Exploring Proportions: Comparative Visualization of Categorical Data

Harald Piringer[*]
VRVis Research Center, Vienna, Austria

Matthias Buchetics[†]
VRVis Research Center, Vienna, Austria

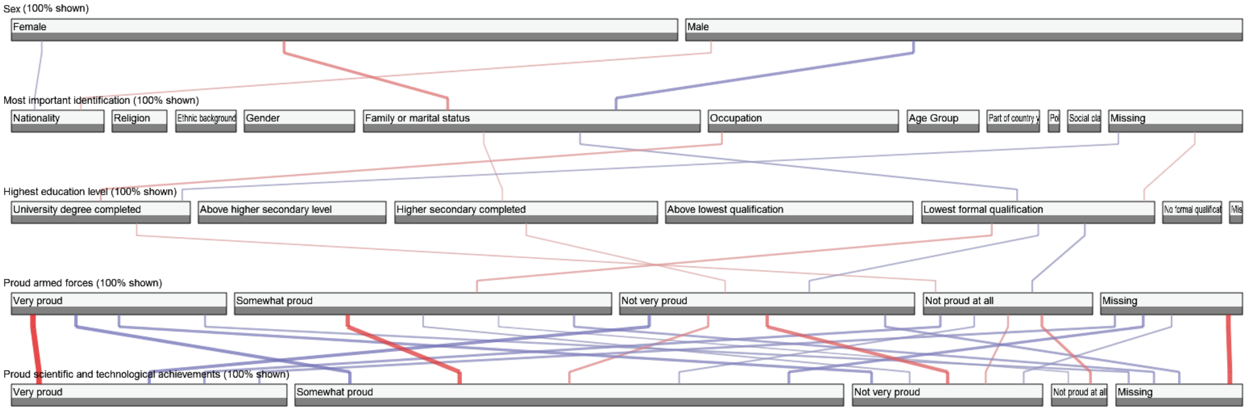

Figure 1: Five dimensions of a social survey. Red and blue lines indicate over- and under-proportional relationships between categories of adjacent axes by visualizing the measure *Lift* [1]. The visualization shows that people being proud of their nation's army tend to be proud about national scientific and technical achievements as well.

Abstract

This poster describes an approach to facilitate comparisons in multi-dimensional categorical data. The key idea is to represent over- or under-proportional relationships explicitly. On an overview level, the visualization of various measures conveys pair-wise relationships between categorical dimensions. For more details, interaction supports to relate a single category to all categories of multiple dimensions. We discuss methods for representing relationships and visualization-driven strategies for ordering dimensions and categories, and we illustrate the approach by means of data from a social survey.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Interaction Styles

1 Introduction

The identification of relationships between categories of different dimensions is an important and common task in statistics. Cross-tabulations [4] are a traditional method for relating two categorical dimensions. They typically provide the absolute and expected joint probabilities as well as both conditional probabilities for any pair of categories. Mosaic displays are a well-known method to graphically represent cross-tabulations [2]. They recursively subdivide space and represent frequency values of categories by areas on the screen. However, despite their usefulness and popularity, Mosaic displays require a ranking of the dimensions and do not scale well with respect to the number of displayed dimensions.

Parallel Sets [3] are an interactive approach to visualize cross-tabulations by combining the layout of parallel coordinates with a frequency-based visualization of the categories as scaled boxes. The width of parallelograms represents the joint probabilities $P(A \cap B)$ for all pairs of categories A, B in adjacent axes. However, joint probabilities themselves provide no information about the proportionality of a relationship. In fact, over- or under-proportional relationships are not easy to perceive from the parallelograms of Parallel Sets.

The goal of our work is to facilitate comparisons in multi-dimensional categorical data by representing over- or under-proportional relationships explicitly. Our approach supports this comparison at two levels: First, the visualization of various measures conveys an overview over pair-wise relationships between categorical dimensions. Second, interaction supports relating a single category to all categories of multiple dimensions. We illustrate the approach by means of social survey data concerning national consciousness and identity.

2 Overview of Proportional Relationships

The basic layout of our approach is motivated by Parallel Sets. Each dimension is represented as an axis representing 100% of the data which is subdivided with respect to the relative frequencies of the categories (see Fig. 1). Multiple dimensions are shown as parallel axes. In contrast to Parallel Sets, however, the user may choose between different statistical measures to be visualized as connections between pairs of categories: Support ($P(A \cap B)$), Confidence ($P(A \cap B)/P(A)$, also denoted as $P(B|A)$), Lift [1] ($P(A \cap B) \ / \ (P(A) * P(B))$), Difference ($P(A \cap B) - (P(A) * P(B))$), and Degree of Independence (DoI) ($P(B|A) - P(B)$). A key distinction between these measures is the sensitivity to category sizes. For example, Lift may generate misleading results for small categories, while Difference may only become large in case of significant support.

