CSCI-SHU 235 Information Visualization

Final project paper—MovieLens Recommendation

Yijing Zhou and Yi Wang

1 ABSTRACT

For this final project, we present a movie visualization system from data exploration to movie recommendation. After conducting data analysis, we build a system to visualize: 1)background of the user group; 2)movie data overview and statistics based on different genres, genders, user occupations and 3)movie recommendation system built upon collaborative filtering recommendation methodology. Through visualization, the study provides a deep understanding of the movie industry and implements the simple recommendation principle behind big movie sites like Netflix.

2 INTRODUCTION

Movies are now an inseparable component of people's daily life. The trend of the movie industry is highly dependent on audiences' taste and preference over time. There are several questions that motivated us to select this topic: What type of movies are among the most popular? How many differences are there among different users' tastes? How does the movie recommendation system work to accurately recommend related movies to users? To answer these questions, we conduct a thorough examination of the user preference and underlying patterns in the US market and also simulate a movie recommendation mechanism using collaborative filtering methodology. The data we employ is the 1M MovieLens data including 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users. The user ratings are recorded among year 2000 to 2003. All the movies are published between 1919 and 2000. Ratings are made on a 5-star scale. We choose this dataset instead of the initial ones to access user backgrounds, which is a key factor in this industry.

3 RELATED WORK

There are many people who already conducted basic data analysis using the MovieLens data and implemented some rudimentary visualizations using Python, R and other statistical packages. Also, there are many open sources that provided the complete steps of movie recommendation system using Python or Java—we actually refer to the one on GitHub to build our recommendation system. However, few of them succeed in visualizing the whole dataset and the recommendation system with all faithfulness, expressiveness and elegance. Hence, we intend to contribute to this topic and make further improvements.

4 OUR METHODS

Our system contains six visualization designs to present our analysis on the movie and user dataset as well as the recommendation system.

4.1 Data Analysis Techniques

Our first step before visualization is to prepare the original raw data using proper data analysis tools. We first conduct data analysis by Python and then load the data into Javascript for visualization. The original data is composed of three parts: Movies, Ratings and Users. The movie data includes approximately 3900 movies with its title, movie ID, released year and genre. The ratings data contains each user's recorded ratings—each user has at least twenty ratings. The users data includes 6040 MovieLens users with user ID, age, state location and occupation.

Basically we employ Python pivot table to explore and consolidate data features to prepare for visualization. For the users data, we calculate the number of total, male and female users as well as the overall average age of users for each of the 51 states. For the movies and ratings data, we calculate the average ratings and number of views. Number of views are calculated as the total number of ratings each movie receive. The calculation is based on the assumption that each user will only rate the movie after they have watched it. Popularity of each movie is evaluated by the number of views and reputation of each movie is evaluated by the average rating.

Besides, we also do analysis on each movie genre. For each genre, we calculate its yearly average rating by gender and number of views by occupation.
In terms of our recommendation system, it is built by neural network using Python. There are basically two methods to build the recommendation system: content-based filter and collaborative-based filter. We choose the second method to visualize. The final data is first outputted as json file and then loaded into Javascript.

4.2 Visualization Design & Techniques

1. We plot an interactive user map (Figure 1) to present the geographic information of the users in our dataset. The color of each state represents the number of users living in it. The darker the color, the more users in this state. It is clear that California, New York and Minnesota have the most MovieLens users. Overall speaking, the number of male users outweighs the number of female users in our database.

