
Towards Enabling Social Analysis of Scientific Data

Juliana Freire

School of Computing
University of Utah
Salt Lake City, UT 84112 USA
juliana@cs.utah.edu

Cláudio Silva

SCI Institute
University of Utah
Salt Lake City, UT 84112 USA
csilva@cs.utah.edu

Abstract

Computing has been an enormous accelerator to science and it has led to an information explosion in many different fields. Future advances in science depend on the ability to comprehend these vast amounts of data. In this paper, we discuss challenges and opportunities for social data analysis in the scientific domain.

Keywords

Social data analysis, scientific data, workflows, provenance

ACM Classification Keywords

H3 [Information Storage and Retrieval]: Information Search and Retrieval, Online Information Services; H5.m. [Information interfaces and presentation]: Miscellaneous.

Introduction

Social Web sites and web-based communities (e.g., Flickr, Facebook, Yahoo! Pipes), which facilitate collaboration and sharing between users, are becoming increasingly popular. An important benefit of these sites is that they enable users to leverage the *wisdom of the crowds*. For example, in Flickr, users, in a mass collaboration approach, tag large volumes of pictures. These tags, in turn, help them to more easily find pictures they are looking for. In the (very) recent past, a new class of Web site has emerged that enables users to upload and collectively analyze many types of data (e.g., Many Eyes and Swivel). These are part of a broad phenomenon that has been called “social data analysis”. This trend is expanding to the scientific domain where a number of *collaboratories* are under development. As the cost of hardware decreases over time, the cost of people goes up as analyses get more involved, larger groups need to collaborate, and the volume of data manipulated increases. Science collaboratories aim to bridge this gap by allowing scientists to share, re-use and refine their computational tasks (workflows). In this position paper, we discuss the challenges and key components that are needed to enable the development of effective social data analysis (SDA) sites for the scientific domain.

Challenges and Requirements for Scientific SDA

To analyze and understand scientific data, complex computational processes need to be assembled and insightful visualizations need to be generated, often requiring the combination of loosely-coupled resources, specialized libraries, and grid and Web services. The heterogeneity of the data, its size, and location, greatly complicate the data analysis pipelines. Whereas existing SDA systems require that data be uploaded to a central location for analysis, in many scientific applications that manipulate large volumes of data, this is not feasible. Furthermore, the visualization process is likely to require staged processing and a larger variety of underlying visualization techniques than what is currently supported by systems such as ManyEyes.¹ Data analysis generates more data (e.g., graphs, visualizations) that add to the overflow of information scientists need to deal with. Ad-hoc approaches to data exploration, which are widely used in the scientific community, have serious limitations. In particular, scientists and engineers need to expend substantial effort managing data (e.g., scripts that encode computational tasks, raw data, data products, images, and notes) and to record provenance information² so that basic questions can be answered, such as: Who created a data product and when? When was it modified and by whom? What was the process used to create the data product? Were two data products derived from the same raw data? Not only is the process time-consuming, but also error-prone. The absence of provenance makes it hard (and sometimes

¹ <http://services.alphaworks.ibm.com/manyeyes/home>

² Provenance refers to all information needed to reproduce a certain piece of data. It is also referred to as audit trail, lineage, or pedigree.

impossible) to reproduce and share results, to solve problems collaboratively, to validate results with different input data, to understand the process used to solve a particular problem, and to re-use the knowledge involved in the data analysis process. In addition, it greatly limits the longevity of the data products—without precise and sufficient information about how the data product was generated, its value is greatly diminished. SDA systems aimed at the scientific domain need to provide a flexible framework that not only enables scientists to perform complex analyses over large data, but that also captures detailed provenance of the analysis process [1].

Analysis Pipelines and Provenance: The Basis for Scientific SDA

Shared pipeline and provenance repositories can expose scientists to computational tasks that provide examples of sophisticated uses of tools. They can also uncover common pipeline patterns that can be re-used to solve different problems. For example, given a set of pipeline patterns with high support in a collection, a recommendation system can suggest a series of modules that are most likely to *fit* the pipeline being developed, like an auto-completion for workflows. In addition, by analyzing how people refine pipelines over time, we can potentially re-use these refinements in different tasks as well as learn effective strategies to solve problems. By querying and analyzing the information in these shared repositories, scientists can leverage the *wisdom of the crowds* to learn by example; expedite their scientific training; and potentially reduce their time to insight. But for this to become reality, we need to give the scientists appropriate and usable tools to explore the data in these shared repositories. In early 2005, we started the

VisTrails project in an attempt to address some of these issues.

The VisTrails System. VisTrails is an open-source scientific workflow and provenance management system [2][3][4]. The system was designed to aid users in exploratory computational tasks, such as visualization, data mining, and simulations, which are adjusted in an iterative process. This is in contrast to previous scientific workflow, such a Kepler and Taverna, and workflow-based visualization systems, such as SCIRun and ParaView, which were designed primarily for automating repetitive processes.

