PREDICTING MOVIE SUCCESS AND ACADEMY AWARDS THROUGH SENTIMENT AND SOCIAL NETWORK ANALYSIS

Krauss, Jonas; Nann, Stefan; Simon, Daniel; Fischbach, Kai; University of Cologne, Pohligstrasse 1, Cologne, Germany, {jkrauss,snann,simond}@smail.uni-koeln.de, kfischbach@wim.uni-koeln.de

Gloor, Peter, MIT, 3 Cambridge Center, Cambridge MA, USA, pgloor@mit.edu

Abstract

This paper introduces a new Web mining approach that combines social network analysis and automatic sentiment analysis. We show how weighting the forum posts of the contributors according to their network position allow us to predict trends and real world events in the movie business. To test our approach we conducted two experiments analyzing online forum discussions on the Internet movie database (IMDb) by examining the correlation of the social network structure with external metrics such as box office revenue and Oscar Awards. We find that discussion patterns on IMDb predict Academy Awards nominations and box office success. Two months before the Oscars were given we were able to correctly predict nine Oscar nominations. We also found that forum contributions correlated with box office success of 20 top grossing movies of 2006.

Keywords: Trend Prediction, Dynamic Social Network Analysis, Online Forum, Internet Movie Database, Oscar Awards

1 INTRODUCTION

It has been widely acknowledged that the "wisdom of crowds" as demonstrated in prediction markets (Surowiecki, 2004, Manski, 2006) is a surprisingly accurate mechanism to predict future trends. Large groups of "ordinary" people are better in predicting trends than a single expert. At the same time, the Web has turned into a major platform for information exchange, thus becoming a mirror of the real world: Millions of volunteers post latest news on Web sites such as Wikipedia, and political blogs such as dailykos and instapundit. In addition people express their opinions in forums and online communities, and tell openly what matters to them. Approaches such as "Netnography" (Kozinets, 2002) make use of this fact for marketing research, proposing analysis of statements of "devotees" and "insiders" in online forums and other Web sites. This paper proposes combining these two ideas, interpreting opinionated discussions and the level of "buzz" about the movie business on the Web as some kind of a prediction market.

Our approach offers an automated, efficient, and cheaper way to tap people's opinion than polling people over the phone. Our method calculates levels of "Web Buzz" by mining discussions in movie-related online forums, combining information about the structure of the social network with an analysis of the contents of the discussion. This paper demonstrates our approach by predicting the success of movies based on the communication in the online community IMDb.com. We analyze its communication patterns in regard to metrics like "intensity" and "positivity". While intensity means the frequency of the subject in discussion, positivity refers to the degree of positive feelings towards a movie expressed by contributors. Thereby we factor in quantitative and qualitative dimensions of discussion allowing us to extract an aggregated community opinion about individual movies.

These measurements are the basis of our two hypotheses. First, we assume that the chances of a movie to win an Oscar can be determined by the communication structure of the IMDb community. Second, we speculate that there is a relationship between the communication intensity about a movie and the performance of the movie at the box office.

The remainder of this paper is organized as follows. Section 2 surveys current research pertaining to movie success and the influence of word-of-mouth in online communities. Section 3 develops a methodology to measure structural properties of online communities and to predict the success of a movie from these properties. Sections 4 and 5 apply our method to the online movie discussion forum IMDb.com.

2 RELATED WORK

There have been different approaches to examine the potential determinants of movie box office success. Most of the studies conclude that movie critics play a significant role for the success or failure of a film (Terry & Butler & De'Armond, 2005). Eliashberg and Shugan (1997) distinguish two possible perspectives on the role of critics: The influencer and the predictor perspective. From the first perspective critics are opinion leaders who influence their audience and, consequently, the box office performance of movies. The predictor perspective suggests that critics might be predictors of performance but not necessarily causing it. Dodds and Holbrook (1988) conducted an analysis where they compared influence and the effect of an Oscar nomination and movie critics on the success of a movie at the box office. Pardoe (2005) focused on models predicting nominees or winners at the Academy Awards.

Awad, Dellarocas and Zhang (2004) analyzed the influence of online movie ratings on box office success and developed statistical models based on these ratings to forecast movie revenues. Furthermore, they examined the relationship of traditional consumer communication, such as infomediary (professional critics), and online word-of-mouth versus offline word-of-mouth. They used the Internet Movie Database (IMDb, http://www.imdb.com) as their main source for the online data and determined the correlation between infomediaries and online word-of-mouth as well as infomediaries and offline word-of-mouth. Eventually, they came to the conclusion that online word-of-mouth has great potential for growth and an increasing number of consumers will use online rating and online review sites as the Internet becomes more pervasive. Surveying current critical issues in the motion picture industry, Eliashberg, Elberse and Leenders (2006) suggest further research relating Internet resources and movie consumption as well as box office sales.

