IMDb Movie Data Visualization

Boqing Zheng, Eric Zhang, Ruiqi Chen

Introduction

As time progresses, people have an increasing number of choices for movie-watching. As a result, various video rating sites have emerged. The IMDb movie website is one such site, which offers detailed information about movies. This information is incredibly valuable for movie enthusiasts, researchers, and industry professionals seeking to gain a better understanding of movie culture and audience preferences. Our data visualization tool will provide an intuitive and interactive representation of the IMDb movie dataset. By referring to our visualization, users can make informed decisions about which movies to watch, analyze market trends, and gain a deeper understanding of movie culture. We hope to explore the relationship between different variables of IMDb films and the feature of IMDb users, specifically working on the following questions:

1. What is the number of films in different genres or countries along the time?
2. What is the correlation between different film categories?
3. What are frequent words in film titles?
4. What are average IMDb scores of films by different countries?
5. What is the distribution of different film ratings by country, year, and film length?

Method

Dataset collection and cleaning

Our mainly data are scraped from the IMDb's website[1], and we only focus on the data scraped from IMDb website for other film and TV sites have adopted anti-crawler mechanisms.

We used the python programming language for data scraping, including libraries called request[2] and beautifulsoup[3] used to access and parse web data. With these packages, we scraped the HTML code of the IMDb web page and extract our data from it. Our dataset is constructed by information of movies' titles, years, ratings, genres, runtimes, IMDb's ratings, meta scores and votes, and geography. Based on the data scraped by request, we filter out different required information with keywords such as div, class, span, etc. For instance, we use *'span', class_ = 'certificate'* as keywords filter for film-ratings search. We also use *regularization*, *split()*, and *replace()* to cut raw mixed data into several pieces of valid information. Because we want to crawl a valid, complete database with nine dimensions of information, we take the first stage of data cleaning. After getting the information of top 11561 movie titles, we discard the data with incomplete information directly. For example, movies with unknown regions and unknown years are discarded. In the end, we get the initial dataset, composed of the 11561 data scraped and the 5002 data left after cleaning, which is *top_5000_movies.csv* file.

After we get the raw csv file of the top 5000 movies, we use Python to clean our data with Pandas[4] libraries for data processing. In this process, we fetched useful information for each

---

[1] "IMDb."

[2] "Requests · PyPI."

[3] "Beautiful Soup Documentation — Beautiful Soup 4.9.0 Documentation."

[4] "Pandas Tutorial."

specific plot to research questions, such as titles, genres, and countries, etc., and then used these filtered data to generate our final target csv files that can be visualized.

Analysis and visualization

To answer our visualization question, we choose to use Timeriver, heatmap, word cloud, map and parallel plot. These plots can fully display the characteristics of time series information, geographic information, as well as the relationship and distribution of various variables to answer the research questions.

The next step is to visualize our results by using d3js and Python. For the map, we chose to use Python's Plotly library. For the rest, we create our visualizations with d3js, referring to online resources, such as d3.js.Gallery[5] and d3 resources on GitHub[6].

After the completion of visualizations, we interpret the IMDb user behavior and the film industry and culture by observing the feature of data.

Results

1. Timeriver of number of films in different genres and countries along the time

Our Timeriver are shown in Figure 1 and 2. The horizontal axis of this graph shows the year of the film, while the width of each color section along the Y-axis shows the number of films, with different colors representing the country/genre of the film.

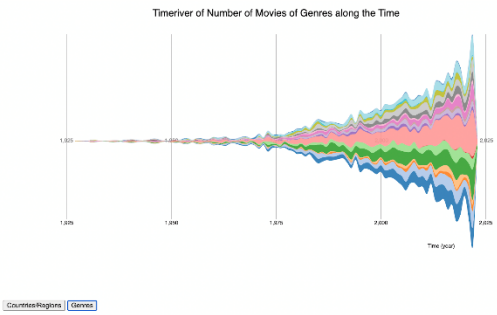

Figure 1 (genres along the time).

