a powerful technique for large data visualization. Using query-based visualization, users can quickly and intuitively explore very large, highly complex multivariate data. But although human-domain knowledge is always a driving factor, query-based visualization is commonly guided by trial and error and is often ad hoc. In addition, ad hoc approach becomes extremely cumbersome and time consuming when the exponentially large growth of multivariate time-varying simulation data is considered. In this domain, more efficient user tools with versatile automation are necessary.

Motivated by the elegance and power of domain specific programming language interfaces, we have developed two query language interfaces for visualizing features in large, multi-variate, time-variant datasets. The first was developed to enable the user to express features in terms of statistical properties of local spatio-temporal neighborhoods. Clauses on relative terms can be used to avoid requiring overly detailed, and hence overly constrained, specifications. The second is a temporal pattern language directly modeled after *regex* (regular expression). Wildcards provide a powerful way to represent the existence of non-topical events, whether temporal or multivariate. Queries containing wildcards are automatically expanded to a number of actual range queries that can be extracted concurrently in order for a systematic exploration of the data. Both languages are tightly integrated into the visual analysis process. We apply these language frameworks to a recent climate modeling simulation and describe a mechanism making query processing scalable and of low latency.

Outline

- Introduction
  - Brief motivation from the perspective of climate modeling research
- Related Work
  - Query-driven visualization
- Spatial Querying
  - Neighborhood-based querying
  - Bin selection and predicate clauses
- Temporal Querying
  - Textual pattern matching
  - Syntax, interpreter and scalable query servers
- Scalable Back-End Servers

Contribution

This segment of the tutorial surveys the latest research trends of query-driven visualization. A few known methods to build a scalable server to enable high-throughput queries in datasets are also discussed in this talk.

Information-theoretic Methods for the Visual Analysis of Climate and Flow Data

Heike Jänicke

Abstract

The amount of scientific data is doubling every year. Due to the general availability of high-performance computers, not only the number of performed simulations and data records increases, but also the size of each data set. Scientists generate data at finer resolutions in space and time, and employ many different coupled variables resulting in data set sizes of several giga- or terabytes (1,000 gigabytes). Hence, compression techniques are required to reduce the amount of data to the essential and support scientists in getting the big picture.

Information theory is a powerful tool to automatically achieve such high compression, while retaining the crucial information in the data. Different abstraction layers have been proposed to define features in a purely mathematical way: A core technique is local statistical complexity which identifies those regions in a data set that locally contain a lot of information about the data set's fluid dynamics. To allow for more insight into the local dynamics, causal states partition the domain into different classes of behavior. They encode the local dynamics of unsteady flow and give a cause and effect relationship between a positions past and future. Summarizing all causal states in a data set and analyzing the transitions between them, we resolve an ϵ-machine that can be represented by a directed graph. ϵ-Machines are a static representation of the temporal dynamics in a time-dependent data set and allow for the identification of different patterns present in the data. The feature level is reached when incorporating additional information into the ϵ-machine to guide the user's attention. Such methods allow, for example, for the investigation of the temporal evolution of the flow and for automatic feature tracking based on the encoded local dynamics. As features can be either defined on a per position or per particle basis, we show how the information-theoretic concepts are applied to Eulerian and Lagrangian flows in order to detect relevant structures.

Outline

- Introduction
- Related Work
- Information-theoretic Tools for Data Compression and Analysis
  - Causal states
  - Local statistical complexity
  - ϵ-Machines
  - Complexity measures in Eulerian and Lagrangian flows
- Mathematical Feature Specification and Analysis

Contribution

All techniques that will be presented in this part of tutorial have been successfully applied to different application domains such as fluid flow and climate simulation. Using the automatic techniques, we could identify well known important structures as well as interesting new formations in a purely mathematic way.

Importance-Driven Data Analysis and Visualization

Chaoli Wang

Abstract
As we have reached a stage where data easily overwhelm the limits of human comprehension, it is imperative to identify and present the most relevant information. This part of tutorial focus on new techniques that identify data features and highlight important regions for cost-effective data evaluation, viewing and understanding.

