Analytic Provenance Dataset

Sina Mohseni*
Texas A&M University

Eric D. Ragan†
Texas A&M University

ABSTRACT

We present an analytic provenance dataset to study analyst's behavior and thought process through interaction logs. We conducted a series of user studies involving exploratory data analysis scenario with textual data. Interactions logs, think-alouds, videos and all other coded data in this study are available online for research purposes. Analysis sessions are coded in sub-tasks based on user think-alouds, video and audios captured during the studies. Dataset includes logs from 24 participants working on three different analysis scenarios. We selected data analysis tasks with sufficient complexity and scope to allow the exploration of various topics and hypotheses.

1 INTRODUCTION

Visual analytic tools assist analysts with exploratory inspection of large amounts of data to identify, understand, and connect pieces of information. At a meta level, understanding analysis processes is important for improving tools, communicating analysis strategies, and explaining the evidence. *Provenance* for data analysis tracks the history of the analysis, including the progression of findings, interactions, data inspection, and visual state [3]. Analyzing user interactions and data provenance reveals more information about analysis process, helps in understanding how the user discovers the insights and unfolds analysis steps.

Designing analytic provenance visualizations tools requires analysis records. We conducted a set of user studies using text analysis scenarios to provide a provenance test data. A text explorer tool (Fig. 2) logged user interactions (interaction provenance) along with documents ID (data provenance) associated with each action. Data logs, videos, and audio were captured from 24 analysis sessions. Think-aloud transcript and insight coding are done in all studies. Captured videos, user interaction logs and ProvThreads visualization for all studies are available on-line at research group web page [1] for research purposes.

2 CAPTURING USER INTERACTIONS

To complete the analysis task, participants used a basic visual analysis tool (see Figure 2 on next page. The tool supports spatial arrangement of articles, the ability to link documents, keyword searching, highlighting, and note-taking. When loading the data in our document explorer tool, each document starts as collapsed with only its title visible. Users could "open" any document by double-clicking the title bar or by clicking a dedicated button on the document's title bar, and this would expand the document to a window containing the text of the document. The document could be collapsed back to the title in the same way. Within an open document, users could highlight text by selecting it, right-clicking, and activating a menu item. When a window has highlighted text, the window could be "reduced to highlight", which would hide all text in the document except for the highlighted content. At the beginning of the study, documents were arranged in the left screen without a specific order or grouping. Users clicked and dragged documents, freely re-arranging documents in the workspace. They could also create editable notes windows in the same workspace. When using the *search* functionality, both matching words within windows and the windows themselves were highlighted. Users could also draw connection lines across document windows, which created a line to denote relationships visually.

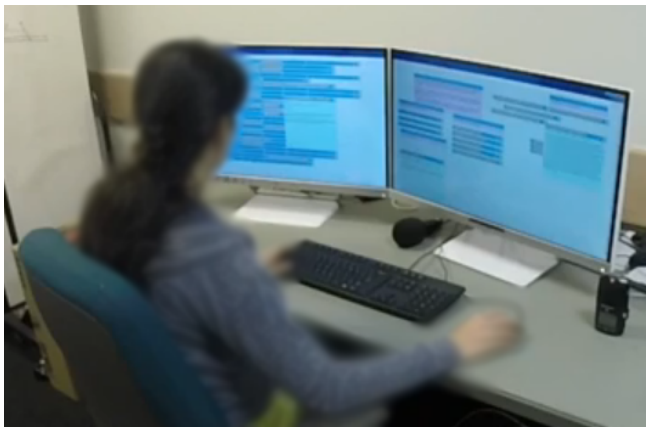

Figure 1: The document explorer tool and workspace used for the analytic tasks study.

All user interactions at a rudimentary level like mouse movements and clicks are captured during the study using the text explorer. Later we transform basic data log recorded from explorer tool to nine type of user actions, see Table 1. We associate analytical reasoning with different interactions available to the user, and later use it to modify the topic models.

Based on prior observations (e.g., [1, 2]) that mouse input can correspond with informational attentional. We use hovering the mouse over new document titles as users intend to explore new information. Hovering mouse over document text shows reading interaction of the articles.

Table 1: List of user interactions events and captured details

Interaction	Time	Duration	Document ID	Content
Open documents	x	x	x	x
Read documents	x	x	x	x
Move documents	x	-	x	x
Brush document titles	x	x	x	x
Search keywords	x	-	-	x
Highlight text	x	-	x	x
Bookmark documents	x	-	x	x
Connect documents	x	-	x	x
Create new notes	x	-	x	x
Writing notes	x	-	x	x