With VisTrails, we introduced a series of operations and intuitive interfaces that simplify common tasks in the scientific exploration process: scientists can easily navigate through the space of workflows created for a given exploration task; visually compare workflows and their results; explore large parameter spaces; query and refine workflows by analogy. These operations and interfaces are made possible by VisTrails' change-based provenance model, which uniformly captures changes to parameter values as well as to workflow definitions. Besides the lineage of data products, this model also captures information about how workflows evolve over time---workflows are treated as first-class data products. In exploratory tasks, it is important to understand what the differences between workflows are, especially if multiple people are collaboratively exploring data. Figure 1 shows the visual difference interface provided by VisTrails. A visual difference is enacted by dragging one node in the history tree onto another, which opens a new window with a difference workflow. Modules unique to the first node are shown in orange, modules unique to the second node in blue,

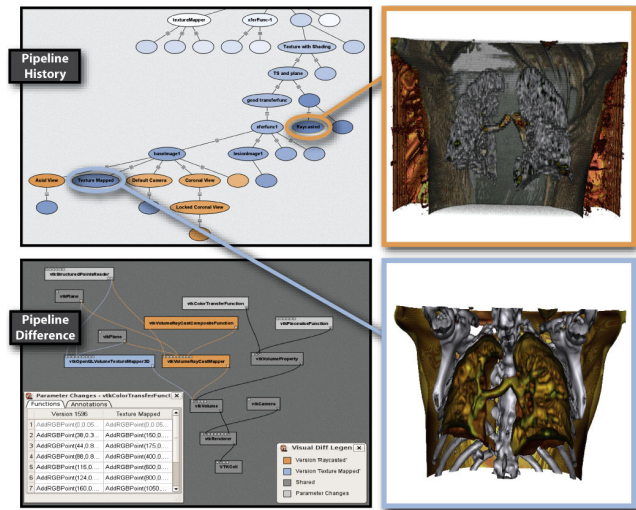


Figure 1. An example of exploratory visualization for radiation treatment planning. Complete provenance of the exploration process is displayed as a history tree with each node representing a workflow that generates a visualization. The visual difference interface allows users to correlate the differences between data products (the images on the right) with the differences between the workflows used to derived the data products.

modules that are the same in dark gray, and modules that have different parameter values in light gray. The ability to compute differences between pipelines can be combined with the analogy mechanism [4]: the difference between two pipelines can be *applied* to a third pipeline, like a patch. This mechanism lets naive users to modify workflows without having to directly edit their definitions, and it has the potential to lower the barrier of adoption for workflow-based systems, which are notoriously hard to use. VisTrails also provides a query-by-example interface which allows users to construct as complex, structure-based queries (e.g., find workflows that resample a data set before

extracting an isosurface) by example, using the same interface used to build workflows [4].

Ongoing and Future Research. In the context of the VisTrails project, we are developing infrastructure to enable social analysis of scientific data.

- Provenance analytics. The problem of mining and extracting knowledge from provenance data has been largely unexplored. By analyzing provenance data, scientists can debug their tasks and obtain a better understanding of their results. Mining this data may also lead to the discovery of patterns that can potentially simplify the notoriously hard, time-consuming process of designing and refining scientific workflows.
- Usable web-enabled interfaces. By weaving services together, it is possible to construct complex applications such as scientific workflows and Web mashups, which both automate repetitive tasks and ensure result reproducibility. However, constructing these applications is a non-trivial task, especially for users who do not have programming expertise. This problem is compounded for exploratory tasks, where the application needs to be iteratively refined. We are working on a new framework for manipulating collections of services and workflows, based on a workflow manipulation language (and visual interface) that naturally represents operations that are common in exploratory tasks.
- Information management infrastructure. With the growing volume of raw data, pipelines and provenance information, there is a need for efficient and effective techniques to manage these data. Besides the need to handle large volumes of heterogeneous and distributed data, an important challenge that needs to be

addressed is usability: Information management systems are notoriously hard to use. As the need for these systems grows in a wide range of applications, notably in the scientific domain, usability is of paramount importance [5][6].

Acknowledgements

This article describes work being done in the VisTrails project. We thank all team members for their contributions. This work was partially supported by the National Science Foundation, the Department of Energy, and IBM Faculty Awards.

References

- [1] S. Callahan, J. Freire, E. Santos, C. Scheidegger, C. Silva, and H. Vo. Managing the evolution of dataflows with VisTrails. In IEEE Workshop on Workflow and Data Flow for Scientific Applications (SciFlow), 2006.
- [2] The VisTrails Project. <http://www.vistrails.org>.
- [3] L. Bavoil, S. Callahan, P. Crossno, J. Freire, C. Scheidegger, C. Silva, and H. Vo. VisTrails: Enabling Interactive Multiple-View Visualizations. In IEEE Visualization 2005, pages 135–142, 2005.
- [4] C. E. Scheidegger, H. T. Vo, D. Koop, J. Freire, and C. T. Silva. Querying and creating visualizations by analogy. IEEE Transactions on Visualization and Computer Graphics, 2007.
- [5] L. Haas. Information for people. ICDE Keynote talk, [http://www.almaden.ibm.com/cs/people/laura/Information For People keynote.pdf](http://www.almaden.ibm.com/cs/people/laura/Information%20For%20People%20keynote.pdf), 2007.
- [6] H. V. Jagadish. Making database systems usable. SIGMOD Keynote talk, <http://www.eecs.umich.edu/db/usable/usability-sigmod.ppt>, 2007.