First, I will present a reduced-reference approach to volume data quality assessment which extracts important statistical information from the original data in the multi-scale wavelet domain. Using only the extracted information as "carry-on" features, we are able to identify and quantify the quality loss in reduced or distorted versions of data. Second, I will introduce an importance-driven solution for uncovering the dynamic behaviors of time-varying data. This is achieved by analyzing the importance of data blocks in the joint feature-temporal space using the concepts from information theory. The analysis yields an importance curve for each block characterizing its temporal behavior. Clustering all importance curves effectively classifies the time-varying data for importance-driven visualization. Finally, the concept of importance can also be drawn from domain knowledge and used in data reduction. I will describe an algorithm that utilizes a reference feature to partition and compress time-varying data.

Outline

- Introduction and Motivation
- Volume Data Quality Assessment
  - Wavelet hierarchy and subband statistics
  - Feature representation
  - Quality assessment
- Importance-Driven Visualization
  - Entropy-based data importance
  - Importance-driven techniques
- Application-Driven Compression
  - Reference features and data prioritization
  - Compression, decompression and visualization

Contribution
This part of the tutorial highlights how effective data viewing and understanding can be achieved by leveraging the important aspect of data. The solutions presented point out a promising direction for coping with the large data problem facing computational scientists and visualization researchers.

Comparative Visualization and Transfer Functions for Time-Varying Data

Jonathan Woodring

Abstract
To visualize and analyze the time-varying data, several techniques are introduced that move beyond animations and side by side comparison to reason about temporal trends, due to human perception limitations such as short-term visual memory, change blindness, and the inability to make precise quantified reasoning. Visualization of time-varying data involves classification (transfer functions) to be able to specify and find spatio-temporal patterns in the data. To generate transfer functions for time-varying data, similar temporal trends are used for data classification, which is realized through classification via time histograms and classification via analysis of temporal trends. Visualization presentation techniques are introduced that can aid in the visualization of time-varying data by comparative analysis. Multi-scale temporal spreadsheet views aid in the comparison between time steps by quantifying value differences over time at different time scales. Chronovolumes and 3D composition visualize time-varying data in such a way that it can precisely describe spatio-temporal trends in a single image through comparative analysis.

Outline

- Introduction and Motivation
- Comparative Visualization for Time
  - Chronovolumes and composition
  - Multi-scale temporal spreadsheet
- Time-Varying Transfer Functions
  - Time histograms
  - Trend clustering and sequencing

Contribution
This segment of the tutorial discusses methods for time-varying transfer function generation and the use comparative visualization in time-varying data. Transfer function generation in general is a difficult problem, and even more so for time-varying data. We describe two methods for time-varying data transfer function specification: a manual and semi-automatic method through time histograms and a semi-automatic method via analysis of the data. Second, animations and side-by-side comparison of frames is lacking for being able to quantify and describe similarities and differences that happen over time in data. We describe several comparative methods that makes it easier for a user to precisely see what is changing over time through: multi-scale filtering, quantification of values in a spreadsheet view, and a comparative rendering technique that compresses multiple time steps into one dataset.

Tutorial Notes

The tutorial notes will consist of the description of the tutorial, copies of the slides for each talk, and an extensive bibliography including specific references used in the tutorial as well as a general selection of relevant references.

Instructors

The background of each instructor is listed in alphabetical order.

Jian Huang
University of Tennessee, Knoxvile

Jian Huang is an associate professor in the EECS Department at the University of Tennessee, Knoxville. His research focuses on large data visualization, multivariate visualization, and parallel, distributed and remote visualization. He received a BEng in electrical engineering from the Nanjing University of Posts and Telecom, China in 1996. During his subsequent graduate study at the Ohio State University, he was awarded an MS degree in computer science and an MS degree in biomedical engineering, both in 1998, and a PhD degree in computer science in 2001. Dr. Huang's research has been funded by NSF and DOE. Dr. Huang is a recipient of DOE Early Career Principal Investigator Award in 2004.

Relevant Publications

Wesley Kendall, Markus Glatter, Jian Huang, Tom Peterka, Robert Latham and Robert Ross, Terascale Data Organization for Discovering Multivariate Climatic Trends, *SC'09*, November 2009, Portland, OR.

C. Ryan Johnson and Jian Huang, Distribution Driven Visualization of Volume Data, *IEEE Transactions on Visualization and Computer Graphics*, 15(5):734-746, 2009.

