Capturing, Visualizing, and Sharing the Process of Data Analysis

Ritesh Agrawal[1], William Pike[2]
The Pennsylvania State University, Regional Visualization and Analytics[1]
Pacific Northwest National Laboratory[2]

We live in an age of data abundance, but also in an age where data is often incomplete, time-varying, ambiguous and distributed across different sources. Data analysis and information production from such vast distributed digital data sources is often a complex and an iterative process. While, much work has been done on supporting data analysis, research on helping an analyst to effectively communicate and share this iterative process of data analysis is still lacking. Hence, in this research we extended Scalable Reasoning Systems to capture and visually represent an analyst's cognitive thought process as he/she distills information.

Developed at Pacific Northwest National Laboratory, Scalable Reasoning Systems (SRS) is an online visual analytic environment. SRS provides a visual interface to analyze sets of documents, assign roles (such as hypothesis, fact, conclusion, etc.) to these sets of documents, and graphically connect them to build a logical argument. Each set of documents along with their associated role is called as an artifact. We extended SRS to track an analyst's interactions with the system and use these interactions and the associated server-side queries to infer human-cognitive tasks. The inference of cognitive tasks is made possible by the use of a mapping scheme between server-side queries and cognitive tasks. The mapping scheme was developed in conjunction with the developers of SRS. Once identified, these cognitive tasks and data transitions are stored in a hybrid database schema. The hybrid database schema is a combination of a "Snapshot" and a "Transition" database schema. Finally, for an easy understanding and effective communication of the data analysis process, the stored cognitive tasks and data transitions are visually represented using a "simile" timeline view (Fig 1.).

Currently, our system is in its nascent state and requires much more research for its wider use in data analysis and visual analytics. At the theoretical level, there is a need for a comprehensive taxonomy of cognitive tasks and transformations between low-level human-machine interactions, such as mouse movements and keystrokes, and human-cognitive tasks. Such a taxonomy will enable developing an independent module of our logging system that can be connected with any other visual analytic environment. At the database level, there is a need for a database schema that can seamlessly handle transitions in data, visualization, and cognitive tasks. There are many suggested ways to classify data, visualizations, and cognitive tasks. However developing a robust and generic logging systems from these classifications remains uncertain. At the visualization level, new and innovative ways are required to represent data analysis process. Current timeline view has limited capability to show complex data analysis process and does not scale up for long and complex data analysis processes.

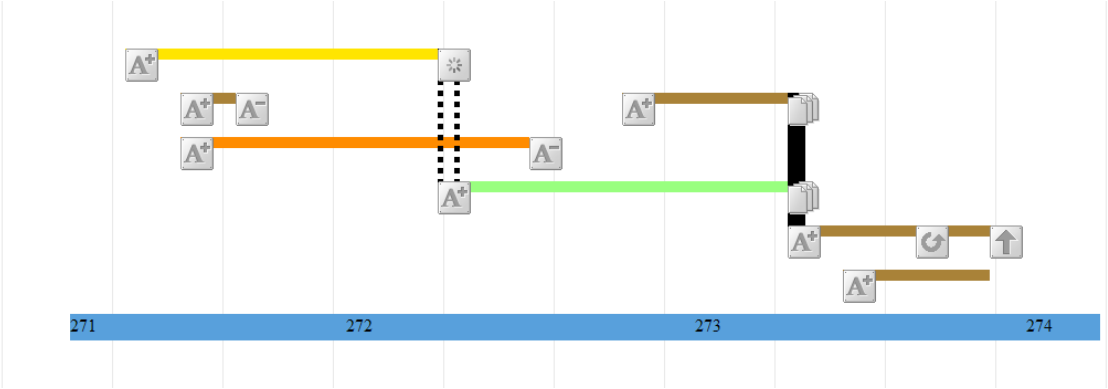

Figure 1: Visualizing the Process of Data Analysis

In the above figure, icons are referring to various cognitive tasks such as create an artifact, merge artifacts, delete an artifact, etc. Horizontal lines are presenting different roles of an artifact such as hypothesis, fact, assumption, etc.