IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 22, 2016, accepted June 12, 2016, date of publication June 14, 2016, date of current version July 7, 2016.

Digital Object Identifier 10.1109/ACCESS.2016.2580911

A Peek Into the Future: Predicting the Popularity of Online Videos

SHUXIN OUYANG[1], CHENYU LI[2], (Student Member, IEEE), AND XUEMING LI[2]
[1]School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Beijing University of Posts and Telecommunications, Beijing 100876, China

Corresponding author: S. Ouyang (ouyangshuxin@gmail.com)

ABSTRACT Predicting the popularity of online videos is an important task for the service design, advertisement placement, network management, and so on. In this paper, we tackle the challenge head-on by casting the popularity prediction problem into two consecutive tasks: online video future popularity level prediction and online video future view count prediction. We first predict the future popularity levels of online videos, based on a rich set of features and effective classification technique. Then, according to the popularity level transitions, we build specialized regression models to predict the future view count values. We validate our approach on the exhaustive dataset of a leading online video service provider in China, namely, Youku. The experimental results show that comparing with two state-of-the-art baseline models, our proposed method can significantly decrease the relative prediction errors of 32.25% and 19.82%, respectively. At last, we also discuss the model setup and feature importance of our method. We believe our work can provide direct help in practical for the interested parties of online video service, such as service providers, online advisers, and network operators.

INDEX TERMS Online video service, video popularity prediction.

I. INTRODUCTION

Among the various kinds of Web 2.0 services, online video service is currently the dominating one on the Internet. As a white paper of Cisco System notes [1], in terms of bytes the video traffic accounted for 70% of all the Internet traffic in 2015, and will be up to 82% by 2020. Every second, nearly a million minutes of video content crosses the network, and it will take an individual over 5 million years to watch all the videos across the Internet each month. With such large volume of video content competing the limited user attention and time, it is not surprised that ecosystem of online videos turns into the asymmetric *rich-get-richer* effect. A small fraction of the videos attracts most of the user interest, whereas the vast majority are barely noticed by users [2].

Hence, predicting the future popularity of online video content, which is measured by view count, is of great importance for a number of contexts. For the service providers, such knowledge can help them design more effective information services, such as video ranking, video searching and recommendation schemes [3]. For the online advertisers, by identifying the next rising star of online videos, they can better place their advertisements and estimate the revenue in advance [4]. And for the network operators, with the estimations of incoming video views, delivery network resources (e.g. bandwidth and cache servers) can be proactively allocated to avoid the potential bottlenecks [5]. Moreover, in a quite different context, the future popularity of online videos is of great interest for the opportunistic communications among mobile devices [6]. In such resource-constrained environments (e.g. limited bandwidth and storage), predicting hot videos can provide help for the strategies of content delivering, caching and replicating.

In this paper, we study the video popularity of Youku (*www.youku.com*), a leading online video service provider in China. Our work is based on the data of 200,714 videos collected from Youku for 30 consecutive days. With these data, We tackle the challenge of predicting future popularity of online videos by proposing a two-stage prediction method. First, we estimate the future popularity level of a video based on a rich set of features and effective classification technique. Then, according to the transition from the observed popularity level to the predicted popularity level, we build specialized regression models to predict the precise value of future view count for the video. We test our method on a real-world dataset and compare the prediction performance with two state-of-the-art baseline models of online video

popularity prediction. The experimental results show that our method leads to significant reductions in the prediction errors, reaching up to 32.25% and 19.82% over the baseline models, respectively. We further analyze the potentials and limitations of different detectors in the popularity level prediction. In addition, we investigate the importance and utility of each feature and feature group, and shed light on the key impact factors related to the video popularity. The main contributions of our work are summarized as follows:

- We release a large-scale, long-term and up-to-date online video analysis database to the public.[1]
- We provide in-depth analysis of the importance of different features in the online video future popularity level prediction.
- We propose a two-stage method to predict the future popularity of online video.
- The proposed two-stage online video future popularity prediction model gets state-of-the-art results and outperforms conventional methods.

The rest of this paper is organized as follows: We firstly outline the related work in Section II. Then in Section III, we briefly describe our dataset and provide a characterization of the video popularity distribution. The details of popularity prediction, including the model, the experiment, and the discussion, are provided in Section IV, Section V and Section VI. At last, we conclude our paper in Section VII.

II. RELATED WORK

A. POPULARITY LEVEL PREDICTION OF ONLINE CONTENT

Jamali and Rangwala [7] extracted several user-based and comment-based features for Digg stories, and trained different classification methods to predict the future popularity level with those features collected during the first ten hours of a Digg story. Lee *et al.* [8] introduced a widely used survival analysis method (Cox proportional-hazards regression) into the problem of online content popularity prediction, and detected whether a DPreview or MySpace thread would still be popular (with comments more than a threshold) after a certain period of time. Tsagkias *et al.* [9] addressed the problem of predicting the comment volume for online news story as two consecutive classification tasks. They first identified whether a story would receive a comment, and then based on the results they predicted the comment volume to be low or high. Vallet *et al.* [10] collected cross-system features from YouTube and Twitter websites, and used gradient boosted decision tree method to classify the YouTube videos into popular, viral, or both categories. Roy *et al.* [11] also extracted information from Twitter, and associated popular topics to YouTube videos. They proposed an algorithm to describe the social transfer and further identified the YouTube videos that would experience sudden bursts of popularity.

