Box-Office Opening Prediction of Movies based on Hype Analysis through Data Mining

Ajay Siva Santosh Reddy
St. Francis Institute of Technology

Pratik Kasat
St. Francis Institute of Technology

Abhiyash Jain
St. Francis Institute of Technology

ABSTRACT

Box-office performance of a movie is mainly determined by the amount the movie collects in the opening weekend and Pre-Release hype is an important factor as far as estimating the openings of the movie are concerned. This can be estimated through user opinions expressed online on sites such as Twitter which is an online micro-blogging site with a user base running into millions. Each user is entitled to his own opinion which he expresses through his tweets. This paper suggests a novel way to mine and analyze the opinions expressed in these tweets with respect to a movie prior to its release, estimate the hype surrounding it and also predict the box-office openings of the movie.

Keywords

Twitter, hype, micro-blogging, opening, movies

1. INTRODUCTION

It is quite difficult to forecast the success of a particular movie before its release. Therefore, predicting the box-office success of a particular movie has been a difficult task for many industry analysts. Such unpredictability of the film industry makes the movie business one of the riskiest endeavors for investors to take in today's competitive marketplace. Most analysts have tried to predict the total box-office success of motion pictures after a movie's initial theatrical release with some level of success. This paper hopes to devise a method to give an idea about the business, a movie would do based on its pre-release hype. Hype is one such factor which drives a layman to a theatre to watch a movie and this decides the openings of a movie that is the occupancy of theatres playing that particular movie. To determine the hype surrounding the movie which is an indirect approximation of the excitement the people have towards that movie andcreates a need to turn to the areas where the activity of the people is highest online, that is the social networking sites. In today's shrinking world, you can see the boom in the use of social media like Twitter. On these social networking site people communicate about every potential area of interest, including, Movies, food, sports, music, etc. It is estimated that there are over 900 social media sites on the internet. But this paper focuses on the site 'Twitter'.

Twitter is an online social networking and micro-blogging service that enables its users to send and read text-based messages of up to 140 characters[10], known as "tweets". It was created in March 2006 by Jack Dorsey and launched that July [10].The service rapidly gained worldwide popularity, with over 500 million active users as of 2012[10], generating over 340 million tweets daily and handling over 1.6 billion search queries per day. Since its launch, Twitter has become one of the top 10 most visited websites on the Internet, and has been described as the SMS of the Internet. Unregistered users can read tweets, while registered users can post tweets through the website interface, SMS, or a range of apps for mobile devices. As of July 2012, the average number of tweets sent per day was 140 million. People are not just viewing the content and they are also socializing and tweeting about like they like or not. As a result of the time spent on the social media, the businesses are intrigued to predict the success of movie largely on the basis of interests of the people on the social networking sites. The use of social media sites has grown significantly, and this fact is being recognized by film business investors, including film distributors. In response to this, businesses are considering social media as an deciding factor of their various ventures. Every Tweet generates a new data point, a new bit of information that may be of value to film industry as well as other businesses. This information might help a movie in regenerating revenue.

In this study, the site Twitter is used as a source for input data where millions of tweets pertaining to the particular movie could be find out. A variety of opinions are expressed on this forum which holds a lot of value for the owner of the object on which the opinion has been expressed. These tweets would be analyzed by the model prior to the movie release and converted into a graphical structure which would be easy to evaluate. Moreover, it will help film-makers to predict the success of their movie weeks before the release. The data model described in the paper basically accumulates the tweets posted by the users relating to particular movie on the Twitter website and converts them into a graphical structure for easy interpretation of data. Moreover, it is not simply based on the number of tweets itself but rather takes into account, the number of distinct users. A ratio between the number of users and the tweets are calculated using a formula described in the later section. This ratio gives the accurate approximation of the hype and thus the success of movie by comparing it with the number 1. The closer the value is to the number 1, the better will be its success at the box office. Similarly, if value is less close to number 1, it is least likely to be successful one.

2. LITERATURE REVIEW

Literature Review on predicting success of new Movies be classified based on the type of forecasting model employed: (i) econometric/quantitative models-those that explore factors that influence the box-office receipts of newly released movies (Litman, 1983; Litman& Kohl, 1989; Litman&Ahn, 1998; Neelamegham& Chintagunta,1999; Ravid, 1999; Elberse&Eliashberg, 2002; Sochay,1994) and (ii) behavioral models-those that primarily focus on the individual's decision-making process with respect to selecting a specific movie from a vast array of entertainment alternatives (De Silva, 1998; Eliashberg&Sawhney, 1994; Eliashberg et al., 2000; Sawhney&Eliashberg, 1996; Zufryden, 1996). These behavioral models are based on a hierarchical framework, where these important behavioral traits of consumers are combined with the econometric factors in developing the forecasting models [1].