Research regarding trendsetters (Clark & Zboja & Goldsmith, 2007, Valente, 1996) is often associated with the concept of social network analysis (Wasserman & Faust, 1994). One prominent concept is the one of information cascades (Bikhchandani & Hirshleifer & Welch, 1992, Anderson & Holt, 1996, Anderson & Holt, 1997, Bikhchandani & Hirshleifer & Welch, 1998) which explains convergent behaviour patterns and therefore holds potential to identify trends and trendsetters. However, other experiments showed only limited validity of the concept being applied to different laboratory setups (Huck & Oechssler, 2000, Hung & Plott, 2001). Trendsetters have also been of great interest for quite some time in the field of marketing where Myers and Robertson (1972) discuss the importance of "opinion leadership". Connected to opinion leaders is the concept of social contagion which describes the spreading of behavior patterns in a community (Burt, 1987, Crandall, 1988, Rodgers & Rowe, 1993, Kretschmer & Klimis & Choi, 1999). Yet, contagion of opinion does not necessarily result from social influencers, also marketing actions can induce the spread of a certain opinion (Bulte & Lilien, 2001). While IMDb has been frequently used as a basis to predict movie success by other researchers (Eliashberg & Sawhney, 1996, Jensen & Neville, 2002, Pardoe, 2005, Simonoff & Sparrow, 2000, Dellarocas & Awad & Zhang, 2007, Kaplan, 2006), little research has been done so far in using communication behavior and social network structure of an online community as a determinant of movie success at the box office and as a predictor for Oscar nominations.

Although Awad, Dellarocas and Zhang (2004) base their model on movie ratings of an online community, they do not make use of further information which could be retrieved through an analysis of the patterns of communication in that community. Our approach enhances prior research by taking into account social network structures in an online community and by measuring discussion content rather than movie ratings.

3 OUR APPROACH

Predicting real world events based on the communication structure and contents of online word-of-mouth networks is a rapidly emerging field (Patak et. al, 2007, Ganiz & Pottenger & Yang, 2007). This paper contributes to this field by using methods of social network analysis and web data mining to run a model for forecasting movie success.

We chose two Web data sources for our research; namely IMDb, and the "Box Office Mojo" (www.boxofficemojo.com) webpage. In our analysis we focused on the message board community of IMDb. This community exclusively discusses movie and theater related topics and has over 4 million users (Big Boards, 2007) making it the biggest online movie community. With at least 15 million monthly unique U.S. visitors in 2007 (Compete, 2008), IMDb considerably outperforms other online movie communities in terms of its traffic. Additionally, amongst the biggest online movie communities IMDb is the only one having a dedicated subforum for discussion of topics related to the Academy Awards. As mentioned above, IMDb – with the message board community being an integral part – has been subject of extensive research in the past and has also gained wide recognition in public, e.g. by being labelled as one of the "25 Sites We Can't Live Without" in 2007 (Time Magazine, 2007).To measure box office success, movie release dates and movie show times we used Box Office Mojo. In our work this information was compared to the IMDb message board communication structure. A social link between two participants in a forum is constructed if an answer to a message is also an answer to all messages previously posted. Our analysis is based on all posts in a forum from December 2005 to December 2006.

Based on the communication retrieved from the IMDb message boards we applied a model consisting of three relevant components: Discussion intensity, positivity and time. While intensity and time seem to be easy to analyze, the degree of positivity expressed in the discussion requires a more sophisticated approach. Various authors used the degree of positive emotions expressed in communication for deriving insights about the topic of their analysis (Bales, 1950, Bales & Cohen & Williamson, 1979, Gottman & Rose & Mettetal, 1982, Echeverria, 1994, Losada & Fredrickson, 2005). Their results show that discussion positivity can be a key factor for analysis. We will describe our application of intensity, positivity and time in the next section.

The research and application of the model was done with the Condor social network analysis tool, formerly called TeCFlow (Gloor et al., 2003). Condor creates visual maps, movies and many graph metrics of relationships related to social networks by mining web site link structures, online forums and e-mail networks. For example, Condor can create graphical static link views of the communication between users in a web forum and calculates the actor contribution index (Gloor et al., 2003), which delivers clues about the relevance and importance of key actors contributing to the communication. For this paper we make use of Condor's two main features: Firstly, it allows analyzing continuous temporal changes in communication structures in a web forum. Secondly, it supports content analysis of terms being used in forum communication, which also can be displayed graphically in a static or dynamic view.

For further comparison of our results we used the online version of the Linguistic Inquiry and Word Count (LIWC, www.liwc.net) software which offers features to rate textual inputs according to their emotional properties.

4 ACADEMY AWARDS FORECAST BASED ON COMMUNITY COMMUNICATION

The goal of our first experiment was to pick likely candidates for the Oscar Academy Awards, given end of February 2007, based on an analysis of the forums on the IMDb concluded end of December 2006.

As our first hypothesis suggests, we assume that a correlation between the Academy Awards presentation for a particular movie and the communication about that movie in the IMDb forums exists. We speculate that communication intensity and quality of the discussion about a particular movie are indicators of a movie being nominated for an Academy Award. While it would be very hard to also predict in what category a movie would receive an award, we will show that we are able to predict if a movie will be a candidate for an award.

As the basis of our analysis we used the "Oscar Buzz" forum, which is a subforum of IMDb. In this forum topics related to the Academy Awards are being discussed by the IMDb community. This forum has a high frequency of readers and message posters (500 to 1000 posts per day).