[1] https://github.com/lichenyu/Datasets/tree/master/Youku_Popularity_151212_151221

B. POPULARITY NUMBER PREDICTION OF ONLINE CONTENT

Szabo and Huberman [2] observed a strong linear correlations between the log-transformed long-term popularity and log-transformed early popularity. Based on this characteristic, they proposed a simple log-linear model to predict the future popularity of online content. This method was verified on various kinds of online content: YouTube videos [2], Digg stories [2], online news articles [12]–[14] and etc. In [2], the daily view count increments were treated equally. It is not difficult to notice that treating daily increments equally is not accurate. Pinto *et al.* [15] further modified the log-linear model, and proposed a multivariate regression model by assigning different weights to the daily increments in view count during the early observation period. After that, Vasconcelos *et al.* [16] used an ordinary least squares (OLS) multivariate linear regression model based on user, venue and content features, to predict the micro-review popularity of a location based social networking platform, Foursquare. Li *et al.* [17] analyzed how the popularity of Youku videos evolved over time, and used multivariate regression models with different parameters to predict the future popularity of the videos with different popularity evolution patterns. In addition to these regression based methods, there are also efforts towards building models to predict online content popularity using other techniques such as reservoir computing [18], time series analysis [19] and hidden Markov model (HMM) [20].

Our study complements these existing works by considering the popularity level prediction and the future view count prediction conjointly. We cast the prediction of video popularity into a classification task and a regression task, and propose a two-stage prediction method. Moreover, we introduce the notion of popularity level transition into building specialized regression models, and eventually gain significant improvements of the prediction performance.

III. PRELIMINARIES

A. DATASET

The data used in this paper were collected from a leading online video service provider in China, namely Youku. Youku archives more than 500 million monthly active users and 800 million daily video views [21]. It provides a comprehensive type of online video service, including both user-generated content (UGC) and video-on-demand (VoD). To study the popularity of Youku videos, we crawled and tracked the **video meta-data** and the **user meta-data** for a set of videos via Youku Open API [22]. The **video meta-data** includes the static and dynamic properties of a video, such as: title, description, category, duration, published time, tags, stream type, copyright type, public type, source name, view count, comment count, favorite count, up count, down count, user ID and etc. And the **user meta-data** includes the statistic information of the video uploader, such as: registration time, account status, video count, video view count, favorite count,

playlist count, status count, follower count, following count and etc.

The data were gathered with a two-step procedure: 1) First, on a certain initial date we crawled the video list from the portal's "*most recently uploaded videos of the day*" section. Video and user meta-data of these newly uploaded videos were retrieved at the same time. 2) Then, we tracked the popularity of these videos for 30 consecutive days since the publication date, by retrieving their video meta-data and extracting the updated view counts every day. Thus, for each video in our dataset, we could obtain its video meta-data, user-meta-data and the consecutive view counts for a month.

In the first step, we collected video lists from December 12th, 2015 to December 21st, 2015, and got 10 sets of the daily newly uploaded video lists. During the second tracking step, 48,171 videos were either deleted by the uploaders or blocked by Youku. We excluded these videos from our dataset. Eventually, our dataset consists of 200,714 videos uploaded by 130,403 uploaders and watched 2,024,950,569 times by users.

The Youku API also provides several other kinds of video lists, such as "*most viewed videos of the day*" or "*most commented videos of the day*". In our analysis, we choose to collect the most recently uploaded videos, in order to avoid sample bias [2], [23] and get a complete history of view count since the publication date.

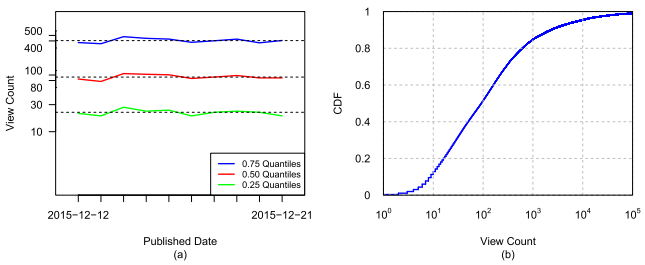


FIGURE 1. (a) Day 30 view count quartiles for the videos published on different date. (b) CDF of popularity on the 30th day for all the videos.

