Visualizing Relationships among Categorical Variables

Seth Horrigan

Abstract—Centuries of chart-making have produced some outstanding charts tailored specifically to the data being visualized. They have also produced a myriad of less-than-outstanding charts in the same vein. I instead present a set of techniques that may be applied to arbitrary datasets with specific properties. In particular, I describe two techniques – Nested Category Maps and Correlation Maps – for visualizing, analyzing, and exploring multi-dimensional sets of categorical and ordinal data. I also describe an implementation of these two techniques.

Index Terms—information visualization, questionnaires, multi-dimensional data visualization, statistical analysis, treemaps

1 Introduction

Many surveys, both professional and amateur, are based on banks of questions presented in a questionnaire. These surveys may be created and distributed by major institutions or they may be impromptu constructions from students using tools like SurveyMonkey [1], Zoomerang [2], or QuestionPro [3]. Information collected by institutions like the Pew Charitable Trust in their annual Pew Internet Survey undergoes a great deal of educated, thorough analysis. Statisticians can make entire careers from analyzing the results of this data and subsequently drawing and publishing conclusions based on the data. Marketing researchers will collect information about potential customers or reviews of new and existing products using questionnaires – be it online, in malls, on street corners, or by random digit telephone dialing.

Basic social science statistics such as pair-wise correlations, chi-square tests of independence, and analysis of variance (ANOVA) can reveal vital information hidden beneath the distribution of answers.

Unfortunately, this data is seldom presented in a format that makes visual exploration simple. Researchers customarily have specific correlations they expect and they confirm or disprove their hypotheses by testing the empirically obtained numeric values against the expectations. Cross tables of raw sums constrained by responses on related variables can reveal much information to the highly trained eye, and statistical packages such as STATA and SPSS provide simple ways to issue these queries, thus providing a limited degree of interactive exploration [4, 5]. The increase in processing power of personal computers has allowed such comparisons to be rendered on-demand in near real time. Still, in all these cases, the data is seen as banks of row upon row of numbers and text.

1.1 Analysis

The communities built around this data have become highly skilled at analyzing these numbers and running the proper tests to find out the information they expect as well as occasionally finding unexpected results that warrant further study. Unfortunately, many potential interesting comparisons may go completely ignored simply for lack of a skilled analyst with the time and motivation to thoroughly explore the dataset.

This problem is compounded when one considers as well the staggering number of surveys conducted by non-experts using ready-made tools like SurveyMonkey. Such sites provide very simple aggregation of numbers according to question, which allows unskilled investigators to identify basic trends in response, but offers little or none of the more interesting comparison of interrelation among responses (see Fig. 4). Happily, in most cases the data collected via the online tool can be exported to common spread-sheet applications such as Microsoft Excel, or in commonly shared formats like Comma Separated Value files for analysis later. When the data is collected through secondary agencies or directly via paper questionnaires it will likely also be recorded and distributed in spreadsheet formats that could be analyzed given the proper tools.

Seth Horrigan is with the Berkeley Institute of Design in the Department of Electrical Engineering and Computer Science at the University of California at Berkeley, E-Mail: eomer@cs.berkeley.edu.

1.1.1 Textual

Many of the questions on such questionnaires are open-ended, "free-response" inquiries. Answers to such questions are notoriously difficult to analyze and categorize. Often analysts will sort through them searching for keywords, or subjectively categorizing each response. If the number of respondents is small enough, humans can manually parse all individual responses and present their own subjective evaluation of the responses in aggregate, but as the number of respondents grows, this becomes an increasingly daunting task.

With the growth of the internet, the question of visualizing large corpora of computerized text becomes ever more important. Research in this area has produced very useful techniques like Word Trees, ThemeRiver, and TextArc [6, 7, 8]. ManyEyes, in particular, provides an interface for employing such techniques to visualize arbitrary datasets [6]. Applied correctly, such textual visualization methods can be used to visualize and explore the results of free-response survey questions - an invaluable tool when the number of responses grows far too large to analyze manually.

1.1.2 Interval

Due to the complexity of interacting with, summarizing, exploring, and quantifying large numbers of free-form textual responses, when the expected number of respondents is large, survey designers often attempt to construct the survey in such a way that the responses can be easily represented numerically and analyzed using the statistical methods mentioned earlier. Certain types of inquiries, such as the respondent's age or the number of hours spent weekly washing dishes, lend themselves to numerical definition. These interval variables allow robust interaction and aggregation. Their continuous nature lends itself to representing the values using simple two-dimensional encodings like scatterplots and line graphs that rely on position according to a specific x-y grid (see Fig. 1). Such interval variables allow analysts to quickly identify groupings along the continuum of possible variables. For example, they may identify that, although respondents can specify any number of hours a week for dishwashing, they generally grouped themselves into around 3 hours or around 6 hours with the number corresponding to the respondents' age.

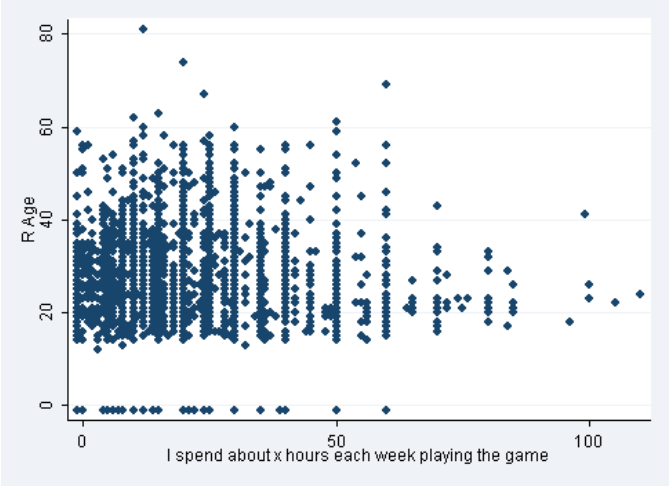

Fig. 3. Scatterplot of two interval variables, as produced by STATA

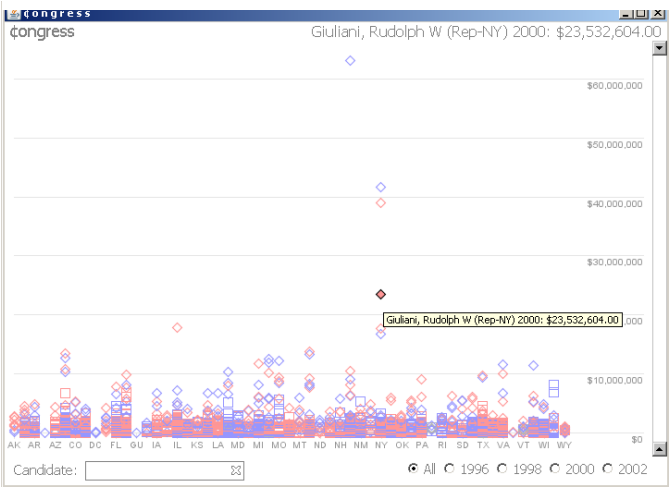

Fig. 2. Scatterplot comparing state with campaign contributions, using color to encode political party and shape for office