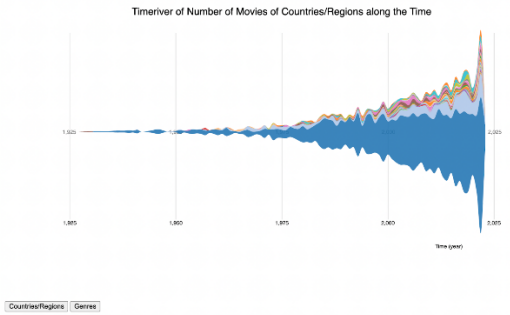

Figure 2 (regions along the time).

In terms of interaction, as shown in Figure 3, the graph can be switched to the number of movies in different countries or genres over time by clicking the button below. Hovering the mouse highlights the selected section to see trends in detail. When the mouse is on the river, you can zoom in and out by sliding the scroll wheel/pad.

---

[5] "The D3 Graph Gallery – Simple Charts Made with D3.Js."

[6] "Gallery · D3/D3 Wiki."

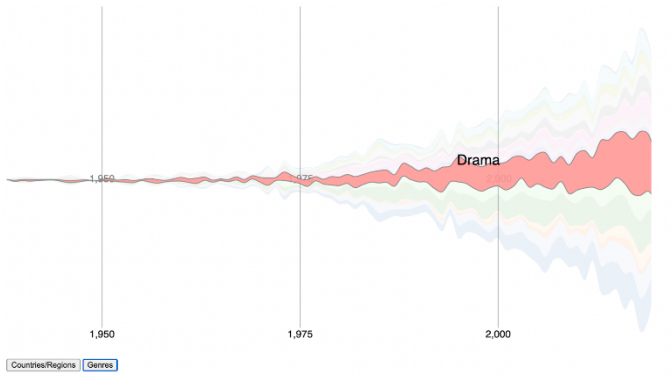

Figure 3.

Through our observation, we found the following:

1. American films account for more than 80%, followed by the UK (dark and light blue in Figure 2): This indicates that the number of American films on IMDb far exceeds that of other countries, which may be because USA is more developed in the world film industry, or may be the users of IMDb are mainly concentrated in USA.
2. In recent years, the total number of films on the site has soared: this is probably due to advances in film-making technology and the development of the film industry. It may also be because IMDb users pay more attention to recent films.
3. The number of Drama films accounts for a large proportion nearly all the time: this indicates that probably Drama films in all eras are popular and have a high resonance with audiences, maybe by the reflection on society.
4. The number of Comedy, Action, Crime and Adventure films exploded after 1975: This indicates that these types of films probably received more attention in market after 1975, and technological progress and cost reduction also made these types of films easier to produce and promote.

2. Heatmap of the correlation between movie genres

Figure 4 is a heatmap that shows the correlation between different movie genres. Basically, since a movie might have different genres such as action, adventure etc., if two genres appear in one movie, we add 1 to correlation. Applying this method to all movies we can get a correlation matrix of movie genres to plot the heatmap. The x and y axis represent the genres of movies and the color scales represent the correlation between genres. The interactions include zooming the graph, dragging the graph, and mousing-on to see the exact number of correlations, as shown in Figure 5.

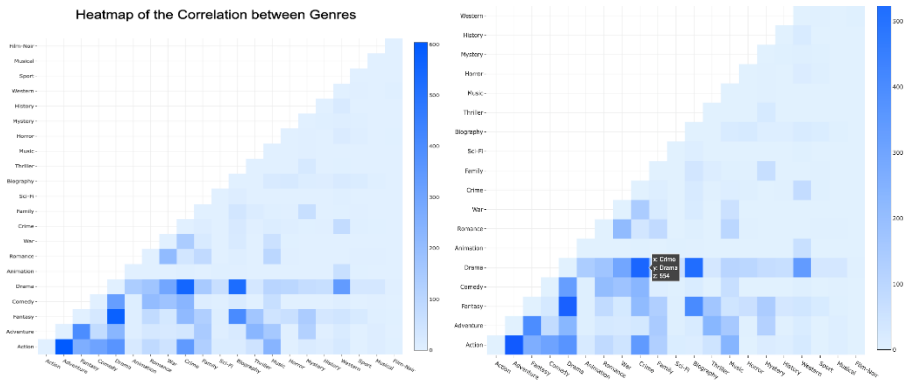

Figure 4. **Figure 5.**

From the heatmap, we can observe that:

