ProvThreads: Analytic Provenance Visualization and Segmentation

Sina Mohseni, Alyssa Pena, Eric D. Ragan*
Texas A&M University

Abstract

Our work aims to generate visualizations to enable meta-analysis of analytic provenance and aid better understanding of analysts' strategies during exploratory text analysis. We introduce *ProvThreads*, a visual analytics approach that incorporates interactive topic modeling outcomes to illustrate relationships between user interactions and the data topics under investigation. *ProvThreads* uses a series of continuous analysis paths called *topic threads* to demonstrate both topic coverage and the progression of an investigation over time. As an analyst interacts with different pieces of data during the analysis, interactions are logged and used to track user interests in topics over time. A line chart shows different amounts of interest in multiple topics over the duration of the analysis. We discuss how different configurations of *ProvThreads* can be used to reveal changes in focus throughout an analysis.

Index Terms: Information interfaces and presentation

1 Introduction

Visual analytics tools assist analysts with exploration of large amounts of data to identify, understand, and connect pieces of information. *Provenance* for data analysis tracks the history of the analysis, including the progression of findings, interactions, data inspection, and visual state [5]. Our research is motivated by the need to support meta-analysis of analytic provenance by researchers and designers to better understand analysts' strategies, to improve analysis tools, and to design effective training programs for data analysts. Analyzing user interactions and data provenance can reveal information about the analysis process, help in understanding how the user makes discoveries, and explain different analysis strategies.

In exploratory data analysis, it can be difficult to keep track of the different thoughts and topics considered during analysis, and analysts do not want to have to interrupt their thinking and work flow to annotate their thought process. Prior work has shown that interaction history can be highly effective for understanding analysis behaviors (e.g., [1–3]). However, full interaction logs are often long and verbose. Methods for summarizing and visualizing provenance are needed to provide a high-level overview that can be understood more quickly and easily. We aim to summarize the analysis process automatically using only the system logs from user interactions with data analysis software, thus avoiding the need for supplemental comments from the analysts. Summarization can be done by dividing the analysis process into smaller meaningful segments where each segment represents a stage of the analysis.

2 Method

In this work, we present a method to segment and visualize analysis history of an exploratory text analysis. To do so, we first need a set of analytic provenance data, then a method to segment the provenance data into smaller stages, and finally the generation of a visualization.

*e-mail: sina.mohseni, mupena17, eragan@tamu.edu

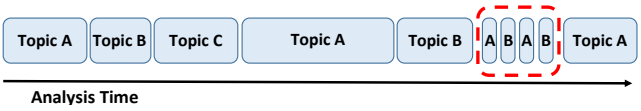

Figure 1: A basic example of how users interact with different topics during an analysis. The length of a topic segment indicates the time spent interacting with data related to a particular topic. A single long span for one topic corresponds to a focused inspection of one topic, while a burst of multiple short segments (circled in red) might represent a consolidation of topics.

2.1 Analytic Provenance Data Capture

To collect provenance test data to design and demonstrate our approach, we conducted a set of user studies using text analysis scenarios from the VAST Challenge datasets (2010 MC #1, 2011 MC #3, and 2014 MC #1). Three different datasets were used to help assess the robustness and reliability of the design across datasets. We ran 24 study sessions where participants performed the exploratory analysis. To complete the analysis task, participants used a basic visual analysis tool that supports spatial arrangement of articles, the ability to link documents, keyword searching, highlighting, and note-taking. Anonymized versions of the captured provenance records, user interaction logs, and ProvThreads visualization for all studies are available online [1] for research purposes.

2.2 Analytic Provenance Segmentation

To segment temporal data, different features and methods can be used to produce meaningful segments, and the selected features largely depends on the nature of the data and the reasons for segmenting. In our research, we aimed to segment interaction history in a way that corresponds to stages of human analytic thinking. Through careful review of the captured videos and transcripts of think-aloud comments from the user studies, we studied the times where the participants changed their goals or topics of investigation. We observed that changes happen when users start looking for new information and connections to support a hypothesis, when they search for new evidence after discovering an insight, or when they continue searching for new clues. Topic-change behavior also reveals intuitions about analyst's strategy. For example, longer periods of time spent on a single topic is indicative of top-down analysis, whereas instances of multiple short, successive topic changes demonstrates bottom-up analysis behavior. Figure 1 shows how topic changes could be used to infer analyst strategy.

Thus, our method uses the interaction history to automatically infer interests and topic changes over the duration of the analysis. As other researchers have found, identifying user intentions and reasoning form interaction data can be effective [2, 4], but it is difficult to achieve with concise and accurate representations.

