Visualizing nominal sequential data

Maria Frank
PhD Candidate, Stanford University
473 Via Ortega #280, Stanford, CA, 94305
mfrank0@stanford.edu

Flavia Grey
PhD Student, Stanford University
473 Via Ortega #280, Stanford, CA, 94305
fgrey@stanford.edu

ABSTRACT

In this paper we describe a visualization interface for complex, nominal sequential data. Many visualizations focus on continuous sequential data or time series. However, we identified a need to allow for the representation of nominal data associated with time in different scales as well as other forms of sequential categorical or nominal data. Through this interface, researchers can develop an intuitive understanding of existing patterns in large data sets and be able to develop hypothesis around the data set at large and subsequently focus research on patterns of interest. The tool is designed as a web application enabling users to upload their own data and explore it through various interactive functions.

Author Keywords

Big data, visualization, nominal data, sequencing.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Big Data is not only created by individual unrelated points of information but often through data collection over time, different location, or other forms of sequences [2][3].

This type of sequential Big Data can come in the forms of nominal, ordinal, and continuous data. Continuous and ordinal data are frequently visualized by showing the relative change in value between two sequential points. These visualizations often use line or bar graphs indicating the different sequences. Physiological sensor data is one current example in which values are captured on a continuous scale over time. End users and researchers then receive line graphs of the relative changes in the value of the sensor to inform analysis and interpretation. Often, the sensor data then is aggregated qualitatively to inform the user about an average state. This type of qualitative evaluation though, is not readily shown as a sequence of time aggregated for longer periods to see bigger behavioral patterns. And it is still difficult to visualize longitudinal data comprehensively [5].

However, there is also nominal data that can be connected in a sequence and over time [4]. As we collect more and more data and create Big Data of nominal type we have to develop new ways of visualizing these types of data in a condensed and interpretable format that allows to see sequences and patterns over time and in a grounded theory approach develop questions and hypotheses based on initial data insights. Visualization research has largely ignored these types of data. As Big Data becomes bigger and bigger we need to ways of aggregating data, and thereby turning the individual continuous data streams into combined interpreted nominal data streams for further analysis.

In this paper we present FrickViz, a visualization interface that allows for the representation of aggregated discrete dimensional categories in sequence. The visualization will enable researchers to identify patterns in very complex data.

The pattern identification will allow to answer questions such as: How is the data distributed? Are there continuous blocks of a certain category, or are the categories interspersed? What do the transitions between categories look like? Are these transitions consistent throughout the data set? Are there "transitional" categories, that are always between a certain pair? Is the distribution of categories homogeneous or are there different "stages" in the data? Are there specific patterns towards the end or the beginning of a data sample?

CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA.

The general identification of such patterns in big data allows the researcher to identify and define moments of interest as a first step of data analysis. The identification can be understood as a narrowing of the research focus to be able to handle large data sets and focus the general research efforts.

We employ a linear scanning approach to create an interactive interface visualizing complex data in an intuitive graphic display.

RELATED WORK

Most sequential data we encounter and see represented is of ordinal or continuous nature. This data is easily and frequently represented as line charts showing the relative change of the variable value over the course of the sequence.

The most common approach for nominal, time-series data is a Gantt chart [8]. However, this approach provides information about the general composition of the data, for example, which category appears most often, but it makes it hard to see what data in a fine granularity, e.g. on at a second-by-second level. And creates large visualizations that can only accommodate very limited amounts of data. (Figure 1)

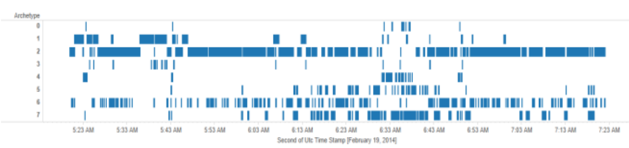

Figure 1. Example of a Gantt chart.

In Gantt charts, transitions between stages are intransparent and hard to interpret which complicates the recognition of even unusual patterns. This problem is worsened by the general structure of a Gantt chart, requiring to fit the data horizontally in our screens in one continuous band. This results in difficulties observing details, which are smaller the smaller the user screen becomes. Hence, the user needs to decide between having a very general overview without enough detail or a detailed view without enough context. Our approach creates a compact version of the Gantt chart where all the categories are presented in one block that dynamically adjusts to the screen size. Our visualization avoids horizontal compacting by looping around and creating several lines in a vertical distribution.

