I J C T A, 9(41) 2016, pp. 629-640 ISSN: 0974-5572

IS MOVIE A BOX OFFICE SUCCESS? ANALYZING SEARCH QUERIES TO PREDICT COMMERCIAL SUCCESS

Maharshi Vyas[*], Keval Shah[*] and Nitin Upadhyay[**]

Abstract: Movie production is a risky and time consuming task. significant resources are invested in creating, editing, distributing and showcasing a movie across the world. In this industry, the most common and high-priority question to be answered is "whether a particular movie with the specified cast in the selected genre will be a box office success or not?". This will allow the stakeholders associated with the movie to explore the opportunity space in promoting, branding and repositioning the strategies to make gain commercial success while maintaining the overall gist of the movie plot and other core features. In recent days, viewers are extensively using search engines to know about the features of a movie such as genre, cast, studio etc. These platforms in turn provide more reliable insights on the generated search engine data that can be traced to societal trends and behaviour. Therefore the massive data generated from the search engines has widened the perspective of the market research and analysis. In this paper, authors provide movie analytical framework to predict its success based on the the available massive search query data.

Key Words: ***Market Research, Market Prediction, Big Data Analytics, Movie Analytics, Machine Learning.***

1. INTRODUCTION

Amid 2009-2015, Hollywood annually generated a revenue of at least $10 billion [7]. The global film industry shows healthy projections for the coming years. In this industry, for a given project large amounts of resources are gathered from various producers around all parts of the world. It is estimated that the average movie now costs nearly $100 million after including production and marketing expenses.It becomes very important to achieve the first glimpse of the tentative movie performance in the market. For example, a famous movie series fast and furious, 5th movie of which was not released majorly in China. Sixth installment of the movie series earned $67 million in China which inspired them to release seventh edition widely in China which resulted in earning $390 million dollars in China which was higher than US, the origin country. Marketing and theatre releases cost millions of dollars, they are expected to generate revenue and add to the revenue a movie has targeted to achieve, so it is of utmost important how this limited amount of

[*] *Computational Science, DA-IICT, Gandhinagar, Gujarat, India*
me.maharshi.vyas@gmail.com , *kevalds51@gmail.com*

[**] *Information technology,Goa institute of Management, Goa, India*
nitin@gim.ac.in

money is used. Market research and market prediction is a must if comprehensive marketing strategies are to be applied. Prediction of a success of a movie is a wide area which depends upon a lot of factors being a reason for a less accurate prediction but practised by researchers with a great amount of variations in methods to be applied.

This type of prediction is attempted before by Mestyan, Kertesz [11]. Their predictive model is based on collective activity of all type of users online. They predicted popularity of a movie by activity of editors and viewers of corresponding movie on Wikipedia. Recently, Taneja [16] tried to predict the success of a film at the box office by looking into the past combinations of an actor, actress, director and revenue earned by those combinations at box office. A movie success prediction model is made by prithvi [17] also, where he uses features like IMDB data, ratings of actors and directors, budget, popularity if trailer etc. to predict. These are most commonly used attributes for movie predictions. Some have also used MPAA rating, number of theatres to release, IMDB rating, Genre, No. of voters, metascore, Tomato User ratings and Tomato-Meter [3,6]. There has been a great amount of work for analysis of search engine queries which can provide extensive social applications. Choi and Varian [4] has depicted that the query data from the Google trends[8] is significantly useful to forecast the near-term values of economic indicators. Google queries are used to detect influenza epidemics in various areas in USA by Jeremy [12]. Preis et al. showed that weekly transaction volumes of ‘‘S&P 500 companies’’ and weekly Google search volumes of company names are highly correlated [11].Search query volumes also showed that for information about preceding and following years, a country’s GDP and the predisposition of its inhabitants to look forward are highly correlated [12].

This paper targets to provide a systematic movie analytics process and a technique to predict how much a particular movie will gross in a selected market or a country. This will allow the key stakeholders of the movie such as producers and distributors to scale the project accordingly. Understanding search engine queries will comprise of much larger data, it can be helpful to predict a movie success, which is supported by results here.

The rest of the paper is structured as follows: In section 2, related work is presented. Section 3 describes a framework for data collection and analytics. In section 4, detailed analysis, discussion and recommendation to the stakeholders are covered. Finally, section 5 provides contributory conclusion and pointers on future scope of the research work.

