Summer 6-26-2019

Prediction of Broadcast Volume and Analysis of Influencing Factors About Online Videos

Tong Shao
School of Economics and Management, Tongji University, Shanghai, 200092, China

Ruodan Xie
School of Economics and Management, Tongji University, Shanghai, 200092, China

Prediction of Broadcast Volume and Analysis of Influencing Factors About Online Videos

Tong Shao[1], *Ruodan Xie*[2*]
School of Economics and Management, Tongji University, Shanghai, 200092, China

Abstract: With the development of the Internet and the continuous expansion of the online video market scale, it is increasingly important to accurately predict broadcast volume before the launch of videos. On the one hand, it can provide investors and producers with recommendation scheme for video shooting. On the other hand, it can fully understand users' preferences and find videos which potentially to be popular. In allusion to problems of low predictive accuracy and lack of practical application value in video prediction research, this paper, based on the related data of online video website TED, obtains a prediction model with high-precision broadcast volume through feature selection and model fusion. Further, it analyzes the impact of video themes, the number of languages and official events, which can provide some reference for investors and producers of different scales before the videos go online.

Keywords: broadcast volume, data mining, TED

1. INTRODUCTION

The prediction of broadcast volume is a kind of estimation of the unreleased video, which can provide valuable reference for the investors and producers of video. Along with the development of the Internet, online video market scale is increasing fast. Large network video services system will have tens of millions of daily broadcast volume. For example, the total broadcast volume of Facebook reaches 3 billion times in a day, the video uploaded by YouTube video website is about 300 hours long every minute, and Tencent video generates nearly 1 billion volume by several million of users every day. With the explosive growth of video data, accurate prediction of video broadcast volume is becoming more and more important.

For service providers, the popularity of videos and prediction of broadcast volume are of great help to their future content filtering, video ranking and the design of recommendation system, which would help users find videos with more potential value more easily[1]. For advertisers in online marketing, predicting the next rising Internet star precisely will maximize revenue through better advertising[2]. Meanwhile, network operators can also take the initiative to manage bandwidth requirements and deploy cache servers in the popular video content distribution network in advance[3].

Online video website TED is a nonprofit organization dedicated to spreading ideas, which often in the form of short and powerful talks. TED was started in 1984 as a networking event for the elite, which originally opened to 1,000 people and was expensive. In order to spread great ideas more broadly, TED began offering free online TED talks in 2006. The TED talk brings technology, entertainment and design together, covering almost every topic from science, business to global issues in more than 100 languages. Meanwhile, TED also hosts independent TEDx events to help share ideas with communities worldwide.

From the global TEDx community to the TED-ed series, all of them are driven by the goal: how can we best spread great ideas? How do we maximize our influence? To be more precise, an important measure of influence is the broadcast volume. Accurate prediction of broadcast volume of online video of can help improve service efficiency, enhance video quality and enrich video resources. It can effectively improve user experience, as

* Corresponding author. Email: philzhou@tongji.edu.cn (Zhongyun Zhou)

platform can reasonably schedule and combine videos to achieve higher communication effectiveness.

2. LITERATURE REVIEW

Previous literature has lots of elaboration on the prediction of movie box office. For movies, box office can measure the influence and popularity of a film, while for many online free videos, the broadcast volume can be compared to the box office to a certain extent, which represents its influence.

At present, there are two main prediction methods for movie box office, one is the prediction based on movie content, and the other is the prediction based on user-generated content. User-generated content can be further divided into research based on user feedback and related social media communication.

2.1 Prediction based on movie content

Barry Litman[4], a pioneer in the prediction of movie box office revenue, evaluated 697 films screened in the United States in the 1980s, replaced the film box office with the rental income. He used the regression analysis method to obtain the three major factors influencing the box office: creativity, release and marketing ability. Although Litman's model can enlighten the thinking of subsequent researchers, its actual prediction accuracy may not be ideal. Sharda[5], who took the lead in proposing a new prediction method in big data era, regarded the prediction problem as a classification problem, used the neural network method to divide movies into different categories, but the prediction accuracy of this model is not very ideal. L Zhang, J Luo, S Yang[6] in 2008, Zheng Jian, Zhou Shangbo[7] in 2014 and Li Yi and Wang Xiaofeng[8] in 2018 all made predictions based on meta data related to films. One of their common focuses is to improve the prediction accuracy. The realistic significance of prediction models with high accuracy is very strong. Therefore, this paper carries out feature processing and model fusion to achieve higher accuracy.