B. CHARACTERIZATION OF POPULARITY

In this part, we provide a characterization of the popularity distribution for the newly uploaded videos in our dataset. First, in Fig. 1 (a), we plot the 1st, 2nd and 3rd quarter quantiles of the popularity on the 30th day since publication, for the video sets uploaded on each day in our dataset. The average quantiles for all videos are also plotted in the figure with the dotted lines. All the quartile curves keep steady across published dates. It indicates for the video sets uploaded on different dates, their popularity's distribution since the publication date is quite similar. Hence, a general analysis can be performed on whole video popularity dataset.

Fig. 1 (b) shows the cumulative distribution function (CDF) of the view counts on the 30th day for all the videos in our dataset. It can be noticed that the user attention among videos is allocated in an asymmetric way, with most of the videos are barely noticed whereas a few of videos achieve very large view counts. For instance, 51.50% of the videos are watched for less than 100 times, and 664 videos are with only 1 view during the 30 days. In contrast, 4.53% of the videos are watched for more than 10,000 times, and the most popular video (VID equal to "XMTQxNDUxMzUyNA==") is watched for as many as 38,590,588 times.

TABLE 1. Summary of video popularity levels.

	View Count Range	% of Videos	Description
Level 1	$[0, 100)$	51.50%	Extremely unpopular.
Level 2	$[100, 1000)$	33.29%	Relatively unpopular.
Level 3	$[1000, 10000)$	10.68%	Relatively popular.
Level 4	$[10000, 38590588)$	4.53%	Extremely popular.

Since the popularity of different video varies substantially, we try to use different view count ranges (i.e. popularity levels) to cluster videos with different popularity. As the popularity distribution is highly skewed, the division size of different levels will not be equal. In fact, the unpopular level should cover most of the videos, while the popular level should highlight the videos with extremely large view counts. In our analysis, we split the popularity into 4 non-overlapping ranges, as shown in Table 1. The range values of each popularity level are about an order of magnitude larger than those of the previous level, and each level accounts for around 50%, 35%, 10% and 5% of the total videos, respectively. By using this criterion, we can assign the videos into different types with different importances and values for the service provider and network operator.

IV. POPULARITY PREDICTION

In this paper, we study the problem of predicting the near-future popularity (i.e. view count) of a newly uploaded video. More specifically, for a given video v, after monitoring its popularity for a short period of time from the publication date to an *indicator* date i, we try to predict its future popularity on a *reference* date r. In our analysis, we set $i = 7$ and $r = 30$, same as [2] and [15]. We cast the problem into two consecutive tasks: **future popularity level prediction** and **future view count prediction**. First, we try to identify the future popularity level of the video on the reference date. This problem can be framed as a multi-class classification task. Then, based on the result of the first stage prediction, we forecast the precise value of the future video popularity. This problem is framed as a regression task. The overview of our prediction system is demonstrated in Fig. 2.

A. FUTURE POPULARITY LEVEL PREDICTION

To tackle the popularity level prediction task, we introduce 69 features across five different aspects: video property, user statistic, content topic, textual analysis and historical popularity. Table 2 summarizes all these features.

Video Property Features capture the static properties (e.g. duration, category) and the publication information (e.g. published time, uploading client) of the videos.

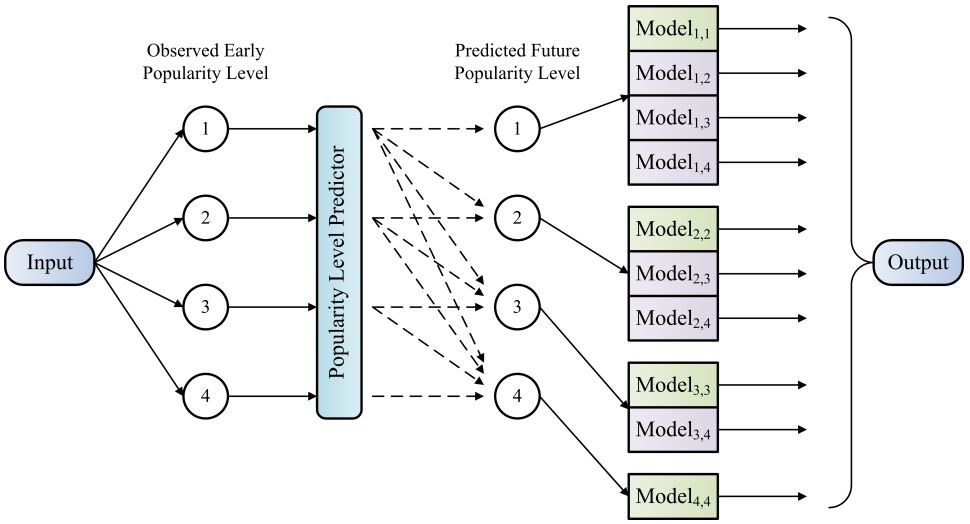


FIGURE 2. Overview of the prediction system.

User Statistic Features capture the account information (e.g. registration time, whether a VIP), publication activities (e.g. the number of uploaded videos) and social influence (e.g. the number of followers) of the video uploader.

