Movie Data Visualization: User Experience, Interactivity, Color.

Vivian Young, Carol Ho
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
Email: {vmyoung, tingyunh}@andrew.cmu.edu

***Abstract*—Existing ways to access movie data exist in varying and fragmented locations on the internet. Lack of readability and visualization for these movie data resources prevent users from gaining insight into their favorite movies, producers, and actors. We have created a comprehensive one-stop-shop movie data application with the intention of visualizing and compiling data in a user-friendly and unique way in order to provide movie lovers with a novel perspective on data in the film industry.**

I. Introduction

With the ongoing expansion of streaming services and increasing prevalence and variety of content across platforms, the demand for movies and the talent behind them is increasing more than ever. But with the constant release of new content, the following questions are often raised:

How do people choose which movie to watch? What aspects of content are most compelling to long-time fans and new viewers?

After several informal interviews, we found that in order to choose new content, most people turn to movie reviews, online film data websites, or look through popular movies made by producers / categorized by genres / starring actors they're already familiar with.

Popular movie review websites like IMDB and Rotten Tomatoes provide movie-centric reviews and data that may help viewers make decisions on whether to watch something in particular. But when it comes to information on directors and actors, most of the data on the internet is focused away from their professional performance and instead composed mainly of a mixture of attention-grabbing headlines, and their latest relationship status. And although many audiences place their trust in movie critics for professional and seemingly intelligent opinions, biased suggestions may prevent the audiences from making informed and impartial decisions. With this problem in mind, our data application aims to provide a people-centric perspective that allows users to learn about the directors/actors through data insights of their works and professional accomplishments. By interacting with our platform, the user can explore the attributes and themes of a directors/actors' works and receive movie suggestions through novel visualizations.

II. Related Works

The two most popular websites for movie reviews are Rotten Tomatoes and IMDB. Both of these platforms show movies, directors, and actors, alongside a multitude of other data. A director's/actor's profile on these sites commonly list related films which are usually ordered by popularity or release date. Although comprehensive, it is almost impossible for the users to learn about the directors/actors from the attributes of the work. Even if the attributes are provided, they are listed in a table format, making it hard to discern trends or distributions across all of the work. We further looked at some websites that offered more people-centric and more readable information. For example, wikipedia serves as a central hub for biography and awards but it is purely text-based.

For our movie colors visualization, we did not have too many online sources to reference as this sort of visualization is rather new. Because we do incorporate "movie recommendations", Netflix and other streaming services provided a baseline for how to display similar content. It was common to see similar movies displayed first by a chosen cover image / poster alongside its title, brief synopsis, and some more information on the director, actors, etc. The algorithm for selecting these cover images (as seen on Netflix) is not made publicly available but one can assume is based on gigabytes of user data that we do not have access to. Due to these reasons, we decided to utilize the color palettes of movie posters instead to illustrate similarity in order to achieve a different and novel effect.

With these findings, we decided to design our platform to provide a unique people-centric perspective with data-driven insights that allow users to learn about the trends and distributions of various aspects in these bodies of work and expose our audience to an innovative visualization of similarity.

III. Methods

The two major movie APIs we used are OMDB and themoviedb. The former provides information such as genre, reviews, ratings, release date, actors, and producers whereas the latter provides specific cast information such as actor biographies and other movies in which they are featured. We first used the people API from themoviedb for search results and joined the data of movies with movie detail API from themoviedb for genres, revenue, budget, and runtime, and joined data from the OMDB API for IMDB ratings and box office performance. Unfortunately, there were many missing

values in the revenue, budget, and box office columns so we ultimately decided to only show movies that had these three data values and used bar charts to show the distribution instead of using line charts to show the trendlines. Almost all of the works have runtime and IMDB rating value so we used a line chart to visualize the trends and inserted a 0 for any missing data values. We then calculated the average amount of revenue, budget, box office, and runtime of the producers/actors' work while excluding the missing data points.

For the movie colors visualization, we first created a Django application and used OpenCV + Python to create color palettes from frames of each movie trailer. This implementation was great but hosting the app on Heroku resulted in a multitude of issues (mostly regarding missing OpenCV dependencies in a non-local application). Because we were unable to resolve this issue, we migrated to a Node.js web application instead (also hosted by Heroku) and utilized ColorThief.js to create color palettes from the movie posters of similar movies found by using themoviedb API. Visualizing the color palettes in the work of directors and actors was a similar process, differing mostly in the fact that much more data cleaning and sorting was required. For example, the top five movies of the input person are chosen by first filtering out any TV shows, movies without release dates, and movies without posters. Then, we sorted the movies by popularity to choose the "top five" and created color palettes with these movie posters. Furthermore, for the dominant movie colors over time graph, we needed to sort the movies chronologically and then choose ten distributed movies from the entire array of movies. In other words, in an array of movies sorted chronologically of size 20, we would choose the movies at index 1, 3, 5, etc. until ten movies are found and then utilize ColorThief.js to pull the dominant colors of their respective movie posters.

IV. Results

This platform is consists of five parts, the biography, three charts, and the movie colors visualization application. After the user input the director/actor's name, the correspondent results will show up below.