2.2 Prediction based on user-generated content

User-generated content includes research based on user feedback and related media communications. The former is to add user feedback for box office forecast. In the blog box office prediction model built by Gilad Mishne and Natalie Glance, they compared the quantity of word of mouth and the tendency of emotional evaluation of word-of-mouth in box office prediction. This study adopted the opening week date, the opening week box office and the opening screen data of each film from IMDB. As a result, the highest correlation coefficient reaches 0.614[9]. While in the news reports of box office forecasting model, put forward by Wenbin Zhang and Steven Skiena in 2009, they used the film's budget, the number of screens, the box office of first week and others to forecast the box office with K-NN model, further improving the prediction accuracy[10].

In addition, film reviews are also a major indicator closely related to the box office. Timothy King studied the relationship between the score and the total box office of the United States films in 2003 and found that the score of films on more than 1000 screens was positively correlated with the total box office[11]. Suman Basuroy systematically revealed how film reviews would affect the box office, and studied the influence and predictive power of reviews of film[12]. Peter Boatwright[13], Anindita Chakravarty[14] and others respectively revealed the influence of different film critics on the film's box office and the film's audiences.

The research on related media communication focuses on the correlation between user behavior and box office. Li. H studied the influence of video's inherent attraction and potential transmission, and predicted the video heat trend based on the previous broadcast volume and potential friend sharing rate[15]. Asur reflected users' interest in video by using the heat of users' community themes on social media Twitter and predicted the potential correlation between social themes and video heat[16]. Szabo[17] found a similar pattern with Asur. He used Digg and YouTube data to conduct quantitative processing on user behavior in different periods and

accurately predicted the trend of video broadcast volume over time. A great characteristic of box office forecasting which can be found from the literature review is that the predictions are based on the user generated data, such as the opening week box office, grading, user emotions and so on. A large part of the prediction often occurs after the release of the film. At this time, the film has become a finished product. In other words, for investors and producers, the scope of adjustment is often limited in marketing, film scheduling etc. As a result, it is still a valuable question how to accurately predict broadcast volume before its release.

Based on above research, we dedicate to use the characteristics and information of online video themselves to more accurately predict the broadcast volume of video, to analyze the important characteristics affecting the broadcast volume, and provide better suggestions for investors to invest in video and producers to choose films.

3. DATA UNDERSTANDING

The source of dataset is a dataset of TED released by Kaggle, including 2,550 pieces of data by 21st Sep. 2019 from the official website of TED with recording and producing information of a TED talk.

From the descriptive statistics of histogram, we find that the normality of duration of speech and broadcast volume is bad. Also, there are partial missing values such as broadcast volume, occupation of speaker. It is also shown in the boxplot that variables including duration of speech, number of languages, type of TED event, main speaker and occupation of speaker have strong correlation with broadcast volume. Partial charts are as follows.

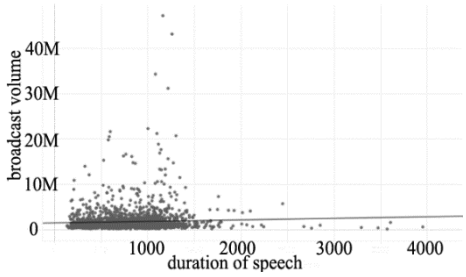

Figure 1. Scatter plot of duration of speech

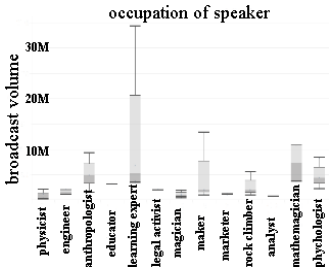

Figure 2. Partial box chart of occupation of speaker

4. DATA PREPARATION

4.1 Data Cleaning

A total of 10 rows with missing values were found after recounting, including 7 rows from occupation of speaker, 1 row from type tag and 3 rows from broadcast volume. We deleted the data with missing value of occupation of speaker and type tag as they are textual variables and far-fetched to replace the blank with mode. As for the broadcast volume, we filled the blank with mean value.