To analyze the communication in the Oscar Buzz forum we ran a series of Condor queries, with data from November and December 2006. From the resulting list of terms we extracted the top 25 movies that were discussed in the subforum. We then counted the number of times they were mentioned as well as the time span from their release date to December 15th, 2006.

We based our computation of the chance of a movie being nominated for an Academy Award on three factors. The three factors consist of two temporal frequency indices, the "Intensity Index" and the "Positivity Index" as well as a temporal noise factor, the "Time Noise Factor".

The Intensity Index measures the degree of communication intensity about a specific topic. It is calculated for each movie separately. The Intensity Index is a normalization of the "numbers of mentions" on a scale of 0 to 1. The index is calculated by dividing each value by the highest value of the compared movies (table 1). By identifying this index we followed the approach of Frank & Antweiler (2004) who found a significant correlation between the amount of messages being posted about stocks in finance-related online forums and their volatility. Although this study deals with a different subject, there are similarities in terms of the underlying technology and the communication patterns of online communities. Therefore, we assume that the more a movie is talked about in the community the higher is its chance to receive a nomination for an Oscar. This fact is acknowledged by our model by comprising the numbers of mentions in form of the Intensity Index as a component of the model.

The second index measures the quality of the communication about a certain movie, in particular how positive the communication about this movie was. To calculate the Positivity Index we used the content processing function of Condor for finding out if the discussion about the movie was associated with positive terms. These terms have been determined by ranking potential phrases in regards to their betweenness centrality with Condor's content processing function. The highest ranked terms then became our actual positivity phrases: "win," "nominate," "great," "good," "award," "super," "oscar," and "academy". We selected those terms because they show that the discussion about a particular movie is carried out under positive aspects regarding its Oscar nomination and they are the most important positive phrases in aspects of betweenness centrality. Our method follows the "bag-of-words" concept, which basically means that the order of words in a document can be neglected (Aldous, 1985). This approach makes no direct use of grammatical structure. In previous research it has been found that only a small increase in accuracy is gained by attempting to exploit grammatical structure in the algorithms (Frank & Antweiler, 2004). However, there are cases where this approach might lead to a wrong result: If a negation of a positive term is used in a forum post (e.g. "not a good movie") our method will still give it a positive rating. In the future we plan to further adapt the sentiment analysis algorithm in order to exclude these cases; however in this project our results show that even this simple approach leads to a good prediction quality.

This approach is similar (to a degree) to the one which is used by the developers of the software LIWC who determine the positivity of a text through comparing it with a dictionary (LIWC, 2007). When comparing our positivity index with LIWC using a random sample of IMDb posts, we found significant correlation between LIWC and positivity index ($R=0.56$, $p<0.01$).

Intensity and Positivity Index are not fully independent: the number of positive terms mentioned in context with a movie will increase with the number of messages about this movie. However, it is also possible that a movie will be talked up in a negative context. To prove this we would also need to incorporate a “Negativity Index”. This will be a necessary extension for further research.

Movie	Intensity Index	Positivity Index	Time Noise Factor	Oscar Model
Apocalypto	0,05	0,15	0,02	0,15
Babel	0,46	0,30	0,16	0,30
Blood Diamond	0,24	0,24	0,02	0,24
Bobby	0,19	0,20	0,07	0,20
Borat	0,24	0,24	0,13	0,24
Departed	1,00	1,00	0,22	1,00
Dreamgirls	0,52	0,45	0,00	0,45
Flag of our Fathers	0,29	0,20	0,18	0,20
James Bond: Casino Royale	0,21	0,18	0,09	0,18
Little Children	0,61	0,37	0,22	0,37
Little Miss Sunshine	0,50	0,28	0,45	0,28
Open Season	0,35	0,23	0,24	0,23
Pirates of the Caribbean	0,17	0,14	0,51	0,14
Pursuit of Happiness	0,16	0,13	0,00	0,13
Stranger than Fiction	0,31	0,15	0,11	0,15
Take the Lead	0,39	0,51	0,80	0,51
Thank you for Smoking	0,14	0,15	0,87	0,15
The Break Up	0,20	0,15	0,62	0,15
The Devil wears Prada	0,16	0,13	0,53	0,13
The Nativity Story	0,23	0,15	0,04	0,15
The Prestige	0,11	0,15	0,18	0,15
The Queen	0,59	0,41	0,24	0,41
United 93	0,55	0,24	0,73	0,24
V for Vendetta	0,09	0,14	0,87	0,14
When a Stranger calls	0,25	0,15	1,00	0,15

Table 1. Factor values of top 25 movies.

An interesting insight of our positivity analysis using Condor is that the terms “oscar”, “win” and “nomin” always build a ring structure in the communication about the movie. This means that these three terms are mostly mentioned together.

