MOVIE SUCCESS PREDICTION SYSTEM

PROJECT

Submitted in fulfilment for the JComponent of ITE2006 –

Data Mining Techniques

CAL COURSE

in

B. Tech – Information Technology

by

JEBARAJ M (15BIT0354)

Under the guidance of

Prof. Thippa Reddy

SITE

School of Information Technology and Engineering

Winter Semester 2016-17

Movie Success Prediction using Data Mining

Abstract:

In this paper I developed a mathematical model for predicting the success class such as flop, hit, super hit of the movies. For doing this I have to develop a methodology in which the historical data of each component such as actor, director, writer and producer that influences the success or failure of a movie is given is due to weightage and then based on multiple thresholds calculated on the basis of descriptive statistics of dataset of each component it is given class flop, hit, super hit label. Based on the weightage of historical data of each film crew the movie will be labeled as super hit, hit or flop. This system helps to find out whether the movie is super hit, hit, flop on the basis of historical data of actor, actress, music director, writer, director, marketing budget and release date of the new movie. The factors such as actor, director, writer, producer and marketing budget historical data of each component are calculated and movie success is predicted. This mathematical model helps to find out the review of the new movie. Due to this model, user can easily decide whether to book ticket in advance or not.

Keywords: *data mining, classification, IMDb, movies, films, hierarchical, class label.*

1. Introduction:

The Indian movie industry produces the maximum number of movies per year at 1000/year, higher than any other country's movie industry. The IMDb is an excellent resource to find detailed information about almost any film ever made. It contains a vast amount of data, which undoubtedly contains much valuable information about general trends in films. Data mining techniques allows us to predict the success of a future film given select information about the film before its release. So we will get the data from IMDb and predict the movie success by applying the mathematical concept. The data in the IMDb database has been distributed in the form of several text files with the data contained in them in a variety of formats. The main challenge is to collect them all and apply cleaning and integration on them to obtain the relevant data. Once this is done, we will use Data Mining Tools(WEKA) to attempt to classify the general rating of upcoming films based on the historical data collected.

2. Literature Review:

* M. Saraee et al.,2004 [13] proposed a method of selectively choosing the relevant data sets from IMDb over a total of 49 files. Java application has been used for cleaning and integration of the files and then storing them in databases. Finally, Microsoft Envisioner and Microsoft Excel has been used for the actual data mining process.

* Philipp Omenitsch [14] has used simple Machine Learning techniques including Twitter Sentiment Analysis to generate an overall review based on interactive inputs from others.

* This paper [2] details the analysis of the Internet Movie Database (IMDb), a free, user-maintained, online resource of production details for over 390,000 movies, television series and video games, which contains information such as title, genre, box-office taking, cast credits and user's ratings.

* Cook, K., Grinstei et al. [9] has used Twitter Sentiment analysis to predict Box-Office outcomes.

* Al-Masoudi [8] et al. has analyzed interactively selects training data and prediction weights based on appropriate analytic visualizations.

* Fazzion, E [4] has implemented movie prediction using Machine Learning and Visual Analytics.

3. Methodology:

The required data for Preprocessing (using WEKA Tool) has been taken from IMDb. Moving on, to further decide the importance of a crew member in deciding the success of a film, we determined the hierarchical importance amongst them. Now to calculate the output class label, we assigned different weightage to each crew member obtained from classification (using WEKA Tool).

Our next issue was to calculate how much weightage each crew member deserved and assigning it to them, but once that was resolved we only had to predict the class label based on the historical data of each crew member. We will be using the following criteria for predicting the class label,

$7.5 <= x < 10$	Super Hit
$5 <= x < 7.5$	Hit
$1 <= x < 5$	Flop

4. Results & Discussion:

After pre-processing I found out the following hierarchical contribution to a movie's success,

Writer>Director>Producer>Actor>Others

Film Crew	**Super-Hit Possibility**
Writer	67%
Director	56%
Producer	36%
Actor	25%

Now I found out the individual weightage of each of the crew member as follows ,

Writer	(0.67/1.84)*10 = 3.6
Director	(0.56/1.84)*10 = 3.0
Producer	(0.36/1.84)*10 = 2.0
Actor	(0.25/1.84)*10 = 1.4

Finally I got the final result as ,

Fim Crews	**Max**	**Super Hit**	**Hit**	**Flop**
Writer	3.60	3.24	2.67	0.90
Director	3.00	2.70	2.22	0.75
Producer	2.00	1.80	1.48	0.50
Actor	1.40	1.26	1.04	0.35
Total	10	9	7.4	2.5