2. RELATED WORK

Data scientists have used search engine data to create models that predict the various features that may be used as a judging criteria for movies. A neural network was trained by Sharda and Delen [6] to process pre-release data, and classified movies into nine categories from “flop” to “blockbuster” in accordance with their predicted net gross. For test samples, neural network classifies only 36.9% of the movies correctly, while 75.2% of the movies are at most one category away from where they should be. Same sort of work is done by Darin and Minho [11] with some positive and negative attributes in the model. Nithin [3] has predicted movie success from features such as IMDB and rotten tomatoes data, actor, director, writer and genre.

Mahesh[5] have built a linear regression model that combines meta-data and text features from reviews by critics before release to predict revenue. Mishne and Glance [18] used sentiment analysis techniques to pre-release and postrelease blog posts about movies and showed higher correlation between actual revenue and sentiment based metrics. Zhang and Skiena [19] used a

news aggregation system to identify entities and obtain domain-specific sentiment and used the aggregate sentiment scores and mention counts of each movie in news articles as predictors. So a lot of text processing is done on pre release posts to predict movie success but it may not fit the original scenario accurately always.

Depending on the user requirement, one can make predictions based on factors such as movie critic ratings, user ratings, box office collection and cast popularity. Deniz et al., [1] has predicted IMDB [10] movie rating using the Google trends which provides access to Google search engine data. The authors in their work have implemented movie prediction by considering movie viewers as primary users. Lee [2] entirely focuses on movie performance in markets based on metrics such as movie query search index, number of theaters and ratings. Their primary source for data collection is the Google trends [8] query database. Both the authors have used linear regression for modelling.

It can be seen that the state-of-art-literature lack in considering the entertainment sector with the massive search data and finance aspect. This paper will extend the work of Lee [2] to include country-wise analysis along with improvisation in some methods and addition of other practical factors that define a movie popularity and success. The overall analysis is achieved by considering proposed analytical framework.

3. FRAMEWORK - DATA COLLECTION AND ANALYTICS

In this section, the framework for data collection and analytics is described. The idea is to collect the data from publicly available data that revolves around the parameters that govern the commercial success of a movie. We adapt certain techniques to compare two movies and extract essential features accordingly. In the analytics part, we work upon the processed data to create efficient and accurate models. Here, the weights are assigned to the parameters to achieve real world scenarios.

3.1 DATA COLLECTION, CHARACTERISTIC, SCOPE

3.1.1 Relative Popularity Index and Benchmarking

Google trends provides a search popularity index(SPI) of a Google query in a specific time period and geographical region. This index is based on the query share, i.e. the total number of searches for a particular query divided by the total search volume seen in the specified geographical region for the given time period. Next, it performs another round of normalization and maps the highest and lowest values with 0 to 100 integers. Thus Google shows the normalized relative search statistics. The normalization constant varies from query to query. More information for the same is provided in choi and Varian [4].

$$SPI = \frac{relative\ \ frequency\ \ of\ a\ query}{span\ of\ relative\ \ frequency} \quad (1)$$

The limitation of the SPI is that it does not allow one to compare the popularity of two different queries using the trends data only. Two movies may have SPI 100 for same geolocation but may have huge difference in volume of overall queries. Deniz [1] has fixed the problem using AdWords data which shows the exact amount of frequencies for a given query. But there is no

surety if Google uses the same data for Google trends and Google AdWords. This often leads to large false predictions. Lee [2] does not provide any solution to this. It may lead to incorrect predictions. In this paper, the authors have overcome the limitations and has proposed a new index, Relative Popularity Index (RPI). Comparison between two queries done in Google trends shown in Fig. 1 have different normalization constants for different time intervals and geolocations. But RPI defined here gives a proper insight to the volume of queries.

RPI_x is equivalent to the RPI of x.

$$RPIx = \frac{SPI\ of\ x\ when\ x\ is\ compared\ \ to\ y}{SPI\ of\ y\ when\ x\ is\ compared\ \ to\ y} \quad (2)$$

From equation 1, it results to:

$$RPIx = \frac{relative\ frequencies\ of\ x}{relative\ frequencies\ of\ y}$$

Hence, two different movies when compared to a common movie, may have same SPIs, but the one with more frequencies will have more RPI. Also $RPI_x/RPI_z = RPI_x$ when compared to z. This property takes care of relative volumes of queries. Ultimately all the hits of a movie are compared with a common benchmark movie, now calculated RPI is the proper index for comparing movie queries.

