Evaluating the Effect of Visualizing Dimensions Search Space on Exploratory Visual Data Analysis

Abeer Alsaiari*
University of Illinois at Chicago

Andrew Johnson
University of Illinois at Chicago

ABSTRACT

In exploratory visual data analysis, analysts constantly investigate different subsets of data. The cost of deciding what to explore next "Gulf of Goal Formation" [1] is a major component of interaction costs in information visualization. When a team of analysts collaborates using multiple devices to work on an analysis task, the decision cost can be higher due to short-term memory and the recency effect. Analysts in collaborative settings need to understand what was investigated by the team and what was left. Visualizing the dimensions search space can communicate to the team what dimensions have been investigated (and in what combination) and what were left. We conducted a between-groups study to evaluate the effect of visualizing the dimensions search space. Our results indicate that visualizing dimensions search space reduces the decision cost and positively affects the rate of goal formation.

Keywords: Exploratory Visual Data Analysis, Collaboration, Multi-device Environments.

1 INTRODUCTION

Exploratory visual data analysis is an iterative process that involves formulating new questions and inspecting data of interest. Supporting exploratory visual data analysis is essential especially when multiple analysts work together. Analysts need to understand what courses of analysis were investigated and what were left. Prior research has presented several methods to guide analysts during exploratory visual data analysis. One direction of these approaches is using visual cues that assist and orient users in their analysis [2][3][4]. Willet et al. [2] and Sarvghad et al. [3] enhanced visualization controls with embedded visual cues to enable users to understand the navigation of the data space by revealing each dimension's frequency of investigation. Sarvghad and Tory [4] used two different visual representations, Circos and Treemap, to visualize the frequency of investigation and co-mapping of each attribute. These approaches focus on visualizing dimensions' frequency of investigation and co-mapping with other attributes. While these are important features, they lack revealing information about what dimensions' data coverage were investigated when analysts pivot the analysis between different subsets of the data space. In this study, we evaluate the effect of revealing information about what dimension's data space coverage were investigated (and in what combination) and what were left. We conducted a between-groups study where half of the groups used a baseline visual analysis tool and the other half of the groups used a full version of the tool enhanced with a visualization of the dimensions search space.

* aalsai3@uic.edu

2 VISUALIZING DIMENSIONS SEARCH SPACE

We selected the parallel set as the visual representation for visualizing the dimensions' search space. This selection was based on a few design goals we want to achieve in a visual representation. These design goals are the ability to: (a) visualize the current dimensions space coverage that the analysis is pivoted to, (b) visualize what parts of each dimension have been investigated so far and what were left, and (c) reveal information about each dimension's frequency of investigation and the co-mapping information of its data spaces. Parallel Set is a visualization approach to visualize categorical data which can be thought of as the parallel coordinate with the addition of the "proportional" component. They show distributions over categories. Although we only have two categories of information to visualize, which are here the current and past dimensions' space coverage, they can provide a simple and intuitive way to visualize such data. Figure 1 shows the detail of the designed visual representation.

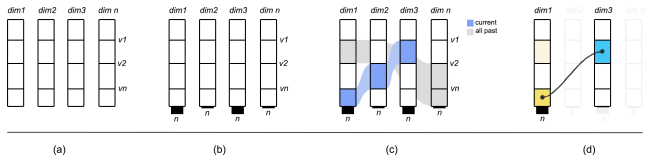

Figure 1: Parallel Set: (a) Each dimension is represented by a line-set divided into several blocks representing its values distribution (categorical or numerical). (b) Each dimension is attached with a bar representing the frequency of investigation of this dimension. (c) The current dimensions that the analysis is pivoted to will be stacked to the left showing the current coverage of the data space (in blue) and all past data spaces will be combined into one category (in grey) and send to the back. (d) Clicking on a dimension shows co-investigation of data spaces (appearing together in a chart).

3 STUDY DESIGN

We conducted a between-groups study to evaluate the effect of visualizing dimensions' data space coverage and co-investigation. The study contained two conditions: *baseline* and *full* versions where half of the groups used a baseline visual analysis tool and the other half of the groups used a full version of the tool enhanced with a visualization of the dimensions search space.

3.1 Participants

We recruited 30 participants as 10 groups of 3. Participants were 25 male and 5 female students between the ages of 21 and 35 years old. They participated in the study for the duration of 43min-2hrs.

3.2 Apparatus

In the *baseline* version, participants used PolyVis [5], a visual data analysis tool designed for cross-device collaboration.

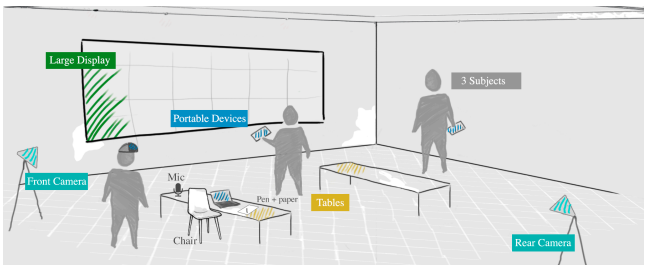

Figure 2: Illustration of the study setup

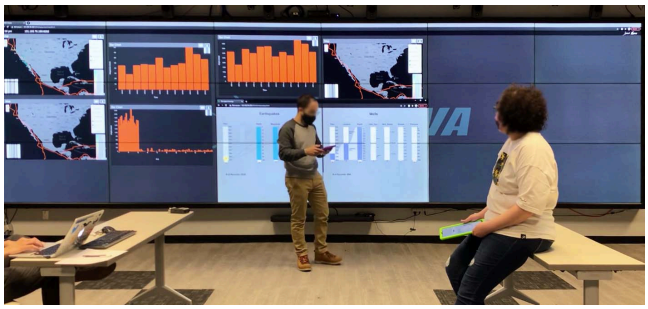

Figure 3: Participants examining a set of created visualizations on the large display. The middle window shows the visualization of the deimensions search space.

