2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology

Improving Movie Gross Prediction Through News Analysis

Wenbin Zhang and Steven Skiena
Department of Computer Science
Stony Brook University
Stony Brook, NY 11794-4400 USA
Email: {wbzhang, skiena}@cs.sunysb.edu

***Abstract*—Traditional movie gross predictions are based on numerical and categorical movie data from The Internet Movie Database (IMDB). In this paper, we use the quantitative news data generated by *Lydia*, our system for large-scale news analysis, to help people to predict movie grosses. By analyzing two different models (regression and k-nearest neighbor models), we find models using only news data can achieve similar performance to those using IMDB data. Moreover, we can achieve better performance by using the combination of IMDB data and news data. Further, the improvement is statistically significant.**

I. Introduction

The movie industry is of intense interest to both economists and the public because of its high profits and entertainment nature. An interesting question is to forecast pre-release movie grosses, because investors in the movie market want to make wise decisions. Traditionally, people predict gross based on historical IMDB data analysis regarding specific characteristics, e.g., the movie's genre, MPAA rating, budget, director, number of first-week theaters, etc., but with somewhat limited success. Nevertheless, recent publications ([1], [2], et al.), have shown the media's power on forecasting financial market like stock prices, volatilities, or earnings. Considering the encouraging results, it is reasonable to infer that news has predictive power for movie grosses as well. We are unaware of any previous attempt to apply linguistic analysis to movie gross prediction. Therefore, here we focus on improving movie gross prediction through news analysis.

Our primary goal is to prove that we can give better pre-release prediction of movie grosses if we use news data, because commercially successful movies, actors, or directors are always accompanied by media exposure. Our experiments use *Lydia* ([3], *http://www.textmap.com*), a high-speed text processing system, to analyze news publicity and output movie news data, and then to help our movie gross prediction. In this paper, we do not use any post-release data in the following experiments, and all the predictions are out-of-sample predictions. In practice, our approach provides a feasible and more accurate estimation regarding the investment worthiness for some pre-release investors and almost all the post-release investors.

The contents of this paper are organized as follows. First, we will review related work briefly. Second, we will describe the movie data sources, both traditional movie data and news data, and give a correlation analysis. We then set up different models with traditional movie data, movie news data, and their combination respectively as well as evaluate their performance. Finally, we conclude that we can improve traditional movie gross prediction through news analysis.

II. Related Work

Different people work on movie gross prediction from different perspectives. Most previous work ([4], [5], [6], et al.) forecast movie grosses based on IMDB data with regression or stochastic models. However, their models either work poorly or need post-release data in order to make reasonable prediction, which are not acceptable in practice. For example, Sawhney and Eliashberg [6] claimed that their model works pretty well by taking the first three weeks of gross data as input, but admitted that it is much more difficult to give shape estimation for either model parameters or gross if they don't have any early stage movie gross data. Although the post-release models are also useful in some situations, pre-release models are of more practical importance.

Moreover, there has been substantial interest in the NLP community on using movie reviews as a domain to test sentiment analysis methods, e.g., [7], et al. Basically speaking, they apply information retrieval or machine learning techniques to classify movie reviews into some categories and hope to produce better classification accuracy than human being. The classification categories are like "thumbs up" vs. "thumbs down", "positive" vs. "negative", or "like" vs. "dislike". Pang and Lee [8] gives a detail review in this domain. However, to the best of our knowledge, news and sentiment analysis has not been previously studied as a predictor of movie grosses. In addition, Mishne and Glance [9] show that movie sales have some correlation with movie sentiment references, but they neither build prediction models or show the value of the correlation because they think the result is not good enough for accurate modeling.

III. Movie Data and Correlation Analysis

There are two kinds of movie data used in this paper, movie specific variables and movie news data. The movie specific variables are collected from traditional movie websites like IMDB, but the movie news data is obtained from *Lydia*. We need to analyze the correlation between movie grosses and traditional movie variables or news variables, and then let it guide us to set up reasonable models for movie gross prediction. The correlation between variables is measured by both strength and significance. The strength is further evaluated by the correlation coefficient r, while the significance can be verified by t-test $t = r\sqrt{\frac{N-2}{1-r^2}}$, in which r is the correlation coefficient, and N is the sample size. In this paper, we use significance level 0.05 to test the statistical significance because 0.05 is conventionally a standard threshold to evaluate significance although in our experiments most of the results are statistically significant to an even stricter level.

