Télécom Paristech

IGR204 Report

Visualization

Movie Focus

Auteurs:
Anna Arias Duart
Bernard Lupiac
Júlia D'Ávila
Ming Yan

October 27, 2017

Contents

List of Figures

1 Introduction

The goal of this project is to create a fully-functional interactive visualization of a dataset. This project was made based on a multidimensional dataset of about 1600 movies with properties such as length, main actor and actress, director and popularity.

The visualization detailed below was a result of applying the concepts learned in this course and thinking about the data chosen and what the user would like to do with it.

2 The data set

The dataset has 1659 films, these films are from the 20^{th} century. They are described with 10 dimensions. Each film has:

1. Its release year, which is a quantitative variable (example : 1928, 1929, 1930, etc.)
2. The duration given in minutes, which is also a quantitative variable (example : 120, 121, 140, etc.)
3. The full name of the title which is a nominal variable (example : "After hours", "Raging Bull", "King of Comedy", etc.)
4. The genre of the film, which is a nominal variable. They are classified into 16 types: action, adventure, comedy, crime, drama, fantasy, horror, music, mystery, romance, science fiction, short, war, western and westerns.
5. The name of the main actor, which is a nominal variable (example : "Robert De Niro", "Danny Glover", "Jean-Marc Barr", etc.)
6. The name of the main actress, which is a nominal variable (example : "Jessica Lange", "Rosanna Arquette", "Kin Sugai", etc.)
7. The name of the director of the movie, which is also a nominal variable (example : "Akira Kurosawa", "Alfred Hitchcock", "Martin Scorsese", etc.)
8. The popularity of the film, which is a quantitative value from 0 to 100.
9. If the movie has any award, the value will be "Yes" and if not "No", so it is binary variable.
10. The last dimension is the name of the cover photo, which is a nominal vairble (example : "NicholasCage.png", "seanConnery.png", "Bergman.png", etc.)

3 The target users

The target users of this visualization are people who like the cinema and are interested in exploring the relationships that exist between movies. People have usually favorite actresses, actors or directors. This visualization will allow them to find movies in which they are probably interested since they are performed by the same actor, actress or directed by the same director.

3.1 Visualization tools

Intuitivity is the most important concept for those tools. Based on the pre-attentive processing, different colors are used for each genre to allow an easy differenciation.

1. First visualization : the main page.

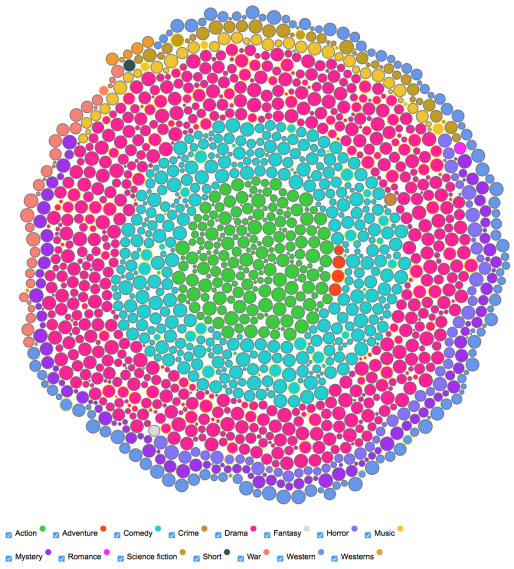

Figure 1: First visualization : the main page

In terms of data, in the first visualization (see Figure 1) a tri-variate representation is used, three variables of each data case are used:

- Genre, represented by color.
- Popularity, represented by the size of the circle.
- If the movie has any award, it would be represented by a yellow outline, color with which gold is commonly associated.

A dynamic query (see Figure 2) is offered, that allows filtering the genres shown. The visualization starts with all genres shown, the user can remove them one by one, until only one remains.

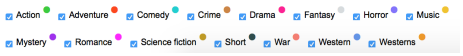

Figure 2: Dynamic query

As there are many movies, a zoom tool is offered, that allows zooming if needed (see Figure 3).

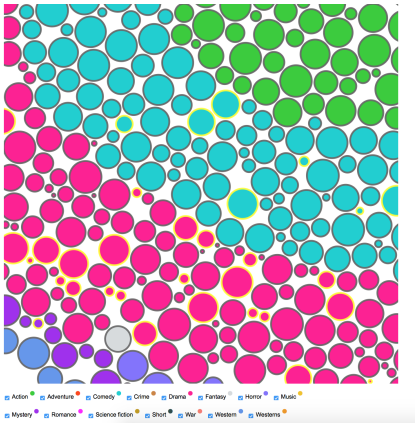

Figure 3: Zoom

For further information about a movie, the user can pass his mouse over any movie, and a tooltip will appear showing the following information:

- The full name of the title.
- The release year.
- The genre of the film.
- The length given in minutes
- The name of the director of the movie.
- The name of the main actress.
- The name of the main actor.
- The popularity.