Then we deleted the outliers of each column. The types of outliers are mainly as follows:

- In the pre-process of rating of speech, it was found that some values didn't correspond to the attributes resulting from dislocation.
- After the conversion of filming date according to the UNIX time rule, there exist some abnormal values which are earlier than the foundation of TED.
- Irregular values (e.g. 'health', 'A3C', etc.) were found in the column of broadcast volume.

8 rows containing outliers were deleted due to outliers, and 8 rows were deleted due to missing value. After deletion, the remaining sample contains 2,534 rows of data.

4.2 Data Transformation

4.2.1 Standardization

There are three types of numerical variables that can be directedly standardized: duration of speech, number

of languages and broadcast volume. We normalized the three variables to be within the threshold value of 0 to 1.

$$x = \frac{(x-min)}{(max-min)} \quad (1)$$

4.2.2 Discretization

The relative values of 1,2,3,4 are determined according to the mean value of broadcast volume corresponding to each categorical variable. TED event, occupation of speaker and main speaker are standardized in this way.

The filming and publishing date are represented by UNIX timestamp. We use the datetime library to convert them into a more readable and calculable form. Following one-hot coding is carried out, which means the normality transformation can be avoided after one-hot coding.

4.2.3 Normality Transformation

The transformation formula should be determined according to distribution shape of variables. If skewness is 2-3 times of its standard deviation, square root value can be considered. Or the skewness is more than 3 times of the standard deviation, using logarithm is more favorable. We applied this principle to all numerical variables.

4.2.4 Construction of New Variables

- Preparation time (pretime)

Preparation time=Publishing time – Filming time. It is speculated that the longer the preparation time is, the higher quality of TED is, which may symbolize the higher broadcast volume.

- Publishing time and filming time in month form

We transformed the two variables into month form and classified data by month. From the distribution of TED talks in different months, it is said that number of talks in February is most, while January and August are few, which suggests the publishing time of speech may influence the broadcast volume of TED talk.

Then we generated 24 new variables according to month, respectively 12 months of publishing date and 12 months of filming date processed by one-hot coding, representing whether filmed in x month and whether published in x month. (flim1, flim2…flim12; publish 1, publish 2…publish12)

- Dismantling of related talk (ave_rt_duraion, ave_rt_broadcast volume)

Related talk, a textual variable, includes id, url, main speaker, title, duraion and broadcast volume of 6 related talks. We extracted duration and broadcast volume of 6 related talks and got mean value, deriving 2 new attributes, namely average duration of related talks and average broadcast volume of related talks.

- Transformation of type tag

Each talk has 6 type tags and the number of different type tag amounts to 569. It is hypothesized that the type tags of TED talk with a broadcast volume more than 10 million reflected the social hot spots and audience's preferences. Thus, we assigned high weights to rows which contained these tags and add it up to generate a new variable. The partial weight distribution of type tag with broadcast volume more than 10 million is as follows:

Table 1. The partial weight distribution of type tag

Type tag	Weight	Type tag	Weight	Type tag	Weight
'psychology'	9	'culture'	4	'mind'	2
'science'	7	'health'	3	'poetry'	2
'culture'	5	'humor'	3	'productivity'	2
'technology'	5	'motivation'	3	'relationships'	2
'work'	5	'social change'	3	'self'	2
'TEDx'	5	'storytelling'	3	'sex'	2
'entertainment'	4	'choice'	2	'society'	2
'happiness'	4	'depression'	2	'time'	2

The new variable: type_tags_zong, which value is added up by 6 original type tags, representing the hot degree of type of this TED talk. Normalization was conducted.