A. Traditional Movie Data and Correlation Analysis

Traditional movie data is available at *http://www.imdb.com* and *http://www.the-numbers.com*. We wrote a spider program and downloaded data for all movies released from 1960 to 2008. Table I summarizes the relationship between some important movie variables and grosses by providing the corresponding correlation coefficients. The most important movie variables include numerical variables like budget or opening screens, and categorical variables like source or MPAA rating. Another important variable is genre. IMDB defines 19 genres, and in our experiments, we find some genres like "*Action*" and "*Adventure*" are positively correlated with grosses, while others genres like "*Biography*" and "*Documentary*" are negatively correlated with grosses.

978-0-7695-3801-3/09 $26.00 © 2009 IEEE
DOI 10.1109/WI-IAT.2009.53

Movie Variables	Categories	Movies	Corr	p-Value
Budget	All	1500	0.672	**<0.001**
Opening Screens	All	1500	0.647	**<0.001**
Release Date	Holiday	640	0.132	**<0.001**
	Non-holiday	860	-0.132	**<0.001**
MPAA Rating	G	41	0.103	0.261
	PG	201	0.128	**0.035**
	PG-13	500	0.154	**<0.001**
	R	646	-0.221	**<0.001**
	NC-17	17	-0.049	0.426
Source	Sequel	127	0.224	**0.006**
	Not Sequel	1373	-0.224	**<0.001**
Origin Country	USA	1191	0.141	**<0.001**
	Not USA	309	-0.141	**0.006**

Table I: Correlation Coefficient of Movie Variables versus Movie Grosses. The bold numbers show the corresponding correlations are statistically significant at a 0.05 significance level.

Entities	Duration	G (Pre)	G (Post)	B (Pre)	B (Post)
Movie	1 week	0.707	0.781	0.497	0.480
	1 month	0.672	0.779	0.463	0.474
	4 months	0.629	0.749	0.437	0.455
Director	1 week	0.494	0.602	0.311	0.389
	1 month	0.371	0.495	0.218	0.389
	4 months	0.192	0.317	0.117	0.078
Top 3 Actors	1 week	0.640	0.726	0.476	0.528
	1 month	0.569	0.683	0.448	0.477
	4 months	0.493	0.618	0.413	0.424
Top 15 Actors	1 week	0.646	0.725	0.533	0.595
	1 month	0.575	0.686	0.477	0.530
	4 months	0.511	0.618	0.415	0.433

Table II: Correlation Coefficient of Logged Pre-release News Article Counts versus Logged Grosses under various scenarios. The rows indicate what kind of entities are examined in terms of what kind of duration, i.e., 1 week, 1 month, or 4 months. The columns indicate the correlation is for gross (G) or budget (B) in terms of pre-release(or post-release) article counts.

A movie's gross may vary significantly, from very high to very low values. In this paper, we will pay more attention on high-grossing movies because they can generate more revenue and have more media exposure.

B. News Data and Correlation Analysis

Movie news data is generated form the *Lydia* system, which does high-speed analysis of online daily newspapers. The input of *Lydia* includes the coverage of around 1000 nationwide and local newspapers. One difficulty for movie news analysis is title matching, which causes lots of false positives or false negatives during entity identification phase. For example, *Lydia* may fail to identify certain movies' name like "15 Minutes", "Pride", "Next", "Interview", etc. correctly. Our solution is to filter out these "bad" data before our actual analysis. Eventually, we get a data set size of 498 movies, and we divided these movies into two parts - 60% as the training set and the rest 40% as the predicting set.

Lydia generates an entity database. For each entity, the *Lydia* data includes daily article counts, daily frequency counts, as well as daily sentiment (both positive and negative) counts in seven categories: *General*, *Business*, *Crime*, *Health*, *Politics*, *Sports*, and *Media*.