There has been prior work on analyzing the correlation between blog and review mentions and performance by SitaramAsur and others [2]. Almost all of them have used meta-data information on the movies themselves to perform the forecasting, such as the movies genre, release date, the number of screens on which the movie debuted, and the presence of particular actors or actresses in the cast. The correlations observed for positive sentiments are fairly low and not sufficient to use for predictive purposes. Apart from the fact that they are predicting ranges over actual numbers, the best accuracy that their model can achieve is fairly low.Another work on Sentiment and Social network analysis has been done by Krauss, Jonas and others [3].The paper analyses message board community of IMDB and prepares 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.

Another work on Classification of Movies using the Data Mining Technique has been done by S. Kabinsingha and others [4]. In this paper the data mining technique is applied to perform classification of movies. In the prototype model, the movies are rated into PG, PG-13 and R. The data are divided into training and testing set with 4 fold cross validation. Among all various attributes of movies like actors, actress, directors, budget, genre, producers, etc., the total number of selected attributes is 8 which depend mainly on the genres of the movies and the words used in the movies. This corresponds to the decision used by most of the film rating organization. The prototype model is created based on the decision tree.

The work on comparing several techniques for learning statistical models in machine learning and data mining has been done by David Jensen and Jennifer Neville [5]. It shows the comparison of the several data mining techniques that have been developed for relational data that include probabilistic relational models (PRMs) (Friedman, Getoor, Koller, and Pfeffer 1999),Bayesian logic programs (BLPs) (Kersting and de Raedt 2000), first-order Bayesian classifiers(Flach and Lachiche 1999), and relational probability trees (RPTs) (Jensen and Neville 2002). In each of these cases, both the structure and the parameters of a statistical model can be learned directly from data, easing the job of data analysts, and greatly improving the fidelity of the resulting model. Older techniques include inductive logic programming (ILP) (Muggleton 1992; Dzeroski and Lavrac 2001) and social network analysis (Wasserman and Faust 1994).The paper employed a relational probability trees (RPTs) to learn models that predict the box office success of a movie based on attributes of the movie and related records, including the movie's actors, directors, producers, and the studios that made the movie.

There has been some prior work on analyzing connections on Twitter byChunyan Wangthat studies social interactions on Twitter to reveal that the driving process for usage is a sparse hidden network underlying the friends and followers, while most of the links represent meaningless interactions. It has been examined Twitter as a mechanism for word-of-mouth advertising. They considered particular brands and products and examined the structure of the postings and the change in sentiments. Galuba et al proposed a propagation model that predicts which users will tweet about which URL based on the history of past user activity. Predicting the future is not an easy task especially when the predicted phenomenon is abstruse and can only be characterized with partial information (Karakan and Koç, 2008). In movie industry, most still believe that the success of movie can only be predicted in the artistic nature of the product. However, given the investment taking place in the movie industry, the need for efficient decision support tools and models are necessary. A general lack of research that describes successful implementation of these decision models suggests that there are challenges as well as opportunities for researchers in developing predictive analytics for this area. This paper describes a way forward in filling such gap.

It is evident that these are post release reviews of a movie. This paper presents a way to estimate pre-release movie hype in a different way. In this study, the concept of distinct number of tweets by a user is done to accurately predict success of a movie at the box office.

3. WORKING

The approach is quite simple. First,findthe number of tweets pertaining to a movie by using a web crawler. These tweets are collected on per hour basis. The first factor contributing to the hype is given by the total number of tweets pertaining to the movie. Finding out β, which is the number of relevant tweets per second

$$\beta = No\,of\ relevant\ tweets\ per\,\mathrm{sec}\,ond\,\mathrm{K}\ [1]$$

Then find the number of distinct users who have posted the tweets. The number of distinct users can be calculated by counting user-id of the users. This process starts one week before the release of the movie. The following formula is used for calculating the hype factor (α):

$$\alpha = \frac{No\ of\ distinct\ users}{No\ of\ tweets\ by\ all\,users}\mathrm{K}\ [2]$$