The movie selected for benchmarking should have minimum variance across the time we have considered for all queries. The other case where it is accepted is when the variance is such that its average value when comparing a query nullify its variance, because again there is no absolute concept for variance too. A sample variance may be much higher when compared to rare queries and may be negligible when compared to widely popular queries.

The authors have selected the benchmark movie to be "*The Shawshank Redemption*" and benchmark personality as Charlie Chaplin for RPIs of actor and director queries. Note that a benchmark should not have equal and constant frequencies for all the countries we are taking, the selected movie and personality ensure the check. In the current research work different model for a different country have been built and trained, so the trained weights will adjust themselves as the frequencies and RPI vary. It will vary equally for everything for that particular country. Though, the frequencies of the benchmarks are nearly equal among all countries taken here.

3.1.2 Feature Selection:

In the current research work, we have selected following features to predict earnings of movies among different countries. The features which does not have any trustworthy data source providing separate data for different countries are not taken into consideration, for e.g. facebook page likes. Rather than generating inconsistencies due to use of common features, we have used features which are separate for each country.

1. Movie Query RPI(M): This is relative popularity index as defined above of a movie query compared with the shawshank redemption. Mostly at the time of release, a movie gets its peak value, it is followed by decay which is 3 to 4 times longer than its rise as shown in Fig. 1a. There are some cases when there is no proper decay, these are the movies which last long in the theatres e.g. *The Frozen* as shown in Fig. 1b. In these cases, we have taken the weeks during which the movie is still in theatres, neglecting further queries.

Now that each time we are looking at the average values of RPIs for the target country for the decided time interval. It can be see in Fig. 1b, popularity of the movie *Batman vs Superman* was nearly double than *The Frozen* but still it has earned 872 millions compared to 1.276 billion of *The Frozen*. The popularity in few months before the movie release affects only the starting weeks of a movie. The total gross depends more on how long the movie stays in cinemas. *The Frozen* has overcome *Batman Vs Superman* in terms of Total Gross and average. This is opposite to method used by Lee [2]. So a movie query RPI uses average of SPI over time t, where t = min(t where SPI_t <=(1/10) Peak SPI Value in that region, t for which movie lasted in theatres).