Based on above raw counts, we evaluated the accumulated news references for the first week (*1-week* data), the second week through the 4th week (*1-month* data), and the 5th week through the 16th week (*4-month* data) period before the release of movies respectively. Our correlation analysis includes the evaluation of the media coverage in terms of four different entities - movie titles, directors, top 3 actors, and top 15 actors. Table II shows the correlation analysis of logged pre-release news reference counts versus logged grosses or budget under different scenarios. Table III shows the correlations between movie grosses and sentiment counts in seven categories.

Scenarios		Gen	Busi	Crim	Heal	Poli	Spts	Media
1week	Pos	0.692	0.666	**0.418**	**0.520**	0.615	0.684	0.695
	Neg	0.665	0.564	**0.594**	**0.624**	0.565	0.444	0.513
1mth	Pos	0.665	0.651	**0.401**	**0.520**	0.603	0.669	0.675
	Neg	0.650	0.579	**0.580**	**0.616**	0.564	0.466	0.507
4mths	Pose	0.625	0.626	**0.370**	**0.497**	0.561	0.635	0.643
	Neg	0.608	0.544	**0.541**	**0.557**	0.531	0.438	0.490

Table III: Logged Movie Grosses versus Logged Pre-release Positive (Pos) or Negative (Neg) Sentiment Counts in Seven Sentiment Categories, in terms of movie title coverage. The bold numbers show that positive references are better correlated with grosses than negative ones except for "*Crime*" and "*Health*" categories. One reason is that a movie may be more attractive due to excess violence.

1) *Movie Grosses versus News Reference Counts*: Some significant observations from our experiments are as below.
   - Article counts vs. Frequencies: Grosses have higher correlation with article counts than with total entity references.
   - Grosses vs. Budget: News references are more highly correlated with grosses than budgets.
   - Pre-release and Post-release References: Table II shows that the post-release data correlates with grosses better than pre-release data.
   - Time Periods: The 1-week data has the strongest correlation, and the correlations of 1-month data and 4-month data decrease accordingly.
   - News Entities: Director references have the least correlation with grosses; movie titles and top actors have better correlations with grosses (or budget).
   - Seven Sentiment Categories: "*General*" and "*Media*" sentiment counts have the highest correlation with grosses among all the seven sentiment categories.
   - Negative References vs. Positive References: From Table III, we can see that positive references are better correlated with grosses than negative ones for all sentiment categories except "*Crime*" and "*Health*".
   - Low-grossing Movies vs. High-grossing Movies: For low-grossing movies, the news references for top 3 actors are better gross predictors than those for top 15 actors. For high-grossing movies, we have the opposite conclusion.
2) *Movie Grosses versus Derived Sentiment Statistics*: Based on raw sentiment references, we derive several sentiment measures, including *polarity, subjectivity, positive references per reference, negative references per reference, and positive-negative differences per reference*. They are defined as the follows.
   - $polarity = \frac{pos\_refs}{total\_senti\_refs}$
   - $subjectivity = \frac{total\_senti\_refs}{total\_refs}$
   - $pos\_refs\_per\_ref = \frac{pos\_refs}{total\_refs}$
   - $neg\_refs\_per\_ref = \frac{neg\_refs}{total\_refs}$
   - $senti\_diffs\_per\_ref = \frac{pos\_refs - neg\_refs}{total\_refs}$

   We do not list the detailed pairwise correlation table here, but the result shows the correlations between grosses and all these five statistic variables are not strong ($<= 0.3$). However, correlation coefficients for several of them, such as *polarity, negative references per reference, and positive-negative differences per reference* are still statistically significant at a 0.05 significance level. We also notice that article count, frequency, positive frequency, and negative frequency

are highly correlated each other. To avoid multicollinearity, our prediction model preferably use only one of them. We can also use some derived sentiment indexes, because they are not highly correlated with raw news references, and thus will give us some new information other than the raw counts.

IV. Prediction Models and Comparison

Two basic modeling methodologies used in this paper are regression and k-nearest neighbor classifiers. Regression models forecast grosses by a regression equation. By contrast, k-NN models identify the most "similar" movie of the target movie from the training set by examining their similarities, because we think that "similar" movies should have similar grosses.