To enhance the estimation of hype it is necessary to consider the reach of a particular tweet by including the follower count of a particular user who referred to the movie in his tweet if the count is above a certain thresh-hold value (τ).The follower-count factor can also be a considered a factor to ensure the reach factor is also included in the determination of the hype. The reach factor (σ) can be given as

$$\sigma = \frac{follower\ count - \tau}{follower\ count}\mathrm{K}\ [3]$$

Where τ = average no. of followers per all the users who tweeted. σ can be scaled down to a scale of 0.1-1 with 0.1 being assigned to the thresh hold value assuming cases where the follower count being more than 10 times the thresh hold value as a rare case and assigning it the value 1.

The final hype can be given as:-

$$Hype = \sum \frac{\alpha + \sigma(scaled\ down)}{2}\mathrm{K}\ [4]$$

for the factors calculated each hour.

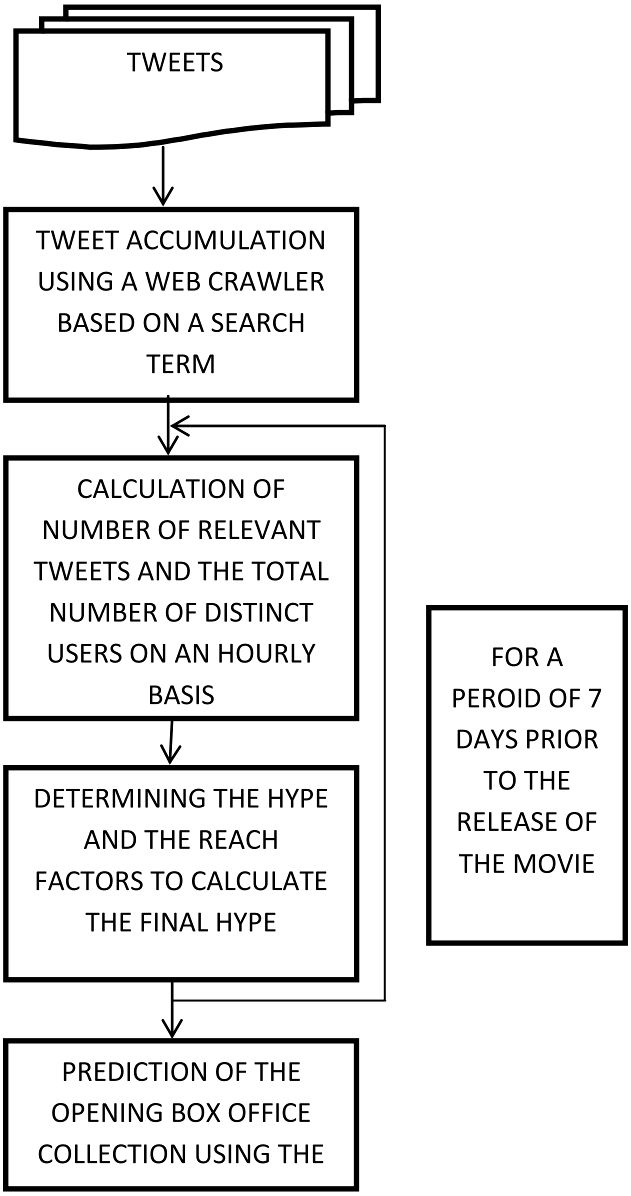
TWEETS
TWEET ACCUMULATION USING A WEB CRAWLER BASED ON A SEARCH TERM
CALCULATION OF NUMBER OF RELEVANT TWEETS AND THE TOTAL NUMBER OF DISTINCT USERS ON AN HOURLY BASIS
DETERMINING THE HYPE AND THE REACH FACTORS TO CALCULATE THE FINAL HYPE
PREDICTION OF THE OPENING BOX OFFICE COLLECTION USING THE
FOR A PEROID OF 7 DAYS PRIOR TO THE RELEASE OF THE MOVIE

Figure1. Model to mine the tweets and calculate the hype

The hype factor (α) gives values which may be an integer or decimal. Therefore, the concept of ratio is used to calculate the hype and thus the chances of movie success. The estimation about hype can be made as follows:

1) Count the number of tweets posted by the user pertaining to movie title.

2) Count the number of distinct users who has posted the tweet specified in step1.

3) Calculate the hype using the formula stated above.

4) Compare the calculated value in the step3 with the number1.

