Similarity-Driven Visual-Interactive Prediction of Movie Ratings and Box Office Results

Feeras Al-Masoudi, Daniel Seebacher, Mario Schreiner, Manuel Stein, Christian Rohrdantz,
Fabian Fischer, Svenja Simon, Tobias Schreck and Daniel Keim

Data Analysis and Visualization Group
University of Konstanz, Germany
{firstname.lastname}@uni-konstanz.de

Abstract

We present an approach developed in course of the VAST 2013 Mini Challenge: Visualize the Box Office. We follow a similarity-driven methodology to predict ratings and box office results based on historic data. An array of interactive visualizations allow analysts to explore structured and unstructured data, activate their domain background knowledge, and come up with predictions as a weighted sum of historically observed figures. We describe the workflow, our developed system, present results obtained during the Challenge execution, and discuss our method in light of extension possibilities.

1 VAST Box Office Mini Challenge

The goal of the 2013 VAST Box Office Mini Challenge was to predict movie ratings and opening weekend box office results for premiering movies. The prediction was to rely on the archived information from the Internet Movie Data Base (IMDB, www.imdb.com) and related messages posted on the Microblog service Twitter (www.twitter.com).

2 Process Model for Similarity-Driven Interactive Prediction

Our prediction concept is based on interactively comparing the target movie with most similar or relevant historic data, which is then used as training data. Specifically, the analyst interactively selects the set of similar movies by comparing the cast, genre, and plot information of the target movie with data extracted from IMDB. Efficient selection is supported by a list of candidate movies and cast suggested by the system based on similarity functions. From these, the user narrows down a set of elements to consider based on visual exploration and background knowledge. To this end, drill-down functionality is provided. The prediction is carried out by computing the weighted average of the performance values (ratings and box office results) of the selected training data set. In that, our prediction follows in principle a *k*-Nearest Neighbor prediction model [2], where the number and type of neighbors is interactively determined. The analyst brings in background knowledge by a) forming the selection and b) setting the weights for the different training examples. Finding weights is also assisted by a sentiment-score visualization of relevant Twitter messages, which can be taken into account. Figure 1 illustrates our process model.

3 Visual-Interactive System

We next describe the components of our prediction system that realizes the aforementioned process model.

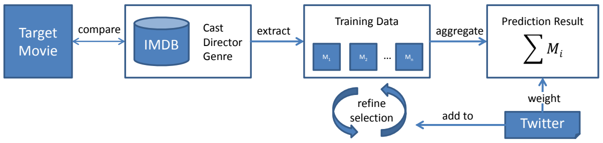
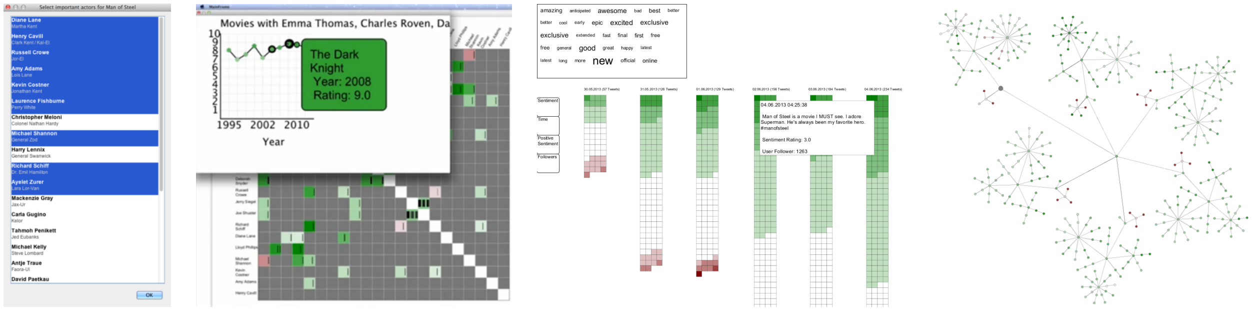


Figure 1: Process model for similarity-driven prediction. The analyst interactively selects training data and prediction weights based on appropriate analytic visualizations.

Data selection. The analyst selects a number of similar movies, actors and directors as the basis for the prediction (see Fig. 2(a)). First, the system computes a number of candidate elements which are then refined by the user. For the selection of candidate actors, a similarity function based on the position of the actor in the credits of the target movie, the average position of the actor in movie credits and the number of information (magazine articles, images and so on) is employed. Detail-on-demand functionality [3] for the actors shows additional information in form of textual descriptions and historic performance time series extracted from IMDB and helps in narrowing down the actor candidate lists. Also, candidate movies are selected in a similar way. For movies, a similarity function based on cast & crew, genre, release year and budget is used. The output is a refined set of movies and actors as a basis for prediction.

