CinemAviz

Charles Perin

CinemAviz

Charles Perin

Abstract— CinemAviz is a visualization tool for the exploration of the Imdb database. It consists of a variant of a scatterplot matrix—with different cell views—and a prediction panel to estimate the user rating and opening week box office of released movies.

Index Terms—Visual knowledge discovery, matrix visualization, linked views.

1 Introduction

Many visual representations exist for the visualization and exploration of multidimensional data. Scatterplot matrix are a popular visualization technique where thumbnails of scatterplots [2] are arranged in a matrix, each dimension of the data being associated a row and a column. The intersection of each row and column defines the two axes of a cell. In each cell of the matrix, data points are dots mapped on the two dimensions of the cell. Visualization tools such as Tableau [5] employ scatterplot matrix for the exploration of multidimensional data.

Because a movie is a multidimensional data point—examples of dimensions are actors, budget and producers—using a scatterplot matrix to represent movies is straightforward. The exploration of a scatterplot being challenging when the number of represented data points is large, we propose several alternative views for the cells of the matrix. CinemAviz is a visualization tool for the exploration of the imdb database [3], designed to predict the user rating and opening week box office (OBO) of movies. The interface consists of 1) an adjacency matrix with several cell view modes for the exploration of the dataset; and 2) a prediction panel where the analyst sets and weights dimensions according to its expertise to estimate the success of a movie.

2 Data

The database consists of two tables. The first one contains the movies released in a specific time interval (from 1990 to today) because older movies would not be pertinent to compare to new ones. For each movie we store several information, such as the budget (converted in US dollars), the OBO and the user rating. The second table is the crew members involved in at least one of the movies. Each movie knows the list of people involved, and each people knows its list of movies. Overall, the data consists of 2713 movies and 236 982 people.

3 Interface

CinemAviz is a client-oriented web application built with javascript and the d^3 library [1]. Once the client has downloaded the data files, the tool runs offline in a modern browser. CinemAviz is designed to compare similar movies—using what we call dimensions of movies—and because movies have a reasonable number of dimensions only, we use an adjacency matrix as the main component of the interface.

3.1 Dimensions

We first select a movie by typing its Imdb ID (Figure 2(a)). The dimensions of the movie appear and the number of movies for each dimension being displayed (Figure 2(b)). We can manually add dimensions with an entry text (Figure 2(c)). This feature was raised as very interesting by one of the analysts we got feedback from during the challenge. Once the dimensions are set up, we can select/unselect them, making the dimensions appear in the adjacency matrix view (Figure 2(d)). Each cell is the intersection of two dimensions, meaning it represents all the movies of the database with these two dimensions.

• *Charles Perin is with INRIA, Université Paris-Sud and CNRS-LIMSI. E-mail: charles.perin@inria.fr.*

We also propose a Budget dimension, which will be associated all the movies having their budget close to the budget of the analyzed movie.

3.2 Cell Visualizations

Different visualizations for the cells are available (Figure 1). With the linechart, barchart and strippedchart views, both the OBO (in red) and the ratings (in blue) are shown, with a different x scale (value) and y scale (number of movies). The color and scale granularity of the visualizations is set using widgets (Figure 2(e)). One may observe the distribution through different views and find outliers or trends for the next steps of the analysis. The last cell visualization is a scatterplot where each dot represents a movie, making the adjacency matrix become a scatterplot matrix. In this view, the x axis is the rating and the y axis the OBO. We also associate to each dot a grayscale according to the number of dimensions the associated movie has in common with the movie we explore. The darkest dots will be very similar to the target movie while lightest ones will share only a few dimensions with it. For instance, when exploring a movie such as "Wolverine", the series of "X-Men" movies appear in dark gray.

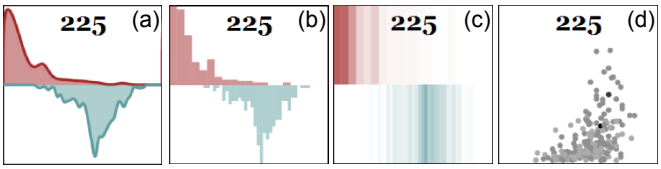

Fig. 1. The same cell visualized using (a) linecharts, (b) barcharts, (c) strippedcharts, and (d) a scatterplot.