1. The top four categories with a high correlation are Adventure - Action, Drama - Fantasy, Drama - Crime, and Drama - Biography, which shows that creators prefer to create movies in these categories and these combinations have a favored status.
2. In terms of individual categories, Action, and Drama are related to more categories than others, which indicates that the films in these categories are highly integrated and diversified. This versatility and flexibility in storytelling can make movies more appealing to different viewers and help them achieve broader commercial success.

3. Word Cloud of the top 200 occurred words in movie titles

Figure 6 is a word cloud of the top 200 occurred words in movie titles, which is generated by filtering out the 200 most frequent words in movie titles. We also removed special symbols, stop words, and uncommon words (II, III, etc.). The frequency of the words is reflected in the shade and size of the text. The interactions of the word cloud include searching the word and mousing-on to see the frequency of the word, as shown Figure 7.

Figure 6.

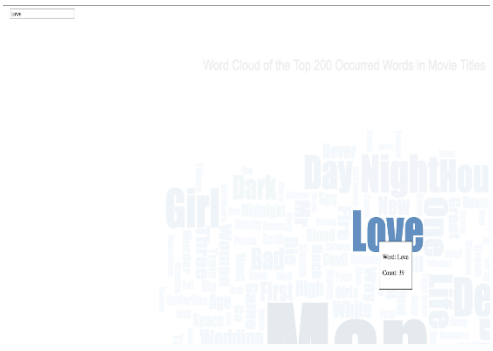

Figure 7.

We can observe that the word “Man” appears most frequently in the movie, followed by “World”, “Love”, “Last”, “Dead”, “Night” and other words. Words like “Man” and “World” may reflect filmmakers' exploration and expression of human experience, society, and

relationships. Words like “Love”, “Last”, “Dead” and “Night” refer to emotional themes like love, death, loneliness, and so on.

4. Map of average IMDb scores by different countries

This is our choropleth map, as shown in Figure 8. It shows the various countries and regions in time. The luminance of colors reflects the average score of the film from that country or region.

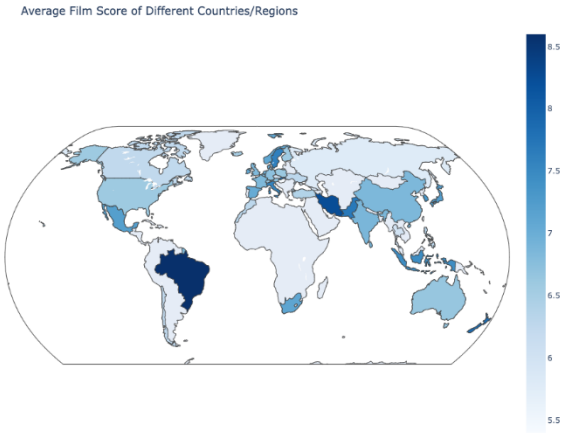

Figure 8.

Figure 9 shows the interaction that the map can be zoomed and panned to observe different regions. Meanwhile, putting the mouse on different regions will display the region name, the average score, and the number of movies in this region.

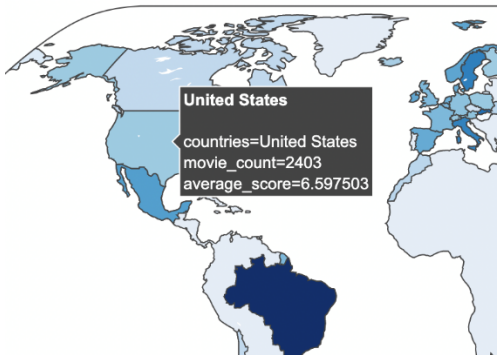

Figure 9.