Content Topic Features: The content topic of a video is closely related to its popularity [24]. In our analysis, we use the presences of certain representative terms in video tags and titles to serve as content topic features. More specifically, for each popularity level, we extract the top 500 terms that most frequently appear in the tags and titles (in the training set). Then, the common terms that appear across popularity levels, as well as the words without actual meanings such as "of" and "very", are removed from the word sets. At last, we keep the top 100 words for each popularity level as the word list, and use bag-of-word (i.e. frequency) features for the videos in the test set as the content topic features.

Textual Analysis Features capture the syntactic and sentiment information of the textual content of the video. Texts in Youku are mainly written in Chinese. We utilize *jieba* [25] to extract the part-of-speech (PoS) features of the video title. We also use *snownlp* [26] to compute the sentiment score of the video title, description, and tags. A sentiment score ranges from 0 to 1, indicating the text from very negative to very positive sentiment.

Historical Popularity Features measure the popularity evolution of a video during the early observation period, including the daily increment of view count, the ratio of daily increment over total view count in the observation period, and other video popularity metrics (e.g. comment, rating and favorite) the video archives at the end of the observation period. Moreover, we describe the popularity evolution pattern of the video by utilizing the frequency of popularity burst in the early observation period. In our analysis, if the daily increment ratio ($\frac{\text{daily increase in view count}}{\text{total view count}}$) is larger that three times of the average ratio (i.e. $1/7$), we consider the video experiences a popularity burst on that day.

For the prediction of popularity level, with those 69 features, we exploit the random forest [27] technique as the classifier.

B. FUTURE VIEW COUNT PREDICTION

In future view count prediction, there are two kind of situations: 1) popularity level stays the same between reference date and observed date, 2) popularity level changes between reference date and observed date.

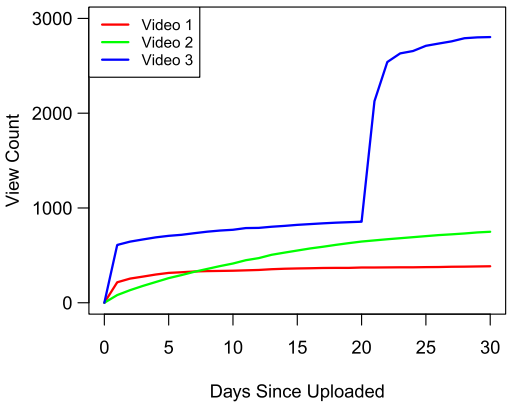


FIGURE 3. Popularity evolutions over time for three different videos.

1) SITUATION 1

If a video's predicted popularity level on the reference date remains the same as its observed popularity level on the indicator date, the most important indicator for future popularity should be the popularity evolution pattern during the early observation period (described by the increases in view count). Take the popularity evolutions of three videos in our dataset for examples, as shown in Fig. 3. For video 1

TABLE 2. Summary of features.

Video Property Features	
1	Video category, one of the 25 predefined types.
2	Video duration, in seconds.
3	Published time, [0, 23] hour of day.
4	Number of video file resolutions supported by Youku.
5	Type of video copyright type, *original* or *reproduced*.
6	Type of video access type, *all*, or *friends*, or *password*.
7	Type of video uploading client.
User Statistic Features	
8	Number of days between the registration date and uploading date.
9	Whether the user is a Youku verified uploader.
10	Whether the user is a Youku VIP.
11	Number of videos uploaded by the user.
12	Total view count of the videos uploaded by the user.
13	Total favorite count of the videos uploaded by the user.
14	Number of playlists created by the user.
15	Number of statuses published by the user.
16	Number of followers.
17	Number of following users.
Content Topic Features	
18–21	Number of tags in the tag list of each popularity level.
22–25	Number of title words in the title word list of each popularity level.
Textual Analysis Features	
27	Length of video title.
27, 28	Number and fraction of Chinese characters in video title.
29, 30	Number and fraction of nouns in video title.
31, 32	Number and fraction of verbs in video title.
33, 34	Number and fraction of adjectives in video title.
35, 36	Number and fraction of adverbs in video title.
37, 38	Number and fraction of prepositions in video title.
39, 40	Number and fraction of numerics in video title.
41, 42	Number and fraction of English letters in video title.
43, 44	Number and fraction of punctuations in video title.
45	Sentiment score of video title.
46	Length of video description.
47	Sentiment score of video description.
48	Number of video tags.
49	Sentiment score of video tags.
Historical Popularity Features	
50–56	Daily increase in view count from Day 1 to Day 7.
57–63	Daily increase ratio of view count from Day 1 to Day 7.
64	Whether a popularity burst exists from Day 1 to Day 7.
65	View count of the video on Day 7.
66	Comment count of the video on Day 7.
67	Favorite count of the video on Day 7.
68	Number of the up rating for the video on Day 7.
69	Number of the down rating for the video on Day 7.

and video 2, their observed and predicted popularity levels are both Level 1. The popularity evolution of video 1 begins with a sudden boom (i.e. burst) and followed by a slow increase, while the popularity of video 2 keeps increasing during the whole tracking period. Hence, although the view counts of video 1 and video 2 on the indicator date are almost the same (330 v.s. 327), the view count of video 2 on the reference date is much larger than that of video 1 (385 v.s. 749). This characteristic was also observed by Pinto *et al.* [15], and they proposed a simple multivariate linear regression model to predict the future popularity of online videos. We keep the multivariate linear regression model for the Situation 1 in our model.

