Visualizing Movie Preference

Varoot Phasuthadol

School of Information
University of Michigan
Ann Arbor, MI 48109, USA

varoot@umich.edu

Working system is available at
http://infovis.varoot.com/

Abstract

Visualizing movie preference is not an easy problem. A person can like many different movies and genres, so it is hard to summarize everything in one shot. Instead, I aimed for the visualization that allows the users to explore the ratings that I gave, compared to the average ratings on IMDb and filtered by genres. The visualization presents the ratings in a 10×10 matrix, with my personal ratings on Y axis and IMDb average ratings on X axis. Users can select the genres to filter the data, and select one cell on the matrix to see the list of movies in that cell. The visualization showed different patterns for different genres, so helped me identify the genres that I like. It also inspires me on how to pick the movies to represent my preference. The closer a movie is to the top-left corner, the stronger my preference for that movie.

1 Introduction

As a film lover, I often get asked about my favorite movies. I always find this question hard to answer. Picking one on two movies as an answer would not be a good representative of what I really like. Providing a comprehensive list of all the movies I like would be too exhaustive and too hard to interpret. Figuring out my Top 10 or Top 20 is also hard because there are many titles to choose from, not to mention having to try to compare them.

Moreover, there is a distinction between the movies general people like, and the movies that I like particularly. For example, a lot of people like Lord of the Rings, telling other people that I like it too does not really help them understand my movie preference.

Thus, a proper visualization would be able to show my personal preference, as well as average person's. The visualization would show an appropriate level of detail, where users can recognize patterns and draw conclusion, while an interaction would allow an access to more details if needed.

2 Data Collection

I have been a regular user of IMDb (The Internet Movie Database) for a long time. I usually go there to check the upcoming movies and rate the movies I have watched. IMDb is one of the most popular websites for movie-goers. I personally have rated more than 200 titles on IMDb so this data would be perfect for this project. The website also allows users to download their data, which include the movie titles, users' ratings,

and average ratings (called IMDb rating). By using the data offered by IMDb, other people can also use this visualization system to plot their movie preference.

2.1 Downloading ratings data

Once logged in on IMDb (www.imdb.com), go to user account by clicking your name on the top right of the site. Click on the link to "view your ratings history". You will see the list of all the titles you have rated so far. Go to the bottom of the page, there will be a link that says "Export this list", right before "Recommended for you" section. Clicking on that link will get you a CSV file containing all your ratings data.

2.2 Converting data for the system

The data downloaded from IMDb is in CSV format. To better suit the design of the system, the file has to be converted to a JavaScript file. I have created a PHP code for this purpose. Running data.php will convert *ratings.csv* in *js* folder to *data.js* which can be read by the system. This conversion process will omit the titles that are not feature films (such as TV shows and short films).

3 Visualization Method

3.1 Variables

There are four variables that I want to capture in this visualization. They are as follows:

1. **Movie**

   Movie is best presented by a title (text). I also include release year, in case there are different movies with the same name. The system does not have to always show movie titles, since this might not be the level of detail that the users need. A link to an IMDb page would also provide users more information to a movie.

2. **Genre**

   Genre is a very important variable in this visualization. A person's movie preference differs mostly from genre to genre. IMDb provides 26 different genres, which are tagged to each movie title. A movie can have multiple genres. For example, *the Lord of the Rings* is an action, adventure, drama, and fantasy film.

3. **My personal rating**

   My rating shows my preference to a particular movie.

4. **Average IMDb rating**

   This variable helps comparing and identifying the difference between my preference and average person's preference.

3.2 Ratings matrix

At first, I thought of calculating the difference between my ratings and average ratings, in order to identify the unique pattern of my preference. However, I could not find a suitable formula to calculate the difference. I wanted to use a simple subtraction (*my rating – average rating*) but plotting this value with my ratings only reveals a trend that the higher my personal rating, the more likely this value is positive. I decided instead of showing calculated value, I should show the raw data, and let the chart shows the trend by itself. The visualization should let the users explore their ratings and see if there is any pattern themselves.

I decided to plot both ratings as a scatter plot, with my personal ratings on Y axis and IMDb average ratings on X axis. Position is one of the best ways to present quantitative data. Y axis offers more intuitive interpretation since higher ratings would be "higher" on top of the chart. My personal ratings are integers between 1 and 10, but IMDb ratings are decimal points. So instead of showing a scatter plot of individual movies, I simplified the chart by rounding the IMDb ratings into integers, thus making it a 10×10 matrix. Grouping data together makes the visualization look more appealing and easier to read. I also put diagonal line to split the chart into two parts, the upper-left which are for the movies that I rated higher than the average ratings, and the lower-right which are for the movies that I rated lower. The distance from the diagonal line also signifies how much I like or dislike the movie, compared to an average person. The more the movie is displayed to the left of the line, the stronger my preference for that movie is.

