International Journal of Engineering Research and Science & Technology

IJERST

ISSN : 2319-5991

www.ijerst.com

Email: editor@ijerst.com or editor.ijerst@gmail.com

Int. J. Engg. Res. & Sci. & Tech. 2023

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 11, Issuse.2,April 2023

A novel approach to predict latest trending videos from OVS using MLK (Multivariate Linear Regression)

MRS.T.GANGA BHAVANI , [1],Joga Parameswari [2],Kona Tulya Sree Simla [3],Avirapu Shyam Vineel [4],Kalangi Dinesh Kumar [5],Amrutha Kanchiraju [6]

Abstract:

Predicting the top-N popular videos and their future views for a large batch of newly uploaded videos is of great commercial value to online video services (OVSs). Although many attempts have been made on video popularity prediction, the existing models has a much lower performance in predicting the top-N popular videos than that of the entire video set. The reason for this phenomenon is that most videos in an OVS system are unpopular, so models preferentially learn the popularity trends of unpopular videos to improve their performance on the entire video set. However, in most cases, it is critical to predict the performance on the top-N popular videos which is the focus of this study. The challenge for the task is as follows. First, popular and unpopular videos may have similar early view patterns. Second, prediction models that are overly dependent on early view patterns limit the effects of other

features. To address these challenges, we propose a novel multifactor differential influence (MFDI) prediction model based on multivariate linear regression (MLR). The model is designed to improve the discovery of popular videos and their popularity trends are learnt by enhancing the discriminative power of early patterns for different popularity trends and by optimizing the utilization of multi-source data. We evaluate the proposed model using real-world YouTube data, and extensive experiments have demonstrated the effectiveness of our model.

Keywords – Popularity Prediction, Top-N popular videos, Cross-domain.

I. Introduction

Popularity prediction of online videos, especially the prediction of the top-N popular videos is of great importance to support the development of online video services (OVSs). From the perspective of better user experience, the ability to identify the top-N popular videos is beneficial to video services, such as caching and recommendation. From the perspective of commercialization, identifying the top-N popular videos helps the video service providers to maximize their profits, as advertisers are more likely to pay more for popular videos. Although many attempts have been made on popularity prediction of online videos [18][14][1][17][4], because most of the videos in an OVS system are unpopular; consequently, models preferentially learn the popularity trends of these unpopular videos to achieve better performance on the video set as a whole. Prediction of the top-N popular videos remains a challenging problem for the following reasons. First, popular and unpopular videos may have similar early view patterns, and this similarity limits the performance benefit of video classification based on early view patterns [6].

ASSISTANT PROFESSOR
DEPT OF INFORMATION TECHNOLOGY
PRAGATI ENGINEERING COLLEGE(A),SURAMPALEM(EAST GODAVARI)A.P,INDIA

Second, existing studies show that the strong correlation between early views and long-term popularity dominates the training of theprediction models. This overdependence on early view patterns prevents models from finding popular videos based on multisource data [8][13]. To address the above problem, we present a novel popularity prediction model named multi-factor differential influence (MFDI) based on multivariate linear regression (MLR). We first enhance the ability of early view patterns to identify different popularity trends. We conduct a large-scale analysis of statistical data of early viewers' attitude-related behavior and the long-term popularity of videos. We find that the increase in the future popularity of videos follows an approximate Rayleigh distribution with respect to the degree of contradiction between early viewers with different attitudes. Based on this discovery, by combining early views with knowledge of early viewers' attitudes, we construct early rating patterns that offer better discriminative power for identifying popularity trend and use these rating patterns to replace early view patterns as the input to the proposed model. Furthermore, we incorporate the popularity of the videos' content on a social network to help the proposed model to discover popular videos and to learn their popularity trends. To overcome the
restrictions on multi-source data utilization, we propose a time aware trade-off mechanism to control the model's relative
dependence on enhanced early patterns and social network data. The time-aware trade-off applies higher decay to earlier
enhanced patterns and correspondingly increases the degree of dependence of the model on social network data over time.