2) SITUATION 2

However, for the videos whose predicted popularity level transits into a larger level (e.g. video 3 in the figure), we find the key impact factor of video popularity turns into the popularity bursts during the middle of video lifetime. These popularity bursts usually correspond to the exogenous effects such as: the video is listed on the first page of Youku website; or the video is retweeted by a famous person on an online social network (OSN) [10], [11]. Since the early popularity evolution pattern has little relation to the final view count in this situation, we use the univariate linear regression model instead of the multivariate model. Besides, we add an intercept in our model to describe the popularity bursts which cause the level transition.

The details of our future view count prediction model are summarized as follows. The popularity level transition of video v is defined as $(L_v(i), \hat{L}_v(r))$, $L_v(i)$ is the observed popularity level of v on the indicator date i and $\hat{L}_v(r)$ is the predicted popularity level on the reference date r. Note $\hat{L}_v(r)$, measured by the (cumulative) view count, should be either equal to or larger than $L_v(i)$. Let $I_v(k)$ denotes the daily increment in view count on the k-th day, and $N_v(k)$ denotes the (cumulative) view count on the k-th day. Clearly, we get $\sum_{k=1}^{i} I_v(k) = N_v(i)$. For a video v with the popularity level transition as t, our method estimates its view count on the reference date $\hat{N}_v(r)$ as:

$$\hat{N}_v(r) = \begin{cases} \sum_{k=1}^{i} \omega_{(k,t)} I_v(k) & \hat{L}_v(r) = L_v(i) \\ \alpha_t N_v(i) + \beta_t & \hat{L}_v(r) > L_v(i) \end{cases} \quad (1)$$

where $\omega_{(k,t)}, \alpha_t, \beta_t$ are the regression parameters of the linear model depends only on the popularity level transition. The optimal parameter values of our model can be optimized by minimizing the prediction errors on a given training set V. In our analysis, we use the mean relative squared error (MRSE) as the criterion to evaluate the efficiency of prediction methods:

$$MRSE = \frac{1}{|V|} \sum_{v \in V} \left(\frac{N_v(r) - \hat{N}_v(r)}{N_v(r)} \right)^2 \quad (2)$$

Let $W_v = \left(\frac{1}{N_v(r)}\right)^2$, the optimization problem can be expressed as:

$$\arg\min_{\omega_{(k,t)}, \alpha_t, \beta_t} \frac{1}{|V|} \sum_{v \in V} W_v (N_v(r) - \hat{N}_v(r))^2 \quad (3)$$

and the optimal model parameters can be computed by simply solving this weighted least squares problem.

V. EXPERIMENT

A. EXPERIMENT SETUP

We evaluate our two-stage prediction method on the Youku dataset, including video meta-data, user meta-data and view

count time series for 200,714 videos. In the experiment, we randomly subsample 50% of the dataset as the training set and treat the rest 50% as the test set. As discussed in the previous section, the multi-class classification model is employed in the popularity level prediction, and the regression model is used in the future view count prediction. The parameters (i.e. $\omega_{(k,t)}$, α_t, β_t) of those models are learned from the training set. And we use the test set to evaluate the prediction performance.

1) PERFORMANCE CRITERION

For the multi-class classification task, we use *macro-precision* and *macro-recall* as the criterion of performance evaluation. The precision of a class (i.e. the popularity level in our problem) is the ratio of the number of videos correctly assigned to the level to the number of videos predicted on the level; while the recall is the ratio of the number of videos correctly classified as the level to the number of videos on the popularity level. The macro-precision (macro-recall) is the average precision (recall) computed across all classes. For the regression task, we use the *MRSE* as the performance criterion, as referred in the previous subsection.

2) BASELINES

We compare our two-stage prediction method with two state-of-the-art baseline models for predicting the popularity of online videos. The first one is a Log-Linear model first proposed by Szabo and Huberman in [2]. It is a univariate linear regression model, using the log-transformed early view count to predict the log-transformed future view count as follows:

$$\hat{N}_v(r) = \exp\left[\ln N_v(i) + \ln r(i, r)\right] \tag{4}$$

where $r(i, r)$ is the model parameter independent of videos. The second baseline model, which was proposed by Pinto *et al.* in [15], is a multivariate linear (Multi-Linear) regression model. It uses the daily increases in view count during the early observation period as the predictive variables, and estimates the future view count as follows:

$$\hat{N}_v(r) = \sum_{k=1}^{i} \theta_k I_v(k) \tag{5}$$

where $\theta_k, k = 1, 2, \ldots, i$ are model parameters independent of videos. All the models are evaluated in the same scenario adopted by Szabo and Huberman [2] and Pinto *et al.* [15]: that is, the indicator date i equal to 7 and the reference date r equal to 30. And the model parameters are all determined by minimizing the MRSE on the training set.