For the computation of the Positivity Index each Positivity Term was given a relevance value for its influence on the discussion. As mentioned above, three terms are strongly linked and always built a ring. Reflecting the “term frequency inverse document frequency” weight (tfidf), this means that those three terms share a great amount of posts and are therefore of great significance (Salton & Buckley, 1988, Gloor & Zhao, 2006). This is why we chose the highest values for those terms and gave lower values to the remaining terms. “Frequency” consists of the number of times a term was associated with a movie. The Positivity Index in table 1 is computed by the following formula:

$$\textit{Positivity Index} = \sum_{\textit{positivity terms}} \textit{relevance value} * \textit{frequency}$$

The resulting Positivity Indices are then normalized on a scale from 0 to 1, which leads to the values in the column “Positivity Index” as listed in table 1. For calculating the Positivity Index we used the weights of the term-to-term relationships that factor in the betweenness centrality (Wasserman & Faust, 1994) values of the related terms. Thus, the weights do not just correspond to the frequency of terms. Through the implicit application of the graph drawing algorithm of Fruchterman and Reingold (1991), which is implemented in Condor, also the “importance” of the terms is measured. This algorithm is used to construct the social network and calculate centrality values of the participating actors, in this case the corresponding terms of the positivity network.

The last of the three factors we used for determining the Oscar Model is a noise factor that takes into account that some movies are older than others. This models the fact that discussion of a movie calms down over time in the message boards. Nelson, Donihue, Waldman and Wheaton (2001) also find strong evidence regarding the industry practice of delaying movie releases until late in the year as it improves the chances of receiving nominations. Therefore we introduced a “Time Noise Factor” to our model. It is being calculated by normalizing the days from the movie release date till December 15th, 2006 on a scale from 0 to 1 for all of the 25 movies. December 15th, 2006 is the date where the latest of the 25 movies being subject to our investigation was released. The values of the Time Noise Factor can be looked up in table 1.

To determine the Oscar Model, our predictor for the probability of a movie getting an Academy Award nomination, each of the previously calculated indices, Intensity δ, Positivity γ and Time Noise λ is weighed by a factor:

$$\mathit{Oscar\ Model} = a * \delta + b * \gamma + c * \lambda + \varepsilon \mid a + b + c = 1$$

We empirically determined the best values for these factors by running all possible factor combinations (with steps of 0.1) against the known Oscar outcome. The results suggest that setting b to 1 and a and c to 0 leads to an optimal solution. Figure 1 shows the plotted curves for the different factor combinations. When applying the Oscar Model to a real world event we included an error term ε. By looking up the actual Oscar winners and nominees for the movies of all factor combinations we minimized ε, what can be expressed by the number of movies that neither received an Oscar nor a nomination. In the optimal setting five out of the top ten movies ranked by the Oscar Model received an Oscar and four received a nomination for an Oscar (table 1).

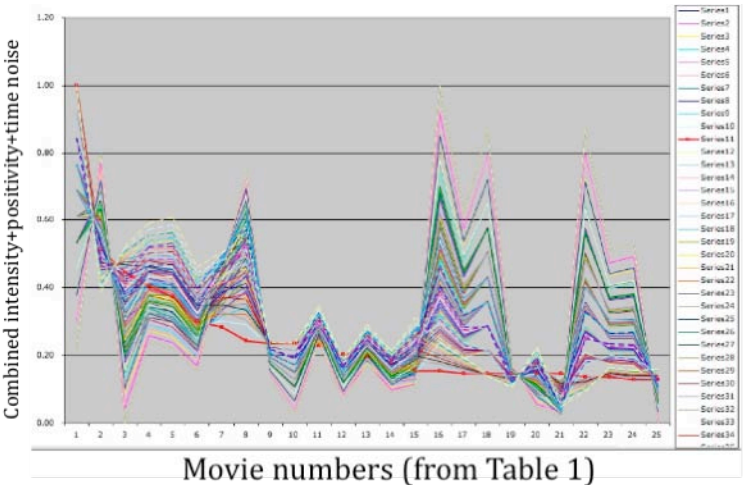
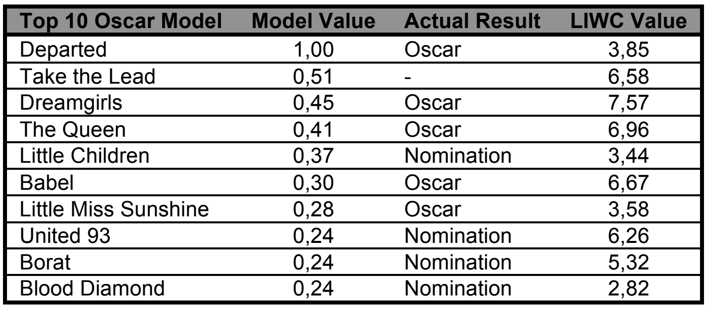


Figure 1. Oscar Model Sensitivity Analysis.

Weighing b with 1 delivered the best result with 9 out of the top 10 movies ranked by the Oscar Model being actual award winners or nominees respectively (red line in figure 1, series 11; for Award winners and Oscar Model values refer to table 2). Interestingly, the best factor combination is therefore the one ignoring intensity and time noise. This comes from the types of users participating in the discussion on IMDb, whom we suspect to be movie buffs and therefore more in line with the opinion of the Academy Awards jury than others.

