FORECASTING INITIAL POPULARITY OF JUST-UPLOADED USER-GENERATED VIDEOS

Changsha Ma, Zhisheng Yan and Chang Wen Chen

Dept. of Comp. Sci. and Eng., State Univ. of New York at Buffalo, Buffalo, NY, 14260, USA
changsha@buffalo.edu, zyan3@buffalo.edu, chencw@buffalo.edu

ABSTRACT

User-generated videos (UGVs) have dominated contemporary social networking sites (SNSs). Forecasting their popularity is of great relevance to a broad range of online services. All existing studies forecast popularity of UGVs using their popularity statistics that are accumulated for a period of time after they are uploaded. Hence, there is always a substantial time lag (days to weeks) before popularity forecast can take effects. However, such a time lag is undesirable for timely popularity forecast as forecasting initial popularity during UGVs' lifetime is vitally important. In fact, we have found in our measurement that the most popular UGVs usually precede others starting from the beginning days and UGVs generally receive the highest attentions during the first few days. In this paper, we present the first exploration on forecasting initial popularity for UGVs at their uploading moment without accumulating their popularity statistics. Specifically, we first design an effective crawler framework to collect the publicly observable features of videos at their uploading moment. We then collect a representative and large YouTube video data set with 318,627 videos. Based on the data set, we select the most relevant features as predictors and design a neural network-based learning model to forecast initial popularity of just-uploaded UGVs. Experimental results validate the effectiveness of the proposed forecasting model and demonstrate the model's benefits for online services such as in-video advertising and video caching.

***Index Terms*—** User-generated videos, video initial popularity, popularity forecast

1. INTRODUCTION

User-generated videos (UGVs) have recently dominated contemporary social networking sites (SNSs) such as Facebook, YouTube, and Vine. Forecasting the popularity of UGVs not only empowers the efficient storage and network resource allocation, which benefits both SNS operators and SNS users, but also brings about better opportunities for third-party applications such as in-video advertising.

This research is supported in part by the U.S. National Science Foundation (NSF) under Grants 0915842 and 1405594; and in part by the National Natural Science Foundation of China under Grant 61550110244.

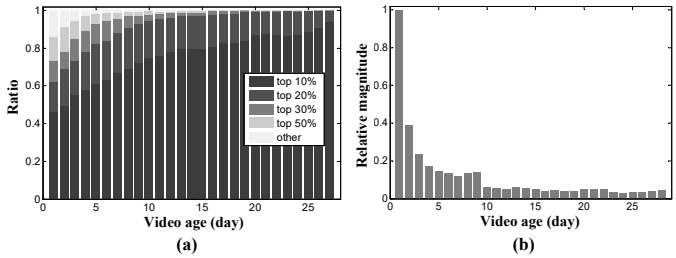

Fig. 1. (a) The change of popularity distribution for the videos who eventually become top 10% popular; (b) the relative magnitude of daily average view counts

Video popularity is usually defined as the relative number of views a video gained during an observation period, most commonly during one single day [1, 2]. According to recent studies, one representative scheme is to directly use the video's early-life statistics as predictors and forecast popularity by regression. In [3], it is proposed to use the historical popularity patterns, i.e., the up-to-date video popularity from the uploading day, to forecast the view counts for UGVs in future days. In [4], video sharing information in the first few days is used to forecast the future popularity. An alternative scheme is based on model fitting. It is proposed in [5] that most of the popularity patterns of UGVs fall into six models. Some other works propose that the long-term popularity patterns of UGVs across their life span can be fitted into a single model constructed in an explainable manner [1, 2]. Future popularity can then be forecasted by fitting the historical popularity patterns into the specific model.

Despite of variations in methodologies, one common feature of all existing works is that they need to accumulate sufficient video popularity statistics once a video is uploaded and they can only start forecasting after a relatively long period. The resulting time lag makes it infeasible to timely capture the *initial popularity*, e.g., the popularity on the first day upon a video is uploaded.

Then, an interesting question would be: *is initial popularity forecast for UGVs meaningful?* To answer this question, we have explored the YouTube video data set we collected and observed that the answer is affirmative because:

1. Initial popularity is highly correlated to future popular-

ity. We study videos that eventually become top 10% popular video after 28 days and show the change of their popularity during the first 27 days in Fig.1 (a). We can see that almost half of these videos have already been top 10% popular on the first day and more than 80% of them are top 50% popular starting from the first day. Besides, the popularity shuffle becomes less significant as time passes.

2. Initial period witnesses the highest view counts compared to future period. We study all videos in the data set and show their daily average view counts from the 1st day to the 28th day in Fig.1 (b). Specifically, we use the number of the 1st day view counts as the reference, and present the relative magnitudes of daily view counts of the following days. The result indicates that the first day has the highest average view counts, and there is a decreasing trend of view counts as time passes. Such a trend also accords with the results in [6].