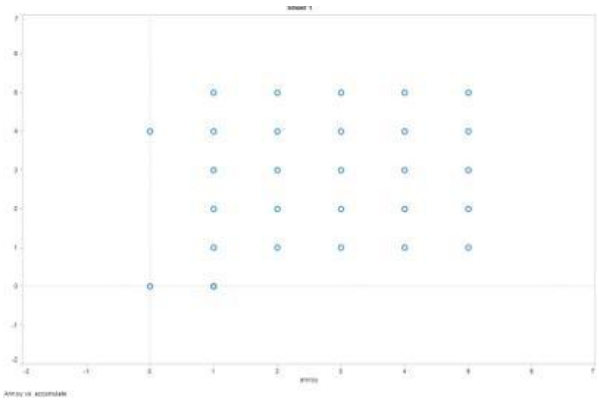

Fig. 4. Scatterplot of two ordinal variables

When one wishes to see more than two variables on a single chart, the task becomes slightly more complex. Scatterplots rely on coordinates along the two axes to encode data. This limitation can be surpassed by employing alternative encodings such as size, value (shading), texture, or shape [9]. Certain of these other methods of encoding data map well onto interval variables, while others do not. For limited ranges, value along a gradient can be useful. Size can encode continuous variables as well, although as size grows, the

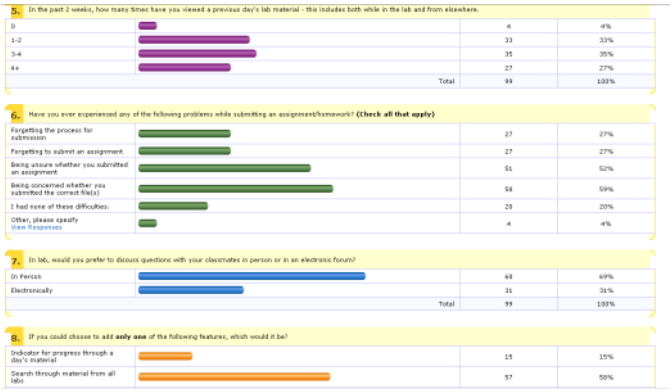

Fig. 1. Bar charts of questionnaire responses to ordinal questions

chance of occluding other data on the plot grows as well. Additionally, while humans are fairly reliable when gauging differences in length, they are generally fairly poor at gauging differences in area or volume [10]. Textures, shapes, and colors, however, do not work well for encoding any sort of continuous value – e.g. if 13 is square and 45 is round, what are 33, 34, and 35? Shapes, colors, and textures can be very useful, however in encoding categorical variables - ordinal and nominal (see Fig. 2).

1.1.3 Nominal and Ordinal

Fortunately, many of the questions found on survey questionnaires are in fact nominal or ordinal. Since answers to categorical variables fall naturally into a finite, usually small, number of possibly categories, they can be mapped directly onto colors, shapes or textures. If the categories have an inherent ordering, that is, if they are ordinal categorical variables, it can be slightly more complex since it is not clear whether green is greater than cyan, or square is less than triangle. In these cases, either the ordinal characteristics may be ignored, or order can be conveyed in alternative methods such as encoding the values at intervals along the spectrum between blue and green.

When the number of variables on the chart exceeds a small number - three in Bertin's lexicon but more commonly around six or seven - it becomes difficult to present and interpret all data simultaneously [9]. Providing interactivity such as zooming, filtering, or varying the display over time, more data can be presented within a single interface. When given a large number of interval variables to visualize, researchers created an impressive myriad of advanced techniques to encode them within a single static graphic. Some of these will be discussed in the related work section later. These methods, though, are specifically designed for interval variables, and often do not work as well when applied to categorical variables.

1.2 Questionnaires

Since questionnaire designers often want specific quantitative results from the surveys they produce, categorical and Likert-type questions are ideal. The Likert scale, designed by Rensis Likert in 1932, offers a range of five values from "Strongly Disagree" to "Strongly Agree" and has been a staple of quantitative social science research for many years [11]. Through extensive use, these scales have become commonplace even outside of social science research [12]. They have also been adapted to display scales other than simple agreement. For example, ranges from "not much" to "always", or "do not enjoy" to "tremendously enjoy". Almost all questionnaire-building tools offer functionality to specify these sorts of Likert-type questions. Because Likert-type questions are not interval variables, since there is not a clearly defined gap between "agree" and "strongly agree" and since answers to these questions are subjectively ambiguous, they cannot be reliably analyzed or visualized using methods designed for interval variables [13]. They