B. FUTURE POPULARITY LEVEL PREDICTION

To choose the proper popularity level prediction model, we compare the performances of several effective classification techniques with the same training and test sets. The candidate methods include the naive Bayes (NB) classifier [28], the k-nearest neighbor (KNN) classifier [29], the support vector machine classifier [30] with linear kernel (SVM-L) and radial basis function kernel (SVM-RBF), the decision tree (DT) classifier [31] and the random forest (RF) classifier [27].

Grid search and 3-fold cross-validation are used on the training set to choose the best model parameters for each method, and the corresponding detection performances on the test set are listed in Table 3. It is clear from the table that the random forest classifier outperforms other algorithms with both the largest macro-precision and the largest macro-recall. This is reasonable since compared to other models, random forest classifier is more suitable for the large-scale data task. The DT method performs slightly worse than RF, followed by the SVM-RBF and the KNN method. The SVM-L method obtains low the macro-precision and macro-recall, only around 50%. And the NB method performs the worst of all, with a macro-recall only 27.88%. Hence, for the first classification stage of our model, we choose the random forest classifier to predict future popularity level.

TABLE 3. Prediction performances of each classification method.

	NB	KNN	SVM-L	SVM-RBF	DT	RF
Macro-Precision	55.47%	84.86%	53.28%	91.80%	94.45%	95.04%
Macro-Recall	27.88%	82.34%	50.43%	88.31%	91.78%	92.57%

TABLE 4. Prediction performances for the future popularity level prediction task.

	Level 1	Level 2	Level 3	Level 4	Macro
Precision	96.14%	95.04%	92.54%	96.43%	95.04%
Recall	98.68%	89.75%	88.07%	93.13%	92.57%

Table 4 shows the per-level and macro-averaged values of the precisions and recalls for the prediction results in our test set. It can be noticed that both the macro-precision and macro-recall are very high (over 95%), verifying the effectiveness of the proposed features and classifier. Generally, the prediction performances for Level 1 and Level 4 are higher than those for Level 2 and Level 3. This is because, for Level 1 videos, their fraction of the total videos is actually very large (around 50% of all videos), and the prediction naturally skews towards them. While for the Level 4 videos, such a high precision and recall indicate these extremely popular videos are very different from other videos, thus easy to distinguish.

C. FUTURE VIEW COUNT PREDICTION

Table 5 shows the MRSE of the prediction results for the baseline methods and the proposed methods. For a better understanding of the model performance, we also separately list the MRSE values produced by each model according to videos with different observed-predicted popularity level transitions. It is apparent from the table that the proposed model outperforms the baseline models with the least MRSE values in both overall and per- popularity level transition cases. The overall MRSE reductions produced by our two-stage prediction model are about 32.25% and 19.82% over the Log-Linear model and Multi-Linear model, respectively.

TABLE 5. Prediction performances for the future view count prediction task.

Popularity Level Transition	MRSE (%)		
	Log-Linear	Multi-Linear	Proposed Method
Overall	5.4594	4.6130	**3.6988**
$1 \rightarrow 1$	3.6677	3.2893	**3.2012**
$1 \rightarrow 2$	18.3506	15.5214	**6.6580**
$1 \rightarrow 3$	95.3032	94.5446	**23.6813**
$1 \rightarrow 4$	99.5592	99.4995	**36.6784**
$2 \rightarrow 2$	3.6307	2.9237	**2.9230**
$2 \rightarrow 3$	21.9495	17.2534	**8.5685**
$2 \rightarrow 4$	95.4621	94.5749	**27.5509**
$3 \rightarrow 3$	5.8836	4.3406	**4.2923**
$3 \rightarrow 4$	30.8680	24.4520	**11.7283**
$4 \rightarrow 4$	6.4426	6.0049	**5.3835**

One important question is where these improvements of prediction performances come from. It can be noticed from the table that i) for the videos with popularity level transition $L_v(30) = L_v(7)$, the proposed method slightly reduces the prediction errors; ii) while for the videos with popularity level transition $L_v(30) > L_v(7)$, significant prediction improvements can be observed; and iii) the larger the popularity level transition is, the more improvements our method gains. This indicates while the baseline models are more biased towards the videos with unchanged popularity level transitions, our model suits all types of the videos. This is because we build specialized models according to different popularity level transitions. Thus, we can bring fewer noise instances into the model training and get more proper model parameters. Another interesting fact we should note is that larger predicted popularity level transitions mean the videos have more potential views in the future. Hence, they are more important for service providers and network operators than the videos with unchanged popularity level transitions. As the prediction performances are extremely good for these kinds of videos, our method is of more application value in practice than the baseline methods.