To evaluate performance (or accuracy) of models, we suppose G is the actual gross and P is the predicted gross, and then we have below evaluation methods:

1) *AMAPE (Adjusted Mean Absolute Percentage/Relative Error)*: $AMAPE = \frac{\sum_{i=1}^{n} |APE_i|}{n}$, where APE (Adjusted Percentage Error) is defined as $\max_{abs}(\frac{G-P}{G}, \frac{G-P}{P})$. The operator "$\max_{abs}$" chooses the element that has the biggest absolute value.
2) *Score of Models*: $Score = \frac{\sum_{i=1}^{n}(100-\min(100,|APE_i|))}{n}$
3) $\alpha\%$ *percentage coverage*: $PC_{\alpha\%} = \frac{Number\ of\ movies\ whose\ |APE| \leq \alpha\%}{Total\ number\ of\ movies\ (n)}$

A. Prediction from Traditional Movie Variables

Traditional movie models are our base models. We build separate models according to budget information availabilities, i.e., "*budget*" and "*nobudget*" cases.

1) *Regression Models (Reg_{budget} and $Reg_{nobudget}$)*: Model Reg_{budget} use variables budget, holiday flag, MPAA rating, sequel flag, foreign flag, opening screens, and genres. Model $\text{Reg}_{nobudget}$ is the same, but with removing budget indicator.
2) *k-Nearest Neighbor Models (kNN_{budget} and $kNN_{nobudget}$)*: The similarity of movies could be measured by "distance", which is further evaluated in a multi-dimensional space. Firstly, we define the distance for each dimension. For example, the distance of two budget value B_1, B_2 are defined as: $dis(B_1, B_2) = \frac{max(B_1,B_2)-min(B_1,B_2)}{min(B_1,B_2)}$. The distance for other variables are defined accordingly. And then we can get the distance formula by regressing the training data set. After this, we find the k movies from the training set which are the k nearest neighbors. In addition, our results show that the k-NN models work poorly when $k = 1$, but they work well enough while $k = 7$.

The performance data shows Reg_{budget} is better than $\text{Reg}_{nobudget}$, which means that budget is capable of improving performance substantially in regression models. The performance of K-NN models strongly depended on the training set size. With additional training data, and the increasing of k (but yet still a small number), the performance of K-NN models will be further improved. For all models, the high-grossing movies are predicted significantly better than low-grossing movies. The overall performance of K-NN models is similar to, but the high-grossing performance is better than that of regression models. If we use regression models for low-grossing movies and k-NN models for high-grossing movies, the best prediction will be expected.

B. Prediction from News Variables

In this section, we will predict movie grosses using news data only. Several models are built as follows.

1) *Regression Models Using News References Only ($nReg_{1w}$ and $nReg_{mov+act15}$)*: Model nReg_{1w} takes three indicators, the pre-release 1-week news article counts in terms of movie titles, top 3 actors, and top 15 actors. By contrast, $\text{nReg}_{mov+act15}$ takes six indicators, the pre-release 1-week, 1-month and 4-month news article counts in terms of movie titles and top 15 actors. The simulation result shows that models nReg_{1w} and $\text{nReg}_{mov+act15}$ have similar performance, and both of them perform better than other news-reference-based models, which means our predictors are chosen properly.
2) *Regression Models Using News References plus Sentiment Data ($nReg_{1w+senti1}$ and $nReg_{1w+senti2}$)*: Based on nReg_{1w}, $\text{nReg}_{1w+senti1}$ adds raw sentiment counts , while $\text{nReg}_{1w+senti2}$ adds derived sentiment statistics like polarity or subjectivity. However, both their overall performance and high-grossing performance have no significant improvements compared to the base model nReg_{1w}, because sentiment counts are highly correlated with the news article counts and thus carry little extra information while regressing.
3) *k-Nearest Neighbor Models ($nkNN_{1w}$, $nkNN_{mov+act15}$, and $nkNN_{1w+senti1}$)*: The three k-NN models use the same indicators as corresponding regression models. The distance of two movies can be easily computed by normalizing the reference or sentiment counts. Surprisingly, the sentiment data in the k-NN models shows some predictive power and the improvement is statistically significant. Moreover, k-NN models have worse overall performance but better high-grossing performance than corresponding regression models. In addition, k-NN models using news data can achieve similar performance with IMDB models, especially for high-grossing movies.