In order to help the user, the movie with the mouse over it will also be highlighted with a green color (see Figure 4).

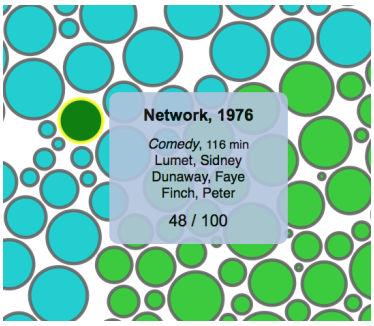

Figure 4: Further information

2. Second visualization : the graph.

   This representation is consistent with the three variables of the first visualization (genre represented by color, popularity represented by the size of the circle and the awards represented by a yellow outline).

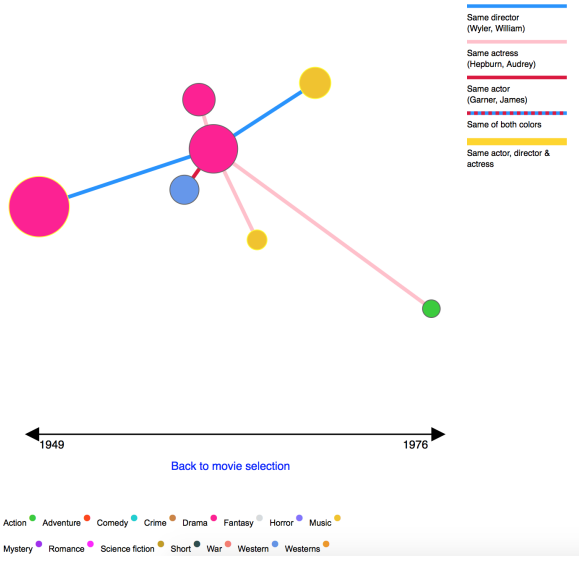

Figure 5: Second visualization : the graph.

But it also adds the possibility of expressing different relationships with the other movies.

- Movies with the same director.
- Movies with the same actress.

- Movies with the same actor.

In this visualization, time is also a variable: the release year of the films is represented. The visualization tool used to represent this variable is the space. If a film is drawn more to the left it means that it is older than one drawn more to the right.

In order to help the user, a time line is represented, to get an idea of when was the release year of the movie, see Figure 5.

Similarly as in the first display, more information about a movie will appear if you mouse over any movie.

Finally, you always have the possibility to return to the first visualization.

4 Representative tasks

To begin with, users want to have a overview of the whole dataset. In other words, The design should give users an impression of scale of the whole dataset at the first sight.

Secondly, details of a certain movie are basic information that users want to get from a dataset. Showing detailed information of a selected movie is also a basic task that should be achieved in the visualization.

Furthermore, the situation in which users want to select a movie to watch from the dataset. In this circumstance, how can users choose a movie that they like in a relativley short time? After thinking about this question, a movie-genre filter was added to help users decide. Size is used to distinguish the popularity of movies. In this way, users can select movies with a high popularity easily.

It is also expected that, apart from getting basic movie information, users may also are interested in the relations between different movies. That is, users want to find out movies with the same director, actor or actress. So it requires this design to have the ability to link different movies.

5 The design

For this design, there are two different representations for the data:

5.1 Main Page

The main page, that shows all movies, see Figure 1-

This page serves for the user to have an overview of the data. Each bubble represents a movie, with the size of the bubble showing its popularity, and its color being the genre.

There are many interactions that the user can make in this page:

- Retrieve: the user can see more specific data from a movie hovering the mouse over it.

- Filter: it is possible to choose exactly the genres that the user wants, using the checkboxes on the bottom of the page.
- Because of the small size of some bubbles with low popularity, it is also possible to zoom with the mousewheel and move everything by clicking and dragging on an empty space.
- By clicking on a movie, the user is shown another representation of the data, explained below.

5.2 Movie Focus

A graph with all movies related to the movie selected, see Figure 5.

This page offers more detailed information for a movie the user has selected. It shows all movies that have an actor, actress or director in common to the one chosen with edges of different colors between nodes. Popularity is still represented by size, and genre by the color. The x-axis represent time: movies on the left were made before movies on the right.