4.3 Data Reduction

4.3.1 Correlation Analysis

Table 2. Thermal force diagram about broadcast volume

Variable	Correlation	Variable	Correlation	Variable	Correlation
ave_rt_viewcount	0.26	film10	-0.025	publish6	0.015
ave_rt_duration	-0.019	film11	-0.029	publish7	-0.062
film1	-0.023	film12	-0.015	publish8	-0.02
film2	0.06	pretime	0.046	publish9	-0.012
film3	0.016	duration	0.05	publish10	-0.0047
film4	-0.047	languages	0.38	publish11	-0.017
film5	-0.0083	publish1	-0.015	publish12	0.0028
film6	0.004	publish2	0.0098	type_tags_zong	0.16
film7	0.018	publish3	0.036	event	0.26
film8	-0.017	publish4	-0.0062	main_speaker	0.51
film9	0.0014	publish5	0.033	speak_occupation	0.44

As is shown in the thermal force diagram about broadcast volume, the variables highly correlated to broadcast volume are main speaker(0.51), occupation of speaker(0.44), number of languages(0.38), average broadcast volume of related talks(0.26), type of TED event(0.26) and type tag(0.16). It is not surprising about number of languages, as it is logically assumed that the more the number of languages is, the larger number of the target audience is. Also, the higher broadcast volume of related talk could lead more people to view relevant video. As for main speaker and occupation of speaker, it is understandable that these two variables have great impacts on the choice of audience.

4.3.2 Method of Feature Selection

Comparing three methods, respectively stability selection, RFE based on logistic regression and RFE based on ridge regression (RFE: Recursive Feature Elimination), we chose stability selection.

Stability selection is a method based on quadratic sampling and selection algorithm. Selection algorithm can be regression, SVM or other similar methods. The principle is to run the feature selection algorithm on different subsets of data and features repeatedly, and finally summarize the result of feature selection such as the frequency of a feature being selected as important feature. Ideally, the score for important feature would be close to 1. Correspondingly, the most irrelevant feature would get zero.

The main idea of RFE is to build a model (such as SVM or regression) and eliminate the best feature, then repeat the process on the remaining features until all the features are traversal. The order in which features are eliminated is the order of significance. Thus, it is a greedy algorithm to find the optimal subset of features.

Running stability selection in python, we got the stability score for selected 33 features as follows:

Table 3. Stability score of 33 features

Variable	Score	Variable	Score	Variable	Score	Variable	Score
duration	1	main_speaker_E	1	film5	0.985	publish9	0.885
languages	1	speaker__occupation_E	1	film3	0.98	film12	0.875
pretime	1	ave_rt_duration	1	film6	0.98	Publish6	0.87
flim1	1	ave_rt_viewcount	1	publish3	0.98	publish1	0.86
film4	1	film7	0.995	publish7	0.965	publish10	0.845
publish11	1	film8	0.995	publish5	0.96	publish2	0.77
type_tags_zong	1	film9	0.995	publish12	0.96	film2	0.75
event_E	1	publish4	0.995	publish8	0.91	film11	0.745
film10	0.7						

According to the results, the stability score of all variables are relatively high, with no variables close to 0,

indicating that they are of much importance and we should keep them into the next step of model construction.

5. MODEL CONSTRUCTION

5.1 Selection of Basic Model

We imported 11 models respectively and selected optimal model by MSE and standard deviation of MSE, using 5-fold cross validation.

Table 4. Test result of different model

Model	MSE	SD	Model	MSE	SD
Ridge	0.051049	0.0089	LinearSVR	0.053269	0.0085
Lasso	0.063639	0.0181	SGDRegressor	0.056887	0.0131
RandomForestRegressor	0.040762	0.0087	BayesianRidge	0.051125	0.0124
GradientBoostingRegressor	0.036194	0.0093	KernelRidge	0.050667	0.0088
SVR	0.055003	0.0085	ExtraTreesRegressor	0.040026	0.0094

We selected RandomForestRegressor, GradientBoostingRegressor and ExtraTreesRegressor to integrate.

5.1 Parameter Adjustment

Since MSE of GradientBoostingRegressor was superior, it was used as the basic model and its parameters were adjusted. Finally, the optimal parameters were got and we used GBR (n_estimators=60, max_depth=3, min_samples_split=100).

5.2 Model Integration

We had tried different weight of the three models above and chose the relative optimal model, the respective weight of GradientBoostingRegressor and ExtraTreesRegressor was 0.8 and 0.2. The MSE of this basic integrated model is 0.036011, which is the baseline of following analysis.