Weber, Alexa and Muller [6] proposed a visualization for time-series information based on spirals (Figure 2). Their approach works well for bringing out periodic behaviors in the data. However, it loses context of the length of specific sequences due to the radial difference of inner and outer circle. Furthermore, the radial representation limits the application of legends, counts, or other additional quantitative data besides the mere image. In addition, radial representations are not an intuitive reading format and hence require additional cognitive resources for the interpreter of the visualization. We believe that our approach showing time in a straight line that wraps, lowers the cognitive load on the reader of the visualization, because it mirrors common text structures. This representation enables the user to include identifying blocks of a certain length of time and thereby supports the detection of non-periodic patterns and behaviors.

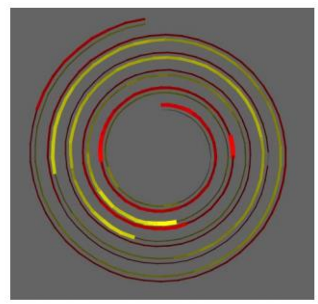

Figure 2. Circular representation of sequential data [6]

The inspiration for the type of visualization we chose comes from the data artist Laurie Frick who makes hand-built work from self-tracking data (Figure 3). She color-codes different activities during her day to visualize the occurrence and duration over various days. She designed an app based on her artwork called Frickbits. While the actual drawing of the data is similar to our proposed design in terms of using bars of color to depict the data, Frickbits is used to create map information [7].

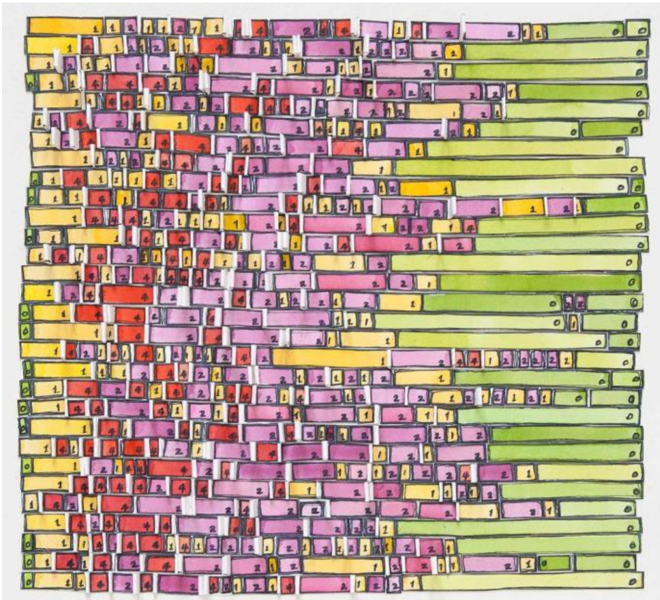

Figure 3. Laurie Frick visualization [7].

For our implementation we were furthermore inspired by the Microsoft SandDance application[1] to visualize data as squared data points that represent individual categories or users (Figure 4). However, SandDance visualized their information in traditional graph types like bar graphs and tree maps that enable the identification of statistically relevant values, but does not allow for the identification of non-statistical qualitative patterns in sequential data.

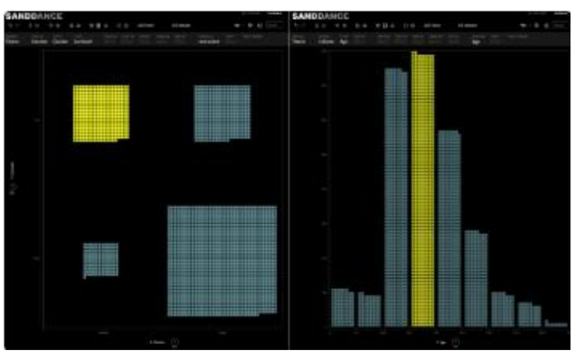

Figure 4. Microsoft SandDance blocks

Not a lot of work has been done on the visualization of nominal data. The related work hence, is scarce and only serves as general inspiration for our visualization tool which will be discussed in the following section.

METHOD

Implementation

The goal of this project was to create an open web tool that allows researchers to upload their own data and create sequenced visualizations of various data sets at their disposal. To accomplish this, the implementation uses Javascript, CSS, HTML, and D3.js. The tool enables direct interaction with the data through the interactive display and multiple adjustable parameters in the display. The data, furthermore, can be filtered and highlighted to lay focus on individual categories.