- Retrieve: the user can still see more specific data from a movie hovering the mouse over it.
- By clicking on a movie on this representation, the user is shozn its own graph.
- Since the positioning of the bubbles can be a bit awkward at times, refreshing the page or clicking on the central bubble will reposition all bubbles in a different way.
- The user can return to the main page by clicking on the "Back to movie selection" text on the bottom of the page.

This designs has aspects in which it excels, and also in which it is lacking:

5.3 Weaknesses

- The main page visualization could have separated clusters for each genre of movie (as in the design document). After many tries, this idea was abandoned for a very similar one.
- This design is clearly not suited for searching a movie in particular: there is no way to see the title without hovering over the movie.
- The graph visualization can be very confusing when there are many related movies, and they can also get out of the boundary of the screen. The best solution would be to scale everything on the screen depending on the number of elements, the radius of their bubbles and their disposition. Unfortunately, this would take too much time to implement and to compute, while also giving users different scales for different movies, which can be confusing.

- In the graph visualization, there is no way to see for example, a movie directed by the actor that played in the chosen movie.

5.4 Strengths

- As the dataset has lots of movies and the bubble size depends on the popularity of the movies, some of them are not so visible (for example, the movies if popularity 1/100) so the zooming aspect of this design allows the user to zoom and navigate easily among the movies.
- Another resource that facilitates the visibility of the films are the check box, which can filter the data according to the user's choice.
- Also if a few genres are selected (for example two), it is possible to see which genre has more movies, see Figure 6.

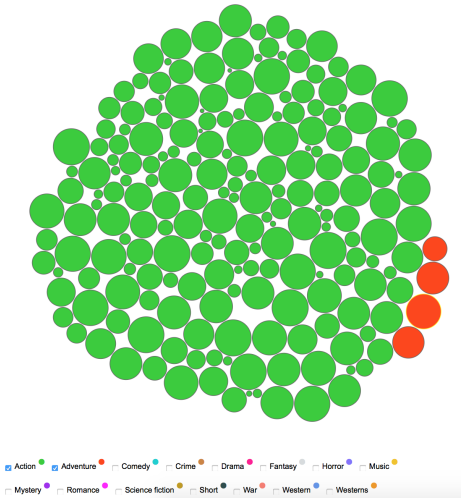

Figure 6: Comparison

- The main strength, because of the graph visualization, is seeing relations between different movies.
- By extension of that last aspect, it is also possible to see when an actor, actress or director was most active, and in what kind of genres they usually participated.
- It is also possible to see errors in the database more easily, for example when an actor has movies made in the '40s, and also in the '2010s (unless it's Christopher Lee). There are a handful of movies from an actor called Tracy Spencer for example that are marked as being made in 1991 in the dataset, even though he died in 1967 (See those data cases here) .

6 Acknowledgments

As we worked as a group, each component of the group was responsible for a specific part of the project. Not everybody knew how to code in D3 or was comfortable doing it, so some parts of the code are less "clean" than others. Besides, we took some fragments of code from parts of the web, each time this was done, we left the source link commented above it.

The first thing we looked for, it was how to create a bubble chart, we saw a lot of examples on the internet but the one we got based on was:
Basic Bubble Chart: How to make a basic Bubble Chart in D3

Then, a zoom feature seemed useful:
Simplest way to add zoom/pan on d3.js (version 3 and 4)

Furthermore, the tooltips of this projects were based on: Simple d3.js tooltips

7 Conclusion

In conclusion, it can be said that Edward Tuft's design principles were achieved in some way:

- Show the data : the data has been shown in different representations.
- Make large data sets coherent : all the movies have been represented on the main page in a coherent and understandable way at the same time.
- Encourage comparison between data : it is possible to compare the number of movies between two genres, for example.
- Reveal the data at several levels of detail : thanks to the filtering option, the number of genres shown can be chosen. Also exist the possibility of seeing the relationship between different films

But as it has been said in the subsection 5.3, this design also has weaknesses.

In addition to Tuft's Principles, this design also satisfies Shneiderman's mantra: "Overview first, zoom and filter, then details-on-demand":

- Overview first : all data cases are shown on the main page
- Zoom and filter : on the main page, it is possible to filter genres and zoom on the bubbles.
- Details-on-demand : more information about a movie can be obtaiend by hovering over it, and all the related movies are shown when it is clicked.

It is, as stated above, not a perfect project due to the technical and time limitations, but it performs the tasks it is supposed to very well. As a final word, we feel that we achieved the goals we set to ourselves in the design document.

8 GitHub repository

The repository is private, but an invite was sent to the user "eaganj".

Link to the repository:
https://github.com/blupiac/blupiac.github.io

Link to see the project directly on a browser:
https://blupiac.github.io/index.html