We evaluate the proposed model using real-world data consisting of videos from YouTube and social network data from Twitter. Our experimental results show that the proposed model outperforms state-of-the-art models, thereby confirming
the benefits of our efforts to improve the prediction performance for the top-N popular videos. The main contributions
of this paper can be summarized as follows:

- We propose a model for predicting the top-N popular videos. By enhancing the ability of early patterns to distinguish among popularity trends and optimizing the model's utilization of multi-source data, we develop a model that achieves the promised performance;
- By using the tags of videos as indicators of their content and jointly training a multi-layer perceptron (MLP) network on the popularity data of videos and their related social content, we estimate the contribution of the popularity of a video's content on a social network to the long-term popularity of the video.

II. RELATED WORK

Since the popularity of online videos has been proven to be predictable through the statistical analysis of large-scale
YouTube data [25][3], numerous related studies have been conducted. Szabo and Huberman (S-H) proposed a content-scaling (CS) model based on log-transformed relations between a video's long-term popularity and its early popularity [18]. Their conclusion is one of the most important foundations of popularity prediction research and has been succeed by many related works [24][1][2]. All of the approaches cited above achieved initial success, but their shortcomings has been uncovered by subsequent research. Pinto and Almeida discovered that videos with similar popularity at a given time may exhibit distinct popularity behaviors in the future. Based on this discovery, they proposed a new multivariate radial basis function (MRBF) model by investigating the view patterns during the early period instead of the cumulative views up to a given time [14]. The MRBF model showed better performance than the models of the SH
paradigm and has been prven to have good extensibility and generalizability by subsequent research [13][7][10][23].

However, existing studies have focused on achieving high prediction performance on the entire video set, whereas the prediction of the top-N popular videos has been largely ignored. Unlike top-N video identification for video recommendation [9][5], the prediction of the top-N popular videos is concerned with the overall popularity in the entire video set rather than being oriented toward individual users. For services such as online advertising, knowledge of the top-N popular videos is of great commercial value for improving the profit-budget ratio. However, the prediction of the top-N popular videos remains a critical problem because the performance of existing models for predicting the top-N popular videos is far worse than their performance on the video set as a whole. This problem is caused by the Pareto distribution of videos' popularity, as most of the views received by a video set are associated with only a few popular videos. Therefore, to reduce the prediction error over the entire video set, models will preferentially learn the popularity trends of the unpopular videos, hence sacrificing prediction

performance on popular videos. Some recent studies have attempted to more deeply analyze the dynamics of video popularity and have related the popularity dynamics to various factors [16][15][21]. Although some of these studies have improved the prediction performance through the leveraging of multiple factors, their experimental results also show that the utilization of multisource data is a critical problem due to the dominance of early patterns in the training of prediction models [4][20]. Enlightening research was performed by Wu and Zhou [22], who modeled the reactions of users and the information cascade as two hidden processes and attempted to fit the evolution of video popularity using a combination of these two processes. Although the Evo model is far from being suitable for real application in popularity prediction, the experimental results of Wu and Zhou's study illustrate that the modeling of additional early features is a feasible way to improve a model's ability to learn different popularity trends. [10–14].

III. MFDI MODEL

A. Problem Statement and Related Definitions

Our task is to predict the cumulative views of online videos at given time tr, based on the observed data from ts < tr, and to retrieve the Top-N popular videos at tr. Such task requires the prediction performance focusing on popular videos rather than the entire video set. Definitions of variables used in this article are listed in Table I.

B. Framework of MFDI

MLR is a widely adopted mathematical model in popularity prediction. The typical framework of MLR based models

combines a regression on early view patterns fvijg ts j=t1 with an optional compensation based on video i's side

information, which can be formulated

$$\widetilde{N}_i(t_s, t_r) = \sum_{j=t_1}^{t_s} \omega_j v_{ij} + \omega_b B(\Psi_i), \tag{1}$$

$$\begin{aligned}\widetilde{N}_i(t_s, t_r) = &\sum_{j=t_1}^{t_s} g(j;\theta)\omega_j v_{ij} f(R_{ij}^p, R_{ij}^n; \Phi) \\ &+ \omega_b B(\Psi_i) \int_{j=t_1}^{t_s} [1 - g(j;\theta)]dj.\end{aligned} \tag{2}$$

C. Enhancing the Discriminative Power of Early Patterns

Similar early view patterns could lead to different popularity dynamics. For potential viewers, the feedback early viewers leave on videos is one of the most important drivers of their viewing decisions and may lead to different viewing dynamics. Therefore, to extract early patterns that better represent the video popularity trend, we intend to combine the early views with knowledge of the early viewers' attitudes. Viewers' attitudes can be reflected through related text, such as comments, and related behavior such as clicking "like" or "dislike" after watching. We choose to incorporate early viewers attitude by leveraging three attitude-related behaviors: "like Count", "favorite", and "dislike Count", where the first two represent positive attitude and last represent negative attitude. We start with the regression of early views vij as expressed in (1). Here, we assume that each early view leads to new views.