A sample of our data with our findings is as follows,

Movie	**Writer**	**Director**	**Producer**	**Actor**	**IMDB Class Label**	**Our Class Label**
Kavalan	Siddique	Siddique	Ramesh Babu	Vijay	Hit	Hit
Thuppaki	A.R. Murugadoss	A.R. Murugadoss	Thanu	Vijay	Super Hit	Super Hit
Nanban	Shankar	Shankar	Shankar	Vijay	Super Hit	Super Hit
Their	Atlee	Atlee	Thanu	Vijay	Hit	Hit
Villu	Sachin Bhowmick	Prabhudheva	Ajay	Vijay	Flop	Flop
Pokkiri	Puri Jagannadh	Prabhudheva	S. Ramesh Babu	Vijay	Hit	Hit
The Bodyguard	Lawrence Kasdan	Mick Jackson	Kevin Costner	Vijay	Hit	Hit
Enthiran	Shankar	Shankar	Kalanidhi Maran	Rajinikanth	Hit	Hit
Sivaji	Shankar	Shankar	Saravanan M	Rajinikanth	Super Hit(7.5)	Super Hit(8.2)
Kabali	Ranjith	Ranjith	Thanu	Rajinikanth	Hit(6.6)	Super Hit(8.2)
Lingaa	Ravi Kumar	Ravi Kumar	Rockline Venkatesh	Rajinikanth	Hit(5.9)	Hit(7.3)
Kuselan	Sreenivasan	P.Vasu	Pushpa Kandasamy	Rajinikanth	Flop(4.7)	Flop(4.9)
Enthiran 2	Shankar	Shankar	Karunamoorthy	Rajinikanth	Still not Released	Super Hit(8.1)
Chandramukhi	P. Vasu	P. Vasu	Prabhu	Rajinikanth	Hit	Hit
Vedalam	Siva	SIva	Aishwarya	Ajith Kumar	Hit(6.3)	Hit(6.6)
Veeram	Jayaraaj	Jayaraaj	Bharathi Reddy	Ajith Kumar	Hit(7.3)	Hit(6.6)
Mankatha	Venkat Prabhu	Venkat Prabhu	Dhayanidhi Alagiri	Ajith Kumar	Super Hit(7.6)	Super Hit(7.5)
Aasal	Yugi Sethu	Saran	Prabhu	Ajith Kumar	Flop(4.0)	Flop(3.9)
Arrambam	Balakrishnan	Vishnuvardhan	A.M. Rathnam	Ajith Kumar	Hit	Hit

5. Conclusion:

The mathematical model that I have derived and implemented enabled us to calculate the hierarchical importance of each of the crew member in determining the success of a given movie based on the historical data. The results of our experiment are in accordance with the IMDb ratings which are accurate as far as the class labels are concerned with the only exception being those cases wherein historical data for a crew member is missing. Our mathematical model gave 100% accuracy.

6. References:

[1] Han, J., Kamber, M., Data Mining Concepts and Techniques, Morgan Kaufmann Publishers: San Francisco, 2001.

[2] Neurosoft S.A., Neurosoft Envisioner, www.neurosoft.gr/products/ envi.asp, 1999.

[3] Thearling, K., Data Mining and Analytic Technologies, www.thearling.com, 2004.

[4] Hamilton et al., Knowledge Discovery in Databases, www2.cs. uregina.ca/~hamilton/courses/831/, 2002.

[5] Attar Software Ltd, Active Data Mining Solutions,www.attar.com /tutor/deploy.htm, 2004.

[6] Labovitz, M. L., What Is Data Mining and What Are Its Uses?, www.darwinmag.com/read/100103/mining.html, 2003.

[7] Nautilus Systems Inc, The Data Mining Process, www.nautilussystems.com/process.html, 1996.

[8] Al-Masoudi, F., Seebacher, D., and Schreiner, M. (2013). Similaritydriven visual-interactive prediction of movie ratings and box office results. http://bib.dbvis.de/uploadedFiles/almasoudi.pdf.

[9] Cook, K., Grinstein, G., and Whiting, M. (2013). VAST Mini Challenge 1: Visualize the box office. http://boxofficevast.org/.

[10] El Assady, M., Hafner, D., Hund, M., Jäger, A., Jentner, W., Rohrdantz, C., Fischer, F., Simon, S., Schreck, T., and Keim, D. A. (2013). Visual analytics

for the prediction of movie rating and box office performance

[11] Fazzion, E., Las Casas, P., Gonçalves, G., Melo-Minardi, R., and Meira Jr, W. (2013). Open Weekend and Rating Prediction Based on Visualization Techniques.

[12] Jäger, A., Hafner, D., and el Assady, M. (2013). Moovis - a visual analytics tool for the prediction of movie viewer ratings and boxoffice.

[13] A data mining approach to analysis and prediction of movie ratings(2004) M. Saraee, S. White & J. Eccleston

[14] http://www.imdb.org/

[15] Mat Kelly, Michael L. Nelson, M. C. W. (2013). Graph-Based Navigation of a Box Office Prediction System

[16] Predicting Movie Success with Machine Learning and Visual Analytics , Philipp Omenitsch(2014)