STORYBOARD: Navigating Through Data Slices with Hierarchical Summary of Visualizations

ABSTRACT

The task of navigating through a large, multidimensional dataset is a common challenge in exploratory analysis. Due to limitations on the number of visualizations that an analyst can examine at one time, the narrow scope of drill-downs can often lead to inductive fallacies. In this paper, we present STORYBOARD, an interactive visualization recommendation system provide safe guarantee during drill-down exploration by picking the proper visualization reference that leads to interesting and informative trends. Given a dataset and the x and y axes of interest, STORYBOARD intelligently explores the lattice of equivalent visualizations across data subsets, and recommends interesting and informative visualizations. The recommended visualizations are then displayed in an interactive dashboard, where the visualizations are organized into a hierarchical layout. Our evaluation study shows that visualization dashboards generated by STORYBOARD are interpretable and leads to higher performance in data analytic tasks compared to the competing baselines.

KEYWORDS

exploratory data analysis, visualization recommendation.

1 INTRODUCTION

The journey of understanding a multi-dimensional dataset often begins with visualizations, exploring the space of attributes, in search of insights. Perhaps the most common route of this journey is to generate visualizations to gain an overview of the data, and drill-down to interesting subsets to generate more visualizations. For example, a campaign manager may be interested in understanding the voting patterns across different demographics (say, race, gender, social class) using the 2016 US election exit polls[1]. A natural first step is to generate a bar chart for the entire population, where x-axis shows the election candidates and y-axis the percentage of votes for these candidates. He then can drill down to specific demographics of interest, say gender based demographics by generating bar charts for female. In this exploration process each drill-down may lead to insights, which derive from the observed visualizations. For example, the drill-down on gender shows that the female demographics follow the trend of overall population, whereas a further drill-down on race shows that the trend of black females differ from that of females. Many existing tools will flag the latter as a potential

[1] https://edition.cnn.com/election/2016/results/exit-polls

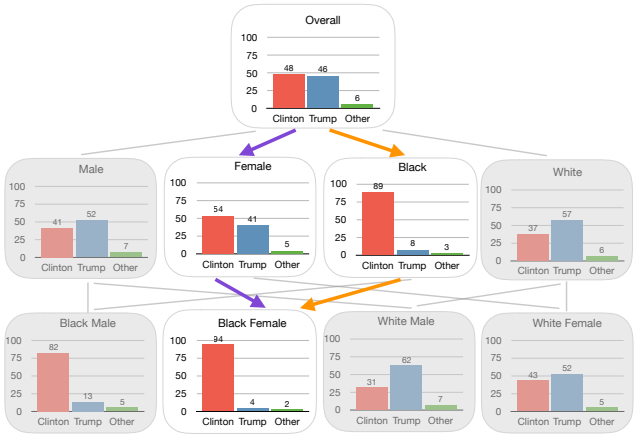

Figure 1: Example data subset lattice illustrating the misleading factor fallacy along the orange path as opposed to the informative purple path.

insight, since the black females exhibit a high deviation w.r.t. an observed parent. The insight, however, is a by-product of the order in which the drill-down has been performed—drilling down on race before gender will explain the trend of black females in light of black demographic. This example demonstrates the case of *drill-down fallacy*, which results from potentially confounding factors/variables not explored in a drill-down path.

Exploring a dataset by selecting a random attribute at each drill-down step analysts become prone to the drill-down fallacy. A naive solution to avoid this fallacy is to explore all potential pathways along the drill-down path. For example, generating and exploring visualizations for both race and gender based demographics, before exploring any of their combinations. Unfortunately, this approach does not scale with increasing number of factors in the drill-down path.

In this paper we develop a tool to help users explore a dataset while avoiding the risk of drill-down fallacy. Our tool automatically identifies the best possible drill-down paths that leads to *informative insights*, and summarizes the paths. The key idea of our tool is *informativeness*—the capability of the observed visualizations to explain the unseen visualizations. Informativeness prevents users from the drill-down fallacy, and helps identify meaningful insights that arise from the irregularities in the data (instead of confounding variables).

Our user study results described in Section 5 shows that this notion of informativeness can guide an analyst towards meaningful insights. The contribution of this paper include:

- Introducing the novel concept of *informativeness* that helps avoid drill-down fallacy in data exploration (Section 3),
- Designing a tool that automatically identifies the best possible drill-down paths based on informative insights, and summarizes those (Section 4),
- Demonstrating the efficacy of our system through a comprehensive user study evaluation (Section 5).

