Interactive Exploration of Multidimensional YouTube Data Using the GPLOM Technique

Seimei Matsusaki, Xi Han, Alark Joshi, and Sophie Engle
University of San Francisco
2130 Fulton Street, San Francisco, CA
[smatsusaki,xhan14,apjoshi,sjengle]@usfca.edu

ABSTRACT

We present an application of the generalized plot matrix (GPLOM) technique for visualizing a multidimensional dataset containing 202 YouTube videos from the "Statistics and Social Network of YouTube Videos" project. The web-based visualization provides users an easy to use interface to explore the data. Using our tool, we found several interesting relationships between the number of views, comments, ratings and stars. Visit git.io/vwDqw for a live demo.

CCS Concepts

•**Human-centered computing** → **Information visualization;**

Keywords

Multidimensional data; generalized plot matrix; linked views

1. INTRODUCTION

Multidimensional datasets are often visualized using techniques such as glyphs, heatmaps, parallel coordinates, scatterplot matrices (SPLOM), and generalized plot matrices (GPLOM) [3]. While the other visualization techniques work well for visualizing continuous quantitative data, the GPLOM technique works well for data that contains categorical as well as quantitative attributes. The GPLOM technique, as introduced by Im et al. [3], displays a scatterplot for visualizing two quantitative attributes, a bar chart for a categorical and a quantitative attribute, and a heatmap for two categorical attributes. For this project, we implemented a GPLOM for visualizing a multidimensional YouTube videos dataset [2]. Figure 1 shows a screenshot.

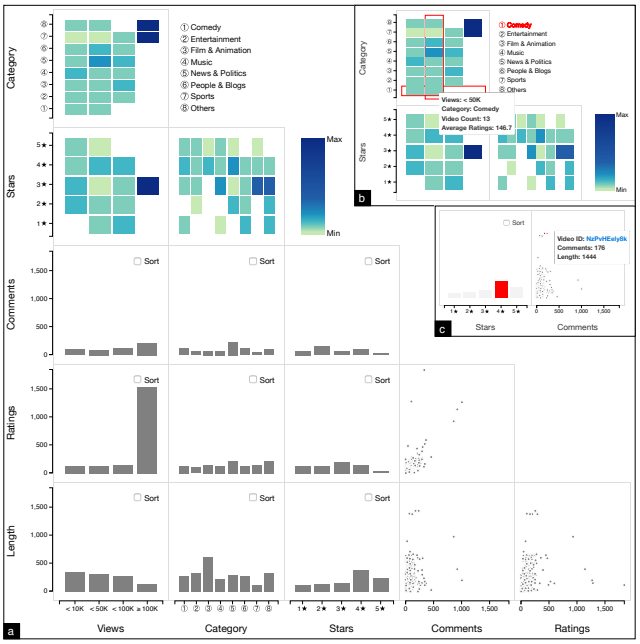

Figure 1: The GPLOM Interface. **(a)** The entire GPLOM interface. Visit git.io/vwDqw for a live demo. **(b)** A heatmap comparing "< 50k Views" on the x and "① Comedy" on the y-axis. The row, column, and category text in the legend is highlighted on hover. **(c)** A bar and scatterplot for "Stars" and "Comments" on the x and "Length" on the y-axis. The bar and scatterplot points are highlighted and a tooltip appears on hover. The sort checkbox orders the bar charts in descending order.

2. DATA CHARACTERISTICS

The "Statistics and Social Network of YouTube Videos" project [2] includes over 100 datasets varying from 100KB to 10MB from Feb 2007 to Sep 2008 made publicly available for academic use. We used a dataset of 202 videos from Mar 2, 2007 for this project. The dataset includes a mix of qualitative and categorical attributes. The qualitative attributes include the video length, total views, total comments, total ratings, average rating, and the number of days between the uploaded and accessed dates). The categorical attributes include the category, such as "Comedy," "Music," and "Sports." We combined "Howto & DIY," "Pets & Animals," "Autos & Vehicles," "Gadgets & Games," and "Travel & Places" into a single "Others" category since there were few videos in these categories. We also binned the total views and average rating into categories. For example, average ratings less than 1 were recoded as a "1 star" videos and average ratings between 1 and 2 were recoded as "2 star" videos.

VINCI '16 September 24-26, 2016, Dallas, TX, USA

ACM ISBN 978-1-4503-4149-3/16/09.