Visual exploration. A number of analytical views helps in both selecting and weighting for the prediction. The *actor co-occurrence view* organizes the list of actors from the target movie in a co-occurrence matrix view, based on joint acting in previous movies. The matrix includes the actor scores and a time series of when the co-acting did occur (see Fig. 2(b)). We also consider Twitter messages as an indication as to which actors or related movies are most frequently discussed and in which sentiment sense. To this end, the *Twitter sentiment view* shows in a day-by-day pixel-oriented view the sentiments of relevant Tweets as filtered by keywords. Again, overview-and-detail views, including Word Clouds and Twitter text views, are employed to help the analyst understand the context of public opinion on the relevant facts, or discover new data elements to include in the training set. Fig. 2(c) illustrates the Twitter sentiment view.

Weighting and Prediction The sum of selected and weighted facts are synoptically shown by the *prediction tree view*. This view is defined as a radial tree. The root of the tree corresponds to the movie to be predicted. The four children represent the main dimensions crew (actors and directors), related movies, related genres, and selected tweet scores. The score of the twitter node is normalized so it is comparable to the IMDb scores. For calculating the score, the ratio of positive to negative tweets is used. The four children comprise all selected elements. Each link in the prediction tree has unit weight as default, and can be interactively changed by the



(a) Selector list (b) Actor co-occurrence view (c) Twitter sentiment view (d) Interactive prediction tree

Figure 2: From left to right: The selector panel allows to chose the training data. The selection is supported by an actor co-occurrence view and a Twitter sentiment view, both of which include detail-on-demand functionality. Prediction is found as a weighted sum in the prediction tree.

analyst based on exploration using the aforementioned visualizations. The final prediction values results as a weighted average of the ingoing training data elements. All tree nodes are color-coded by their respective values (ratings, and box office result, respectively; note that for box office prediction, we just consider related movies and genres, but not crew). We also note that instead of the radial tree layout, the prediction tree can also be visualized in form of a TreeMap. Fig. 2(d) illustrates the radial tree layout.

4 Results

We have implemented the above described approach in a fully interactive system (see also our corresponding demonstration video [1]). Over the course of the VAST challenge, we participated in 17 movie prediction tasks. On average, it took about 30 minutes to reach a rating and box office prediction, using our tool and our teams' joint background knowledge in the movie sector. We found that the tools' exploration capabilities sparked many discussions and supported the group decision process. While the prediction process is basically open-ended, we found that after finding about 20 elements we were convinced to have found a good basis for the prediction.

Summarizing our weekly results, we found that on average our prediction was 0.55 stars / 28 millions close to the true figures as realized by the respective movies. We note that we observed variance in the precision of obtained results. For example, in *The Internship* we came close (0.1) to the realized user rating. On the other hand, we found it rather difficult to predict with our tool the class of *non-animated comedies* as we observed on average, 0.8 stars of error in our respective predictions in these movies. The reasons for this behavior could be diverse. We speculate that comedies are difficult to predict based on past movies, because of the fact that slight differences in humor, actor combination, or target audience can change the perceived quality of such movies substantially. Another explanation may be limited background knowledge of our team with this class of movies.

The weekly feedback provided by the challenge reviewers was very positive and supportive. The approach to let the user manipulate the data and giving the user feedback about how the result comes about was well received. Reviewers also found the tool to be easily understandable and promising.

5 Discussion

Our prediction approach is based on a quite simple model: A weighted average of the results of the most similar movies, crew, and genre movies. Our approach starts by automatic suggestion of candidate training data based on similarity computation. Using default weights, fully automatic prediction is possible. However, the full potential of our tool is accessed in interactive application. The set of visual exploration views allows to refine the selection and weighting decisions. In fact, we see the tool as driven mainly by user interaction. We consider that the exploration facilities allow to bring in analyst background knowledge in an efficient way. Twitter data is included in a rather ad-hoc nature, currently. The Twitter sentiment view allows to explore sentiments towards actors, movies etc. and can be used to fine-tune the weights or discover associations of users to other related movies which the analyst may not have thought about.

Our prediction model does not include, e.g., seasonal effects, changes in trends over time, or possible non-linear relationships among the training data. On the other hand, the linear model is highly intuitive and can even be visually comprehended and communicated from the prediction tree. Our preliminary results indicate that even this simple linear prediction model can yield rather good predictions results. We see several interesting research directions. Inclusion of more advanced prediction models such as non-linear regression or Neural Network approaches could improve over the linear model. Also, the current use of Twitter data is rather ad-hoc and we have not found a deterministic workflow how to include the Twitter results. NLP-type text processing, reliably extracting actor and movie names, etc. could help in including Twitter sentiment information in an even more structured way. Conceptionally, it is a challenging question to assess how good interactive prediction can perform as compared to fully automatic prediction. The conduction of an respective study is pending and considered interesting future work.

References

[1] F. Al-Masoudi, M. Schreiner, D. Seebacher, and M. Stein. IMDb prediction tool demonstration video. `http://www.youtube.com/watch?v=M2gLhe6cQQA\&feature=youtu.be`, 2013. [Online at YouTube; accessed 08-August-2013].

[2] J. Han, M. Kamber, and J. Pei. *Data mining: concepts and techniques (third edition)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2011.

[3] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, VL '96, pages 336–, Washington, DC, USA, 1996. IEEE Computer Society.