Markus Glatter, Jian Huang, Sean Ahern, Jamison Daniel, and Aidong Lu, Visualizing Temporal Patterns in Large Multivariate Data using Textual Pattern Matching, *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1467-1474, 2008.

Markus Glatter, Colin Mollenhour, Jian Huang, and Jinzhu Gao, Scalable Data Servers for Large Multivariate Volume Visualization, *IEEE Transactions on Visualization and Computer Graphics*, 12(5):1291-1299, 2006.

Chaoli Wang
Michigan Technological University

Chaoli Wang is an assistant professor at the Department of Computer Science, Michigan Technological University. He received the PhD degree in computer science from The Ohio State University in 2006. From 2007 to 2009, he was a postdoctoral researcher at University of California, Davis. His research focuses on large-scale data analysis and visualization, high-performance computing, and user interface and interaction.

Relevant Publications

Chaoli Wang, Hongfeng Yu, and Kwan-Liu Ma, Application-Driven Compression for Visualizing Large-Scale Time-Varying Volume Data, *IEEE Computer Graphics and Applications*, to appear, 2009.

Chaoli Wang, Hongfeng Yu, and Kwan-Liu Ma, Importance-Driven Time-Varying Data Visualization, *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1547-1554, 2008.

Chaoli Wang and Kwan-Liu Ma, A Statistical Approach to Volume Data Quality Assessment, *IEEE Transactions on Visualization and Computer Graphics*, 14(3):590-602, 2008.

Heike Jänicke
Swansea University

Heike Jänicke is currently a postdoctoral researcher in the Visual and Interactive Computing group, Swansea University, Swansea, Wales, UK. She received the MSc (Diplom) degree in computer science specialising in medical computer science from Universität Leipzig, Leipzig, Germany, in 2006. For her research of information-theoretic methods in visualization, she was awarded a PhD degree in 2009 from the same university. Her research interests include methods from statistics and information theory, analysis of structure and self-organization, data mining, visualization of unsteady, multivariate data, and visualization of flow, weather and climate data.

Relevant Publications

Heike Jänicke and Gerik Scheuermann, Steady Visualization of the Dynamics in Fluids Using $\epsilon - Machines$, *Accepted for publication in the Special Issue on Knowledge-assisted Visualization in Computers and Graphics*, 2009.

Heike Jänicke, Information Theoretic Methods for the Visual Analysis of Climate and Flow Data, *PhD thesis, University of Leipzig, Germany*, 2009.

Heike Jänicke, Michael Böttinger, Xavier Tricoche, and Gerik Scheuermann, Automatic Detection and Visualization of Distinctive Structures in 3D Unsteady Multi-fields, *Computer Graphics Forum*, 27(3):767-774, 2008.

Heike Jänicke, Alexander Wiebel, Gerik Scheuermann, and Wolfgang Kollmann, Multifield Visualization Using Local Statistical Complexity, *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1384-1391, 2007.

Jonathan Woodring
Los Alamos National Laboratory

Jonathan Woodring is a Ph.D. graduate from The Ohio State University, completing his degree in June 2009. He currently works at the Los Alamos National Laboratory for the visualization team in the high performance computing research division. His current research focuses on ParaView/VTK, open source software and science, visualization on high performance computing, data intensive computing, comparative visualization, distance visualization, and information visualization.

Relevant Publications

Jonathan Woodring and Han-Wei Shen, Semi-Automatic Time-Series Transfer Functions via Temporal Clustering and Sequencing, In *Proceedings of Eurographics/IEEE VGTC Symposium on Visualization (Eurovis'09)*, 2009.

Jonathan Woodring and Han-Wei Shen, Multiscale Time Activity Data Exploration via Temporal Clustering Visualization Spreadsheet, *IEEE Transactions on Visualization and Computer Graphics*, 15(1):123-137, 2009.

Jonathan Woodring and Han-Wei Shen, Multi-variate, Time Varying, and Comparative Visualization with Contextual Cues, *IEEE Transactions on Visualization and Computer Graphics*, 12(5):909-916, 2006.

Jonathan Woodring, Chaoli Wang, and Han-Wei Shen, High Dimensional Direct Rendering of Time-Varying Volumetric Data, In *Proceedings of IEEE Visualization Conference*, pp. 417-424, 2003.