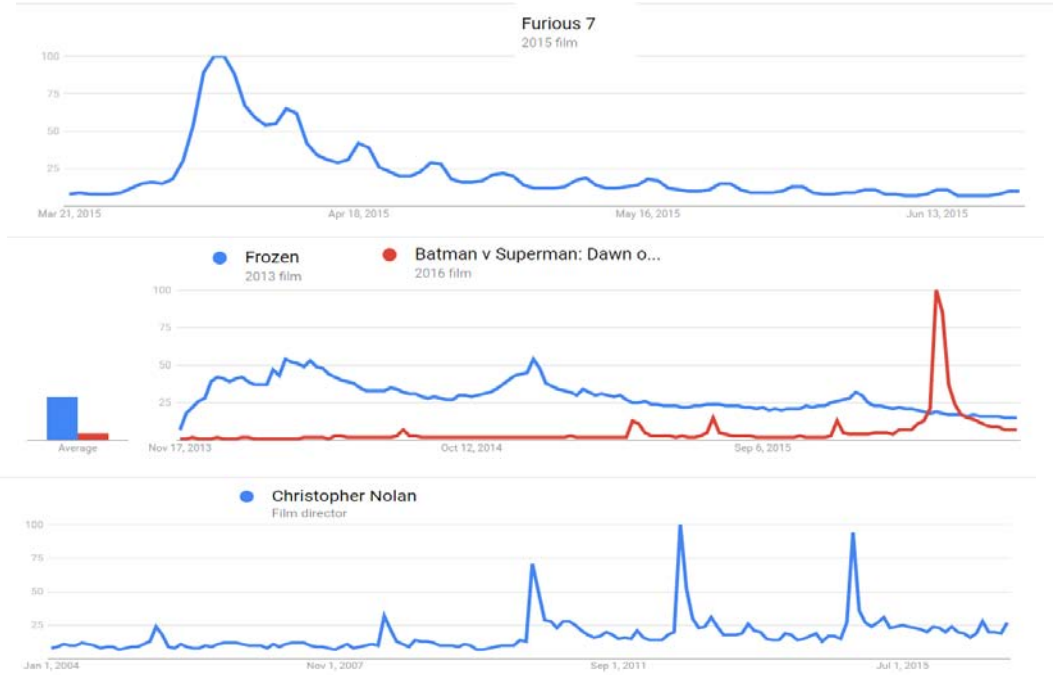

Figure 1. (a) Frequencies of a movie query across a time interval of 4 months, 20 days before the movie release and rest afterwards. (b) Shows a comparison between 2 search queries mapped from 0 to 100 with maximum and minimum frequencies of either query being compared. (c) Shows query frequencies for a well known film director. It has a spike whenever his movie is released. There is an increase because of his popularity and increase in number of Google users also.

2. Director Name Query RPI(D): Director name query RPI which is average of only few weeks before the movie release when its RPI starts to rise.

3. Actor #1 and Actor #2 Name Query RPI(A1, A2): Average of RPI of name queries of Actor#1 and Actor#2 as described in the IMDB which is average of only few weeks before the movie release when its RPI starts to rise. There is a rise in the search queries for directors and actors also when their movie gets released which will help in more accurate prediction. As you can see in Fig. 1c, there is an increase whenever *Christopher Nolan's* movie is released. Noteworthy point is this curve has a peak value at 2012 (*The Dark Knight* Release), before it there wasn't a

huge use of Google, so it has some lower peaks, but afterwards in 2014 (exactly at the time of *Interstellar Release*), he wasn't as popular as in 2012(exactly at the time of *The Dark Knight* Release), and *interstellar* had a lower box office compared to *The Dark Knight Rises.*

4. Genre#1, Genre#2 and Genre#3 Query RPI(G1, G2, G3): Popularity of genre of a movie in a country contributes in overall success of movie in that country. Average of RPI for queries of Genre#1, #2 and #3 around one year of release because release of a movie doesn't affect a popularity Genres.

5. Number of theatres released and number of weeks spent in cinemas(T, W).

Note that we have taken each feature and collected data for different countries we are looking upon.

3.1.3 What to Predict?

Deniz [1] has predicted IMDB movie rating with Google trends. Now, IMDB rating can vary vastly for equally popular movies according to its individual traits. So there can be huge variations with different factors like social media and reviews analysis. This data is not proper for that analysis. Nithin et al., [3] has predicted movie success from IMDB data. The work limits into considering and predicting about popularity and buzz for a movie. Lee[2] has predicted the overall gross with just movie search query of previous months(SPI), rather it can work best for opening weekends only and thus limit the overall prediction about commercial success.

In this research work, the data for a movie is collected for countries which form the major market for hollywood. These are US, UK, Australia, France and Germany. The data gives us a relative idea of popularity of a movie for a certain time and higher popularity leads to higher average attendance at cinemas. Multiplying average attendance with average price ticket for that country, and multiplying the result with the total number of theatres where movie is released will give us the total amount the movie will earn from the given country. Again, we are training different models for different countries and these multiplications are handled by each model independently. So there is no need to specify the average ticket prices for each country. Proper thing to predict is the buzz for the movie in people of a country, which as described directly leads to total gross in that country. Moreover, our prediction for each country will be more accurate than overall prediction of Lee [2] which is supported by results. Section 4.2 describes how a country wise analysis is more helpful. For example, *deadpool* was not released in China, when compared to furious 7 which was released there, *furious 7* earned $320 million dollars in China which results in huge difference of global income between both movies though having nearly same RPIs. But both movies had comparable gross when compared for USA which can be seen using RPIs for USA only.

Response Variable Movie Income(I): Total gross of movie for each decided country. We have taken natural log(I) to convert the movie income into a scale easy to predict upon.