A. The genres bar chart

The chart shows how the works are distributed in each genre. The audiences can hover on the bar to get the lists of movies situated in that genre. Since the relationship between film and genres is one-to-many, the audience may find one movie in multiple genres. The goal of this chart is to show the distribution of genres across the director/actor's work. (Fig. 1.)

B. The box office/revenue/budget charts

The box office/revenue/budget charts shows a financial analysis of each work. Depending on the checkbox selection, the graph can be presented individually or with multiple layers. The layering between box office/revenue/budget visualizes the difference, making it easier to compare the cost structure, and easier spot the profit/loss. (Fig. 2.)

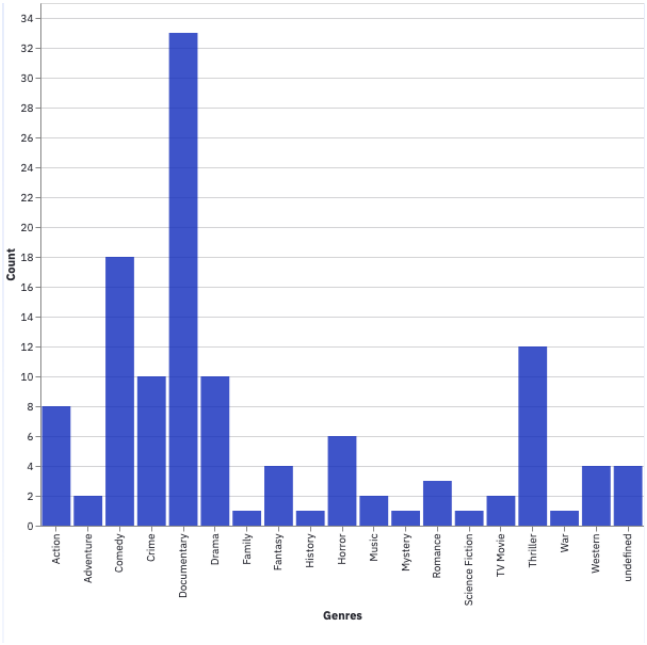

Fig. 1. Quentin Tarantino's works in Genres.

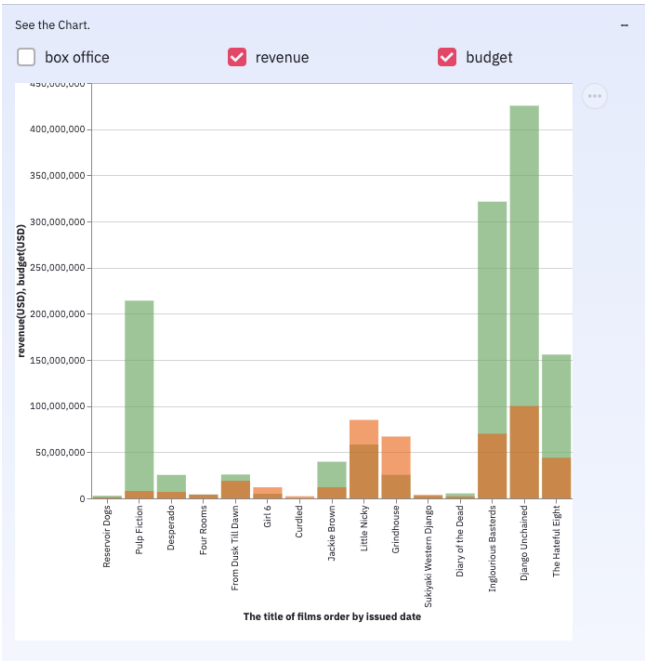

Fig. 2. Budget/revenue of Quentin Tarantino's works.

C. The runtime/IMDB rating line charts

The runtime/IMDB rating line charts aim to provide insights into trends across all works. The audiences can click on the checkbox to compare the average numbers of the works or hover on the data points for numbers, years, and movie names. (Fig. 3.)

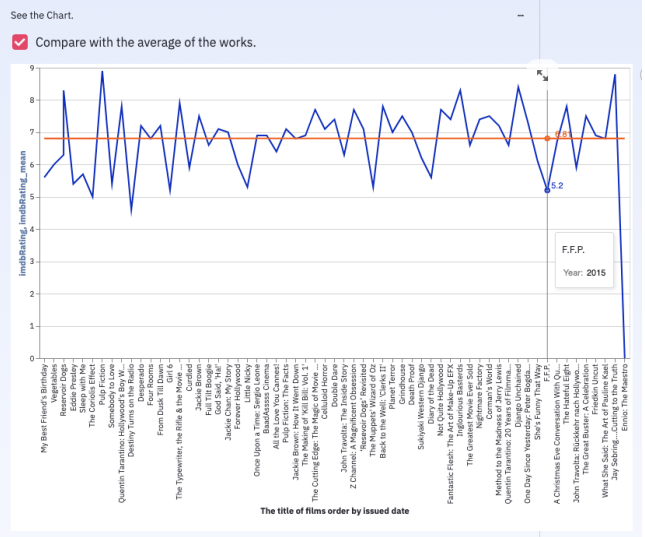

Fig. 3. IMDB ratings distribution of Quentin Tarantino's works.