Through observation, we can conclude:

1. Some regions with high average scores tend to have fewer films on the site, such as Brazil and Iran: this may be because there are not enough film samples on IMDb to represent the country's film industry, leading to large fluctuations in average scores. It could also because IMDb is not well known in these countries, resulting in fewer films being registered on the site.
2. The reason why some countries with the highest number of films have lower average scores may be because their film industries are relatively mature and a mature film industry may be more focused on commercial success than necessarily pursuing artistic value, and therefore may be more inclined to produce high-grossing commercial films.
3. Many countries have no or a few films on IMDb. Some regions’ styles are different from the tastes of mainstream audiences on IMDb, resulting in a smaller number of films on the site. For example, China produced more films for the local market.

5. Parallel plot of the distribution of film-ratings by countries/regions, years, and lengths

In tour parallel plot, Y-axis shows the year, the film-rating, the country/region, and the length of the movie. Each line represents a movie, and each color represents a rating (Figure 10). The intersection of each line with the Y-axis indicates the value. Due to the multi-regional film-rating system displayed on IMDb, we classify all films into three categories, Adults (orange), Teenagers (blue) and Children (green).

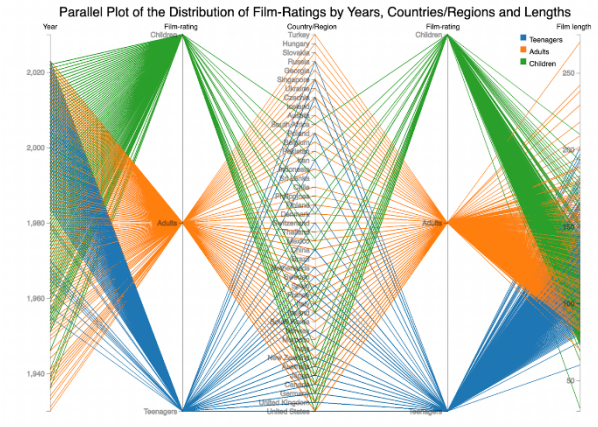

Figure 10.

In terms of interaction shown in Figure 11, the filtering can be achieved by placing the mouse over the lines of a specific color, showing only the lines of this color for a better view of its distribution. Meanwhile, the number of movies of the selected rating will be displayed around the mouse.

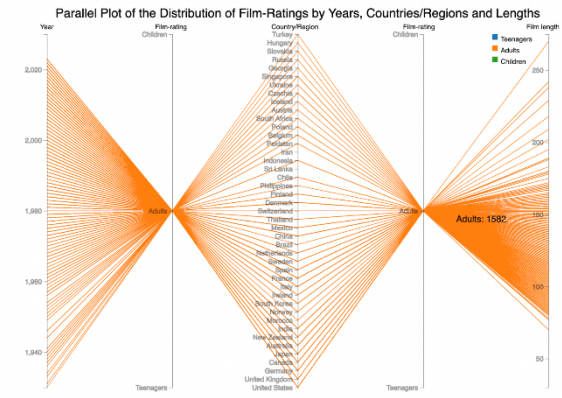

Figure 11.

Through observation, we found that:

1. In terms of the year, films rated to Teenagers tend to concentrate after the 1970s, while the distribution of others is relatively even. This may reflect the cultural and social changes of this period, in which films pay more attention to the needs and interests of teenagers.
2. In terms of countries/regions, movies rated to Adults have appeared in all countries, while films rated to Children and Teenagers distribute sparsely: this may reflect that adults are the main group of moviegoers, and there are the different demands and market scales of movies for teenagers and children.

3. In terms of length, films rated to Adults have a large span of time, while those of Children and Teenagers are relatively shorter: this may be because the interests and demands of adult audiences are more diverse, while those of young audiences have limited tolerance to the length of films.