3.3 Dimensions Weighting

Once we have explored the different dimensions using the matrix view, we select the ones we estimate to be of interest by clicking on the dimensions' headers. For each selected dimension, a slider is created (Figure 2(f)) to weight the dimension. Sliders can be set between 0 and 1000 and their initial value is 500. The analyst's expertise is crucial for this weighting process which may vary a lot, and so the result. This is a subjective process and a limited knowledge of the dimensions would ensure bad results. Six other sliders, named $[1-6]D$, are used to weight the movies according to the number of dimensions they have in common with the explored movie (Figure 2(g)). Note that the $6D$ slider is actually a $6D+$ slider. These sliders are particularly useful when the analyzed movie is for instance the new opus of a series, and basically as soon as a movie looks very similar to some others.

3.4 Estimation Views

When modifying the sliders' values, the OBO view and the rating view (Figure 2(h,i)) are updated. These views consist of one or several linechart, according to the current mode, which can be the plot of each dimension or the average of the dimensions. Weighting the sliders makes the shape of each associated dimension' linechart as well as the average linechart change. The x axis is from 0 to 10 for the rating and from 0 to the maximum value in the dataset for the OBO. The y scale is the number of movies for each value in the x scale. Because an actor will have fewer movies than a genre for instance, the views

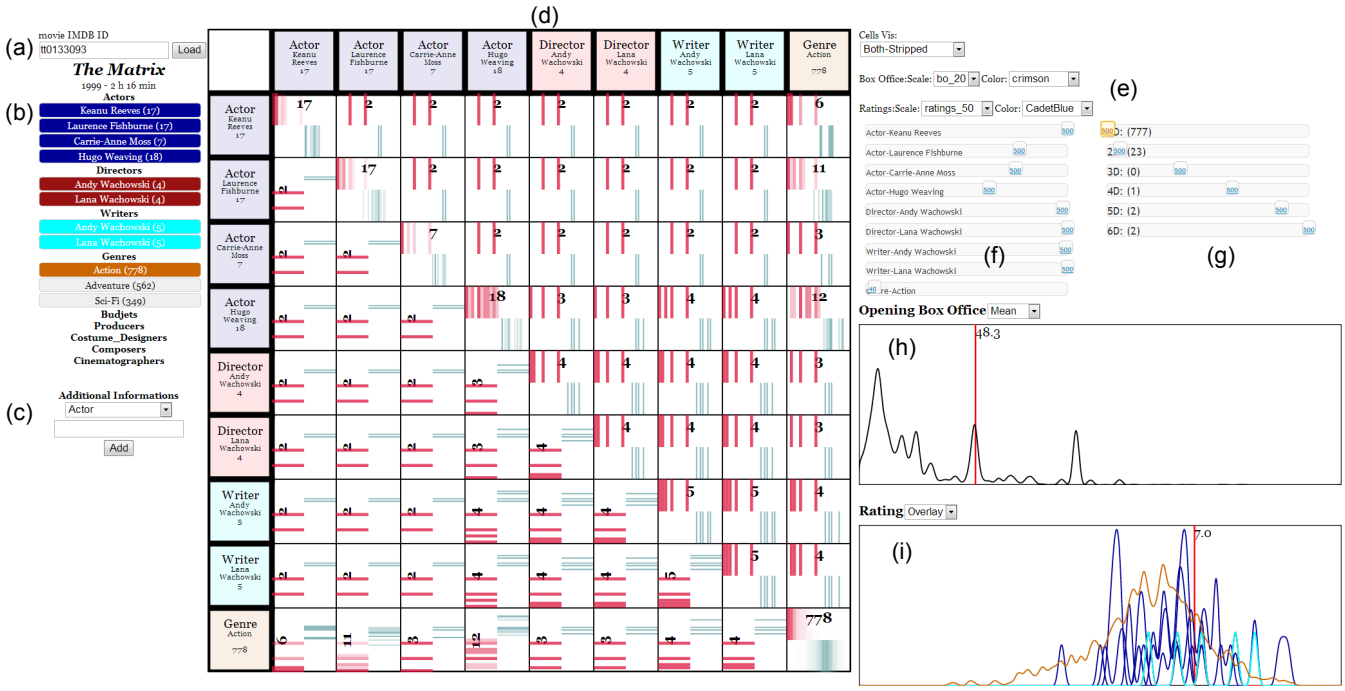

Fig. 2. CinemAviz interface: (a) Imdb unique ID input; (b) dimensions of the explored movie; (c) additional dimensions input; (d) matrix view; (e) matrix view visualization options; (f) sliders to weight each selected dimension; (g) sliders to weight movies according to the number of dimensions they have in common with the explored movie; (h) opening week box office prediction view; (i) user rating prediction view.