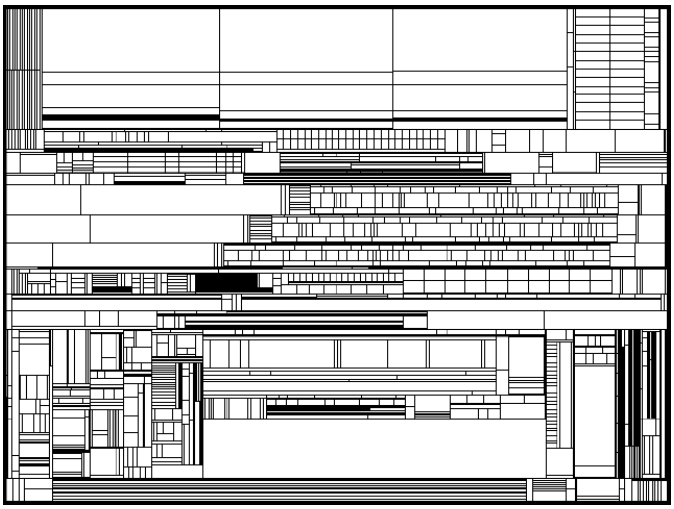

Fig. 5. Treemap of file system before squarifying

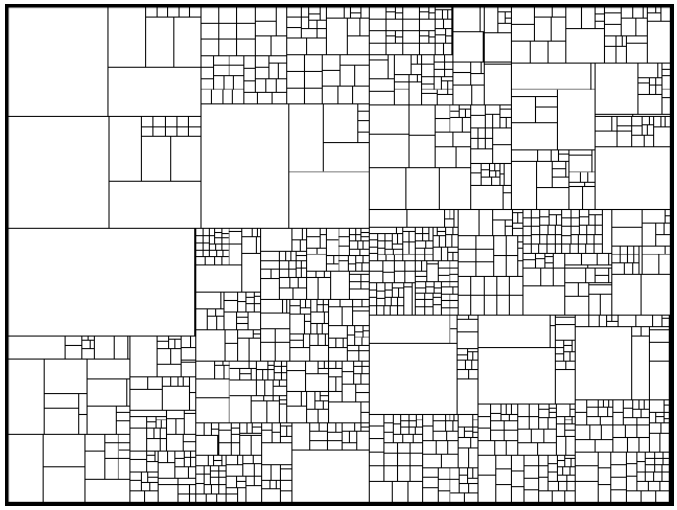

Fig. 6. Treemap of file system after squarifying

are, however, ordered variables. These Likert-type ordinal-polytomous questions and nominal-polytomous or dichotomous questions, such as gender, race, or yes/no questions, make up a large portion of many questionnaire surveys. The data from categorical variables can be summed by category and compared or manipulated numerically.

Such categorical data can sometimes be visually compared with interval variables quite well (see Fig. 2), but when applied to two categorical variables, positional encodings like scatterplots fail to convey much information (see Fig. 3). For analysts to visually investigate relationships among categorical variables, alternative methods of data visualization are necessary. Certain of the survey tools mentioned earlier provide some data visualization for these types of questions by constraining the view to a single variable at a time. This allows even untrained individuals to perceive within-variable trends (see Fig. 4). Such limited visualizations do not offer any support for data exploration, nor do they illustrate relations among the variables. Viewers can guess possible relationships by observing multiple bar charts sequentially, but this is hardly optimal.

2 SOLUTIONS

In the words of Daniel Keim, "For data mining to be effective, it is important to include the human in the data exploration process and combine the flexibility, creativity, and general knowledge of the human with the enormous storage capacity and the computational power of today's computers [14]."

Visual data exploration is especially useful when little is known about the data set or when the expected results are vague. In this case, the ability to see the data and how it interrelates allows humans to identify interesting trends to either confirm or explore further. Although testing for statistical significance within a dataset will likely eliminate the chance that any correlations identified are due to random chance, unexpected results should usually be taken as grounds for further exploration, not as results to report. A well designed survey will clearly confirm or refute a hypothesis; normally the hypothesis should be formed before analyzing the data, not derived from the experimental results. That said, visual exploration does allow humans to quickly form hypothesis about the data based on their perception of its qualities, and these hypotheses can be integral to drawing important conclusions that would otherwise go wholly unnoticed.

Visual exploration leverages the cognitive abilities of humans to fill the gaps that automatic data mining or statistical machine learning cannot. It also allows the researchers to think critically about the significance of specific relations in a way that computers currently cannot.

I sought to address the question of how best to visualize many categorical variables and their interrelation. I have designed two distinct visualization methods for categorical variables and have built a system that I call Survis to demonstrate the concepts.

Underlying the design of each is the so-called Information Seeking Mantra, "Overview first, zoom and filter, and then details-on-demand [15]." Initially, each of the visualizations provides an overview of the data to allow the observer a chance to assess the data as a whole and identify important or relevant trends. Where ever possible, the system provides further details about each component of the visualization via tooltips, and also allowing the user to filter and zoom to specific areas of interest. Throughout the visualizations, the implementation attempts to provide a sense of "location" within the dataset so that analysts can navigate to and from items of interest.

2.1 Nested Category Maps

Treemaps are a form of stacked display created by Ben Schneiderman in the early 1990s [16]. The initial motivation was to visualize the usage of disk space on personal computers. The idea is to provide a means of visualizing a tree hierarchy in a space-constrained layout. The design splits the screen space into ever-smaller rectangles as it traverses down the tree. In the end, this produces a visualization of the tree encoding the position within the hierarchy using size and position. Lower nodes are nested within higher nodes so the size of any rectangle represents the number of descendants it has, and all descendents are contained within the rectangle of the ancestor (see Fig. 5).

Over the past decade and a half, various parties have made improvements to this initial idea. In particular, by "squarifying" the rectangles of the map it improves the human readability of the map substantially, especially in perceiving the actual hierarchy of elements (see Fig. and Fig.) [17].

Nested category maps employ the structure of squarified treemaps in order to visualize relationships among various categorical variables. They allow the analyst to specify the desired hierarchy and with which to order the non-hierarchical data, allowing the analyst to see the composition of answers as they relate to the other variables visualized. However, treemaps are just that: a "planar space-filling map" of a tree [16]. In order to use that structure visualize non-hierarchical data, it is necessary to first impose an artificial hierarchy on it.

