Exploiting History to Reduce Interaction Costs in Collaborative Analysis

Ali Sarvghad and Melanie Tory

Abstract—When analysts work in a distributed fashion, they need to understand what their collaborators have done and what avenues of analysis remain uninvestigated. Although visualization history has the potential to communicate such information, the common representations are often limited to sequential lists of past work. Such representations do not make it easy to understand the analytic coverage of the dimension space (i.e. which dimensions have been investigated and which have not). This makes it difficult for an analyst to plan their next steps, particularly when the number of dimensions is large. In this paper, we propose representing the prior analysis from a dimension coverage perspective. Dimension view provides a unique perspective that can facilitate exploratory analysis by enabling analysts to easily identify what dimensions have been examined and in what combinations. We hypothesize that addition of this view to common representations of visualization history will reduce cognitive and interaction costs by helping the analyst to discover data subsets to explore. We studied the effects of this view on a distributed collaborative visualization process. Our findings show that providing views of the dimension and data space reduces time required for identifying and investigating unexplored regions and increases the accuracy of this understanding. In addition, providing these views results in a larger coverage of entire dimension space.

Index Terms—Analysis History, Distributed Collaboration, Information Scent, Dimension Space Coverage.

1 Introduction

In exploratory analysis, analysts need to constantly formulate new goals and decide on subsets of data to investigate. Based on Lam's framework [1] of interaction costs, the "Gulf of Goal Formation" (i.e. process of articulating a new question to ask and a goal to pursue) is one the three main contributors to the overall cost structure of visualization. This might even be even a greater cost in a collaborative setting. Analysts need to understand what their collaborators have done, which aspects of the data set have been explored and what might be good new directions to examine. Therefore, "goal formation" bears the extra costs of understanding what has been done in the past. Information scent (i.e., navigational cues derived from data and/or meta-data) have been used to reduce the costs of goal formation [2]. A common source of information scent in collaborative context is system-logged or user recorded information such as the history of a document's edits or comments left by users. In particular, state-based analysis history (i.e., recorded states containing information about each individual visualization created by users) can be used to provide cues to help users understand what previous work was done. Yet, the full potential of history in providing information scent has been underexplored. Willet et al. [2] used social data analysis history to augment UI widgets (e.g., slider with a histogram) with information about the aggregated investigation of data values. While effective in helping users acquire knowledge of what data values have been the center of prior attention and what has been left out, it provides no information with regards to the "dimension space". It is rather difficult to answer questions such as, "What data dimensions have been explored? In what combinations?" This problem can grow exponentially with the number of data dimensions.

Because visualization history modules record all of the past work done by an analyst, they ought to be able to provide this information. However, existing representations of history (typically a list of past visualization states) poorly support understanding what dimensions have been investigated, in what combinations, and with what frequencies. We propose adding Dimension view to history tools for providing information scent regarding the coverage of dimension space. We hypothesize (H1) that dimension view will reduce the cost of "goal formation" by reducing the time and increasing the accuracy of discovering less explored dimensions, and (H2) it will result in better overall coverage of dimension space. We performed two user studies to evaluate H1 and H2. Results showed that users acquired more detailed information about the analytic coverage of dimension space in considerably shorter times and participants were much more likely to ask questions considerably different from the initial analysis (i.e., they paid more attention to dimensions and data that the previous analyst had neglected).

2 Related Work

In the context of asynchronous collaborative analysis, Wattenberg et al. [3] hypothesized that providing visual cues into the past exploration of data values may encourage people to analyze uninvestigated dimensions. In their prototype tool, investigated time series were in grey, in contrast to colorful uninvestigated ones. Although their design helped one to discover uninvestigated data, it fell short of fully exposing the investigation of dimension space. Closest to our research, Willet et al. [2], provided information scent by incorporating visual cues into common interface widgets such as radio buttons and sliders to help users identify underexplored values. This data-centric approach can help users to identify less explored data values when the analysis is built around small set of dimensions (e.g., only investigating the job trends across gender over the last 100 years). However, it is unable to support gathering relational and quantitative information about dimension space coverage, which is important when dealing with a multi-dimensional data set and an analysis task that spans the entire dimension space.

3 Dimension View

Dimension view (Figure 1A) provides information scent into the exploration of dimension space. Essentially, this view enables a user to 1) easily distinguish unexplored data dimensions and 2) to discover investigated dimension combinations. In the current version of our tool, we designed this view as a treemap with a squarified layout in which each cell represents a data dimension. Initial rendering of the view enables users to instantly discover investigated / uninvestigated dimensions and the focus of prior work. We use redundant encoding with greyscale and size to convey this information. Size and greyscale represent the relative mapping frequency of a dimension: large dark grey rectangles represent the most frequent dimensions and small white rectangles represent dimensions that were never investigated. We refer to *mapping* as encoding a dimension in a visualization (in the previous analysis session) by an element such as position, size, shape or color. Each cell is also labeled with the dimension name and mapping frequency (e.g., "City[5]"). The redundant encoding using size and greyscale enables users to gather top-level information very fast. For example, with a glance at Figure 2, a user can understand that the two dimensions at the left (Sales and Profit) have been the main focus in prior analysis.