2 PROBLEM FORMULATION

In this section, we first describe how analysts explore the lattice through drill-downs and introduce a common fallacy that arises when analysts have limited time and attention to examine all possible factors that contribute to the observed visualization. Then, we discuss how to resolve the problem of finding informative references along a drill-down path.

Research in visualization storytelling shows that people prefer hierarchically structured visualizations with increasing levels of aggregation [1–3]. In order to find meaningful insights, analysts often drill-down to explore data at different levels of granularity by adding one filter at a time. For each data subset that he encounters, he may want to visualize the distribution of measure values through a bar chart. When analysts perform drill-downs, they naturally formulate their expectation based on the last visualization that they observe, known as the 'parent'. The parent can be obtained by removing one filter constraint from the current visualization in context (known as the 'child' visualization). For example in Figure 1, the visualizations Female and Black are the parents of the Black Female visualization.

As the analyst is drilling down by adding one filter at a time, the analyst is prone to be misguided by child visualizations that highly deviate from one of its parents, overlooking other potential factors that may explain the seemingly-anomalous behavior. We refer to this phenomena as *drill-down fallacy*, as this type of fallacy arises from the inductive nature of the drill-down operation. We demonstrate this fallacy with an example from the 2016 US Elections exit polls dataset. As shown in Figure 1, an analyst can either arrive at the Black Females visualization by going through the purple path or the orange path. At random, if the analyst went down the purple path, he may be surprised at how much the Black Female voting behavior differs drastically from the vote distribution for females. This behavior can be explained if the analyst went down the orange path, where he sees the proper reference (vote distribution for Black) that explains the behavior of the Black Female distribution. While such fallacies can be prevented if the analyst browses through all possible parents of any visualization that he observes in the dataset, the prohibitively large number of visualizations and limited memory and attention of analysts make this task impractical.

Due to these challenges, our goal is to develop a mechanism that would *provide safe guarantee by picking the proper informative parent* as a reference when analysts navigate through the space of data subsets. To model the informativeness of an observed parent in the context of an unseen visualization, we characterize the capability of the parent in predicting the unseen visualization. An observed parent is *informative* if its data distribution closely follows the data distribution of the unseen child visualization, since the visualization helps the analyst form an accurate mental picture of what to expect from the unseen visualization. Specifically, we formulate the informativeness of an observed parent V_i^j of an unseen visualization V_i as the similarity between their data distributions measured using a distance function $D(V_i, V_i^j)$. The most informative parents V_i^* of an unseen visualization V_i are the ones whose data distributions are most similar to the unseen.

$$V_i^* = \underset{V_i^j}{argmin}\, D(V_i, V_i^j) \tag{1}$$

We regard a visualization as informative if its distance falls within a user-defined threshold θ% close to its most informative parent:

$$V_i^{*,\theta} = \{V_i^j : \frac{D(V_i, V_i^*)}{D(V_i, V_i^j)} \geq \theta\} \tag{2}$$

For example in Figure 1, while both visualization Black and Female visualizations are considered parents of the Black Female visualization, only the Black visualization are considered the informative parent of the black female population, for any values of $\theta \geq 11\%$ via the Euclidean distance metric. Note that, our proposed system can work with different distance metrics such as cosine similarity and earth mover's distance. Without loss of generality, we chose to use Euclidean distance metric for the remainder of our paper.

REFERENCES

[1] Jessica Hullman, Steven Drucker, Nathalie Henry Riche, Bongshin Lee, Danyel Fisher, and Eytan Adar. 2013. A deeper understanding of sequence in narrative visualization. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2406–2415. **https://doi.org/10.1109/TVCG.2013.119**

[2] Jessica Hullman, Robert Kosara, and Heidi Lam. 2017. Finding a Clear Path : Structuring Strategies for Visualization Sequences. 36, 3 (2017).

[3] Younghoon Kim, Kanit Wongsuphasawat, Jessica Hullman, and Jeffrey Heer. 2017. GraphScape: A Model for Automated Reasoning about Visualization Similarity and Sequencing. *Proc. of ACM CHI 2017* (2017). **https://doi.org/10.1145/3025453.3025866**