Assuming the data is stored in a tabular format, this is accomplished iteratively. An empty root is established for the tree. Then the values specified for each respondent – each row of the table – are added successively to the tree. The internal structure of the tree corresponds to the values in each column of the row. The permutation of values in each row is considered the path through the tree to that node. As new values in a specific column are found, they are added to the internal nodes at the level corresponding to the column. If the value already exists in the tree, it reused and the

percents	optimize	template	collecting
Extremely Interested	Very Important	Often	A Little
Extremely Interested	Extremely Important	Always	A Little
Extremely Interested	Very Important	Often	Not At All
Somewhat Interested	Very Important	Seldom	Some
Somewhat Interested	Very Important	Seldom	Some

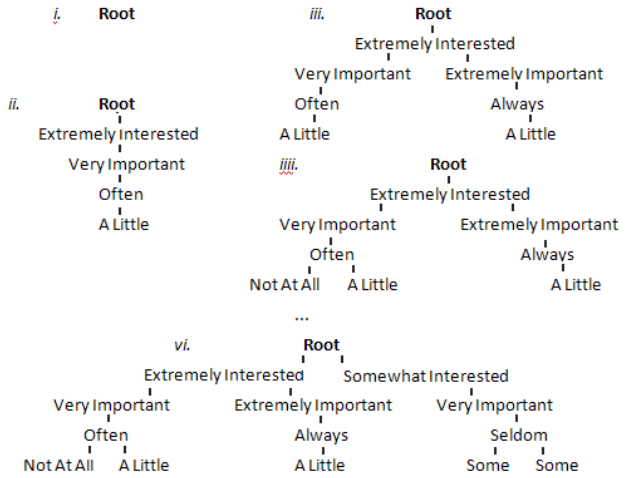

Fig. 5. Construction of tree from categorical data

search down the tree continues. If there are few enough variables, then the last column in the table corresponds to the leaves of the tree, which will be added to the proper parents; else, the iterative process will stop when the specified depth is reached and that column will become the leaves of the tree (see Fig. 7).

As should be apparent from figure 7, the tree structure will lose its visual usefulness unless there are multiple leaves under each node at the second-to-last level. At the point that the paths through the tree deteriorate to having only a single leaf each, the visualization will becomes no more than a series of boxes all the same size, conveying no useful information. Up to that point it is interesting to visualize the distribution.

Experimentally, it appears that given a reasonable distribution of responses, it is most useful to visualize no more than $\log_k n$ - 1 dimensions at once, where k is the degree of the nodes – the number of possible answers to the question – and n is the number of responses. Beyond this number it becomes difficult to distinguish between levels, especially when boxes are overlaid with text describing their content. As such, nested category maps provide two forms of filtering to explore variables not visualized initially.

First, they allow dynamic reordering of variables so that analysts can quickly change the variables currently being displayed. Second, they allow drill-down to further details. When an analyst sees a particular top-level region of interest, he is able to select that as the focus, thus filtering the displayed respondents to only those who answered the top level question as specified. The hierarchy is visually established by the nesting of squares. The nesting is emphasized by decreasing the width of the lines delineating the squares, and decreasing the size and value of the textual labels.

This drill-down provides a new nested category map constrained by the filter. If the analyst sees something interesting, he can then drill down further, rearrange the displayed variables, or return to the previous map. In this manner the nested category maps are also a sort of zooming user interface [15]; however, each level that the analyst zooms in decreases the number of respondents until he either finds that there is only one respondent who meets that particular criteria or all respondents at that level are homogenous. In practice, then, the depth of possible zooming is determined both by the number of respondents and the variety of responses. Too little variety will reach homogeneity at a given level quickly. Too much variety will cause the filter to quickly reduce to only one possible respondent.

There is one key insight that the initial implementation does fully not address. In order to provide the most usefulness, the nested category map should provide a sense of location within the hierarchy at all times. This is accomplished by providing bread crumbs showing the path taken to the current visualization – already implemented, and by maintaining the same overall layout as the analyst drills down – not implemented yet. Each new level of the nested category map should be an expansion of the selected block from the previous level, but at present the implementation lays out the structure from scratch rather than just expanding the block to the whole screen and adding one further layer.

2.2 Correlation Maps

Highlight tables are a concept recently developed by Tableau Software [18]. They behave like heat maps applied to textual tables – using color and saturation to identify the magnitude of values within a cell. Correlation maps overlay the concept of small multiples and basic statistical analysis with the idea of highlight tables to produce a visualization to convey a wealth of information at various levels of detail.

A correlation map is composed of a set of tiles. Each tile offers a comparison between two categorical variables. When applicable, the tiles are colored according to the significance and strength of correlation between those two different variables, hence the name "correlation map." A simple coloring of the squares could convey the correlation between two variables, but this conveys little information in and of itself. Hence, each tile also presents a very small graphical representation of the comparison of the two variables. This graphical tiling can be accomplished in two ways, both of which have advantages and disadvantages.

In the first case, each tile can be represented as a grid of values. Each row corresponds to a particular value along the y axis, and each column corresponds to a value along the x axis. Responses are plotted at the intersection of the values. In this sense, it is similar to a scatterplot. However, as illustrated earlier, a scatterplot fails to distinguish the number of respondents at the intersections of the values. Introducing sufficient jitter can convey the information, but this does not work well in the very small space allotted to the tiles of the correlation map. Instead, the correlation tiles apply the idea of bar graphs. At each intersection, a separate bar plotted, with the height of the bar corresponding to the number of respondents who fit that particular combination of responses. Unfortunately, this relies on the length of the bars to encode the information and the possible variation in the bars depends greatly on the number of possible categories within the tile. If there are three categories along the y axis, the height of the tile is divided into thirds and then the thirds are apportioned to the bars encoding the information. This is reasonable. However, if there are twelve possible categories, the tile is divided into twelve. Supposing a tile height of 50 pixels, and leaving one pixel between each part of the grid for visual delineation, we are left with *(50 – 11) / 12 = 3.25* pixels for each bar. It is difficult to convey much information using length when there are only three pixels to adjust.

An alternative allows the tile to use its full height to encode information but at the cost of constraining the number of categories that can be conveyed. By dividing the tile into only columns instead of a grid of rows and columns, the entirety of the vertical space may be used to encode information. In this case, each row becomes a column, and within that column each column of the former row becomes a sub-column; thus, each value has the full 50 pixels to encode data. Unfortunately, this means that each part of the grid must be laid out linearly. Supposing there are five categories in each of the two variables visualized then we must graph 25 bars (five columns of five bars), and leaving one pixel to visually delineate the outer columns, we are left with *(50 – 4) / 25 = 1.76* pixels to encode each bar. This seem reasonable enough, except that this leaves no