2.3 ProvThreads Method and Design

ProvThreads is designed to visualize the provenance of topic investigation in a way that connects interaction behaviors with data content. In the proposed design, analytic provenance segmentation is done

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

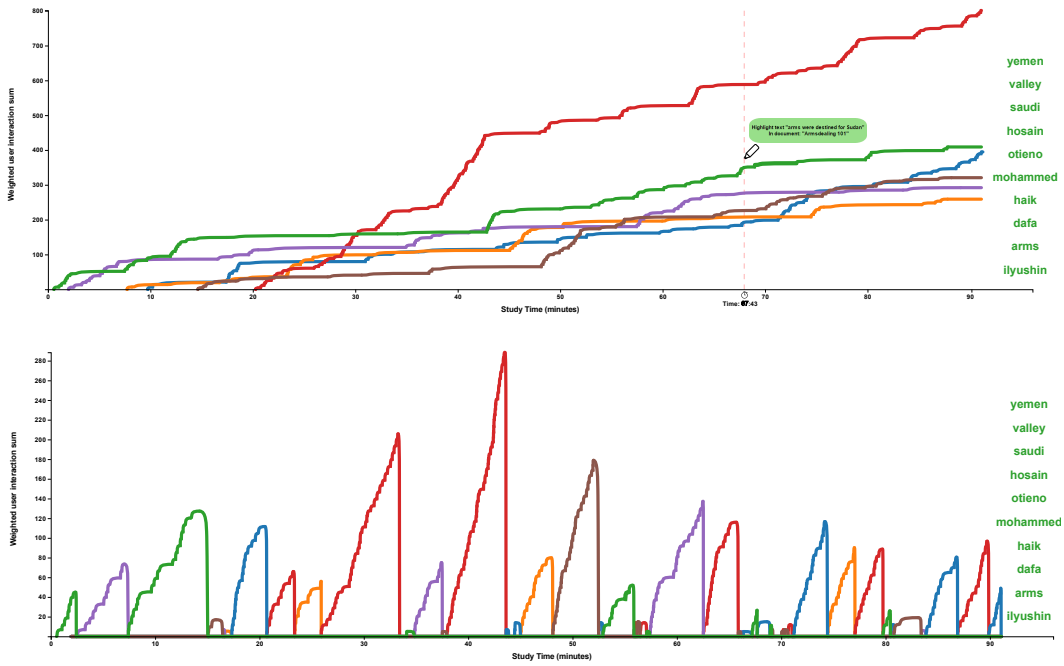

Figure 2: Two views of the ProvThreads visualization from the same user's interaction logs. Each colored thread corresponds to a topic in the text corpus, and the height increases with additional interactions over time. Brushing over a line shows interaction details in a tooltip, and a list of key terms for the associated topic are shown on the right. Top: The *topic coverage* view shows the accumulation of user focus or inspection in each data topic over the duration of the analysis. Bottom: The *topic segments* view groups topics together based on nearby interactions, and topic height resets when the analyst switches to a new topic group.

by input data classification and assigning user interaction to each data topic. In our current implementation, we use topic modeling methods to classify text documents using Latent Dirichlet Analysis (LDA) to assign a topic label to each text document. Although text documents contain a mixture of multiple LDA topics with different probabilities, the current design labels documents with only the most probable topic to help simplify the visual summary. All user interactions from the exploratory text analysis are labeled with the corresponding document's topic. For instance, if the user is opening a document labeled with topic A, writing a note about topic A, or searching a keyword from topic A, then the associated captured user interaction is labeled with topic A as well.

ProvThreads represents the provenance of topic interactions using colored lines that represent different data topics (see Figure 2). A user can interactively brush threads to see details (via tool tips) about the interactions or data focus at different times in the analysis. Additionally, the user can toggle the display of icons the indicate actions for different types of user interactions over time. The threads design allows for visual identification of patterns of attention to particular topics, and it shows the relationship between interactions and topics. It shows for how long and how deep the analyst focuses on different pieces of information. In addition, brushing different topic lines shows the terms associated with each topic (see the word lists on the right side of Figure 2).

ProvThreads supports two visual designs for showing topic interactions: the **topic coverage** view and the **topic segments** view. In the **topic coverage** view, each topic thread will increase in height when the analyst interacts with the corresponding data topic (see Figure 2, top). The main purpose of this design is to track and compare amount of total attention (inferred via interaction) to each topic over time.

The **topic segments** view follows a different design that emphasizes topic changes (see Figure 2, bottom). In this view, topic lines still increase along with associated user interactions, but a new topic thread starts when the user switches to a new topic. To achieve this view, we segment the provenance timeline using a recursive algorithm that combines related topics based on concurrent interactions with data associated with multiple topics. This view simplifies and highlights the analyst's topic of interest at each moment and shows clear transition points when the focus changes.

3 Conclusion

We propose a new method and design to segment and visualize analytic provenance that takes advantage of user interactions with data topics to gives a better picture of an analyst's thought process. We prepared a publicly available analytic provenance dataset, and in this ongoing research, we are exploring the effectiveness of a human-in-the-loop approach that supports interactive topic merging to assist segmentation.

References

[1] T. Blascheck, M. John, K. Kurzhals, S. Koch, and T. Ertl. Va 2: A visual analytics approach for evaluating visual analytics applications. *IEEE transactions on visualization and computer graphics*, 22(1):61–70, 2016.

[2] W. Dou, D. H. Jeong, F. Stukes, W. Ribarsky, H. R. Lipford, and R. Chang. Recovering reasoning process from user interactions. *IEEE Computer Graphics & Applications*, 29(3):52–61, 2009.

[3] D. Gotz and M. X. Zhou. Characterizing users' visual analytic activity for insight provenance. *Information Visualization*, 8(1):42–55, 2009.

[4] R. Linder, A. M. Pena, S. Jayarathna, and E. D. Ragan. Results and challenges in visualizing analytic provenance of text analysis tasks using interaction logs. *Logging Interactive Visualizations and Visualizing Interaction Logs (LIVVIL) Workshop*, 2016.

[5] 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.