3.1.4 Data Characteristics

It can be seen that movie name and related actor, directors queries increase during the release of a movie, but genre popularity doesn't get affected by it. So genre popularity may be very much comparable for a number of movies as there are 3 features for genres. In these cases, features M, T and W will contribute more towards the predictions. Cases when movie names are similar to some regular terms *e.g. The lovers, Home, Interview* can lead to wrong predictions. This will not happen

in our case as we are not using a movie name query, we are using a movie as an entity described on Google. We look for only those queries where a movie is referenced and shown on the page of Google, opposite to the work done by Lee[2]. We have done the same for directors, actors and genres. We are taking number of theatres(T) for each country, which contributes a lot in the prediction for those cases when a movie is popular, but has reached to limited screens, which results in lower earnings and happens frequently for lower budget movies. Number of theatres(T) is a key feature for a movie success, but individual movie characteristics matter a lot when it comes to the total gross. Therefore it can help predict opening weekend gross correctly but to predict total gross, we are taking few weeks after the movie release into consideration as described in section 3.1.1.

3.2 ANALYTIC FRAMEWORK

In this subsection the overall analytic process is described. Initially, the main task is to decide the judging parameters and extract the publicly available data for the same. The accuracy and correctness of the processed data will determine the quality of the model. After prediction of the expected gross, the focus will shift to reassigning the goals of various departments such as branding, marketing, distribution, casting and editing.

3.2.1 Analytic Process

The analytic process proposed in this research work consider three major steps- data collection, data preprocessing, training, prediction and increased profit.

1. Data Collection: The model uses 8 features among which features M, D, A1, A2, G1, G2 and G3 for each movie and for each country are collected from Google trends, and then converted RPIs are used. The box office gross of a movie(I), number of theatres(T) and number of weeks(W) it lasted, again separately for each country, are taken from box office mojo[7]. All of the data is gathered manually so we were able to collect data for 112 movies, from 2013 to 2016 so that variations are less. Most of the movies are released after 2014, which decreases anomalies created due to different ticket pricings and all.

2. Data Preprocessing: After collecting raw data, a python script converts it into a structured data to feed in the model to be used. If a query doesn’t have any entry for our specific country in the csv file from Google, then value 1 will be assigned to RPI, which shows equal popularity for a particular query and its benchmark query. The comparison will remove a country from its list only if the country SPI is 0 for both of the queries compared. Entries with SPI value 0 for only a query or its benchmark query are considered to be 0.5 for either case. If a query has less than 0.5 value of SPI, then it is indexed at 0 by google trends. To compare two RPIs these 0 values have to be replaced by a value which indicates relative lower than 1 and higher than 0.

3. Training: The data after preprocessing is passed through models described in Fig. 2 for training. For an unseen movie, features D, A1, A2, G1, G2 and G3 can be calculated from google trends. For a movie query in training, the average for the time described in section 3.1.2.1 is taken as overall average.For prediction of totally unseen movie, time is taken as described in section 3.2.2. A lot of movies last longer than expected which can prove our assumption wrong but this is the average case.

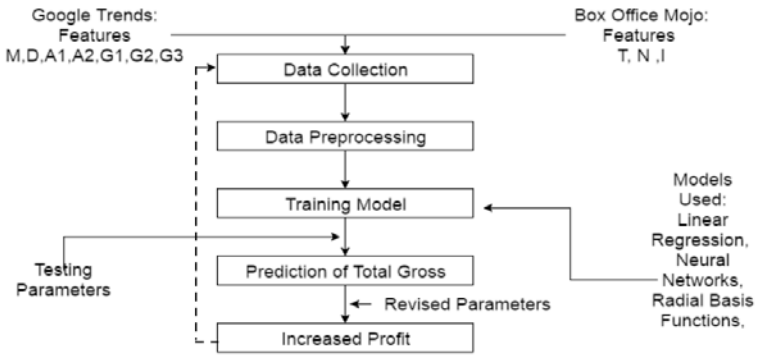

Figure 2. Overall Process from data collection to increasing profit

4. Prediction: Once the trained model is obtained, the testing parameters are set and the prediction for the total gross can me made for a test movie.

5. Increasing Profit with revised Parameters: A movie distributor can revise a parameter and see how much profit a movie will make. Even a small change in the distribution policy and marketing strategy can lead to better results. At this point, targets can be laid for various departments to make the movie an overall commercial success.