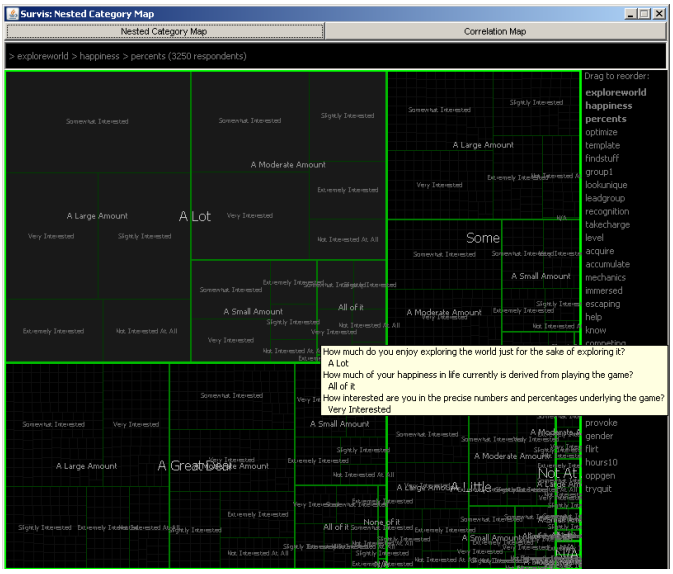

Fig. 8. Nested category map drilled down to show only those who enjoy exploring the world "A Lot" (1262 of the 3250 respondents)

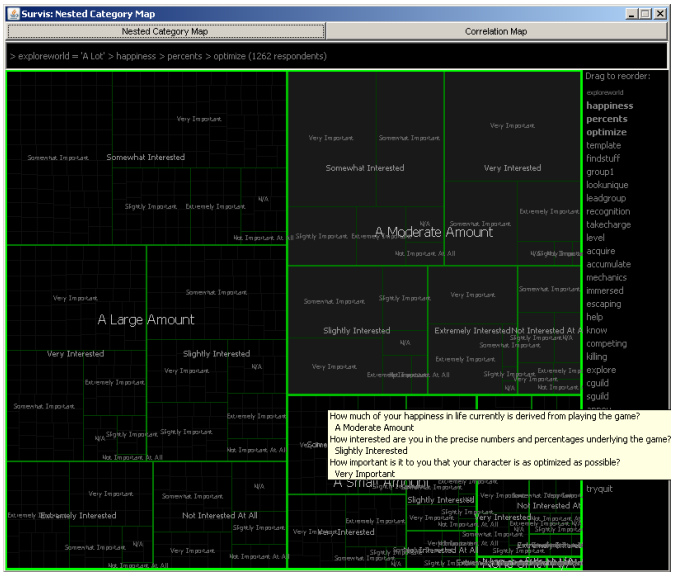

Fig. 9. Nested category map at the highest level

room for another pixel between each bar to improve perception, and it means that it is impossible to visualize more than seven categories on each axis in 50 pixels – *(50 – 6) / 49 = 0.90*. Still for a small number of categories, such as a traditional Likert-type scale, either of these types of tiles will suffice.

In order for the visualization to take advantage of the cognitive benefits of small multiples, though, some other measures are required that tip the balance slightly in favor of the second method. In "The Visual Display of Quantitative Information," Tufte states "small multiples resemble the frames of a movie: a series of graphics showing the same combination of variables, indexed by changes in another variable...the design remains constant through all the frames, so that attention is devoted entirely to shifts in the data [10]." Correlation maps attempt to use this principle to allow viewers to identify similarities and changes in the distribution of data.

In order to focus on the changes between tiles, it is necessary first to scale the tiles in such a way that the changes are predictable and significant. The scaling for a tile depends on the number of possible categories. For example, all 3x6 tiles should be scaled the same. All 6x3 tiles should be scaled the same, but not necessarily the same as the 3x6 tiles. All 6x6 tiles should be scaled the same. As for the scaling itself, the maximum value for any single bar in any tile of the specified dimension should be used as the scaling constant. That is, each bar's length is determined by *h * (n / m)* where h is the height of the tile, n is the respondent count for the column and m is the maximum count from any similar tile. This ensures that all bars will be less than or equal to in length the total possible length for a bar of that type. This also means that a single tile with an irregularly large number of respondents in a single bar can compress all similar tiles. If the tile is subdivided into five rows, this can prevent the grid-based tiles from displaying any useful information. Happily, the linear tiles have the full height of the tile to distribute, meaning that even suppressed, the difference in heights is still apparent even given a small tile space.

The initial implementation of correlation maps uses red and green to encode statistical significance; however, since red-green color-blindness is relatively common in males, perhaps alternative encodings are preferable. The design uses both the red-green hue and the color saturation to encode information. Statistical significance is determined using Spearman's *r* for rank ordered variables. Pearson's *r*, most commonly used when determining statistical significance in social science, assumes two continuous interval variables (although not necessarily ratio variables) and a normal distribution. Spearman's *r* is similar, but specifically accounts for the fact that the ranking of rank ordered variables is not necessarily a regular measure of interval. As such, it produces a correlation co-efficient, *r*, describing the correlation line between any two ordered categorical variables. It can also be applied to dichotomous variables, since any dichotomous variable can be considered ordered.

The correlation co-efficient is actually directional, with a negative co-efficient corresponding to a negative correlation, but correlation maps do not visually encode this difference. Rather, the hue of each tile is determined by statistical significance of Spearman's *r* at $p = 0.01$. That is, tiles are colored green if there is less than a one percent chance that the correlation found is random. Tiles are given a red hue if there is greater than a one percent chance that any correlation found is due to random chance. Most often, p of 0.05, or possibly even $p = 0.10$, is reported, but in the initial implementation, the significance threshold is fixed at $p = 0.01$. This is to reduce the chance of individuals being overwhelmed by color encoding weak correlations. It would also be possible to allow a variable significance threshold for those analysts who would prefer to see weaker (or only see stronger) correlations.

The saturation (intensity) of the color is determined by r^2. Since r is the co-efficient of the correlation line, r^2 is used as a measure of the strength of the correlation. $r^2 = 1.0$ indicates a perfect correlation, and $r^2 = 0.0$ indicates no correlation. As r^2 increases for statistically significant values, the saturation of the tiles moves closer to 0.75. A tile is never saturated to 1.0 since the readability of the graph decreases as the background moves closer to a saturated color. Likewise, as r^2 shrinks further from statistical significance the red saturation of the tile increases towards 0.75. Colors that are closer to the cusp of statistical significance appear in light pastels or nearly white, drawing attention to the extreme values in the chart. Perhaps only statistically significant values should be colored, and the red should be removed from the chart, emphasizing only to the green tiles, but this makes it difficult for analysts to tell what is just barely statistically significant from that which is not even close.