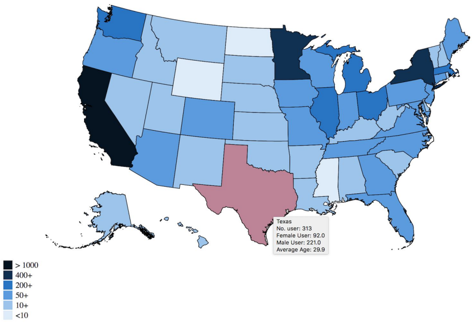

Figure 1

2. An interactive brushable scatterplot and a word cloud plot (Figure 2) are implemented to show overall movie trend across the years. The circle item represents an individual movie and the color represents the genre it belongs to. The X-axis represents the released years and Y-axis marks the average rating or number of views of the movies. We can see that number of films have boosted after 1970s. Overall speaking, comedy, drama and action movie are the top three popular genres among all the 18 genres. On the contrary, in terms of ratings, genres above perform lot worse. Comedy ranks 12 over 18 genres and action ranks 14. Genres with the highest average rating is film-noir, documentary and western, which all share poor popularities. To show a more straightforward and clearer illustration, we employ a word cloud to better demonstrate the popularity of each genre by gender (Figure 3 shows male's preference). Comparing the genders, we find that comedy and drama are of common preference while the males prefer action movies specifically and the females have special taste for romance movies.

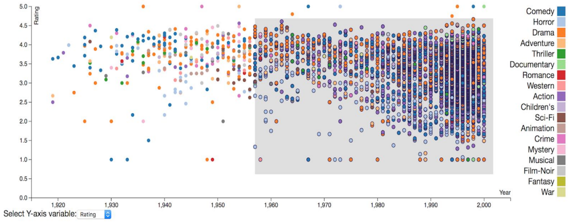

Figure 2

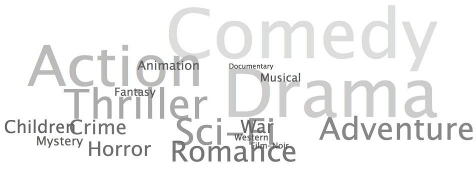

Figure 3

3. For further analysis on each genre's reputation, we use a bar chart to plot the average rating of each genre by gender over year 2000 to 2003 in Figure 4. Overall speaking, the charts show a decreasing average rating users over the years. Female tend to rate higher scores than male.

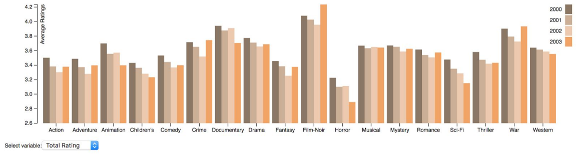

Figure 4

4. Figure 5 shows the distribution of users by occupation across each genre. Each donut pie represents a genre with its genre name printed in the inner circle. The outside sectors illustrate the proportion of each occupation for each genre. The findings from the analysis are: Students and people work in business sectors are two major audience segments of the movie industry. By comparing areas of sectors, we could also draw conclusions such as students have specific preference over animation and children's movies while engineering is the major

target of Sci-fi movies.

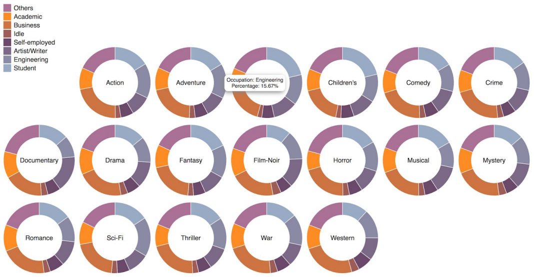

Figure 5

5. A set of radial tree (figure 6) is used to visualize our movie recommendation system. The recommendation system is built upon a collaborative filtering recommendation methodology. A collaborative filtering recommendation methodology takes one movie as an input, find top k similar users who share the same attitude to the movie and retrieve a list of favorite movies of those users. The system will filter the list and return five movies randomly from the list to you as recommendation. In our system, we pick the top 50 popular movie (ranked by the number of view) as the input list and generate recommendations for them respectively. After acquiring the five recommendations based on the original input, the system will generate a second layer of recommendation as a representation of the recursive process. As shown in Figure 6, the root of the tree stands for the input movie, the first layer of the tree is the five movies recommended based on the root. The second layer of 25 movies is generated based on the movies in the first layer.