This shows that movies that are being discussed in a positive way in the sub forum “Oscar Buzz” have a high probability of getting a nomination for the Academy Awards. It further indicates that the users who are participating in the communication in “Oscar Buzz” are movie enthusiasts who value similar criteria in a movie as the Oscar poll does. As shown by an Oscar Index twice as high as the next movie, there is a clear favorite for the Oscar nomination in the IMDb community, namely “The Departed”. The community opinion (reflected by the values of the Oscar Model) is not limited to only a few movies but rather a broad range of movies is being discussed intensively (table 1, Intensity Index). As stated earlier there is indeed a correlation of 0,88 between the intensity of discussion and the Positivity Index , yet it is the positivity index which is the best predictor of winning an Oscar. Moreover, due to the Time Noise Factor being included in our computation, an aggregation of the indices in a multiplicative model does not appear to be applicable. In the worst case such a multiplicative model could lead to highly positive discussed movies receiving an Oscar Model value of 0 if being released on December 15th, 2006. Thereby the Time Noise Factor would be significantly overvalued.



Top 10 Oscar Model	Model Value	Actual Result	LIWC Value
Departed	1,00	Oscar	3,85
Take the Lead	0,51	-	6,58
Dreamgirls	0,45	Oscar	7,57
The Queen	0,41	Oscar	6,96
Little Children	0,37	Nomination	3,44
Babel	0,30	Oscar	6,67
Little Miss Sunshine	0,28	Oscar	3,58
United 93	0,24	Nomination	6,26
Borat	0,24	Nomination	5,32
Blood Diamond	0,24	Nomination	2,82

Table 2. Values of the Oscar Model Vs. Academy Award results.

In order to compare our results with other available methodologies for analyzing the positivity of communication, we repeated the same analysis with the above mentioned LIWC software. However, we found no correlation (R=0.065, non-significant) between the results computed by LIWC and the values of the Oscar Model. A possible explanation might be that LIWC uses a general dictionary as opposed to our customized method of calculating the Positivity Index. Table 2 lists the values of LIWC.

It should be pointed out that there are different categories of Oscar Awards. There are six major ones that people primarily focus on: best picture, best director, and the four acting awards (best actor/actress, best supporting actor/actress). Hard core film buffs may also talk about second-tier awards like best screenplay or editing or music, and the other awards in the more technical arts (Art Direction, Sound, etc.), but these are not typically the subjects of most of the buzz. What we found is that the importance of the awards movies got corresponds to the level of buzz. “Babel” with a lower value for the Oscar Model won an award for best score, which is a minor Oscar. By contrast “Departed” with the highest value won two major awards (Best Picture and Best Director) and also two important second tier awards (Best Editing and Best Adapted screenplay). “Little Miss Sunshine,” which won for best actor and best original screenplay, the first a slightly more prominent award than the second, but still not in the same rank as best picture and best director, also has a value for the Oscar Model slightly higher than “Babel,” but much lower than “The Departed.”

The results of this first application of our approach encouraged us to apply the same model to the prediction of a movie’s box office success. We will describe the procedure of adjusting the model for this application in the next section.

5 CORRELATION BETWEEN MOVIE SUCCESS AND COMMUNITY COMMUNICATION

Based on our findings that intensive and positive online forum communications are predictors for Oscar success, we applied the same insights to predict commercial success of not yet launched movies. To study movie success at the box office we chose the IMDb sub forum “Previews & Reviews”. As our metrics of financial success we analyzed the US movie box office rankings of 2006, which we obtained from Box Office Mojo. The major success criterion of a movie we used in this analysis is its gross sales at the box office in 2006. We concentrated on twenty films, which prevailed in the community discussion in the “Previews & Reviews” IMDb forum and also showed top ranks in the 2006 gross sales list.

Movie	Intensity Index	Positivity Index	Trendsetter Index	Values of the Buzz Model	Box office success in $
Pirates of the Caribbean: Dead Man's Chest	1,00	1,00	1,00	1,00	423.315.812
Cars	0,62	0,67	0,88	0,68	244.082.982
Superman Returns	0,76	0,67	1,00	0,78	200.081.192
Ice Age: The Meltdown	0,29	0,67	0,75	0,50	195.330.621
Casino Royale	0,49	1,00	1,00	0,75	167.220.102
Over the Hedge	0,51	0,33	1,00	0,55	155.019.340
The Departed	0,95	0,67	1,00	0,87	132.208.177
Borat	0,25	0,17	0,88	0,35	128.488.700
Dreamgirls	0,18	0,33	0,88	0,37	102.266.997
Inside Man	0,56	0,17	0,88	0,51	88.513.495
Monster House	0,33	0,33	0,75	0,41	73.661.010
Underworld: Evolution	0,27	0,50	0,50	0,39	62.318.875
Little Miss Sunshine	0,62	0,83	0,88	0,73	59.863.257
Blood Diamond	0,22	0,33	0,88	0,38	56.576.961
The Queen	0,27	0,00	0,88	0,31	54.581.202
The Prestige	0,29	0,33	0,88	0,42	53.089.891
Apocalypto	0,27	0,50	1,00	0,49	50.866.635
Stranger than Fiction	0,25	0,50	0,88	0,45	40.435.190
Snakes on a Plane	0,29	0,00	0,63	0,27	34.020.814
Friends with Money	0,29	0,17	0,25	0,25	13.368.437

Table 3. Indices and box office gross sales of top 2006 movies.

Our goal was to develop an appropriate metric to measure the communication behavior of the community regarding movies. Therefore, using Condor's content processing capabilities and following our general approach of analyzing communication in regards to intensity and expressed positivity, we created three individual indices that capture the communication patterns of the users in the subforum. We again used Intensity Index and a Positivity Index. A new metric was introduced with the Trendsetter Index, all three indices were combined into a "Buzz Model".