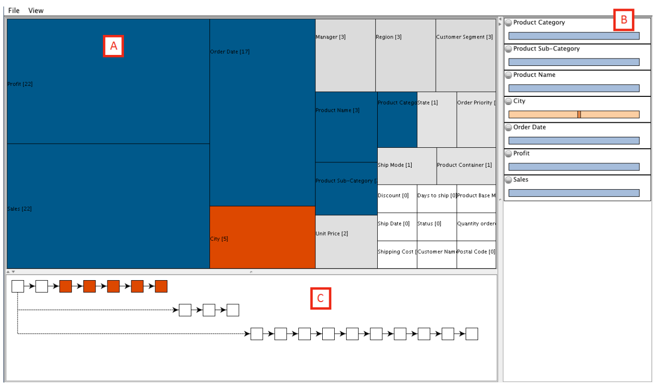

Fig. 1. Prototype: Prior data analysis is represented from three angels, Dimension view (A), Data View (B), and Sequence view (C). In Dimension view, uninvestigated dimensions are grouped together and rendered with white background. Investigated dimensions have initially grey backgrounds. Greyscale and size redundantly encode mapping frequency. When a dimension is selected (as shown here with orange background) other dimensions that have been investigated with this dimension become blue. A user can make multiple selections. Data view provides information scent into the investigation of data values, and Sequence view shows recorded sates in temporal order.

Interacting with Dimension view enables users to discover co-mapping dependencies. When the user clicks on a dimension, the selected dimension's background colour changes to orange and any dimensions that have been mapped in a visualization along with this dimension become blue. Other cells remain unaffected. Color-coding assists the user to immediately recognize related dimensions in the view. For example, selecting City (as in Figure 1A) shows that it has been considered with six other dimensions: Sales, Profit, Order Date, Product Name, Product Category and Product Sub-Category. City was not considered with Unit Price, Customer Segment, and several other dimensions. If required, the user can select multiple dimensions to investigate their co-mapping. This view is interlinked with the other two views: user interactions in Dimensions view propagate to Sequence and Data views. Sequence view (Figure 1C) is a more traditional representation of visualization history showing the temporal order of visualization states. Data view (Figure 1B) represents each dimension as a horizontal bar and uses color coding to reveal which data values have been included in charts more frequently.

4 Evaluation

To assess our hypotheses, we designed and implemented a prototype tool for reviewing history. This prototype represented history from three perspectives 1) a sequential list of all recorded states similar to common linear representation of history, 2) Dimension view that provided on demand navigational view for dimension space, and 3) Data view (similar to [2], with minor design alterations) that provide similar information for data values. Participants of both studies were randomly assigned to a full version of prototype with all three views or a base version that only included the sequential view. For the first user study, evaluating (H1), we recruited 20 computer science students who performed a task that involved answering 11 multiple-choice questions. These questions were designed to examine participants' ability to understand coverage of dimension space (i.e., what was investigated and what was left out) and their ability to gather relational/quantitative information (e.g., was City considered more than Region? Was Profit considered with Returns?). Results showed that full version users were both faster (average of 7.12 (SD=4.5) minutes versus 12.8 (SD=7.9)) and more accurate in answering the questions.

To evaluate (H2), we recruited 20 business students to perform an actual business data analysis task using Tableu software. They were asked to continue the analysis started by their "collaborator". They started by reviewing the collaborator's prior work and they could refer to the collaborator's work via the prototype at any point during the session. We assigned each participant a similarity score (ranging from 0 [Completely different] to 1 [Identical]) comparing their analysis to that of the "collaborator". Similarity was based on the investigated dimension combinations. Results showed that participants with access to the full version of the prototype had significantly lower similarity scores (full version similarity mean=0.33, SD=0.11, baseline mean=0.58, SD=0.21, $t(339) = 9.192$, $p < 0.0091$). In other words, their work was significantly more divergent from the prior work. This result indicates Dimension view's potential to assist in the process of formulating new questions and identifying new subsets of data to investigate.

5 Discussion

Our findings from the two studies show that Dimension view has the potential to reduce the cost of goal formation by 1) reducing the time that is required for an analyst to understand/recall what dimensions have been investigated and in what combinations, 2) increasing the accuracy of attained information, and 3) discovering new subsets of data to investigate. This interactive view can provide visual information scent that facilitates the navigation of information space. This could be critically important in asynchronous collaborative work where there may be limited direct communication between collaborators and sharing work history may be one of the main channels for sharing past work.

6 Future Work

The primary focus of the initial stage of this research was to examine the value of providing Dimension view. Therefore we designed the prototype as a standalone tool separate from a visual data analysis tool. In the next step of this research, we are aiming to redesign and integrate Dimension view into a visual data analysis tool. We plan to evaluate the effects of providing information scent from this perspective on ongoing analysis.

References

[1] H. Lam, "A framework of interaction costs in information visualization.," IEEE Trans. Vis. Comput. Graph., vol. 14, no. 6, pp. 1149–56, 2008.

[2] W. Willett, J. Heer, and M. Agrawala, "Scented widgets: improving navigation cues with embedded visualizations.," IEEE Trans. Vis. Comput. Graph., vol. 13, no. 6, pp. 1129–36, 2007.

[3] M. Wattenberg and J. Kriss, "Designing for social data analysis.," IEEE Trans. Vis. Comput. Graph., vol. 12, no. 4, pp. 549–57, 2006.