5.3 Sensitivity Analysis

Firstly, we deleted the variable with worst performance according to stability selection, namely whether filmed in October. The MSE is 0.036034>0.036011, indicating the model is worse. Then we deleted the two worst variables, namely whether filmed in October and whether filmed in November. The MSE is 0.036071>0.036011, signifying the model is worse. To summarize, the selected 33 variables are superior enough to predict the broadcast volume and shouldn't be eliminated.

6. TESTING AND EVALUATION

6.1 Analysis of result

According to the result of final model, MSE is around 0.036. Qianqian Lv used MSE in the prediction of music trend based on machine learning [18] and his MSE was 0.037. Also, Ling Zheng got MSE of 0.077 in the research of prediction model of film box-office based on neural network[19]. Referred to the previous MSE of others, the accuracy of our model can be considered to reach optimum level. It is worth noting that we rule out variables generated by viewers after the releasing of video in our model, which means our model can accurately predict broadcast volume before the releasing and is of great reference significance for investors and producers.

6.2 Analysis of the relevant characteristics of broadcast volume

From the result analysis of highly correlated features, the features highly correlated with the broadcast volume are the number of languages, type of video, main speaker, occupation of speaker, TED event, average broadcast volume and duration of related videos.

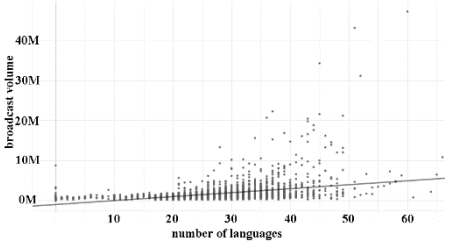

Figure 3. Chart of languages and broadcast volume

First, the influence of number of languages on the broadcast volume is more intuitive. The more languages

there are, the more people can understand the video content, and the broadcast volume will gradually increase.

Secondly, On the one hand, the high correlation between extracted tags and broadcast volume demonstrates the effectiveness of this disassembly and dimensionality reduction method; on the other hand, it also shows the hot spots and topics of different attention. We find that topics with high broadcast volume include psychology, science, culture, society and so on.

Thirdly, TED event, namely video activity has an impact on broadcast volume. We find that activities with the highest TED broadcast volume are TED×Houston, TED×Bloomington, TED Talks Education, Stanford University and TED×MET. It shows that the characteristic activities held by some characteristic universities and cities are easy to attract the attention of the audience.

Finally, the broadcast and duration of the relevant videos influence the broadcast volume. Relevant hot speeches, often can result in more users to watching related speeches. According to the scatter plot analysis, the higher the related video broadcast volume is, the higher broadcast volume is. Rather, there is a non-linear relationship between the duration of relevant video and broadcast volume.

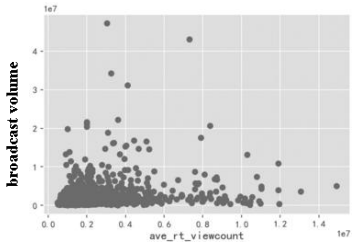

Figure 4. Scatter plot of related video viewing and self-viewing

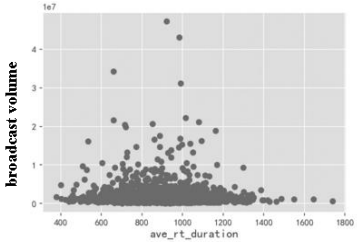

Figure 5. Scatter plot of related video duration and self-viewing

6.3 Expected contribution and future research

For TED investors, a higher number of viewers represents a higher impact and exposure, which is conducive to the promotion of brand awareness. Investors can use our prediction model as a basis to better predict broadcast volume, combining scientific and reasonable evaluation system to decide whether to invest in TED and the amount of investment.

For TED producers, they can better arrange a series of procedures of video selection, production, video arrangement and promotion, launching more popular video with high broadcast volume. Further, for similar speech sites, if it is a new and less well-known producers, the manager can focus on the theme of the video to be produced and the link between the video and the related video. Specifically, selecting the themes with high volume and using existing hot videos to drive new videos to increase the video broadcast volume and website popularity help to reduce costs under certain promotional effects compared with increasing the number of languages and organizing special events. If the producers already have a considerable scale, they can focus on increasing the number of languages and organizing special events. This may have a huge effect on expanding the audience, forming a video brand and further enhancing the influence of the website.