To calculate the Trendsetter Index we first identified users with the highest betweenness centrality values in the sub forum "Previews & Reviews". With a minimum value of 0.03, the betweenness centrality of these 10 identified users was at least 12 times higher than the average betweenness centrality of 0.0025. In a second step we counted for each movie how many trendsetting users were mentioning the movie favorably. The index was then calculated by normalizing the number of participating trendsetters on a scale of 0 to 1, which is in line with the calculation of the Intensity Index. This metric implicitly emphasizes the social aspects of the communication. It weighs the impact of the most between users in the conversation and is an indicator of the importance and influence of trendsetters on the communication about a certain movie in the forum. We speculate that discussion of these trendsetters will likely have a direct impact on the success of a movie at the box office. Table 3 displays all three indices and the values calculated with Condor.

We used a similar formula as for the Oscar Model to determine a combined "Buzz Model" with Intensity δ, Positivity γ, Trendsetter λ, and Error Term ε:

$$Buzz\ Model = a*\delta + b*\gamma + c*\lambda + \varepsilon \mid a+b+c=1$$

To determine the optimal values for a, b and c we ran all possible factor combinations (in steps of 0.1) against 20 top grossing movies in 2006. At values a = 0,5, b = 0,3 and c = 0,2 correlation is 0.75 ($p<0.01$), showing a very strong relationship between the communication intensity/behavior and the box office success of movies. Despite a positive correlation with the Intensity Index (R=0.44) and the Positivity Index (R=0.42), the Trendsetter Index does not become superfluous and obviously contributes to the optimal solution.

While analyzing IMDb.com it was obvious that certain movies were significantly more discussed than others. The question was if there would be a relationship with the financial success of the movie or if the discussion at imdb.com would be independent from the "real" world.

In our analysis we found robust support for our hypothesis that higher movie success correlates with higher communication intensity. A positive discussion about a movie in the forum correlates with higher revenue of the movie at the box office. This means that high positive discussion by trendsetters predicts success of a newly released movie at the box office.

Furthermore, we have seen that the most influential (high-betweenness) users lead the discussion, which indicates that this discussion may have an impact on the result of the movie at the box office. More in-depth analysis shows, however, that this opposite conclusion can not be proven. Our analysis does not tell if "talking up" a movie will guarantee financial success. The IMDb.com community consists of movie experts who are not showing the same attitude towards a movie as the average moviegoer. The value of the Buzz Model of the movie "Snakes on a Plane" illustrates this point. This movie was "hyped up" long before its release throughout the web, yet in the discussion on IMDb.com it received comparably bad press, which shows that IMDb.com users are clearly more differentiated in their perception of the movie than the mainstream user was, and more resistant to attempts of manipulation by movie publishers.

Note that this paper is not focusing on the general discussion of the effects of "Buzz" in a community. This might be subject of a more in-depth analysis of online communities and part of a continuation of this article.

6 LIMITATIONS

Our research is subject to limitations, which, though they do not affect the positive results in this paper, need to be tackled through further adaption of the model. One aspect, which has been mentioned already, is the "bag-of-words" concept. This needs to be resolved through the application of a context sensitive method, which takes into account the actual relation between phrases in the analyzed communication. The quality of the sentiment analysis could be further increased through broader sensitivity analysis of the potential phrases.

The results of last year's Academy Award could be predicted relatively well (though the award category was not predicted). However, results should be scrutinized by applying the same model to future Oscar elections. With respect to the Buzz Model, results could be re-evaluated by applying a multiplicative model. Another point left to discuss is the causality chain: Is movie success determined by forum discussion or does forum communication follow movie success? Our approach only calculates the correlation between these two, yet the underlying reason for the correlation remains unclear.

7 SUMMARY AND CONCLUSION

This paper represents an extension of the research on the influence of online communities on the success of movies. It is addressing two main issues: First it introduces a model to predict Academy Award nominees based on the communication of an online community. It then applies the same approach to examining if there is a correlation between community communication and movie success at the box office. Doing so, we were able to make predictions about real world events based on social networks in an online movie community.

In our first experiment we showed that there is a correlation between the IMDb community discussion and the chance of a movie getting nominated for an Academy Award. Some insights could be gained about the structure and properties of the community in the Oscar Buzz sub forum of IMDb. Oscar influencers are movie buffs who do not necessarily have the same opinions as mainstream movie viewers. With "The Departed" a clear favorite of the forum for getting a nomination for an Oscar was identified 8 weeks before the Oscars were awarded.

In our second experiment we found that a high intensity of discussion about a particular movie at IMDb is a strong indicator of success of that movie at the box office. While not every movie being successful at the box office is actively discussed in the community, every movie, which generates high positive buzz on IMDb appears high in the box office charts. This means that high discussion volume predicts success at the box office, but generating lots of buzz will not help a movie to increase viewing in theaters. For Oscars, just gauging the level of positivity of posts is enough to predict future success. Using a customized dictionary yields better results than a generic positivity measurement tool such as LIWC. For predicting financial success, on the other hand, a more complex model assigning higher weight to trendsetters, i.e. people with central network positions, delivers the best results.