For the non-symmetric measures Confidence and DoI, the category A is assumed to belong to the upper axis. The result for each pair is mapped to the color and the width of an according connection line. For lift, difference and DoI, the hue indicates the "sign" of the relationship. We apply red to over-proportional associations (i.e., lift > 1, difference > 0, DoI > 0). Under-proportional relations are optional, as they often represent the reverse conclusion and thus usually convey little new information (if visualized, they are shown in blue). For support and confidence, there is no sign

[*] e-mail: hp@vrvis.at

[†] e-mail:buchetics@vrvis.at

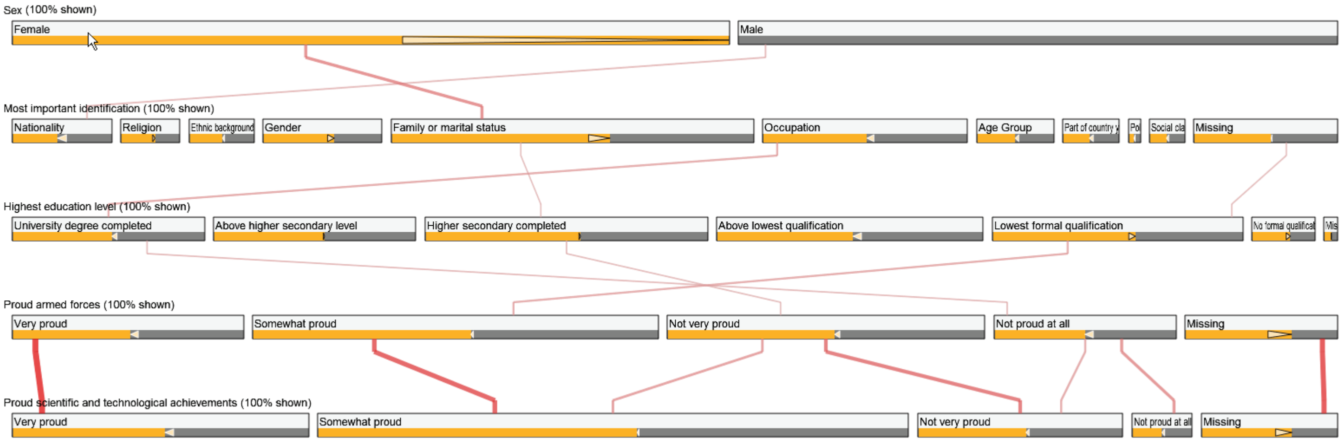

Figure 2: Relating the category ”female” to four other dimensions. Orange bars show the percentages of entries per category and arrows indicate the over- and under-proportional relationships. As one result, women left questions concerning pride unanswered more often than men.

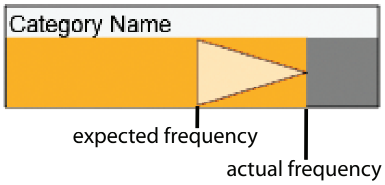

Figure 3: Arrows visualize the difference between expected and actual frequencies.

and lines are drawn in shades of gray. The intensity of the respective color and the width of the line are derived from the distinctness of the association. The result range necessary for this mapping is inherent in all cases apart from positive values of lift, where results greater than 4 are clamped to 4 as a heuristic (results of 4 already indicate very strong associations and only occur for small categories). In order to reduce clutter and to focus on significant relationships, the user may hide results below a certain user-defined threshold. Altering this threshold is an effective way for trading off the amount of displayed information with visual complexity.

Both the order of the axes as well as the order of the categories within one axis are important for an effective visualization. Apart from interactive reordering by drag and drop, various automatic strategies may emphasize different aspects. Concerning the order of axes, one strategy is to put strongly related dimensions next to each other. To do so, we determine the absolute maximal and the absolute average result of the current measure for all pairs of dimensions. These values can be used to select dimensions iteratively for ranking them from top to bottom. An interesting option is to specify the first dimension as a given starting point.

Concerning the order of categories, a visualization-driven strategy is to minimize the length of strong connection lines. More precisely, the user may reorder the categories of one axis at a time with respect to one neighboring axis. As finding an optimal solution becomes too computationally expensive for a growing number of categories, we apply a heuristic algorithm. However, a detailed description of this algorithm is beyond the scope of this poster.

3 Interactive Comparison of Categories

In addition to comparing adjacent axes as descibed above, our approach also supports to relate single categories to all dimensions simultaneously (see Fig. 2). Moving the cursor above any category C displays an orange bar in all other categories. For each category X, the length of this bar corresponds to the percentage of entries of C, i.e., the joint probability $P(C \cap X)$. Drawing such bars is standard but does not provide information about proportional relationships.

Therefore, each category X also visualizes the difference to the expected joint probability $P(C * X)$ as an arrow (see Fig. 3). The direction of the arrow indicates the sign of the proportionality (left means under-proportional, right means over-proportional), and the size corresponds to the distinctness. Due to its simplicity, this technique has turned out to be both intuitive and efficient. Moving the cursor along an axis supports exploring relationships consecutively for all categories within a short time.

4 Discussion and Future Work

We collected first feedback from a sociologist, who frequently analyzes survey data. He considers the fast visual identification of over- and under-proportional relationships in the context of absolute frequencies of categories as a key advantage of our approach. In joint analysis sessions, the visualization provided useful overviews for up to 10 dimensions. The sociologist also considered the arrows intuitive and potentially suitable to improve the identification of relationships in different frequency-based visualizations (e.g., bar charts and pie charts).

As the most important disadvantage, the layout somewhat limits the number of categories per dimension to 10 – 30 (depending on the screen resolution and the distribution of frequencies). Very small categories may also become hard to read. We intend to address these issues in future work by providing means for grouping or filtering categories as well as by adding lenses. Another aspect of future work concerns a more thorough evaluation in different application domains. Finally, while connection lines are now only drawn between adjacent dimensions, an interesting aspect for future work would be to split connection lines by an additional ”active” dimension in a similar way as Parallel Sets.

References

[1] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. In *SIGMOD ’97: Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, pages 255–264, New York, NY, USA, 1997. ACM Press.

[2] M. Friendly. Visualizing categorical data: Data, stories and pictures. In *SAS User Group International Conf. Proceedings*, pages 190–200, 2000.

[3] R. Kosara, F. Bendix, and H. Hauser. Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 12(4):558–568, 2006.

[4] StatSoft Inc. Electronic Statistics Textbook. `http://www.statsoft.com/textbook/stathome.html`, 2007.