3.2.2 Prediction Process and Feature Analysis

Movie queries RPI and number of theatres are the key features considered for the analysis, Fig. 3 shows how a movie income(I) is dependent on features M and T. It may happen that a movie was not famous enough but released on wider scale or it became more famous than expected and was released small scale. To handle these kind of cases, the authors introduced a feature (M)*(T) which can handle both type of cases and results in less variance from the predicted function which results in lower errors.

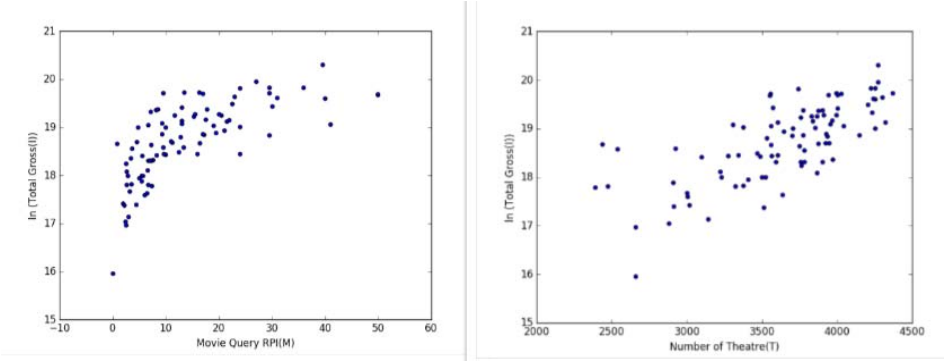

Figure 3. (a)Movie Queries RPI(M)(US) plotted against total income(I).(b)Number of theatres(T) plotted against the same.

Fig. 1a shows performance of a movie after release, general trend of which is: Highest peak on the day of release or the next day, followed by spikes of lower heights at every week around weekend. Our model is trained on movie queries taken after release also,so for the unseen data, we

will predict next 5 weeks from given average and then apply it to our model to predict total income. We will take an average case for each unseen movie which will return to 0 after 5 weeks of the release, with peaks on each friday and then an RPI over whole time period will be taken to predict income(I).

3.2.3 Prediction Methods and Results

To predict the function fitting the Income (I) with given features, we have used different models. We have used Linear Regression, Neural Networks (Multilayer Perceptron) and Radial Basis Function Neural Networks to predict the function. For US, UK, Germany, France and United Kingdom, for each model, we calculated mean squared error(MSE) and absolute mean error(AME). Prediction will be done by taking exponent of the predicted model output. Here are the highlights of the results. Table 1 shows 3 results when M, T and M*T are the key features for the predictions, as compared in table 1.(i,ii). Table 1.(iii) is for different set of test cases from table1.(i), all the other countries have higher errors compared to US. Tolerance for US is around 0.05 and for other countries around 0.20. AME Training error for US is around 0.30 which corresponds to 34% error in the prediction. Taking tolerance into consideration, error for US ranges from 28% to 37%. These errors are much lower compared to Lee[2].

USA	Linear Regression Using	3 features(M,t,M*T)
	Train	Test
MSE	0.191254443	0.3463884
AMS	0.340090706	0.43000679

(i)

USA	Linear Regression Using	All features
	Train	Test
MSE	0.182095835	0.33996533
AMS	0.329346398	0.43054713

(ii)

USA	Linear Regression Using	All features
	Train	Test
MSE	0.211020368311	0.289290594534
AMS	0.349528782352	0.357498925699

(iii)

Germany	Linear Regression Using	All features
	Train	Test
MSE	0.541020368311	0.137126520455
AMS	0.691743761608	0.279214274821

(iv)

Table 1. (i,ii,iii,iv) MSE and AME values for example models

4. DISCUSSION AND FINDINGS

In this section the findings are discussed and overall analysis is presented to cover the contributions of the current research work.

4.1 Performance of RPI vs SPI

Fig. 4 clearly shows that an RPI can match the pattern and predict the income much better than compared to SPI from Lee [2]. Here datasets are different, but overall accuracy of RPI model is much higher than SPI.