VI. FEATURE ANALYSIS

We evaluate the effectiveness of each feature and each feature group in the popularity level prediction. In our analysis, we propose 69 features from 5 different aspects related to online video popularity.

We rank these features based on the node impurity decreases (measured by the Gini index) when building the RF classifier. We find that the most discriminative feature is "*view count on Day 7*", followed by "*view count increase on Day 1*", "*view count increase on Day 2*", "*follower count of the user*", "*total view count of the videos uploader by the user*", "*video category*" and etc. This indicates that i) the popularity evolution pattern during the early observation period, ii) the social influence and publication history of the uploader, and iii) video category play important roles in estimating the video popularity.

We also investigate the importance of each feature group, by running the RF for a number of times and removing one group of features each time. Table 6 shows the corresponding macro-precisions and macro-recalls when we remove video property features (-VP), user statistic features (-US), content topic features (-CT), textual analysis features (-TA) and historical popularity features (-HP), respectively. It can be noticed that the "*historical popularity features*" is the most discriminative feature group. When it is removed, prediction performance experiences a significant drop, with the macro-precision decreasing by 19.91% and the macro-recall decreasing by 32.50%. The "*user statistic features*" is the next discriminative feature group, with reductions of the macro-precision and macro-recall 6.35% and 6.02%, respectively. These observations confirm the feature importance analysis above. And for the rest feature groups, we observe similar slight decreases in macro-precisions (around 5%) and macro-recalls (around 3%), implying that they all contribute to the overall performance. By introducing these features, the popularity level prediction can be further improved.

TABLE 6. Prediction performances when removing each feature group.

	All	-VP	-US	-CT	-TA	-HP
Macro-Precision	95.04%	90.01%	88.69%	90.92%	90.12%	75.13%
Macro-Recall	92.57%	89.73%	86.55%	88.51%	89.35%	60.07%

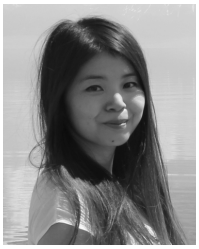
VII. CONCLUSION

In this paper, we proposed a two-stage prediction method to tackle the problem of popularity prediction of online videos. We introduced a rich set of features and utilized effective classification technique to identify the transition from the observed popularity level on the indicator date to the predicted popularity level on the reference date. We built specialized regression models according to the popularity level transitions to estimate the future view count of online videos. Experimental results showed that our method outperformed the baseline models and could reduce the prediction errors significantly. We further revealed the key impact factors of online video popularity, by investigating the feature importance in the classification task. We believe our work presented in this paper can provide valuable help for service providers, online advisers, and network operators. Future work of the paper involves testing the method on other datasets and further improving the classification performance of popularity levels.

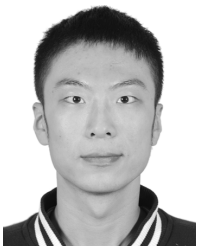
REFERENCES

[1] *Cisco Visual Networking Index: Forecast and Methodology, 2015–2020*, accessed on Jun. 25, 2016. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html
[2] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," *Commun. ACM*, vol. 53, no. 8, pp. 80–88, 2010.
[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.*, vol. 17, no. 5, pp. 1357–1370, Oct. 2009.
[4] A. Ghose and S. Yang, "An empirical analysis of search engine advertising: Sponsored search in electronic markets," *Manage. Sci.*, vol. 55, no. 10, pp. 1605–1622, 2009.