The insights we gathered and the methods we apply could be of value also in the field of marketing science, especially in the field of viral marketing. For example, motion picture studios could optimize their marketing strategy through identifying trendsetters in forums and the internet and then address those with their marketing campaigns. Forum communication analyzed by our methodology could be used as an indicator for early success prediction of an upcoming movie release. These few examples show the practical relevance of our analysis, ideas of connected research are suggested below.

Our experiments can be extended in different ways in future research. An obvious extension would be to increase the sample size by widening the data analysis over longer periods of time and by including other forums. It would be of great interest whether including other forums would entail an even higher correlation or whether those forums would perform worse in terms of predictive qualities, thereby strengthening our perception of IMDb being an expert community. Secondly, it would be interesting to examine whether similar insights could be obtained for other movie genres as well. For example, one could focus on the discussion about TV shows and compare the communication structure in the forums with audience ratings. These approaches could be easily used as an indicator in the movie business to predict which movies, TV shows, etc. would be successful in the future. Thus, IMDb message boards and similar forums could be used as a market research platform for all kinds of movie-related predictions. It might be interesting to apply our approach of quantifying unstructured communication to motion picture business external fields using blogs or newsgroups and trying to make predictions about other real world events based on communication taking place in these groups.

For example, it would be of great interest to apply our Oscar prediction model to other award nominations to test the model with other data sets. We are also currently applying the same model to online investor forums to predict financial performance of selected stocks. Although our approach worked well predicting this year's Academy Awards and movie box office success, it will need much further work to get a more robust proof of its predictive qualities.

8 ACKNOWLEDGEMENTS

We are grateful to Rob Laubacher for helping us understand the dynamics in the movie business and educating us on the differences between „major“ and „minor“ Oscars. We also thank Ken Riopelle for pointing us to the LIWC sentiment analysis software.

9 REFERENCES

Aldous, D. J. (1985). Exchangeability and related topics. In École d'Été de Probabilités de Saint-Flour XIII — 1983, 1–198, Springer, Berlin.

Anderson, L. R. and Holt, C. A. (1996). Classroom Games: Information Cascades. Journal of Economic Perspectives, 10, 187-193.

Anderson, L. R. and Holt, C. A. (1997). Information Cascades in the Laboratory. The American Economic Review, 87 (5), 847-862.

Awad, N. F. and Dellarocas C. and Zhang X.(2004). Is Online Word-of-mouth a Complement or Substitute to Traditional Means of Consumer Conversion. Sixteenth Annual Workshop on Information Systems Economics (WISE), Washington, DC.

Bales, R. F. (1950). Interaction Process Analysis: A Method for the Study of Small Groups. Addison-Wesley.

Bales, R. F. and Cohen, S. P. and Williamson, S. A. (1979). SYMLOG: A System for the Multiple Level Observation of Groups. Free Press.

Big Boards (2007). IMDb statistics. http://www.big-boards.com/board/926, retrieved 2007.

Bikhchandani, S. and Hirshleifer, D. and Welch, I. (1992). A Theory of Fads, Fashion, Custom and Cultural Change as Informational Cascades. Journal of Political Economy, 100, 992-1026.

Bikhchandani, S. and Hirshleifer, D. and Welch, I. (1998). Learning from the Behavior of Others: Conformity, Fads, and Informational Cascades. The Journal of Economic Perspectives, 12 (3), 151-170.

Bulte, C. and Lilien, G. (2001). Medical Innovation Revisited: Social Contagion versus Marketing Effort. The American Journal of Sociology, 106 (5), 1409-1435.

Burt, R. S. (1987). Social Contagion and Innovation: Cohesion Versus Structural Equivalence. The American Journal of Sociology, 92 (6), 1287-1335.

Clark, R. A. and Zboja, J. J. and Goldsmith, R. E. (2007). Status consumption and role-relaxed consumption: A tale of two retail consumers. Journal of Retailing and Consumer Services, 14 (1), 45-59.

Compete (2008), SnapShot of imdb.com. http://siteanalytics.compete.com/imdb.com/?metric=uv, retrieved 2008.

Crandall, C. S. (1988). Social Contagion of Binge Eating. Journal of Personality and Social Psychology. 55 (4), 588-598.

Dellarocas, C. and Awad, N. F. and Zhang, X. (2007). Exploring the Value of Online Product Ratings in Revenue Forecasting: The Case of Motion Pictures. Working Paper, Robert H. Smith School Research Paper.

Dellarocas, C. and Narayan, R. A. (2005). Statistical Measure of a Population's Propensity to Engage in Post-purchase Online Word-of-mouth. R. H. Smith School of Business, University of Maryland, College Park, MD 20742, Working Paper.

Dodds, J. C. and Holbrook, M. B. (1988). What's an Oscar worth? An Empirical Estimation of the Effect of Nominations and Awards on Movie Distribution and Revenues. Current Research in Film: Audiences, Economics and the Law, 4.

Echeverria, R. (1994). La Ontologia de1 Lenguaje. Dolmen Ediciones, Santiago de Chile.

Eliashberg, J. and Elberse, A. and Leenders, M. (2006). The Motion Picture Industry. Marketing Science, 25 (6), 638-661.