DOI: http://dx.doi.org/10.1145/2968220.2968233

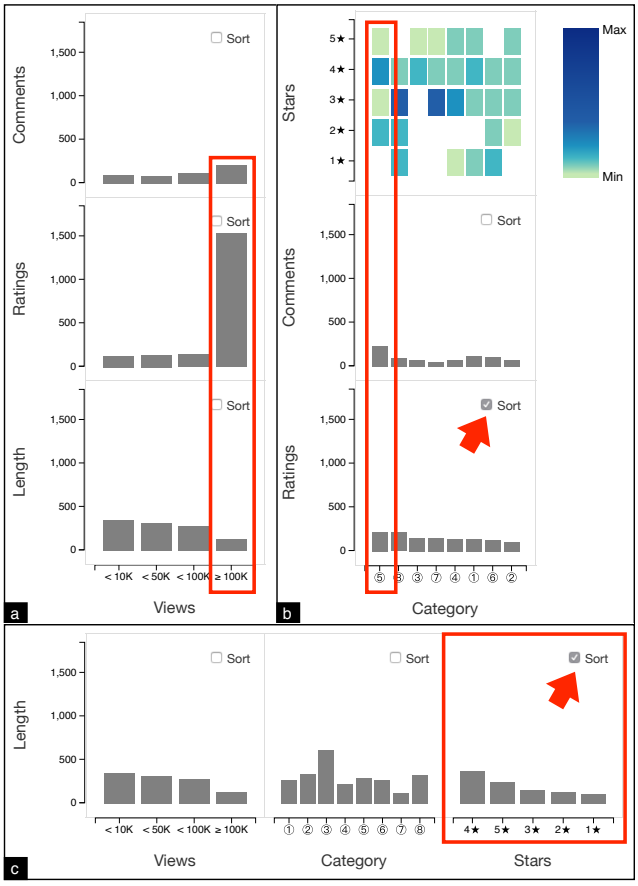

Figure 2: Findings. **(a)** The total comments and ratings increase as the total views increases, but not at the same rate. The total ratings increased more drastically than the total comments for videos with over 100k views. **(b)** The "⑤ News and Politics" category has the most comments and ratings, but does not have the most 5 star ratings. **(c)** As the length of a video increases, the ratings tend to be higher but the total views tends to be lower.

3. GPLOM IMPLEMENTATION

We used the D3.js [1] JavaScript library to create a highly interactive in-browser visualization. We first load the pre-processed YouTube dataset CSV. We then create 15 SVGs for each cell in the GPLOM and link the interactivity of each of these SVGs together. We also create SVGs for the category legend, color legend, and each of the attribute axes for a total of 27 linked SVGs.

Figure 1 shows a screenshot of a user exploring categorical and quantitative data using our GPLOM tool. The GPLOM uses a coordinated multiple visualization [5] approach to facilitate quick exploration of data which works well in our case since the YouTube dataset has many variables. To facilitate seamless interaction, we use the familiar hover paradigm that provides details-on-demand to a user in the form of a tooltip as depicted in panels b and c of Figure 1. Additionally, the user has the ability to sort categories in a bar chart by quantitative data attributes by clicking the sort checkbox as illustrated in Figure 2b and 2c. When the sort checkbox is selected, all of the associated bar charts and heatmaps in that column are reordered, which helps with comparing quantitative attributes in the data.

To help the user explore regions of interest as specified by the user's mouse position, we incorporate focus + context [4] using hue-based highlighting (changing the data elements of interest to red) and by reducing the saturation levels of other regions. Figure 1b and 1c shows the relevant elements highlighted based on the user interaction.

4. FINDINGS

Visualizing the relationships between multiple categorical and quantitative attributes simultaneously allowed us to make three observations about this dataset. First and unsurprisingly, people rate and comment on videos more often as the number of video views increases. However, if the total views exceeds $100k$, the total ratings suddenly increases at a faster rate than the total comments. Second, the "⑤ News and Politics" category has the highest total ratings and highest total comments. However, consider the heatmap for category versus stars. Focusing on the row for 5 stars, we can see this category does not have the most 5 star ratings. If we focus on the column for this category instead, we can also observe that most videos are either 2 or 4 stars. Finally, if we sort by video length, it appears that longer videos tend to have fewer total views but more stars.

5. CONCLUSIONS

In this paper, we demonstrate the use of the GPLOM technique to visualize a multidimensional dataset on YouTube videos. Based on our implementation of the GPLOM technique, we found interesting insights regarding the relationships between several attributes in our dataset. The capability of the GPLOM technique to show both quantitative and categorical data simultaneously, as well as the rich interactivity of our tool, were both instrumental in making these findings. We plan to continue exploring the GPLOM technique with larger datasets from the YouTube project [2] as well as other multidimensional datasets to evaluate the strengths and weaknesses of the technique.

6. REFERENCES

[1] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[2] X. Cheng, C. Dale, and J. Liu. Statistics and social network of YouTube videos. In *Proceedings of the 16th International Workshop on Quality of Service*, pages 229–238, 2008. Data at netsg.cs.sfu.ca/youtubedata.

[3] J.-F. Im, M. McGuffin, and R. Leung. GPLOM: The generalized plot matrix for visualizing multidimensional multivariate data. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2606–2614, 2013.

[4] J. Lamping, R. Rao, and P. Pirolli. A focus + context technique based on hyperbolic geometry for visualizing large hierarchies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 401–408, 1995.

[5] J. Roberts. State of the art: Coordinated & multiple views in exploratory visualization. In *Proceedings of the 5th International Conference on Coordinated and Multiple Views in Exploratory Visualization*, pages 61–71, 2007.