C. Prediction from Combined Variables and Performance Comparison

We have shown that decent models can be built using either traditional IMDB data or news data. Now we build models with the combination of IMDB data and news data, and indeed yield even better prediction results. For example, in the *"nobudget"* case ("budget" is not an input variable), $\text{Reg}_{nobudget}$ is the regression model with IMDB data and it yields only a coefficient of determination R^2 of 0.448, which is almost the same with the result of pre-release model from Simonoff and Sparrow [5]. By contrast, $\text{Reg}_{nobudget}$+nReg_{1w} is the corresponding regression model with IMDB data plus news data, and it achieves a R^2 of 0.788, which indicates a big improvement.

We studied the adjusted percentage error (or residual) plots for IMDB models, news models, and combined models. The results show some movies' grosses are highly overestimated and some others are highly underestimated if we use only IMDB data. However, those highly underestimated or overestimated grosses are eliminated in news models or combined models.

Method	Model	$\text{Perf}_{Overall}$		Perf_{High}	
		Err	Score	Err	Score
IMDB	$\text{Reg}_{nobudget}$	**7.83**	**92.8**	**8.97**	**92.41**
	Reg_{budget}	3.53	96.47	2.03	97.97
News	nReg_{1w}	**8.72**	**92.1**	**4.02**	**96.2**
	$\text{nReg}_{mov+act15}$	10.46	92.07	2.87	97.13
Comb	$\text{Reg}_{nobudget}$+nReg_{1w}	**3.82**	**96.81**	**2.48**	**97.52**
	$\text{Reg}_{nobudget}$+$\text{nReg}_{mov+act15}$	3.79	96.21	2.4	97.6
	Reg_{budget}+nReg_{1w}	2.76	97.24	1.57	98.43
	Reg_{budget}+$\text{nReg}_{mov+act15}$	2.63	97.37	1.54	98.46

Table IV: Performance Comparison of Regression Models for IMDB, News, and Combined methods. Error is evaluated by AMAPE. The bold numbers show the comparison of a group of experiments.

The completed performance data of regression methods for IMDB models, news models, and combined models are listed in Table IV. Compared to pure IMDB models or pure news models, the combined models yield nice performance improvement, which can be indicated by smaller AMAPE and higher scores. As for k-NN models, we have the same result. Our t-test proves the improvement is statistically significant. Table IV also shows that the pure news models have

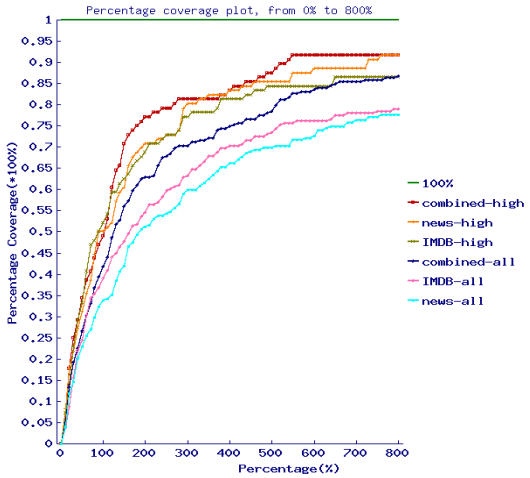

Figure 1: Comparison of Regression Models (*"nobudget"* case). These models use IMDB data, news data and their combination respectively. The combined model works best among all three models, both for overall performance and high-grossing performance.