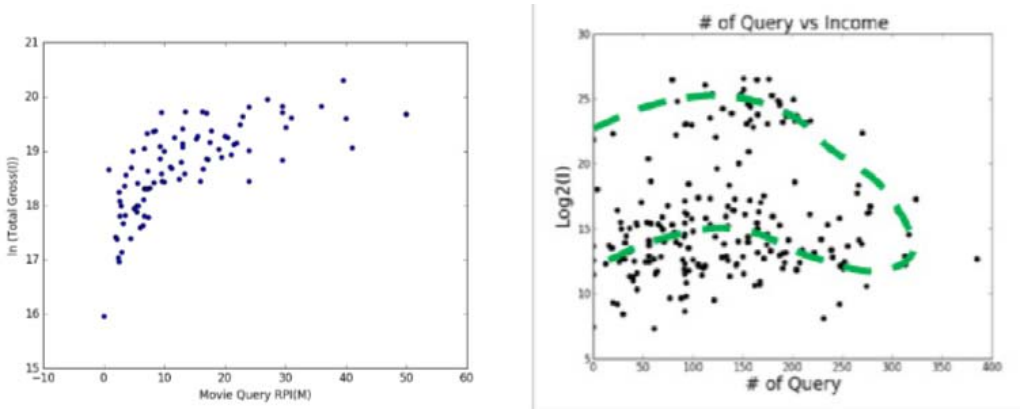

Figure 4. (a) Movie Queries RPI plotted with Income(I)(US) (b) Movie queries SPI plotted with Income from Lee [2].

4.2 Performance of RPI for different Benchmark Queries

Fig. 5 shows how a benchmark query can affect the movie income(I). A benchmark query with higher avg. query shifts the data towards lower values, increasing slope for a linear regression predicted line, which will result in uncertainty in predictions. Fig. 5(a) has some anomalies which may have generated due to a large frequency change in a benchmark query for a certain time. Nearly equal values of frequencies of a query with its benchmark can lead to abnormal behaviours.

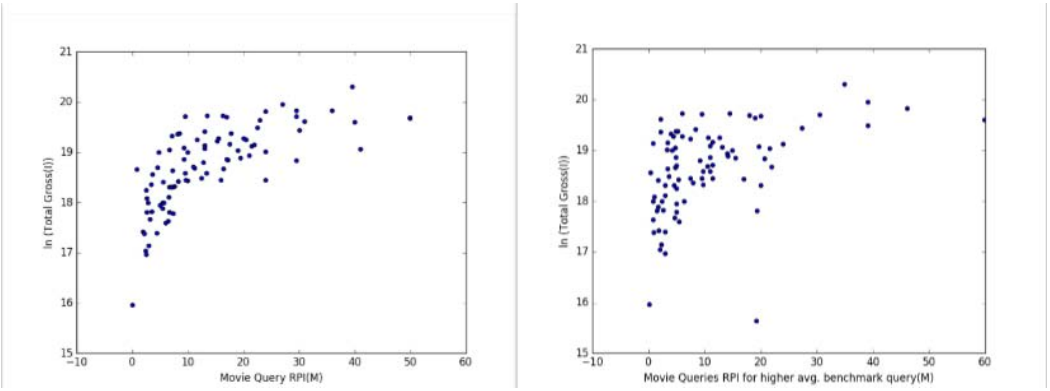

Figure 5. (a) Movie queries(M) vs income(I)(US) for benchmark query *The Shawshank Redemption* which has been used for final results.(b) Same plot for a different benchmark query which has slightly higher volume of google queries for the same time.

The features covered in the current research work resulted into varied impact for different countries. Highest weightage value for a movie queries was in Australia whereas highest director weightage value was for USA. So we can comfortably say that people in USA are more involved in who has directed a movie, same with actor #2 also. It also has highest weightage in the USA.

Weights assigned to each country will be different, according to distribution costs, a movie distributor can vary number of theatres for a country such that it can make the highest profit from given amount of production cost. Fig. 6 compares 3 countries by how much more a movie can make if we increase 100 theatres or screens for a country. The movie will make 1.21 times the money it is making now, if it is released on another 100 screens in Australia. Remember it is a ratio, so we have to consider absolute values to see how much increased profit can be gained and

comparing it with cost of increasing screens for different countries, profit can be maximized accordingly. At the time of release, feature M (showing popularity) and feature T affect the income heavily. So a movie producer should think of increasing marketing in the country or lower the screens for the country and invest that money in increasing screens in a country which is more popular according to the predicted results.