will often be stretched by dimensions with many movies. To give less importance to these dimensions, we visually weight the dimensions by observing the feedback in the views. Estimations are performed using the focus views, but going back and forth with the dimensions weighting step. Moving the mouse will trigger the inspector and display the value at the mouse position; the average value of the weighted dimensions is a vertical red line; brushing in the area makes a selection rectangle appear; and the average value within the brushed area is the vertical orange line at its center.

4 Discussion and Conclusion

CinemAviz is based only on visual exploration and visual decision. As the challenge requested the user to have an important role, we did not use advanced mathematical models and focused on a tool where user's expertise is crucial. Then, to obtain good results with the tool, the analyst needs to have a very good knowledge of cinematography.

We found that making accurate predictions was harder when the movie had only a few similar movies in the database and the tool is for example not well suited to predict independent movies and movies with unknown actors. We obtained good predictions overall, according to the different results and recognitions we got during the challenge. In particular, CinemAviz is reliable for the viewer rating estimation, which is highly dependent on the dimensions of the movie; we finally got a viewer rating average absolute error of 0.7 for 28 predictions, with some very precise results. The OBO estimation was less accurate. Although we had for the July 12 results the best OBO prediction made at this date, we were not always that precise and we also had several bad predictions. We partly explain this because the tool is strongly based on the analysts knowledge; and we have to admit that we are not expert in all kind of movies. Our predictions were often very wrong for movies that we are not interested in. We think that CinemAviz gives accurate results as long as the user knows the topic very well. Only him will know which weight a dimension should have, and this may vary a lot depending on his subjective preferences for actors, genres or directors. We also think that the OBO would be easier to predict using social media data, and it is one of the limitations of our tool. The OBO is not influenced only by the dimensions of the movie, but also for example by the other releases of the week and social events occurring at the same time (vacations, sport events, etc.). The most useful factors were star actors and directors because their casting in a movie may appeal—or the opposite—spectators. Another very important dimension was the budget which highly impacts the OBO, while it is far less the case for the rating. This is explained by the fact that a blockbuster movie will be extremely advertised, with star actors, and often with a very impressive trailer with incredible special effects. However, if spectators can be abused and spend money for a movie even if it is not worth it—they did not see the movie yet—they rate the movie afterwards and many high budget movies end up with bad viewer ratings. Less pertinent dimensions were genre, cinematographer, composer, and costume designer. Indeed, either they were involved in only a few movies, or they have huge numbers, and with various OBO and rating, making them unreliable. Finally, a crucial factor was the number of similar dimensions. The scatterplot matrix as well as the $[1-6]D$ sliders were very useful for this purpose.

Several improvements may enhance the tool such as being able to brush each cell or dimension of the matrix to filter movies and remove outliers with high precision [4]. We may also consider other dimensions such as the length of a movie, the period it was released (we know that during summer, box office scores are often higher), and the production company. Finally, we realized that our tool is really helpful to find movies similar to others although it was not its original purpose. We used it a lot for personal research and discovered unknown movies, based on their similarities with our favorite movies, actors or directors.

The video presenting CinemAviz is available at `http://youtu.be/7274CWLhrTQ`

References

[1] M. Bostock, V. Ogievetsky, and J. Heer. D3 data-driven documents. *IEEE TVCG*, 17(12):2301–2309, Dec. 2011.

[2] W. Cleveland and M. McGill. *Dynamic Graphics for Statistics*. statistics/probability series. Wadsworth & Brooks/Cole, 1988.

[3] Imdb. The internet movie database. `http://www.imdb.com/`.

[4] A. R. Martin and M. O. Ward. High dimensional brushing for interactive exploration of multivariate data. In *Proc. VIS'95*, pages 271–, 1995.

[5] Tableau. `http://tableausoftware.com/`.