*e-mail: sina.mohseni@tamu.edu
†e-mail:eragan@tamu.edu

[1] https://research.arch.tamu.edu/analytic-provenance/

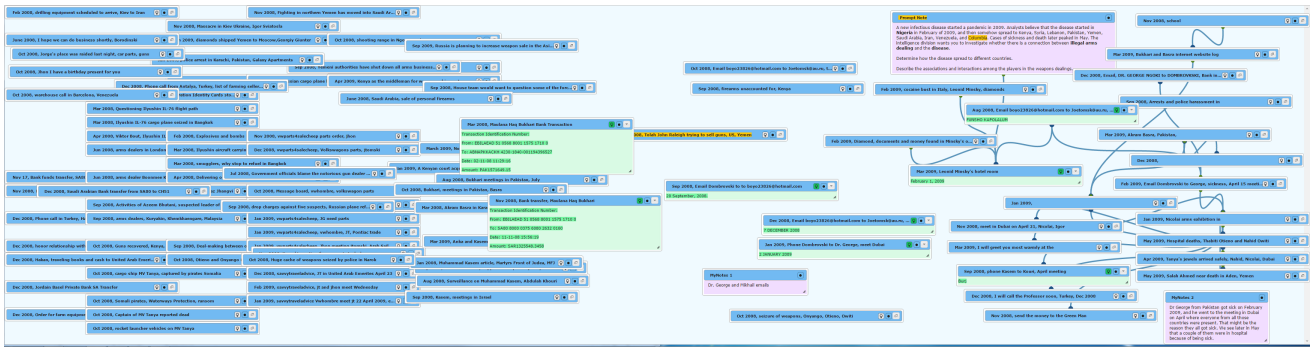

Figure 2: Partial screenshot of the document analysis tool used for collecting provenance data in the user studies. All text documents are listed in a collapsed format with a random order on the left monitor at the beginning of the study. Documents have titles and users are able to drag and displace documents in the explorer tool space. The user can write notes about their thoughts and conclusions, move and link documents and notes to organize information, highlight text to remind important events, and search words to find similarity in other documents.

3 Study Procedure and Data Sets

All participants for each session (6 female, 18 male) were university students from varying majors, and ages ranged from 20 to 30. None of the users were expert in analytic tasks. Participants used the document explorer tool desktop computer with two 27-inch monitors to analyze the data for 90 minutes. At the beginning of study sessions, text explore tool and analysis task were explained to the users. Users had a 15 minutes time to work with the tool and ask questions prior to the start.

Users were asked to work on the analysis scenarios. Each study session involved one of three intelligence analysis scenarios selected from the VAST Challenge data sets [4], a set of synthetically created data sets and analysis scenarios designed to be similar to real-world cases and problems. Specifically, our studies used data from the 2010 mini-challenge 1, 2011 mini-challenge 3, and 2014 mini-challenge 1. All data sets contain various text records such as news articles, emails, telephone intercepts, bank transaction logs, and web blog posts. For example, the 2014 data set involves articles about events and people related to missing individuals and violence related to a protest group in a fictional island. Participants were tasked with gathering information and finding connections between events, people, places and times in the data sets. Text documents varied in length from single sentences up to multiple paragraphs. While all of the data was in plain-text format, some of the documents primarily consisted of numerical data related to financial transactions. The 2010 data set had a total of 102 documents. Due to the larger sizes of the 2011 and 2014 data sets, we reduced the number of included documents to accommodate constraints of user study duration for 90-minute sessions. We used a subset of each data set to limit these two data scenarios to 152 documents.

4 Think-Alouds and Topic Changes

We used a think-aloud protocol to capture participants analysis process and insights during the study. Users were instructed to explain their thoughts and actions during the study. We transcribed users think-alouds by watching the screen-captured videos of each session. Think-alouds transcripts include all user's actions, talks, and time stamps of events. Two of the authors identified times where user changed topic of the investigation. We save all topic changes moments during the exploratory task and code them as topic changing (inflection) points.

For example, P7 working on the third dataset says "I'm looking for these caterers at the executive breakfast" and searches for "caterers". The participant continues reading documents from this search for about 10 minutes. Then the user says "I'm trying to figure out what the government was doing at the company", which is a change in topic of investigation. The user looks through titles and picks a couple of documents about government for about 8 minutes. While reading new documents, P7 finds out about the name "Edward" and is searching for incidents related to this name for the next 4 minutes. There are also moments that user is done with current topic and wants to change the subject. For instance, P3 working on the second dataset says "Let's search for some keywords" after 3 minutes of thinking and taking no actions. Then the user searches for keyword "thread" to find new articles about it. Also, in many cases, topic changing does not include think-alouds, like opening a random document and continuing with that, or writing a note about an old topic, or even returning to an old topic after a while.

5 Online Dataset

You are free to use the Provenance Analytics Dataset for research purposes. Captured videos, user interaction logs, document explorere tool and ProvThreads visualization for all studies are available online at research group web page:

- https://research.arch.tamu.edu/analytic-provenance/

for research purposes. Although all publications which use this dataset should cite following work.

- Mohseni, Sina and Pena, Alyssa and Ragan, Eric D. ProvThreads: Analytic Provenance Visualization and Segmentation. Proceeding of IEEE VIS (2017).

6 Acknowledgements

This material is based on work supported by NSF 1565725.

References

[1] M. C. Chen, J. R. Anderson, and M. H. Sohn. What can a mouse cursor tell us more?: Correlation of eye/mouse movements on web browsing. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '01, pp. 281–282. ACM, New York, NY, USA, 2001.

[2] J. Goecks and J. Shavlik. Learning users' interests by unobtrusively observing their normal behavior. In *Proceedings of the 5th international conference on Intelligent user interfaces*, pp. 129–132. ACM, 2000.

[3] E. D. Ragan, A. Endert, J. Sanyal, and J. Chen. Characterizing provenance in visualization and data analysis: an organizational framework of provenance types and purposes. *IEEE transactions on visualization and computer graphics*, 22(1):31–40, 2016.

[4] J. Scholtz, M. A. Whiting, C. Plaisant, and G. Grinstein. A reflection on seven years of the vast challenge. In *Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors-Novel Evaluation Methods for Visualization*, p. 13. ACM, 2012.