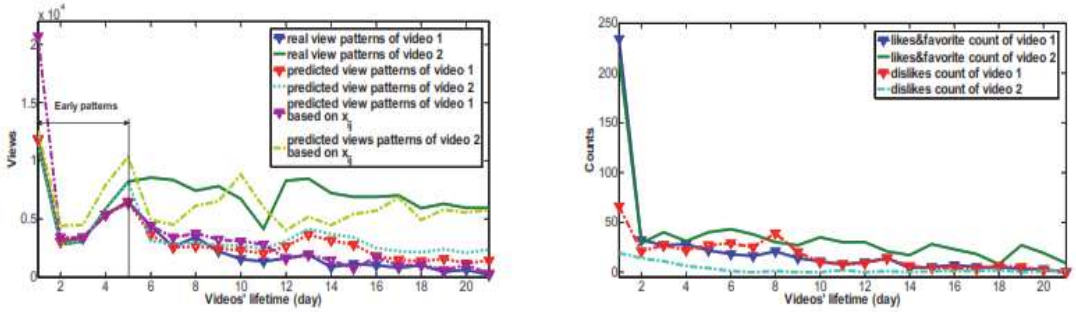
videos	$num_{t_s=5}^{similar}$	mRSE by v_{ij}	mRSE by x_{ij}
9677	4498 (46.48%)	0.2907	0.2214

TABLE II: The influence of x_{ij} on MLR prediction

D. Leveraging Content Popularity on Social Networks

Information related to a videos' content is usually discussed on social networks before the release of the video; and the

popularity of this related information significantly contributes to the subsequent popularity of of the video [11]. The contents of videos on YouTube are reflected by their "tags" and "description". With regard to statistical data, the "tags" of a video usually denote its key content, whereas little content information can be learned from the "description". Hence, we choose to learn the relation between the tags and views of a video based on the popularity of videos and related tweets to estimate the contribution of the popularity of information related to a video's content on a social network to its number of views.



(a) (b)

Fig. 3: (a) Early view patterns, early rating patterns and predicted future views; (b) data on attitude related behaviors for the two videos

E. Using Multi-Source Data with a Time-aware Trade-off

The influence of early view patterns at a given time can be described as follows:

$$\widetilde{v}_{ik} = \sum_{j=t_1}^{t_a} \phi(k-j;\theta)\omega_j x_{ij},$$

IV. EXPERIMENTS AND ANALYSIS

A. Data Preparation

The video data were obtained from YouTube using the YouTube API 3.0 (https://developers.google.com/youtube/v3/).

We obtain the basic information on each video including the "id", "title", "description" and "tags", and the time aware

data including "views", "like Count", "dislike Count" and "favorites" for every 24 hours. A typical example of the data

collected for a video is shown in Table IV. To track the popularity of tweets that shared tags with the collected videos,

we obtained their "retweets" data at the same frequency using Rest API 2.0 1 . Specifically, we first search for videos uploaded over the previous three days and obtained 216,000 different videos. Then, we extracted the tags of each video and created a tag set containing the 9341 most frequently appearing tags (we extracted only the first 5 tags of each video). Next, we used the Twitter API to search for tweets with tags that appeared in the tag set, identifying 3314 tags from 38114 tweets. Then, we tracked the "retweets" of each identified tweet every 24 hours for the next week. We tracked only the "retweets" of the 3734 tweets with the 126 most frequently appearing tags. The crawled tag set covers 20:3% of the filtered videos. Based on the crawled data, we removed videos that received no views on at least one week and those with too few cumulative views. The final video set contains 48,369 videos.

