Managing the Evolution of Dataflows with VisTrails

Juliana Freire
http://www.cs.utah.edu/~juliana
University of Utah

Joint work with:
Steven P. Callahan, Emanuele Santos,
Carlos E. Scheidegger, Claudio T. Silva and Huy T. Vo
Data Exploration through Visualization

- Hard to make sense out of large volumes of raw data, e.g., sensor feeds, simulations, MRI scans
- Insightful visualizations help analyze and validate various hypothesis
- But creating a visualization is a complex process
Visualization Systems: State of the Art

♦ Systems: SCIRun, ParaView
♦ Visual programming for creating *visualization pipelines*—dataflows of visualization operations
  – Simplify and automate and the creation of visualizations
♦ Hard to create and compare a *large number* of visualizations
♦ Limitations:
  – No separation between the specification of a dataflow and its instances
  – No provenance tracking mechanism
  – Users need to manage data and metadata

*The generation and maintenance of visualizations is a major bottleneck in the scientific process*
Example: Visualizing Medical Data

Provenance manually maintained: file naming conventions and detailed notes kept.

Hard to differentiate among visualizations: need to inspect files and notes...
VisTrails: Managing Visualizations

- Streamlines the creation, execution and sharing of complex visualizations
  - VisTrails manages the data, scientists can focus on science!
- Infrastructure for large-scale data exploration through visualization
  - Systematic maintenance of visualization provenance: akin to an electronic lab notebook
  - Interactive comparative visualization
- Not a replacement for visualization systems: provides infrastructure that can be combined with and enhance these systems
- Many important applications. Some ongoing collaborations:
  - Harvard Medical School (radiation oncology); OHSU (environmental observation and forecasting systems); UCSD (biomedical informatics)
VisTrails

Evolving dataflow

Link to video:
Action-Based Provenance: Example

<action date="29 Mar 2006 09:22:56" notes="" parent="829" time="830" user="erik" what="changeParameter">
    <set function="AddPoint" functionId="11" moduleId="2" parameter="(unnamed)"
        parameterId="0" type="float" value="1990"/>
    <set function="AddPoint" functionId="11" moduleId="2" parameter="(unnamed)"
        parameterId="1" type="float" value="1"/>
</action>

...
Action-Based Provenance

- Uniformly captures both data and process provenance
- Records user actions—compact representation
- Detailed information about the exploration process
  - Results can be reproduced
  - Scientists can return to any point in the exploration space
- History tree structure enables scalable exploration of the dataflow parameter space through
  - Macros: re-use actions for repetitive tasks
  - Bulk updates: quickly explore slices of parameter space
VisTrails

Macros

Link to video:
http://www.cs.utah.edu/~juliana/talks/videos/vistrails_macros.avi
VisTrails

Bulk updates

VisTrails

Generating animations

Conclusions

- Identified the problem and proposed the first solution for managing fast-evolving workflows
- Detailed data and process provenance automatically captured
- The VisTrails system
  *Replaces the lab notebook*
  
  *Enables large-scale data exploration through visualization*

  *And scientists can do it!*

- Focus on visualization, but ideas are applicable to general workflows
Current and Future Work

◆ Platform for collaborative visualization
  – Distributed synchronization algorithm
◆ XTrails: support for general workflows
  – Support for Web services (BIRN)
  – Execution over the Grid (Chimera)
◆ GUI---better interaction with history
◆ Mine trails—potentially useful information about good visualization strategies
  – Automate generation of visualizations
Acknowledgements

- This work is partially supported by the National Science Foundation (under grants IIS-0513692, CCF-0401498, EIA-0323604, CNS-0514485, IIS-0534628, CNS-0528201, OISE-0405402), the Department of Energy, an IBM Faculty Award, and a University of Utah Seed Grant.

- We thank
  - Dr. George Chen (Harvard Medical School) for the lung datasets;
  - Gordon Kindlmann (SCI) for the brain data set; and
  - The Visible Human Project for the head.
More info about VisTrails

Google vistrails

Or

http://www.sci.utah.edu/~vgc/vistrails/
VisTrails Architecture

- 15-16k lines of python code
  - Easily integrate components
- Re-use existing free software
  - QT, OpenGL, VTK
VisTrails User Interface

VisTrails Builder

VisTrails Spreadsheet

VisTrails Version Tree