Figure 1: *Ratings matrix showing all data*

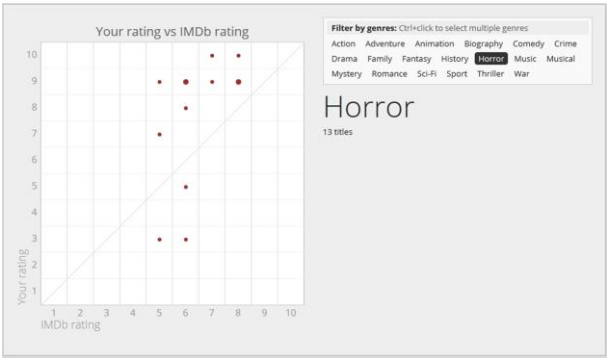

Figure 2: *Ratings matrix showing filtered data by selecting a genre*

Figure 3: *Ratings matrix showing filtered data by selecting multiple genres*

The movies are represented by a solid circle in each cell. The number of movies is encoded with the size (area) of the circle. I chose to encode using area, instead of radius, so that when the visualized is viewed as a whole, the density of the ink in each cell would reflect more accurately to the number of movies *(Figure 1)*.

3.3 Interaction

I chose to use interaction to capture the other two variables (genre and movie). The system would allow the users to explore and drill down the information they want to see. Since a movie can have many different genres, it is good to allow the users to select multiple genres at the same time. The data in the matrix would update to include only the movies that belong to those genres *(Figure 2 and 3)*. The circles also change color to represent different genre. An animation was also added to draw the viewer to the changes in the chart.

Users will be able to see the alphabetical list of movies in each cell by clicking it *(Figure 3)*. This way the users can pinpoint outliers and draw conclusion by themselves.

4 Implementation

I use D3 JavaScript library for visualization. As mentioned earlier, raw data from IMDb is converted to a JavaScript file using PHP script. This conversion is only one-time process and the PHP script is not needed to run the visualization.

HTML5 and CSS are used for standard compliance. Most of the styling is specified in the CSS file. This allows more flexibility for modification, since the code for presentation is separates from the processing.

5 Patterns

One pattern I recognized from the visualization is that I usually rated the movies higher than the average ratings. Most of the movies I rated have the average ratings of 5 to 8 *(Figure 1)*. This is due to the fact that the IMDb ratings are averaged; therefore they tend to be around the middle range. Moreover, I am a regular user on IMDb and I usually check the ratings before I watch a movie. So if the rating is really low (lower than 5), it's less likely that I would watch the film. My personal

ratings, however, are more diverse, especially for the movies that have average ratings of around 5 or 6.

By clicking each genre, I recognized some patterns about my movie preference. For genres like *Comedy*, *Crime*, and *Family*, there are almost no movies that I rated lower than the average ratings (on the lower-right part of the chart) *(Figure 3)*. The circles are also grouped tighter, which means I have stronger preference for them.

For genres like *Horror* and *Mystery*, the circles are divided roughly into two groups, the one above the diagonal line and the one below *(Figure 2)*. This difference shows that for movies in these genres, I either like it or hate it.

6 Future Work

There are some improvements that can be done to this visualization. First, there is no summary of what genres are presented most in my rating history. Instead of showing just the name of each genre, a circle can be added with the size reflecting the number of movies in that genre. The interaction part can also be improved by fading out other genres when a genre is selected, and show only the combination of genres that exists in my rating data. For example, if I select *Family*, *Biography* would fade away since there is no movie in my data that is in both *Family* and *Biography*.

There are also other ways to represent a movie, such as using poster image, or a pie chart representing genres that belong to that movie. This way, it would be easier to glance through the movie list to draw conclusion.

The visualization matrix also inspired me on how to calculate a value that reflects movie preference. Since the diagonal line represents neutral preference compared to an average person, a line that is drawn from the point that a movie is plotted on the chart to the top-right corner of the chart would form an angle with the diagonal line *(Figure 4)*. This angle would represent my preference towards the movie, compared to an average person. Similarly, if the movie is in the lower-right part of the chart, the angle can be calculated using the lower-left corner as a vertex, and make the value negative to signify non-preference.

I actually calculated this value and called it "*affinity index*". This index can be used in further visualization to show the movie that I really like, compared to an average person. It can also be used to summarize my top movies and offer an overview of my movie preference, which is something that is still lacking in the current visualization.

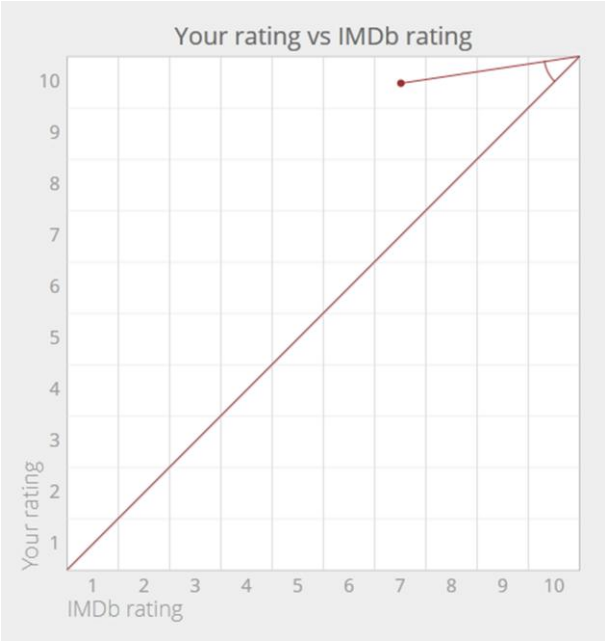

Figure 4: *The angle a movie made to the diagonal line represents my preference towards the movie.*