5) If the value is closer to number 1 then there are high chances of movie generating high opening collections similarly, if the value is farther from number 1 there are less chances of getting a good opening.

As per above methodology, the closer the value is to the number 1, the closer it is to success and more chances of it getting hit at the box office. The reason to use the ratio based approach is to get the closest approximation of hype which would be difficult had it been without the number of users. Number of distinct users is an important factor because hype can be best known through the number of users being interested in particular movie. The success of movie at the box office can be best determined by the ratio of number of users to the number of tweets rather than taking only number of tweets in consideration which would not give the best possible approximation.

This is because the number of tweets is not directly proportional to the number of users. It is difficult to determine the accurate hype by only considering the number of tweets because the number of users is not known .The number of users is very important because ultimately the people will decide movie’s success or failure.

For example, Let the number of users who have posted the tweet be 10 and the number of tweets being posted by them pertaining to a movie be 100. By considering only the tweets, the actual hype would not be determined. So, only after consideration of the number of users actual hype is known. In this case since there are 10 users who have posted 100 tweets, after analyzing the number of users the fact that the hype is much less and movie is less likely to be a successful one is known. Another situation could be 500 tweets posted by only 50 users which prove that considering only tweets is specious.

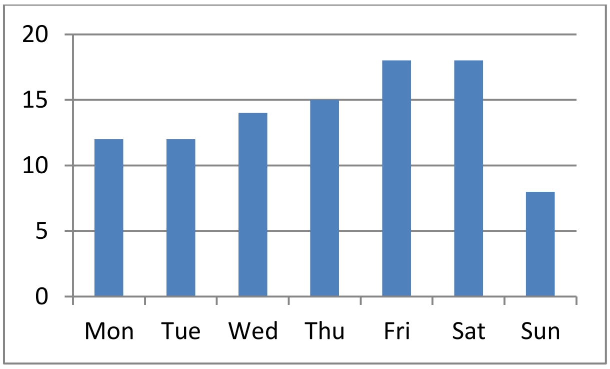

Fig. 2Percentage of tweets containing the term ‘Cocktail’ for each weekday

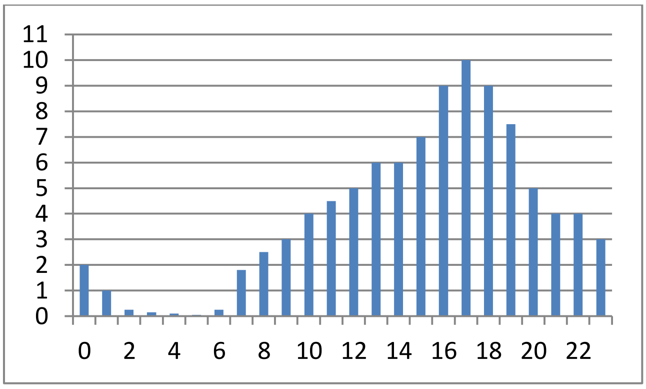

Fig. 3 Percentage of tweets containing the term ‘Cocktail’ for each Hour.

The method gives the approximation of the popularity of movie by using hype based approach estimated using the formula stated before. On applying the formula to the same examples mentioned above. In the first case the hype is 0.1

much farther from number 1 and therefore less likely to be successful at the Box-Office. Similarly, Lets estimate the hype for number of users say 350 and number of tweets being 500. Here the estimated hype is 0.7 much closer to number 1 and thus the movie is likely to be a successful one. Therefore, the approach would correctly predict the popularity of movie by estimating the hype. Thus, the accurate prediction of the movie's success can be made by adopting this novel approach.

The number of tweets per second is another factor which gives us a rough idea about the hype, the movie possesses as it would mean that there are a lot of tweets being posted about the movie which means that the movie is being discussed and anticipated for. Also the reach of the tweet is to be considered as it would mean that that a wide number of followers were subjected to witnessing a tweet posted by the person theyfollow. Withconsideration of these factors the hype is calculated which should be as high as possible to ensure the film a good opening weekend.

The opening weekend collection can be calculated using the hype factor and the knowledge of how many screens the movie is going to release in the occupancy of each movie theatre is analogous with the hype surrounding the movie. The opening box office collection (O) can be predicted as

$$o = \mu * Hype * \varphi \mathrm{K} \; [5]$$

Where

O is the opening box-office collection

µ is the number of shows per day in all screens together for the weekend

φ is the average price of all tickets per screen per show

Comparison with collaborative Filtering techniques

Earlier work presented bySuhaas Prasad showed the collaborative filtering technique in which movie rating were predicted by present and past preferences of the users. In this approach the model predicts how a user will rate a movie based on rating histories of the user along with many others. The collaborative Filtering approach is used which attempts to learn from the past user relationships[9].