[5] Y. Zhou, L. Chen, C. Yang, and D. M. Chiu, "Video popularity dynamics and its implication for replication," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1273–1285, Aug. 2015.
[6] B. Han, P. Hui, V. S. A. Kumar, M. V. Marathe, J. Shao, and A. Srinivasan, "Mobile data offloading through opportunistic communications and social participation," *IEEE Trans. Mobile Comput.*, vol. 11, no. 5, pp. 821–834, May 2012.
[7] S. Jamali and H. Rangwala, "Digging digg: Comment mining, popularity prediction, and social network analysis," in *Proc. Int. Conf. Web Inf. Syst. Mining*, Nov. 2009, pp. 32–38.
[8] J. G. Lee, S. Moon, and K. Salamatian, "Modeling and predicting the popularity of online contents with Cox proportional hazard regression model," *Neurocomputing*, vol. 76, no. 1, pp. 134–145, 2012.
[9] M. Tsagkias, W. Weerkamp, and M. de Rijke, "Predicting the volume of comments on online news stories," in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2009, pp. 1765–1768.
[10] D. Vallet, S. Berkovsky, S. Ardon, A. Mahanti, and M. A. Kafaar, "Characterizing and predicting viral-and-popular video content," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2015, pp. 1591–1600.
[11] S. D. Roy, T. Mei, W. Zeng, and S. Li, "Towards cross-domain learning for social video popularity prediction," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1255–1267, Oct. 2013.
[12] M. Tsagkias, W. Weerkamp, and M. de Rijke, "News comments: Exploring, modeling, and online prediction," in *Advances in Information Retrieval*. Milton Keynes, U.K.: Springer, 2010, pp. 191–203.
[13] A. Tatar, P. Antoniadis, M. D. de Amorim, and S. Fdida, "Ranking news articles based on popularity prediction," in *Proc. Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2012, pp. 106–110.
[14] A. Tatar, P. Antoniadis, M. D. de Amorim, and S. Fdida, "From popularity prediction to ranking online news," *Social Netw. Anal. Mining*, vol. 4, no. 1, p. 174, 2014.
[15] H. Pinto, J. M. Almeida, and M. A. Gonçalves, "Using early view patterns to predict the popularity of YouTube videos," in *Proc. 6th ACM Int. Conf. Web Search Data Mining*, 2013, pp. 365–374.
[16] M. Vasconcelos, J. M. Almeida, and M. A. Gonçalves, "Predicting the popularity of micro-reviews: A Foursquare case study," *Inf. Sci.*, vol. 325, pp. 355–374, Dec. 2015.
[17] C. Li, J. Liu, and S. Ouyang, "Characterizing and predicting the popularity of online videos," *IEEE Access*, vol. 4, pp. 1630–1641, 2016.
[18] T. Wu, M. Timmers, D. De Vleeschauwer, and W. Van Leekwijck, "On the use of reservoir computing in popularity prediction," in *Proc. Int. Conf. Evol. Internet*, 2010, pp. 19–24.
[19] G. Gürsun, M. Crovella, and I. Matta, "Describing and forecasting video access patterns," in *Proc. IEEE Int. Conf. Comput. Commun.*, Apr. 2011, pp. 16–20.
[20] L. Jiang, Y. Miao, Y. Yang, Z. Lan, and A. G. Hauptmann, "Viral video style: A closer look at viral videos on YouTube," in *Proc. Int. Conf. Multimedia Retr.*, 2014, p. 193.
[21] *Youku Tudou Partners With Xiaomi to Accelerate Multi-Screen Ecosystem Development*, accessed on May 20, 2016. [Online]. Available: http://ir.youku.com/phoenix.zhtml?c=241246&p=irol-newsArticle_print&ID=1988630
[22] *Youku Open API*. accessed on May 20, 2016. [Online]. Available: http://cloud.youku.com/docs
[23] Y. Borghol, S. Mitra, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti, "Characterizing and modelling popularity of user-generated videos," *Perform. Eval.*, vol. 68, no. 11, pp. 1037–1055, 2011.
[24] F. Figueiredo, J. M. Almeida, F. Benevenuto, and K. P. Gummadi, "Does content determine information popularity in social media?: A case study of YouTube videos' content and their popularity," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2014, pp. 979–982.
[25] *Jieba: Chinese Text Segmentation*, accessed on Jun. 25, 2016. [Online]. Available: https://github.com/fxsjy/jieba
[26] *SnowNLP: Simplified Chinese Text Processing*, accessed on Jun. 25, 2016. [Online]. Available: https://github.com/isnowfy/snownlp
[27] T. K. Ho, "Random decision forests," in *Proc. Int. Conf. Document Anal. Recognit.*, 1995, pp. 278–282.
[28] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, vol. 25. Englewood Cliffs, NJ, USA: Prentice-Hall, p. 27, 1995.
[29] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *Amer. Statist.*, vol. 46, no. 3, pp. 175–185, 1992.
[30] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[31] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Boca Raton, FL, USA: CRC Press, 1984.



SHUXIN OUYANG is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, China. Her research interests include traffic measurement and classification, mobile Internet traffic analysis, data miningm and interactive learning.



CHENYU LI is currently pursuing the Ph.D. degree with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, China. His current research interests focus on traffic measurement and classification, mobile Internet traffic analysis, cloud computing, and data mining.

XUEMING LI received the B.E. degree in electronics engineering from the University of Science and Technology of China in 1992, and the Ph.D. degree in electronics engineering from the Beijing University of Posts and Telecommunications (BUPT) in 1997. From 1997 to 1999, he was a Post-Doctoral Researcher with the Institute of Information Science, Beijing Jiaotong University. He has been with BUPT since 1999. In 2002, he was a Guest Lecturer with Karlsruhe University, Germany. His current research interests include digital image processing, video coding, and multimedia telecommunication. He has undertaken many state and enterprise R and D projects. He has authored three books and over 50 papers in the field of multimedia information processing and transmission. He is currently a Senior Member of the Chinese Institute of Electrics, and a Senior Member of China Society of Image and Graphics.