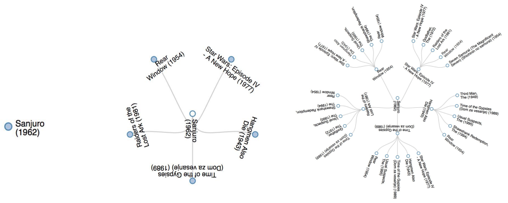
Sanjuro (1962)
Rear Window (1954)
Star Wars: Episode IV - A New Hope (1977)
Raiders of the Lost Ark (1981)
Sanjuro (1962)
Hangmen Also Die (1943)
Time of the Gypsies (Dom za vesanje) (1989)
Third Man, The (1949)
Usual Suspects, The (1995)
Shawshank Redemption, The (1994)
Godfather, The (1972)
Seven Samurai (The Magnificent Seven) (Shichinin no samurai) (1954)

Figure 6

5 USAGE SCENARIO

This visualization system is built upon various interactions. We employ tooltip on the user map (Figure 1) to show the average age and gender distribution for each state. Users can put mouse over each state to see detailed information. In terms of the scatterplot (Figure 2), users can brush an area to zoom in and examine the area they are interested in. The Y-axis also provides two options of “average rating” and “number of view”. User can compare the average rating of movies across years and across genres based on this chart. For figure 4, the Y-axis has the option of total, average female and average male rating across genres and years. The figure can be used to compare different genders’ preference over different movie genres across years. Users can track the change of preference by gender based on the bar’s transition of its height. Figure 5 also uses tooltips to show the exact percentage of the occupation. Compare the proportion of each occupation across different donuts (genres), you can have a view of each occupation segments preference on each genre.

For the recommendation system design, users will be given five movies from the top 50 list randomly and choose whether they like it by clicking on the node. If they click the node of the movie, the tree will spread the first direct recommendation layer and then spread the second layer of recommendation after a few seconds. The system will feed five random root nodes at one time, if the user are not interested or want to explore more, he can click the update bottom and the system will rearrange five different random movies.

6 CONCLUSION & FUTURE WORK

Based on our six visualization designs, we have a deeper understanding of the movie industry and the underlying patterns. It is true that visualization provides a more effective and efficient way to convey information compared with showing pure numbers. In the future, there could be many improvements if more time are allowed. Firstly, we plan to add more up-to-date datasets. Our current data has movies from year 1919 to 2000 and users rating from 2000 to 2003, which is somehow outdated and might not be perfect to capture the trend. Hence, adding more current data points could help with reflecting on the modern movie industry, which will be of higher value and practical meaning. A second improvement could be made to integrate more features into one single visual design. For instance, in figure 5, we plan to use each donut’s size to represent each genre’s total views, which will be more straightforward and comparable. A third improvement could be made to add more interactive designs. For example, instead of merely showing the recommendation tree, we could guide the user through the whole process—first rate their watched movies, and generate a recommendation tree particularly according to their ratings—this will let them understand the system deeper and make it more persuasive. Finally, we plan to compare our recommendation system with those big movie sites or conduct some user surveys to check whether our recommendation algorithms are reasonable. Also, we plan to listen to more opinions and suggestions about our six visualization designs.

ACKNOWLEDGMENTS:

This final project owes big thanks to Professor Nan Cao and Doctor Xianbin Gu.

REFERENCES:

[1]E.Grimaldi.http://towardsdatascience.com/how-to-build-from-scratch-a-content-based-movie-recommender-with-natural-language-processing-25ad400eb243 Accessed: 2018-11-14.
[2]S.Chen.https://github.com/chengstone/movie_recommender Accessed: 2018-11-15.
[3]M. Bostock.https://bl.ocks.org/mbostock/3887051 Accessed: 2018-12-01.
[4]M. Bostock.https://bl.ocks.org/mbostock/3888852 Accessed: 2018-12-01.
[5]M.Chandra.http://bl.ocks.org/michellechandra/0b2ce4923dc9b5809922 Accessed: 2018-12-01.
[6]KoGor.http://bl.ocks.org/KoGor/5685876 Accessed: 2018-12-01.
[7]Coopey.http://bl.ocks.org/ericcoopey/6382449 Accessed: 2018-12-02.
[8]FernOfTheAndes.https://codepen.io/fernoftheandes/pen/pcoFz Accessed: 2018-12-02.