Eliashberg, J. and Sawhney, M. S. (1996). A Parsimonious Model for Forecasting Gross Box-Office Revenues of Motion Pictures. Marketing Science, 15 (2), 113-131.

Eliashberg, J. and Shugan, S. M. (1997). Film critics: Influencers or Predictors?. Journal of Marketing, 61 (2), 68-78.

Frank, Murray Z. and Antweiler, Werner (2004). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. The Journal of Finance, 59 (3), 1259-1294.

Fruchterman, T. M. J. and Reingold, E. (1991). Graph Drawing by Force-Directed Placement. Software-Practice and Experience, 21 (11),1129-1164.

Ganiz. M. and Pottenger, W. M. and Yang, X. (2007). Link Analysis of Higher-Order Paths in Supervised Learning Datasets. In Proc. 5th Workshop on Link Analysis, Counterterrorism and Security, SIAM International Data Mining Conference.

Gloor, P. A. and Laubacher, R. and Dynes, S. B. C. and Zhao, Y. (2003). Visualization of Communication Patterns in Collaborative Innovation Networks: Analysis of some W3C working groups. In Proceedings of the Twelfth International Conference on Information and Knowledge Management.

Gloor, P.A. and Zhao, Y. (2006). Analyzing Actors and Their Discussion Topics by Semantic Social Network Analysis. In Proceedings of 10th IEEE International Conference on Information Visualisation IV06.

Gottman, J. M and Rose, F. and Mettetal, G. (1982). Time-series analysis of social interaction data. Emotion and Early Interaction, 261-289.

Huck, S. and Oechssler, J. (2000). Informational cascades in the laboratory: Do they occur for the right reasons?. Journal of Economic Psychology, 21, 661-671.

Hung, A. A. and Plott, C. R. (2001). Information Cascades: Replication and an Extension to Majority Rule and Conformity-Rewarding Institutions. The American Economic Review, 91 (5), 1508-1520.

Jensen D. and Neville J. (2002). Data mining in social networks. Invited presentation to the National Academy of Sciences Workshop on Dynamic Social Network Modeling and Analysis, p. 7-9, Washington, DC.

Kaplan, D. (2006). And the Oscar Goes to… A Logistic Regression Model for Predicting Academy Award Results. Journal of Applied Economics & Policy, 25 (1), 23-41.

Kozinets, R. (2002). The Field Behind the Screen: Using Netnography for Marketing Research in Online Communities. Journal of Marketing Research, 39, 61-72.

Kretschmer, M. and Klimis, G. M. and Choi, C. J. (1999). Increasing Returns and Social Contagion in Cultural Industries. British Journal of Management, 10 (1), 61–72.

Linguistic Inquiry and Word Count (2007). The LIWC2001 Application. http://www.liwc.net/liwcdescription.php, retrieved 2007.

Losada, M. and Heaphy, E. (2004). The role of positivity and connectivity in the performance of business teams: A nonlinear dynamics model. American Behavioral Scientist, 47 (6), 740–765.

Manski, C. F. (2006). Interpreting the Predictions of Prediction Markets. Economic Letters, 91, 425-429.

Myers, J. H. and Robertson, T. S. (1972). Dimensions of Opinion Leadership. Journal of Marketing Research, 9, 41-46.

Nelson R. A. and Donihue M. R. and Waldman D. M. and Wheaton C. (2001). What's an Oscar Worth?. Economic Inquiry, 39 (1).

Pardoe, I. (2005). Predicting Academy Award winners using discrete choice modeling. In Proceedings of the 2005 Joint Statistical Meetings, Alexandria, VA. American Statistical Association.

Patak, N. and Mane, S. and Srivastava , J. and Contractor, N. (2007). Knowledge Perception Analysis in a Social Network. In Proc. 5th Workshop on Link Analysis, Counterterrorism and Security, SIAM International Data Mining Conference.

Rodgers, J. L. and Rowe, D. C. (1993). Social contagion and adolescent sexual behavior: A developmental EMOSA model. Psychological Review, 100 (3), 479-510.

Salton, G. and Buckley, C. Term-Weighting Approaches in Automatic Text Retrieval (1988). Information Processing and Management, 24 (5), 513-523.

Simonoff, J. S. and Sparrow, I. R. (2000). Predicting movie grosses: Winners and losers, blockbusters and sleepers. Chance 13 (3), 15-24.

Surowiecki, J. (2004). The Wisdom of Crowds. Doubleday, New York.

Terry, N. and Butler M. and De'Armond D. (2005). The Determinants of Domestic Box Office Performance in the Motion Picture Industry. Southwestern Economic Review, 32 (1), 137-148.

Time Magazine (2007), 25 Sites We Can't Live Without. http://www.time.com/time/specials/2007/article/0,28804,1638266_1638253_1638236,00.html, retrieved 2008.

Valente, T. W. (1996). Social network thresholds in the diffusion of innovations. Social Networks, 18 (1), 69-89.

Wasserman, S. and Faust, K. (1994). Social Network Analysis, Methods and Applications. Cambridge University Press.

Wolfers, J. and Zitzewitz, E. (2004). Prediction Markets. Journal of Economic Perspectives, 18 (2)