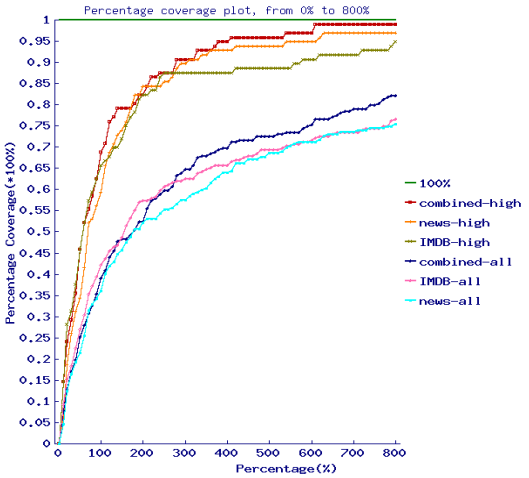

Figure 2: Comparison of k-NN Models (*"nobudget"* case). These models use IMDB data, news data and their combination respectively. The combined model works best among all three models, both for overall performance and high-grossing performance.

similar overall accuracy to, but better accuracy of high-grossing movies than pure IMDB models.

Figures 1 and 2 show the "Percentage vs. Percentage Coverage" comparison (*"nobudget"* case) of IMDB, news, and the combined models. The X-axis shows the $\alpha\%$ percentage, while the Y-axis shows the corresponding $\alpha\%$ percentage coverage. These plots show both overall performance and high-grossing performance of combined models are higher than those of IMDB or news models. We have exactly the same conclusion for *"budget"* case. Furthermore, the comparison of Figures 1 and 2 also shows that regression models work better for overall performance, while k-NN models perform better for high-grossing performance. That is, regression models are more suitable for low-grossing movies, but k-NN models are more suitable for high-grossing movies.

V. Conclusions

We have discussed the correlation of movie grosses with both traditional IMDB data and movie news data, and built models with IMDB data, news data, and their combination respectively. Our experiments proved media's predictive power in movie gross prediction.

Detailed conclusions are as the follows. Firstly, movie news references are highly correlated with movie grosses, and sentiment measures including derived sentiment indexes are also correlated with movie grosses. Secondly, movie gross prediction can be done by either IMDB data, news data, or their combination. Prediction models using merely news data can achieve similar performance with models using IMDB data, especially for high-grossing movies, while the combined models using both IMDB and news data yield the best result. Therefore, news data is proven to be capable of improving movie gross prediction in our analysis. Thirdly, both regression and k-nearest neighbor classifiers can be used for movie gross prediction. With the same indicators, regression models have better low-grossing performance, but k-NN models have better high-grossing performance. Finally, article counts for movie entities are good movie gross predictors. News sentiment data are good predictors for k-NN models, but not good predictors for regression models.

References

[1] W. S. Chan, "Stock price reaction to news and no-news: Drift and reversal after headlines," *Journal of Financial Economics*, vol. 70, pp. 223–260, 2003.

[2] P. C. Tetlock, M. Saar-Tsechansky, and S. Macskassy, "More than words: Quantifying language to measure firms' fundamentals," in *Proceedings of 9th Annual Texas Finance Festival*, May 2007.

[3] L. Lloyd, D. Kechagias, and S. Skiena, "Lydia: A system for large-scale news analysis," in *Proceedings of 12th String Processing and Information Retrieval (SPIRE 2005)*, vol. LNCS 3772, Buenos Aires, Argentina, 2005, pp. 161–166.

[4] A. Chen, "Forecasting gross revenues at the movie box office," *Working paper, University of Washington, Seattle, WA*, June 2002.

[5] J. S. Simonoff and I. R. Sparrow, "Predicting movie grosses: Winners and losers, blockbusters and sleepers," *Chance*, vol. 13(3), pp. 15–24, 2000.

[6] M. S. Sawhney and J. Eliashberg, "A parsimonious model for forecasting gross box-office revenues of motion pictures," *Marketing Science*, vol. Vol. 15, No. 2, pp. 113–131, 1996.

[7] P. Chaovalit and L. Zhou, "Movie review mining: a comparison between supervised and unsupervised classification approaches," in *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*, 2005.

[8] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. Vol. 2, No 1-2, pp. 1–135, 2008.

[9] G. Mishne and N. Glance, "Predicting movie sales from blogger sentiment," *In AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, pp. 155–158, 2006.