The variables visualized are laid out along the two axes and tile at the intersection of the two variables encodes the comparison and significance of the correlation. Correlation tiles are added up to the identity line (see Fig. below). At the identity line, a miniature bar graph describing that variable is displayed - uncolored, since the correlation would always be perfect and thus would indicate nothing. Past the identity line, tiles are not added to the visualization as this would only be an unnecessary and distracting repetition of the comparisons already visualized.

The tiles of the correlation map are initially laid out in the order specified by the input spreadsheet. This choice assumes the variables

Fig. 10. Correlation map at the identity line, showing 3x3, 3x6, and 6x6 tiles as well as bar charts

Fig. 11. Details display for a single 6x6 tile

will be listed in the order presented on the questionnaire and that this represents a logical ordering. The correlation map should allow dynamic reordering of variables and subsequent rearrangement of tiles. This would not visualize any new information, but it may allow analysts to visually group logically similar variables. The initial implementation of correlation maps does not yet support this functionality, as mentioned later.

Like the nested category map, the correlation map is structured to support the information seeking mantra. The initial view provides an overview of the whole of the information. Individuals can then identify specific areas or tiles that interest them and find out more details through tooltips or through detailed, expanded views of the tile data (see Fig. 11). The correlation map should also allow degrees of zooming from a highest level where the tiles contain only color-coding and no graphic representation of the underlying data, to only viewing a single tile; however, the current implementation does not yet support such zooming. The only filtering support offered at present is the ability to move the viewing lens around the tiles to view any square subset of them. The expanded view of the tile, shown when the analyst selects it through a mouse click, does provide a limited method of zooming.

3 Implementation

Survis, shown in figures 8 to 11, is an example implementation of most of the functionality of nested category maps and correlation maps. It is coded entirely in Java. It uses the Prefuse toolkit for parsing data from spreadsheets and for the basic squarified treemap layout [19]. For the correlation map, and most of the other visual functionality, Survis uses the Swing toolkit from Sun Microsystems' Java Foundation Classes. The code is open source and freely available for download.

The example visualizations are constructed from questionnaire data collected in phase 17 of Nick Yee's Daedalus project, comprising responses from 3250 players of massively multi-multiplayer online games. Phase 17 employed a battery of sixty-one questions (fifty-eight categorical ones) addressing issues related to game-play style and relations between game play and personal life.

Survis allows arbitrary banks of categorical variables to be read in and visualized. Three spreadsheets of data are needed to fully construct the visualization. First, the actual data must be provided via spreadsheet in comma separated value format. Second, in order to fully identify the variables, Survis requires a list of the variables including the variable name, the type (used to identify the proper value labels obtained later), and the full text of the question as presented to the respondents. Third, the labels for each possible value within a question must also be provided in a comma separated value spreadsheet. This structure is due to the format in which survey data is normally encoded. The data is imported to and exported from statistical packages using one-word semi-cryptic names of the variables and numeric encoding of the textual answers (e.g. strongly agree = 5 and strongly disagree = 1). Survis accepts the date in this format, but for actual exploration it is very limiting, thus Survis also allows more descriptive labels to be specified for each question and each answer with each question.

Since the full description of many of the answers is too lengthy to display in the limited screen space available, much of that is contained within descriptive Java Swing HTML tooltips. While the tooltips do occlude parts of the visualization when shown, they seem to offer an optimal trade-off between information visibility and data density [10].

A Intel Core 2 Duo processor in a notebook computer with 2 GB of 778 MHz random access memory and a 256 MB NVidia Quadro NVS 140M video card requires 5 seconds to construct and display a nested category map based on 3250 respondents to the 58 categorical questions of phase 17 of the Daedalus Project. The bulk of this time is taken in reading the data from comma separated value spreadsheets into Prefuse's tabular format. It takes just under one second to construct each additional nested category map using that data.

Nested category maps take substantially longer to construct. A tile much be constructed for every comparison between variables – both the graphic and the correlation values; hence, it takes exponential time in $O(m * n^2)$ where n is the number of variables and m is the number of respondents. Using the same notebook computer referenced above, construction of each tile requires approximately 0.16 seconds, resulting in 0.16 * 58 * (58 / 2 + 1) = 278 seconds or just under five minutes. Constraining the display to only 30 variables reduces the time to less than 2 minutes.

Due to this discrepancy between initialization times, Survis displays the nested category map as soon as it becomes available but spawns a separate thread to initialize the correlation map while the analyst interacts with the nested category map.

Certain intended functionality of the visualizations is not yet implemented in Survis. As mentioned above, unlike the nested

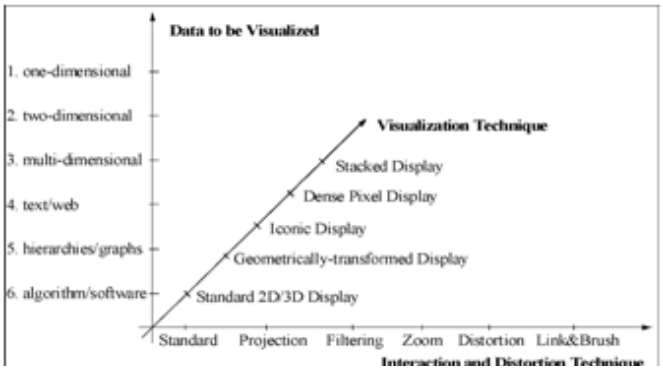

Fig. 6. Data types and corresponding visualizations [14]

category map, correlation maps do not currently allow reordering of the variables. Also, the correlation map does not yet allow zooming. Additionally, the nested category map places labels for each square at each level in the exact center of the square. This method is acceptable if squares are sufficiently large, but when they become small, as when one value at the top level has very few respondents, they can overlap and readability decreases. In the simplest case, this problem can be reduced by placing the top level labels, marking that space as taken, and then placing each subsequent level of labels in the remaining space. There will be circumstances, though, that make it impossible to fit the text of all labels within the allotted space without shrinking the fonts used.