In this paper movie's success is accurately predicted using the ratio of number of users to the number of tweets. Along with this a follower count is also used for more accurate results. The results obtained in this technique are more accurate than collaborative filtering technique discussed above. In this way the model used in this paper outperforms the above technique.

Sample Experimental Result

Considering the following tweets mined using the search term 'SKYFALL' which is the upcoming James Bond film a month and a half prior to the release, tweets such as these were obtained:-

"Couldn't think of a better person to sing the #**skyfall** title song!"

"In #**skyfall**, #JamesBond@007 will appear wearing the steel-on-steel #OMEGASeamaster Planet Ocean 600m 42mm."

" That #**skyfall** trailer gives me goosebumps#awesome"

"New #**Skyfall** trailer with #Adele theme. We think the theme is both appropriate and awesome. You?"

Through this collection of tweets the numbers of relevant tweets per second were calculated to be 27, tweeted by 22 distinct users having an average follower count of 93which can be taken as the thresh-hold.

Hence

$\beta = 27$ using the equation [1].

$\alpha = \frac{22}{27} \approx 0.814$ using the equation [2].

Considering a user with 114 followers, the value of σ is obtained as

$\sigma = \frac{114 - 93}{114} = 0.18$ using equation [3].

Hype is then given as

Hype= $\frac{(0.814) + (0.18)}{2} = 0.497$ using equation [4].

With this data the film and the assumption that the film releases in 1600 screens with $5000 being the approximate mean full house collection of each screen, the film is estimated to earn:-

$o = 1600 * 0.497 * 5000$ using equation [5].

$o = \$3976000$ Per day in the opening weekend

This model predicted the opening weekend collection at the box office for the movie Skyfall , based on the hype factor calculated approximately a month and half prior to the release. A more accurate result can be obtained on performing the analysis a week prior to the release.

4. CONCLUSION

To the best of our knowledge this novel approach used in prediction will guarantee results that are slightly better than any of the known published literature for this problem area. Beyond the accuracy of our prediction results of box-office hype, this model could also be used to predict the success changes using the particular parameters used within the model to calculate the prerelease hype. The accuracy of this model is increased by including new hype Factor and the reach Factor and scaling the factor to a threshold limit due to which using the Hype Based formula can be a more accurate method to predict the opening success of the box office. This is more viable method than present before to calculate the opening success of the movie based on the pre movie release hype.

5. REFERENCES

[1] Ramesh Sharda, DursunDelen, "Predicting boxofficesuccess of motion pictures with neural networks", Department of Management Science and Information Systems, William S. Spears School of Business, Oklahoma State University, Stillwater, OK 74078, USA

[2] SitaramAsur, Bernardo A. Huberman, "Predicting the Future With Social Media," Social Computing Lab HP Labs Palo Alto, California

[3] Krauss, Jonas; Nann, Stefan; Simon, Daniel; Fischbach, Kai," PREDICTING MOVIE SUCCESS AND ACADEMY AWARDS THROUGH SENTIMENT AND SOCIAL NETWORK ANALYSIS," University of Cologne.

[4] S. Kabinsingha, S. Chindasorn, C. Chantrapornchai, "Movie Rating Approach and Application Based on Data Mining", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 1, July 2012.

[5] David Jensen and Jennifer Neville,"Data Mining in Social Networks ," Computer Science Department, University of Massachusetts, Amherst.

[6] M. Saraee, S. White & J. Eccleston, "A data mining approach to analysisAnd prediction of movie ratings", University of Salford, England.

[7] Roosevelt C. Mosley Jr., "Social Media Analytics: Data Mining Applied to Insurance Twitter Posts," Casualty Actuarial Society E-Forum, winter 2012-Volume.

[8] Lyric Doshi, Using Sentiment and Social Network Analyses to predict Opening–Movie Box Office Success,Department of Electrical and Computer MIT, USA, Feb 2010.

[9] Suhaas Prasad," Using Social Networks to Improve Movie Rating Predictions"

[10] Twitter, http://en.wikipedia.org/wiki/Twitter