D. The movie colors visualization application

The application is intended to display similarities between the chosen movie and its recommended movies, show color palettes of movies within an individual's body of work, and visualize the changing of color themes over time. Movie search queries result in color palettes of both that movie and its recommended movies. Producer and actor search queries result in color palettes of their respective selected movies and dominant colors in their movies over time. (Fig. 4.)

Fig. 4. The interface of the movie colors visualization application.

V. Discussion

The genres bar chart shows the distribution of genres across an individual's body of work. This distribution and combination of their most-used genres are often directly tied to their filming/acting style. For example, Quentin Tarantino's work heavily involves documentary, comedy, thriller, action, crime, and drama genres, which relate to his well-known aesthetic. However, the correlation between the genres and the visual / cinematic style may not be as strong if the directors / actors are involved in a diverse set of works. More data is needed to represent the style.

From the box office/revenue/budget charts, the audiences can discern between the financial successes and failures of certain movies but some caution is recommended before any users decide to draw correlations between the data points. Although it may appear that box office, budget, and revenue are solid indicators of the "success" of movies, we found that this data is often not as straightforward as it seems. For example, the total amount of box office is collected by the theaters but only a portion of box office is attributed to the revenue. In addition, total revenue may also be artificially inflated by personal contributions by the directors / actors. Therefore, the goal of the chart is to identify the profitable films and engage the audience's curiosity, prompting users to find out for themselves why certain movies are making more money.

In the runtime/IMDB rating line charts, we can see the distributions of IMDB ratings across the data. As an example, Quentin Tarantino's works released after 2000 consist of more above-average data points. We can seemingly conclude that his works saw an increase in popularity and public reception after the turn of the century. Although we did not notice any trends by visualizing the runtimes of movies, it is still helpful to see these data points when choosing a film to watch.

Our decision to utilize color palettes and dominant color changes over time proved useful as it was evident that some actors and producers are clearly more involved in more colorful works whereas others stay in a more neutral, grey realm. It was visually compelling to see the color palettes of recommended films displayed right next to the original movie but not much can be said for whether or not this will influence users to choose one recommendation over another. Color palettes in an individual's most popular movies and dominant colors in their work over time was also useful in evaluating the visual language of their career but we were aware of many external factors that were likely to skew the data. For example, because our new implementation is pulling color palettes from movie posters instead of the film or the trailer itself, the color data is skewed by the choices of the artist in charge of the promotional materials and thus, this may not be an accurate representation of the actual colors in the film.The film "Mother!" by Darren Aronofsky comes to mind because the beautiful pastels in the poster (by James Jean) intentionally contrasts the darkness and violence of the film.

VI. Future Work

Three directions came up after our validation with the audiences:

A. More insights from the average of the revenue, budget, and box office by comparing with different directors/actors

The average revenue, budget, and box office from individual directors/actors does not provide much insight without other comparisons. A future implementation can allow users to

compare the data from individual directors/actors with the overall average of directors and actors in their fields. For example, the top 10 directors with the highest box office show the difference between the directors' average budget with the average budget of all movies in a decade.

We didn't go in this direction initially because the comparison might not be valid. For instance, there's no insight to compare the budget or revenue between action film directors and documentary directors because the goals of these films are different. The comparison needs to be designed deliberately. Otherwise, it might create biases that subvert our intended use cases. For example, encouraging people to focus on the box office and revenue may incur a myopic viewpoint on films as a venue for profit while ignoring the artistic value and social impact they bring. This is the reason why we decided to focus only on comparing this data only within the directors'/actors' work.

B. Labeling the style of the directors/actors by their work

The word "style" was mentioned multiple times when people talked about what they liked about the director/actor. Multiple aspects shape this style, such as intended genre, cinematography, color themes, and narratives. Given the limited data we had access to, our platform focused primarily on genres and color themes relating to style for the audience to explore. Still, the concept of "style" is fascinating and can be further extended to analyze and understand the work of directors/actors. For example, if script data is accessible, there might be ways to label the script with style, if there are existing data about the characteristics of the roles in the movie, we can better understand their strength and experience in acting better.

C. Accurate color palettes + future color visualizations

Our initial idea for visualizing movie colors by grabbing frames of the movie trailers was unfortunately closer to the most ideal color visualization than our current implementation. Due to the short time we had to complete our project, we had to sacrifice accuracy for functionality. Future improvements include figuring out how to use this more accurate visualization in a web application, more engaging and more artistic visualizations, and expanded use cases for our color palettes, such as in designing promotional materials for future movies and psychological analyses of how the color of films affects the audience's emotional and neurological responses. We have - at the very least - utilized a unique approach to visualizing similarities between recommended movies that has not yet been seen elsewhere and if further research proves this useful, this can be used by streaming services and other content platforms.

REFERENCES

[1] The OMDB API, `https://www.omdbapi.com/`

[2] The Movie Database API, `https://developers.themoviedb.org/3`