Interface design

The FrickViz website has two main components, which the user can chose from up front (Figure 5); to explore existing data sets or to upload their own data.

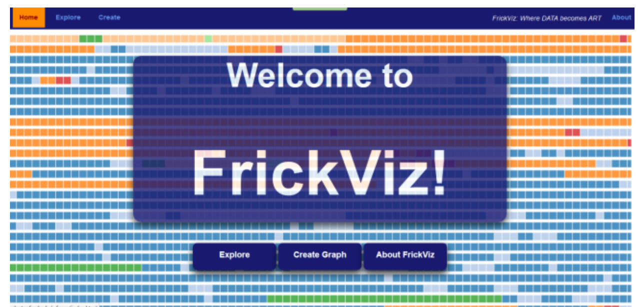

Figure 5. Home page with 'Explore' and 'Create'choices

We integrated sample data sets (Figure 6) that users can explore. Through this they can learn how to interact with data and explore the various options to manipulate the visualizations. Through this the user can gain a better understanding of the design and purpose and thereby prepare their own data set more effectively for their implementation. In addition, it generates a broader understanding of the capabilities to visualize different data sets and the representation of them in this linear fashion, which can inspire new research ideas.

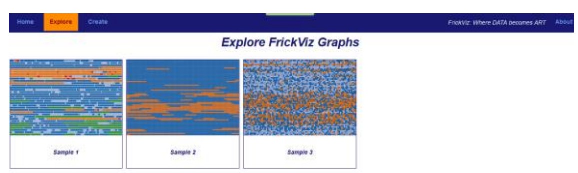

Figure 6. Sample data sets

The initial display structures the data in a simple block, where the dimensions are defined by the screen size (Figure 7). The user has multiple options to structure the uploaded data through numerical input after this. The user can individually define width and height of the rectangle representation, where each rectangle is one data point, one element of the sequence coded through a classification.

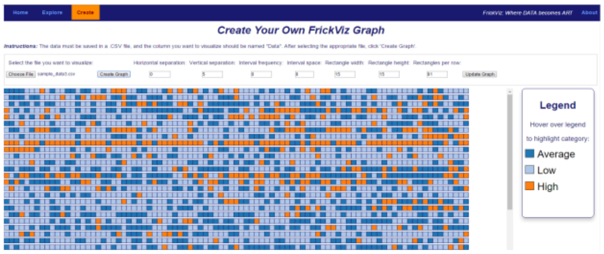

Figure 7. The upload and interaction interface

The user can furthermore define the amount of rectangles per row. Depending on the data set, varying representations are possible, e.g. 60 rectangles per row for a second by second representation, 24, 48, or 72 for a day representation. By allowing for a definition through the user the application allows for a broader variety of data sets in different sequential rhythms, as well as for the experimentation with various sequences to explore different patterns in the data.

Besides the data points specifications, the user can also define visual separations of the data that follow underlying rhythms that might exist in the data set (Figure 8). Thereby intervals and pattern within and across intervals can be highlighted and simplify the visual interpretation by creating reference points.

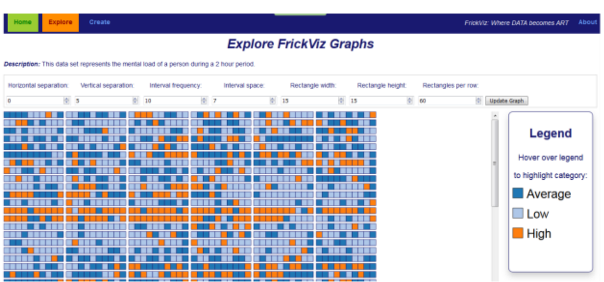

Figure 8. Visual Structuring

After structuring the visual representation, the user has some interactive features available for further investigation. One interaction feature is the highlighting of specific categories, to increase the salience of potential patterns (Figure 9).

The second, currently implemented, interaction is a hover function that provides more detail about the data point.

