Heatmap matrix: a multidimensional data visualization technique

Miguel Mechi Naves Rocha  
School of Technology  
University of Campinas  
Limeira, SP 13484-332  
Email: m156778@dac.unicamp.br

Celmar Guimarães da Silva  
School of Technology  
University of Campinas  
Limeira, SP 13484-332  
Email: celmar@ft.unicamp.br

Abstract—Information Visualization literature presents lots of multidimensional data visualization techniques. We propose a new technique, called heatmap matrix, which aims to show relationships between pairs of categorical or ordinal variables. A heatmap matrix is composed by a matrix of heatmaps that represent frequency of occurrence of values from all possible pairs of variables of a dataset. In this paper we define this technique and illustrate how it may be used for analyzing some scenarios.

I. INTRODUCTION

Information Visualization (InfoVis) research area has been developing lots of interactive visual representations that aim to help users to understand data [1] [2] [3], even when they are part of datasets with high dimensionality and cardinality.

We highlight the role of InfoVis when the dataset to be analyzed is multidimensional. Multidimensional visualization techniques receive increasing interest and visibility from InfoVis researchers, given that they make available fast and simple ways for visual recognition of patterns and trends, and for observation of dataset’s non-obvious characteristics [4].

In our research we are particularly interested on techniques for visualizing multidimensional categorical or ordinal data. We propose heatmap matrix as a visualization technique for this kind of datasets. It combines multiple juxtaposed displays of pairs of categorical or ordinal variables, similar to those used by scatterplot matrices [2], but with heatmaps instead of scatterplots. In this paper, we define how this technique works and exemplify its use in some real world scenarios. An extended version of this work is published elsewhere [5].

This paper is organized as follows: Section II presents some related works on multidimensional visualizations; Section III explains our proposed technique; Section IV presents the tool that implements this technique and its use to analyze real-world examples; and Section V concludes our paper and points out future works.

II. RELATED WORKS

Our technique has some similarities with other classic multidimensional visualization techniques. We highlight dimensional stacking, scatterplot matrix and parallel sets.

Dimensional stacking [6] is a 2D or 3D projection of a dataset, with embedded groups of dimensions. We focus on 2D dimensional stacking, which splits the plane in rectangles. Given $I_1$, $I_2$, $I_3$ and $I_4$ as independent variables, and $D$ as a dependent one, the external X and Y axes are mapped to $I_1$ and $I_2$, respectively. Each cell of this main graphic has an internal graphic that maps its X and Y axes to $I_3$ and $I_4$, respectively. The cells of the internal graphics are painted according to $D$. When a dependent variable does not exist, a default color is used. Scenarios with more than four independent variables should use other groups of internal graphics inside the $I_3 \times I_4$ graphics. Similarly to dimensional stacking, our technique uses the concept of matrices inside a matrix, but with a different visual mapping, as presented in the following section.

A scatterplot matrix (or SPLOM) [7] shows simultaneously multiple views of a dataset. This matrix represents a dataset with $n$ variables in a $n \times n$ table. Each cell of this table has a scatterplot whose axes map two variables of the dataset. The use of brushing technique enables users to select a set of points in one scatterplot, which highlights related points in the other views. Our technique also uses a $n \times n$ table. However, SPLOM is focused on quantitative variables, whereas our technique focuses on categorical and ordinal variables.

Parallel sets technique [8] is inspired by parallel coordinates plot (PCP) [9]. PCP represents each tuple of a dataset as a polygonal line. On the other hand, parallel sets technique represents frequency of occurrence of values in tuples. It fits a multidimensional scenario with categorical variables. Due to value aggregation, it is able to process datasets with high cardinality without occlusion problems that are recurrent in PCP. Parallel sets technique provides one axis for each categorical variable, and parallelograms represent the relationship between the categories of each pair of neighbor axes, given by frequency values. This technique is similar to our own because it presents frequency of occurrence of values. However, the presented frequencies are related only to pairs of variables that are neighbor in the visual mapping. On the other hand, our technique presents frequencies related to all possible pairs of variables, as depicted in the following sections.

III. HEATMAP MATRIX TECHNIQUE

A heatmap matrix is a visual structure that represents a multidimensional dataset. It is inspired on scatterplot matrix and heatmap techniques.
This structure’s visual mapping is as follows. A heatmap matrix that represents a $n$-variable dataset has $n \times n$ cells, which are heatmaps. The column labels of each heatmap contain distinct values of a variable $V_i$. Similarly, the row labels of each heatmap have distinct values of a variable $V_j$. Each cell belonging to a heatmap represents the frequency of occurrence of the pair of values $(v_i, v_j)$ in the dataset, where $v_i \in V_i$ and $v_j \in V_j$. Colors or numerical values represent the frequency values.

IV. RESULTS

Fig. 1 exemplifies our technique with the Titanic dataset. This dataset refers to the tragic accident of the Titanic passenger ship in April 14, 1912. The dataset has 2202 tuples, and its variables are: passenger’s age (child or adult), sex (male or female), class (first-class, second-class, third-class, or crew) and if the passenger survived or not. A red-white-blue palette indicates high, mean and low frequency values. For example, the cell in the row Sex and column Survived at the external

1These figures correspond to the most up-to-date versions of the heatmap matrix. Rocha elaborated them after completing his undergraduate course, taking into account modifications suggested by the reviewers of this paper.

A. Heatmap Matrix Tool

We developed a tool that implements our technique (Figs. 2 and 3). It works as follows. First, the user must indicate a CSV file in the data path section (top-left corner of the tool, see Fig. 2b) and press button “Generate heatmap matrix”. The tool draws a heatmap matrix of the input dataset in the center of the screen. Another option is the “Original Data” tab, which shows the loaded dataset as it is.

B. Limitations

The proposed technique has some limitations. A first limitation is the number of variables to be plotted in the same screen (SPLOM has a similar limitation). A user that wants to visualize a dataset with high dimensionality in this technique must choose to show only a subset of the dataset variables. Besides, the total number of distinct values of all dataset variables impacts the size of the visualization or its resolution. Both limitations may hamper user understanding.

Two other limitations are the absence of a color legend to improve readability, and the absence of support to quantitative variables.

References


Acknowledgment

This work was supported by São Paulo Research Foundation (FAPESP) (grant number #2015/14854-7).
(a) Overview.

(b) Detailed view of the top-left corner of the tool.

Fig. 2. The Heatmap Matrix Tool.

Fig. 3. Heatmap Matrix Tool visualization of the homicide dataset.