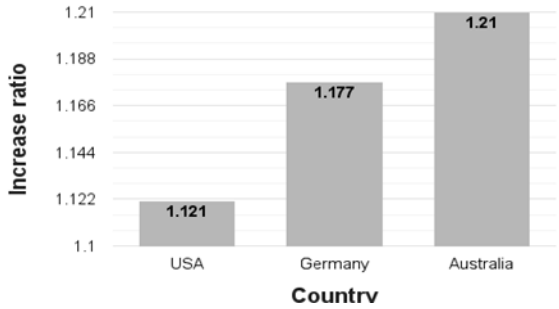

Figure 6. Chart showing increase in income after increasing 100 screens in a particular country

5. CONCLUSION

The authors have proposed a systematic framework - analytic process and technique by which different key stakeholders of the movie will benefit at large. For example, a movie distributor will be able to decide the distribution of a movie around the world. The agenda is to achieve maximum profit for given movie release. The proposed Relative Popularity Index (RPI) allow one to compare the popularity of two different queries using the trends data. RPI gives proper insight to the volume of queries. Throughout the experiment, new parameters have been designed and utilized to predict box office success of a movie. In future research work, authors would like to target the inclusion of the marketing, branding and distribution strategies for the prediction of the commercial success of a movie.

References

[1] Deniz Demir, Olga Kapralova, Hongze Lai, "Predicting IMDB Movie Ratings Using Google Trends," Dept.Elect.Eng, Stanford Univ., California, December, 2012

[2] Chanseung Lee1 and Mina Jung2 Prediction movie success from search query using support vector regression methods.

[3] Nithin VR, Pranav M, Sarath Babu PB, Predicting Movie Success Based on IMDB Data International Journal of Data Mining Techniques and Applications. ISSN: 2278-2419

[4] Choi, H., Varian, H., (2012). Predicting the present with google trends. Economic Record, 88 (s1), 2–9

[5] Mahesh Joshi, Dipanjan Das, Kevin Gimpel, Noah A. Smith, Movie Reviews and Revenues: An Experiment in Text Regression.

[6] Ramesh Sharda, Dursun Delen, Predicting box-office success of motion pictures with neural networks.

[7] Box Office Mojo, 2016. Web. 25 Dec. 2016. <http://www.boxofficemojo.com/>

[8] Google trends, 2016. Web. 25 Dec. 2016. <https://www.google.co.in/trends/>

[9] The Internet Movie Database. IMDb.com, Inc, 2016. Web. 25 Dec. 2016. <http://www.imdb.com/>.

[10] Darin Im, Minh Thao, Dang Nguyen, "Predicting Movie Success in the U.S. market," Dept.Elect.Eng, Stanford Univ., California, December, 2011

[11] Marton Mestyan, Taha Yasseri, and Janos Kertesz. “Early prediction of movie box office success based on wikipedia activity” PLoS ONE, 8(8), 2013.

[12] Jeremy Ginsberg et al. “Detecting influenza epidemics using search engine query data” Nature, 45(9), 2009

[13] Chanseung Lee and Mina Jung “Predicting Movie Incomes Using Search Engine Query Data”International Conference on Artificial Intelligence and Pattern Recognition (AIPR), 2014

[14] Preis T, Reith D, Stanley HE (2010) Complex dynamics of our economic life on different scales: insights from search engine query data. Philosophical Transactions of The Royal Society A 368: 5707–5719.

[15] Preis T, Moat HS, Stanley HE, Bishop SR (2012) Quantifying the advantage of looking forward. Sci Rep 2: 350.

[16] Harsh Taneja,Anupam Dewan ,Vineet Bhardhwaj "PREDICTING BOX-OFFICE SUCCESS OF MOVIES IN THE U.S. MARKET" International Journel of Science and Research. Volume 5 Issue 10, October 2016,ISSN (Online): 2319-7064

[17] Prithvi Sharan S , "Movie Success Predictor" Indian Journel of Applied Research, Volume 6,Issue 6 ,June 2016, ISSN- 2249-555X

[18] Mishne and N. Glance. 2006. Predicting movie sales from blogger sentiment. In AAAI Spring Symposium on Computational Approaches to Analysing Weblogs

[19] Zhang and S. Skiena. 2009. Improving movie gross prediction through news analysis. In *Proc. of Web Intelligence and Intelligent Agent Technology*.