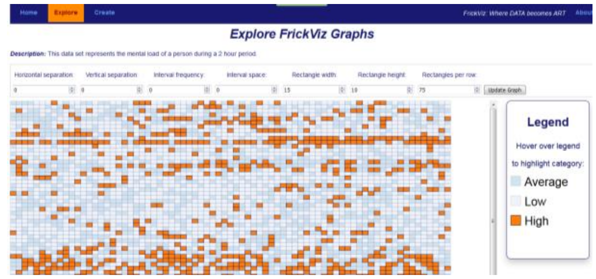

Figure 9. Interaction - highlighting

Testing

Initial testing with users showed a quick understanding of the tool with minimal instructions. This shows the intuitive interaction with the interface.

User feedback was focused around enabling data manipulation of the original data set. As well as additional axis information and detail zoom interaction that allow the user to see more information on the individual data point.

DISCUSSION

The data visualization tool presented in this paper allows researchers to display aggregated discrete dimensional categories in sequence. The visualization enables researchers to identify patterns in very complex data in a grounded theory approach. As big data becomes more prevalent we need new ways to visualize components of it and through this enable the investigator to formulate new questions and identify new qualitative patterns, that otherwise might maintain unanalyzed.

Due to the vast variety of potential sequential data sets, we created a tool that allows the researcher to specify blocks of data based on underlying data patterns. The tool thereby can be used to visualize time series from second, day, year, through century or millennial data. But furthermore it allows for the representation of non-time series data such as DNA sequences, music tunes, literary or musical themes, and many potential other applications.

This web-tool, even though dynamically adjusting to different screens, still allows for a bigger data sets the bigger the physical screen and resolution. However, the current interface with adjustable data points allows for more specification and through this visualization of individual data points. The wrapping of the sequence furthermore limits the visualization capabilities to some extent. However, the ability to change the wrap point by adjusting the number of displayed data points per row, enables the user to change the sequences and through this see various potential different wrapping layouts and thereby patterns.

In further steps we want to enable the web tool to also visualize ordinal and continuous data in a qualitative overview and abstract it into a classified map. The user then will be able to manipulate the data by defining the amount of categories or levels and the related boundaries.

The aggregation of categories in an existing data set to see different structures will be another step for further development of the tool. Often categories can be clustered, which the researcher then can use to apply different lenses to their analysis.

We furthermore want to enable the visualization of multiple columns in direct parallel displays of data sequences that allows visual comparison. This could be data of different individuals, data from sequences collected at different times, or any other type of comparison between two or more sequences.

The next implementation of this tool will also contain a save function, as well as a share function. Thereby researchers can share their visualizations and receive feedback on them, create a history of inquiries and export the visualizations for other purposes.

We furthermore want to create a zoom and selection function that enables the user to zoom into specific aggregated parts of the data to explore underlying patterns that might exist in a multilayered data set such as minutes in hours, hours in days, and days in weeks. Thereby we enable the user to explore different layers of data naturally by identifying bigger patterns, focus on these, and go deeper into the actual data set.

There is more work to be done in the effort to create a comprehensive data visualization tool. The herein presented tool is a start to fundamentally change the way we look at sequential nominal data.

REFERENCES

1. Microsoft SandDance. Online at: http://research.microsoft.com/en-us/projects/sanddance/
2. Bade, Schlehtweg, Miksch: Connecting Time-Oriented Data and Information to a Coherent Interactive Visualization. Online at: http://ifs.tuwien.ac.at/~silvia/wien/vu-infovis/articles/bad_chi2004.pdf
3. Aigner, et al. (2008). Visual Methods for Analyzing Time-Oriented Data. Online at: http://ieeexplore.ieee.org/Xplore/defdeny.jsp?url=http%3A%2F%2Fieeexplore.ieee.org%2Fstamp%2Fstamp.jsp%3Ftp%26arnumber%3D4359494%26tag%3D1&denyReason=-133&arnumber=4359494&productsMatched=null&userType=inst
4. Quantitative Development Methodology Syllabus, Pennsylvania State University. Online at: https://quantdev.ssri.psu.edu/avada_workshops/workshop-01/
5. Cubism, Square Inc. presentation. Online at: https://bost.ocks.org/mike/cubism/intro/
6. Weber, M., Alexa, M., & Muller, W. (n.d.). Visualizing time-series on spirals. *IEEE Symposium on Information Visualization, 2001. INFOVIS 2001*. doi:10.1109/infvis.2001.963273, Online at: http://cs.lnu.se/isovis/courses/spring07/dac751/papers/TimeSpiralsInfoVis2001.pdf
7. Frick, Laurie. Online at: http://www.lauriefrick.com/
8. Gantt chart. Online at: http://www.gantt.com/