In the *Full* version, participants used PolyVis integrated with a visualization of the search space in a separate window as shown in Figure 3. The window of the search space is placed in the middle of the large display to help analysts form their next goal of the analysis. PolyVis was used to collaboratively create visualizations for analysis. When a new visualization is created, the visualization of the search space gets updated to reflect the new state of the dimensions coverage.

3.3 Setup and Data Capture

The study was conducted in a room approximately 10.61 by 5.59 meters, equipped with a high-resolution large display. Overall display size is approximately 7.3 by 2.05 meters at a resolution of 11,520 by 3,240 pixels. Other devices were placed on a table in the middle for use during the study: one MacBook Pro (macOS Sierra, 2.4 GHz Intel Core i5), one 8" Samsung - Galaxy Tab A (32GB, Android 9 (Pie)), one 10" Samsung - Galaxy Tab A (64GB, Android 9 (Pie)), and one Microsoft HoloLens 1 (Windows Mixed Reality OS, Intel 32-bit (1GHz) CPU, 2 GB RAM). Systems usage logs were collected from all deployed devices. We wrote a script to capture all created visualizations and selected attributes. The study was video recorded using two cameras, one showing the full room from behind and one showing the participants' interaction with the large display from the front. In addition, a microphone was placed at the table for audio recording. The setup is pictured in Figure 2.

3.4 Datasets and Tasks

Each group completed two tasks, with focused and open questions. In the first task, participants were given focus questions that can be answered by creating one or two visualizations. We opted for the focus questions to be a practical tutorial on how to use the system. Participants were then asked to explore the earthquake events and wells injection activities and identify trends/observations in the data using two geoscience datasets. The first dataset contained information about earthquake incidents in Oklahoma and California from the years 2000 to 2010. The Wells dataset contained information about the fracking activities in Oklahoma and California also from the years 2000 to 2010. The earthquake dataset was provided courtesy of http://service.iris.edu/ and the Wells injection dataset was provided courtesy of http://www.occeweb.com/.

4 FINDING

Lam [1] presented a framework that comprises the decision cost as a major component of interaction costs in information visualization. It bears the cost of "finding a data subset to explore" and "choosing amongst interface options". The visualization of the dimensions search space played a central role in data selection and attributes co-investigation. The visualization of each dimension's data space coverage facilitated the selection of the next course of analysis. In addition, forming attributes co-investigation mostly took place by referring to the list of dimensions visualized on the big display. To measure the effect of visualizing the dimensions search space on reducing the decision cost, we calculated the rate of producing views for each condition. The higher the rate of views generation, the less the cost of goal formation. First, we counted the number of created visualizations by each group. Full version groups created an average of 43.6 views (SD = 18.15), versus 20.4 (SD = 13.07) for baseline version groups. A tow-tail independent t-test showed that full version groups generated more views than baseline groups ($t = 2.3198$, $df = 8$, $p = 0.0489$ at $p < .05$). To eliminate the effect of sessions' time and calculate the views' generating rate, we divided the number of created views by the session's time. Full version groups created an average of 0.7880 views per minute (SD = 0.1987), versus 0.4360 views per minute (SD = 0.1781) for baseline version groups. A tow-tail independent t-test showed that full version groups generated views at a higher rate than baseline groups ($t = 2.9498$, $df = 8$, $p = 0.0184$ at $p < .05$) which indicates a reduction in decision cost.

5 CONCLUSION AND FUTURE WORK

We presented a result from a study to evaluate the effect of visualizing dimensions search space on decision cost. Our results showed a positive effect in increasing the rate of views' generation which indicates a lower decision cost. In future work, we aim to study how visualization of dimensions search space can increase the breadth of the analysis. By presenting this work to the visualization community, we aim to get feedback and discuss how we can define and measure the breadth and depth of the analysis. In addition, we are looking for feedback on how we can improve the design of visualizing the dimensions search space.

REFERENCES

[1] Lam, Heidi. "A framework of interaction costs in information visualization." *IEEE transactions on visualization and computer graphics* 14.6 (2008): 1149-1156

[2] Willett, Wesley, Jeffrey Heer, and Maneesh Agrawala. "Scented widgets: Improving navigation cues with embedded visualizations." *IEEE Transactions on Visualization and Computer Graphics* 13.6 (2007): 1129-1136.

[3] Sarvghad, Ali, and Melanie Tory. "Exploiting analysis history to support collaborative data analysis." *Proceedings of the 41st Graphics Interface Conference*. 2015

[4] Sarvghad, Ali, Melanie Tory, and Narges Mahyar. "Visualizing dimension coverage to support exploratory analysis." *IEEE transactions on visualization and computer graphics* 23.1 (2016): 21-30.

[5] Alsaiari, Abeer, et al. "PolyVis: Cross-Device Framework for Collaborative Visual Data Analysis. " *IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2019.