Discussion

The followings are our evaluations:

1. **Timeriver to show the relationship between the number of films in different time.**

   User tasks: Characterize Distribution, Sort

   Analysis: Our timeriver presents quantitative data on the number of years and films in the table, as well as categorical data on film categories/countries. Users can see how movies change over time.

   Future improvement: Brushing could be used to select what parts of the area to enlarge and change the area of axes without changing positions.

2. **Heap map to show the degree of correlation between different movies categories.**

   User tasks: Correlate, Retrieve Value

   Analysis: The heatmap gives people an intuitive result of the correlation between different genres. Users can easily find whether there is a strong correlation between two genres by looking at the saturation of the color.

   Limitation: This only shows the relations between two genres and may not identify whether a single genre is a majority among all the movies.

3. **Word cloud to show high-frequency words in film and television titles.**

   User tasks: Find Extremum, Filter, Retrieve Value, Sort

   Analysis: The word cloud provides the top 200 frequent words in movie titles. Users can see the most frequent words in movie titles and their frequencies.

   Limitation: The usage of space seems not so good. The interaction for seeing the whole word cloud may cause inconvenience when users try to find the position of the word after searching.

4. **Map to show the average scores of films in different countries.**

   User tasks: Retrieve Value, Characterize Distribution, Sort

   Analysis: Our map shows the categorical location and the quantitative number and score of movies in the table. Users can see the scores of movies in this region by observing the colors of different areas on the map and clicking on different areas to see more detailed information.

   Limitation: IMDb is mostly used by English users, so some specific film and television situations of many countries are not well shown.

   Future improvement: Increase our data set based on more websites.

5. **Parallel plot to show the relationship between film-rating, countries/regions, and year.**

   User tasks: Characterize Distribution, Correlate

   Analysis: Our parallel plot shows the quantitative and categorical data of rating, year, country, and duration in the table. Users can see the distribution of movie ratings on other variables by looking at the distribution of different color lines on other axes, while placing the mouse over a color to see only that color.

Limitation: Due to the over-plotting, we did not explore the relationship between different variables besides film-ratings.

Future improvement: Explore the relationship between other variables through more marks and channels to deal with over-plotting.

Conclusion

In summary, our five data visualizations explore the relationship between multiple variables of movie data on the IMDb website, including year, film-rating, country, number of movies, movie genres, score, etc. Based on the above visualization results, we clearly observed the laws behind IMDb movie data, and we answered our inquiry questions through interpretations to plots. We hope our findings will provide a perspective for people to understand the audience preferences and the development and culture of the film industry from the data of IMDb, one of the most popular film rating websites.

Roles and responsibilities

Proposal: Ruiqi Chen, Jinjia Zhang, Boqing Zheng

Data scrape and clean: Top 5000 movies detailed information dataset (Ruiqi Chen), Top 250 TV shows dataset (Boqing Zheng), Top 250 movies dataset (Jinjia Zhang)

Visualization: Timeriver (Boqing, Zheng, Ruiqi Chen), Parallel plot (Boqing Zheng, Jinjia Zhang), Heatmap (Jinjia Zhang, Boqing Zheng), Map (Boqing Zheng, Ruiqi Chen), Word cloud (Jinjia Zhang, Ruiqi Chen)

Interim/Poster/Dashboard/Report: Boqing Zheng, Jinjia Zhang, Ruiqi Chen

Reference

"Beautiful Soup Documentation — Beautiful Soup 4.9.0 Documentation." Accessed February 19, 2023. https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

"Gallery · D3/D3 Wiki." Accessed March 8, 2023. https://github.com/d3/d3/wiki/Gallery.

IMDb. "IMDb: Ratings, Reviews, and Where to Watch the Best Movies & TV Shows." Accessed February 19, 2023. https://www.imdb.com/.

"Pandas Tutorial." Accessed March 8, 2023. https://www.w3schools.com/python/pandas/default.asp.

"Requests · PyPI." Accessed March 8, 2023. https://pypi.org/project/requests/.

"The D3 Graph Gallery – Simple Charts Made with D3.Js." Accessed March 8, 2023. https://d3-graph-gallery.com/.