In order to better predict the broadcast volume before going online, we only consider the characteristics of the video itself. With the video going online, user evaluation and the performance of competitors in the same period will have a further impact on the broadcast volume, and broadcast volume may show dynamic changes. Therefore, we still need to collate the follow-up data, add time series considerations to further study the factors affecting video broadcast volume, so as to excavate more general rules of broadcast volume change.

ACKNOWLEDGEMENT

This research was supported by associate professor Zhongyun Zhou, Tongji University.

REFERENCES

[1] Chenyu Li, Xueming Li, "Characterizing the service usage of online video sharing system: Uploading vs. playback", Wireless Personal Multimedia Communications 19th International Symposium on, pp. 280-286, 2016.

[2] Yan Wu, Qiujian Lv, Yuanyuan Qiao, Jie Yang, "Linking Virtual Identities across Service Domains: An Online Behavior Modeling Approach", Intelligent Environments (IE) 2017 International Conference on, pp. 122-129, 2017.

[3] Yanglong Sun, Le Xu, Yuliang Tang, Weihua Zhuang, "Traffic Offloading for Online Video Service in Vehicular Networks: A Cooperative Approach", Vehicular Technology IEEE Transactions on, vol. 67, no. 8, pp. 7630-7642, 2018.

[4] Litman B R. Predicting Success of Theatrical Movies An Empirical Study[J]. The Journal of Popular Culture, 1983, 16(4): 159-175.

[5] Sharda R, Delen D. Predicting box-office success of motion pictures with neural networks[J]. Expert Systems with Application, 2006, 30(2): 243-254.

[6] Li Zhang,Jianhua Luo,Suying Yang. Forecasting box office revenue of movies with BP neural network[J]. Expert Systems With Applications,2008,36(3).

[7] Zheng Jian, Zhou Shangbo. Prediction modeling of film box office based on neural network [J]. Computer applications,2014,34(03):742-748. (in Chinese)

[8] Li Yi, Wang Xiaofeng. Research on box office prediction of domestic films based on BP neural network [J]. Modern computer (professional edition),2018(24):16-20. (in Chinese)

[9] Mishne, G., & Glance, N. S. (2006, March). Predicting Movie Sales from Blogger Sentiment. In AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs (pp. 155-158)

[10] Zhang, W., & Skiena, S. (2009, September). Improving movie gross prediction through news analysis. In Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01 (pp. 301-304). IEEE Computer Society.

[11] King, T. (2007).Does film criticism affect box office earnings? Evidence from movies released in the US in 2003. Journal of Cultural Economics, 31(3), 171-186

[12] Basuroy S., Chatterjee S, & Ravid, S. A. (2003). How critical are critical reviews? The box office effects of film critics, star power, and budgets.Journal of Marketing, 67(4), 103-117.

[13] Boatwright, P., Basuroy, S., & Kamakura, W. (2007).Reviewing the reviewers: The impact of individual film critics on box office performance. Quantitative Marketing and Economics, 5(4), 401-425.

[14] Chakravarty, A., Liu, Y., & Mazumdar, T. (2010). The differential effects of online word-of-mouth and critics' reviews on pre-release movie evaluation.Journal of Interactive Marketing, 24(3), 185-197.

[15] Li H, Ma X, Wang F, et al. On popularity prediction of videos shared in online social networks[C]. Acm International Conference on Conference on Information & Knowledge Management. New York: ACM, 2013:169-178.

[16] Asur S, Huberman B A. Predicting the Future with Social Media[J]. Proc of Wiiat, 2010, 7(2):492 - 499.

[17] Szabo G, Huberman B A. Predicting the popularity of online content[J]. Communications of the ACM, 2010, 53(8): 80-88.

[18] Qianqian Lv. Prediction of music trend Based on Machine Learning[D]. Lanzhou University, 2017. (in Chinese)

[19] Ling Zheng. research of predction model of flim box-office based on neural network[D]. Beijing University of Posts and Telecommunications, 2018. (in Chinese)