4 Discussion

These visualizations are potentially very useful tools for analyzing specific types of multi-dimensional datasets. The exponential time required to construct the correlation map makes it less useful for datasets where the number of respondents is very large or for questionnaires where the number of questions asked is very large. The exponential time will not be a significant issue though when analyzing smaller datasets of the sort usually constructed using tools like SurveyMonkey.

The correlation map also has the potential for suggesting spurious correlations. At present, there is no method for determining from data values whether a categorical variable is ordered or unordered. The correlation map thus assumes that all variables will be ordered and produces pair-wise correlations using Spearman's *r*. If a variable in the questionnaire is unordered, the correlation tile will still indicate correlation or lack thereof, even though such a comparison makes no sense. In such cases, the advantages of tiling small multiples remains, but the use of color to identify interesting comparisons may be diluted by false correlations.

5 Related Work

In the early 2000s, Daniel Keim presented a summary of visualization and visual data mining techniques by data type in [20] and [14] (see Fig. 12). In this work he referenced a diverse set of advanced techniques, many of which are related to this work. Some of which are mentioned below, along with other techniques and systems that have been developed since.

The Grand Tour is one of the earliest examples of interactive dynamic projections. In this idea, Asimov attempts to create plots of two-dimensional projections of all interesting comparisons within a multi-dimensional data set [21]. Like correlation maps, these projections are exponential in the number of dimensions and thus intractable with very high numbers of dimensions. A modification of the idea was the basis of the ScatterDice system presented in the IEEE InfoVis 2008 best paper [22]. It presented interactive animated methods for exploring a multi-dimensional data using a matrix of scatterplots; however, since scatterplots decrease severely in

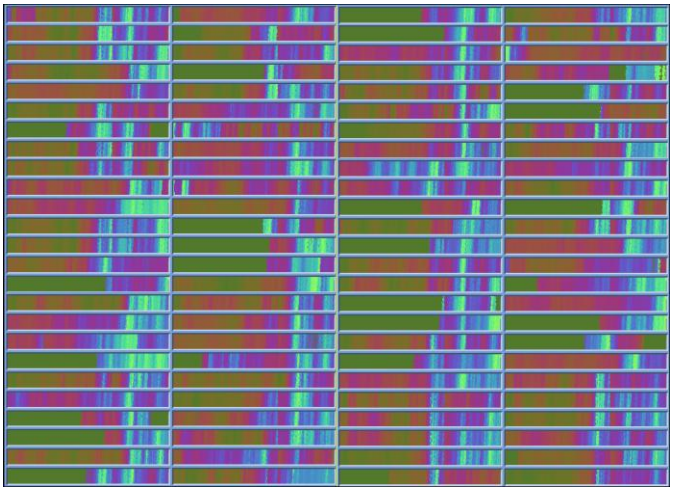

Fig. 7. A dense pixel array: recursive pattern technique [14]

usefulness when visualizing categorical variables, the system is of limited worth in visualizing non-interval variables.

Many Eyes is a web-based system that allows users to upload data, create interactive visualizations, and discuss those visualizations [6]. It incorporates a wide variety of visualization techniques that can be applied according to the composition of the data. Although it is not focused on generating new techniques for visualizing the information, it has already served as a launching platform for various new textual visualizations and it can offer a useful set of tools for visualizing the responses to many types of surveys – questionnaire or otherwise. It allows interval data to be displayed using bubble charts, geographic maps, and many common types of graphs like scatterplots, histograms, and bar charts. It offers stack graphs to display independent categories and their numeric contribution to the whole over time. It also provides well-established techniques like squarified treemaps, and newer visualizations like tag clouds and Wordles.

Systems like Polaris (later Tableau and VizQL), MGV, and Spotfire provide similar services to Many Eyes, but for single users or limited collaborative intranets [18, 23, 24, 25, 26]. Polaris in particular offers a robust selection of visualization techniques designed especially for query, visualization, and analysis of multi-dimensional databases.

Within systems like Polaris, Many Eyes, and MGV and in other custom prototypes, advanced techniques have been demonstrated for visualization of many types of multi-dimensional data. Geometric transformations like parallel coordinate projections and Hyperslice and iconic displays like Chernoff's faces each offer interesting, if somewhat unintuitive ways to visualize multi-dimensional data [27, 28]. Probably most relevant to the problems addressed in Survis are stacked displays, such as Worlds-within-Worlds, and dense pixel arrays, such as VisDB [29, 30].

Treemaps are one form of stacked displays that were incorporated directly into the design of nested category maps. Dimensionally stacked displays could also be very useful [31]. Although they generally convey less information about each of the elements within the graphic, they could allow many more dimensions to be visualized simultaneously. Dense pixel arrays would also allow more data to be encoded in the same space. Although difficult to interpret initially, they can clearly convey information on a very large number of variables as well as relations among them using very little space [30]. The major downside to dense pixel arrays is that they do not lend themselves to interactive exploration, and it is very difficult to select any specific detail for further exploration since each variable is encoded as one single pixel.

The rank-by-feature framework integrated into the Hierarchical Clustering Explorer (HCE) also bears strong resemblance to the idea of correlation maps [32]. This system provides a way to visualize, using a triangular grid of colored squares, relationships between any two variables in the dataset. The system is also specifically designed to visualize and explore datasets with many distinct variables. While the design is tailored to visualizing interval variables, the third version of the software provides a variety of pair-wise statistical tests that can be applied as desired, not simply correlation values. Details on the results of each can be displayed using scatterplots, line graphs, and histograms. Survis offers complementary functionality to HCE for specifically visualizing ordinal and categorical datasets.

6 Conclusion and Future Work

This work provides a first step in an area of information visualization that has been largely overlooked. While significant progress has been made in visualizing highly multi-dimensional datasets of interval variables - specifically ratio variables with clearly defined zero points - these techniques have not often been applied or adapted to categorical data. Some of the techniques, such as dimensional stacking or rank-by-feature frameworks might be very useful with just slight modifications. Still, there may be other, undiscovered methods that are impossible or worthless for interval data and yet highly relevant to categorical data. Further investigation is warranted.