B. Evaluation Metrics and Experimental Settings

The experiments reported here consist of two parts: performance evaluation and result analysis for the performance

evaluation, we choose two metrics. The first is the widely adopted mean relative squared error (mRSE).

$$mRSE = \frac{1}{M^v}\sum_{i=1}^{M^v}\left(\frac{\widetilde{N}_i(t_s,t_r) - N_i^*(t_r)}{N_i^*(t_r)}\right)^2,$$

C. Result Analysis:

To begin our result analysis, we compare the contributions to the prediction performance. We retain each of the three components in the proposed model individually to assess the resulting mRSE reduction compared with the performance

with all three components removed. The result is shown in

Top-N	MFDI-$B(\Psi_i)$	$B(\Psi_i)$	MFDI-$g(j;\theta)$	MFDI
mRSE@10	0.2023	0.2914	0.1801	0.1686
mRSE@20	0.1907	0.3019	0.1659	0.1528
mRSE@30	0.1782	0.3087	0.1493	0.1339
mRSE@40	0.1585	0.3116	0.1366	0.1251
mRSE@50	0.1431	0.3283	0.1252	0.1159
Overall	0.1018	0.8983	0.0964	0.0883

Analysis of the contributions and the performance of the MFDI model

V.CONCLUSION

In this article, we have investigated the problem of top-N popular video prediction and have proposed a novel MFDI prediction model. The proposed model predicts the top-N popular videos by enhancing the ability of early patterns to identify different popularity trends and by optimizing the model's utilization of multi-source data. Experimental results obtained using real-world data demonstrate that the proposed model outperforms other models, including the state-of-the-art model. This article is our initial study on popularity prediction for Top-N popular videos. To the best of our knowledge, this study is the first popularity prediction research to focus on top-N popular videos. Our study still has room for improvement. Possible improvements include leveraging additional related early features and discovering more precise mathematical correlations between the attitudes of early viewers and future popularity trends. For example, in this study, the early viewers' attitudes are inferred from only the three explicit behaviors factors; however, early viewers' attitudes may also be reflected in many implicit ways. If more data related to early viewers' attitudes or similar features could be well modeled, they would be helpful for further improving the model's prediction performance, especially on the top-N popular videos..

REFERENCES :

[1] M. Ahmed, S. Spagna, F. Huici, and S. Niccolini. A peek into the future: Predicting the evolution of popularity in user generated content. In Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, WSDM '13, pages 607–616, New York, NY, USA, 2013. ACM.

[2] P. Bao. Modeling and predicting popularity dynamics via an influencebased self-excited hawkes process. In Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16, pages 1897–1900, New York, NY, USA, 2016. ACM.

[3] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. Analyzing the video popularity characteristics of large-scale user generated content systems. IEEE/ACM Trans. Netw., 17(5):1357–1370, Oct. 2009.

[4] J. Chen, X. Song, L. Nie, X. Wang, H. Zhang, and T.-S. Chua. Micro tells macro: Predicting the popularity of micro-videos via a transductive model. In Proceedings of the 2016 ACM on Multimedia Conference, MM '16, pages 898–907, New York, NY, USA, 2016. ACM.

[5] Z. Deng, M. Yan, J. Sang, and C. Xu. Twitter is faster: Personalized time-aware video recommendation from twitter to youtube. ACM Trans. Multimedia Comput. Commun. Appl., 11(2):31:1–31:23, Jan. 2015.

[6] W. Ding, Y. Shang, L. Guo, X. Hu, R. Yan, and T. He. Video popularity prediction by sentiment propagation via implicit network. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15, pages 1621–1630, New York, NY, USA, 2015. ACM.

[7] F. Figueiredo, F. Benevenuto, and J. M. Almeida. The tube over time: Characterizing popularity growth of youtube videos. In Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM '11, pages 745–754, New York, NY, USA, 2011. ACM.

[8] W. Hoiles, A. Aprem, and V. Krishnamurthy. Engagement and popularity dynamics of youtube videos and sensitivity to meta-data. IEEE Transactions on Knowledge and Data Engineering, 29(7):1426–1437, July 2017.

[9] D. K. Krishnappa, M. Zink, C. Griwodz, and P. Halvorsen. Cachecentric video recommendation: An approach to improve the efficiency of youtube caches. ACM Trans. Multimedia Comput. Commun. Appl., 11(4):48:1–48:20, June 2015.

[10] C. Li, J. Liu, and S. Ouyang. Characterizing and predicting the popularity of online videos. IEEE Access, 4:1630–1641, 2016.