The techniques used in Survis are especially tailored to recognizing relationships among categorical variables; however, they only allow simultaneous comparison of a limited number of variables. Correlation maps do illustrate pair-wise correlations of a large number of variables but do not provide insight into possible intervening variables. There may be ways of adapting the same idea to include categorical ANOVA tests among multiple variables, either at run-time according to the user's demands or automatically when the map is constructed. Analysts may also wish to compose columns and view the resultant interrelations.

Visualization methods like the Grand Tour incorporate ideas of "interestingness" in deciding which variables to visualize. Nested category maps and correlation maps defer judgment on the interestingness of comparisons and instead opt to allow individual exploration of all possible comparisons and subsequent individual judgment of interestingness. Even so, it would likely be worthwhile to experimentally determine which aspects individuals find most useful in visualizing the data and order the display of the variables to emphasize these details.

In order to validate the usefulness of these visualizations, they must be tested with actual people exploring actual data sets. Changes, recommendations, and missing functionality to support analysis can then be identified and created. Design decisions about structure, color, size, and shape – the various details of design and implementation – can then be confirmed or revised.

Animated transitions can be invaluable in maintaining a sense of orientation through transitions [33, 34]. In visualizations like nested category maps, this sense of position or orientation is very easy to lose through drill-down and expansion up. Future work should also include using animations and maintaining similar layout throughout hierarchical exploration.

References

[1] Online Survey Software. http://www.questionpro.com/.

[2] Online Surveys - Zoomerang. http://www.zoomerang.com/.

[3] SurveyMonkey.com - Powerful tool for creating web surveys. http://www.surveymonkey.com/

[4] SPSS The predictive analytics company. http://www.spss.com/.

[5] STATA: Data Analysis and Statistical Software. http://www.stata.com/.

[6] L. Nowell, S. Havre, B. Hetzler and P. Whitney. "Themeriver: Visualizing thematic changes in large document collections,"." *IEEE Transactions on Visualization and Computer Graphics*. 2001.

[7] W. B. Paley, "TextArc: Showing Word Frequency and Distribution in Text." IEEE Transactions on visualization and computer graphics. 2002.

[8] F. B. Viégas, M. Wattenberg, F. van Ham, J. Kriss, M. McKeon. "Many Eyes: A Site for Visualization at Internet Scale." *IEEE Transactions on visualiztion and computer graphics*. 2007.

[9] J. Bertin , *Semiology of graphics.* University of Wisconsin Press, 1983.

[10] E. R. Tufte, The Visual Display of Quantitative Information. 2nd Edition. Cheshire, Connecticut: Graphics Press LLC, 2006.

[11] R. Likert, "A Technique for the Measurement of Attitudes." *Archives of Psychology*, no. 140 (1932): 1–55.

[12] J. Dawes, "Do Data Characteristics Change According to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales." *International Journal of Market Research* 50, no. 1 (2008): 61-77.

[13] E. Babbie, The Basics of Social Research. Thomas Wadsworth, 2005.

[14] D. A. Keim, "Information Visualization and Visual Data Mining." *IEEE Transactions on visualiztion and computer graphics*. 2002. 100-108.

[15] B. Schneiderman, "The eye have it: A task by data type taxonomy for information visualizations." Visual Languages. 1996.

[16] M. Bruls, K. Huizing, and J. J. van Wijk. "Squarified Treemaps." *Proceedings of the Joint Eurographics and IEEE TCVG*. 2000.

[17] B. Tversky, J. Morrison, M. Betrancourt. "Animation: Can It Facilitate?" *International Journal of Human-Computer Studies* 57 (2002): 247-262.

[18] Tableau Software. The Art of Visualizing Survey Data. 2008. www.tableaucustomerconference.com/files/TCC08-CS-eLearningGuild-The-Art-of-Visualizing-Survey-Data.ppt.

[19] J. Heer, S.K. Card, J.A. Landay. "Prefuse: a toolkit for interactive information visualization." *Proceedings of the SIGCHI conference on Human factors*. 2005.

[20] D. A. Keim. "Visual exploration of large databases." *Communications of the ACM*. 2001. 38–44.

[21] D. Asimov, "The grand tour: A tool for viewing multidimensional data." *SIAM Journal of Science & Stat. Comp*. 1985. 128–143.

[22] N. Elmqvist, P. Dragicevic, J.-D. Fekete. "Rolling the Dice: Multidimensional Visual Exploration using Scatterplot Matrix Navigation." *IEEE Transactions on Visualization and Computer Graphics*. 2008. 1141-1148.

[23] D. Tang, C. Stolte, and P. Hanrahan, "Polaris: A system for query, analysis and visualization of multi-dimensional relational databases," *Transactions on Visualization and Computer Graphics*, 2001.

[24] J. Abello and J. Korn, "Mgv: A system for visualizing massive multi-digraphs," *Transactions on Visualization and Computer Graphics*, 2001.

[25] P. Hanrahan, "VizQL: a language for query, analysis and visualization," *International Conference on Management of Data,* 2006

[26] C Ahlberg, "Spotfire: an information exploration environment," *International Conference on Management of Data,* 1996

[27] H. Chernoff, "The use of faces to represent points in kdimensional space graphically," *Journal Amer. Statistical Association*, vol. 68, pp. 361–368, 1973.

[28] J. J. van Wijk and R.. D. van Liere, "Hyperslice," in Proc. Visualization '93, San Jose, CA, 1993, pp. 119–125.

[29] S. Feiner and C. Beshers, "Visualizing n-dimensional virtual worlds with n-vision," *Computer Graphics*, vol. 24, no. 2, pp. 37–38, 1990.

[30] D. A. Keim and H.-P. Kriegel, "VisDB: Database exploration using multidimensional visualization," *Computer Graphics & Applications*, vol. 6, pp. 40–49, Sept. 1994.

[31] J. LeBlanc, M. O. Ward, and N. Wittels, "Exploring ndimensional databases," in Proc. Visualization '90, San Francisco, CA, 1990, pp. 230–239.

[32] J Seo, B Shneiderman, "A rank-by-feature framework for interactive exploration of multidimensional data," *Information Visualization*. 2005.

[33] J. Heer, G. Robertson. "Animated Transitions in Statistical Data Graphics." *IEEE Transactions on visualiztion and computer graphics*. 2007.

[34] B. Schneiderman, "Tree visualization with treemaps: A 2D spacefilling approach." *ACM Transactions on Graphics*. 1992. 92–99.