However, whether or not the initial popularity forecast is feasible and how to perform the forecast have not yet been studied. To fill this gap, we present the first exploration on forecasting the initial popularity of UGVs upon their uploading, i.e., no video popularity statistics are accumulated. More specifically, we focus on the beginning-day popularity forecast instead of initial-hours forecast in order to accommodate user activity variations at different time of a day. To begin with, we design an effective crawler framework to collect the publicly observable features of videos at their uploading moment. We then build a representative and large data set composed of 318,627 YouTube videos and evaluate the possibility of these features to be used for forecasting the beginning-day popularity. Furthermore, we select the most related ones as predictors and design a neural network-based learning model to forecast the beginning-day popularity for just-uploaded UGVs. Experimental results from the model evaluation shows that the proposed model is able to capture the most popular videos with high precision, and provide practical guidance to online services such as in-video advertising and video caching.

2. DATA COLLECTION

We build our data set by crawling videos from YouTube, which is the world's most popular UGV platform. In this section, we first introduce the YouTube crawler framework we designed and the data format this framework outputs. Then we proceed to present the statistical properties of the data set and demonstrate the representativeness of the data set.

2.1. YouTube Crawler

The proposed YouTube crawler handles three tasks: (1) crawling the target YouTube videos, (2) crawling the potential predictors associated with these videos, and (3) crawling the real beginning-day view counts of these videos (ground truth). The framework of the crawler is shown in Fig. 2. It has four crawling stages.

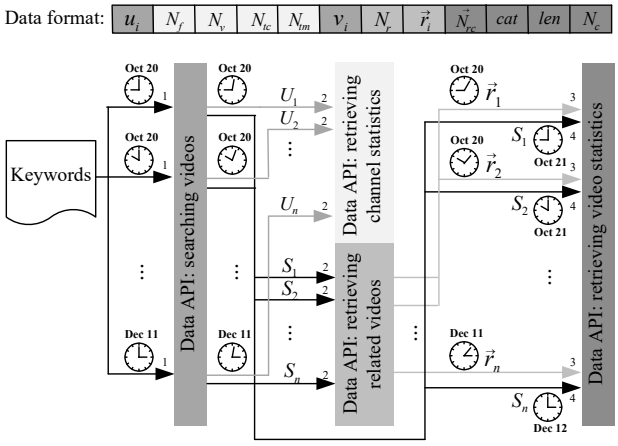

Fig. 2. The crawler framework and the output data format (outputs of different APIs are differentiated by gray scales)

To begin with, we randomly select 150,000 keywords from the Yago lexical ontology [7]. These keywords are then fed as queries to the YouTube data API and retrieve a set of just-uploaded videos within one hour. We record the video ID v_i for each unique video i in the response. Besides, we also record the channel ID u_i of the video's uploader, since we consider the uploader information as a potential factor that impacts the UGV popularity. Therefore, this step results in a video ID set S as well as a channel ID set U.

Then we retrieve all publicly observable features in YouTube associated with a video and a channel. Nonetheless, we avoid collecting sparse features such as location information that are only available for a small set of videos or channels[1]. Specifically, for each channel in U, we request the channel statistics including the number of followers N_f, the number of uploaded videos N_v in the past, and total number video view counts N_{tc} and comments N_{tm} received by the channel. At the same time, we request video statistics for each video in S, including the video category cat, and the video length len, the total number of its related videos N_r and the list of related video IDs $\vec{r}_i$.

At the time we obtain $\vec{r}_i$, we further request the current view counts $\vec{N}_{rc}$ of these related videos as another feature.

Finally, we crawl for the ground truth of what we intend to forecast, i.e., the video view counts N_c after the video has been uploaded for 24 hours.

Note that all the features except for cat and len are time variant. To guarantee that time variant features are collected around the video uploading moment, we adopt parallel pro-

[1]Please refer all available features provided by YouTube data API from https://developers.google.com/youtube/v3/docs

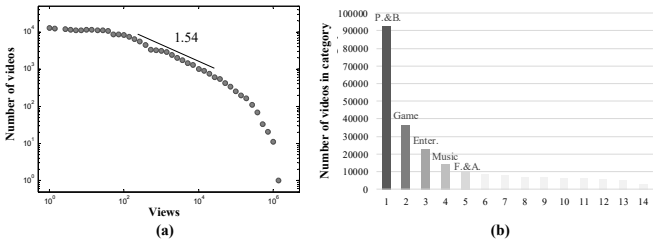

Fig. 3. (a) Popularity distribution; (b) category distribution

cessing to improve time efficiency of the crawler. For the time invariant features cat and len, we simply request them together with N_c.

To build a large and representative data set, we randomly select 14 time points between October 20th, 2015 and December 11st, 2015, and repeat the above process at each time point. The crawler eventually results in 318,627 unique videos with associated features. In addition, we keep track of the video statistic dynamics of 45,204 of these videos for a relatively long period of time (28 days) in order to study the importance of the initial popularity forecast.

2.2. Statistical Properties of The Data Set

We examine the representativeness of our data set from two aspects, i.e., the beginning-day popularity distribution and the category distribution. Existing studies show that the popularity distributions of YouTube videos can be fitted using a power-law distribution with an exponential cutoff [8]. We confirm this trend by showing the number of views a video received plotted against the number of videos falling into that bin in Fig. 3(a). We can see that the popularity distribution shape follows a power-law distribution in the waist and has an approximate exponential decay at large values of views. More specifically, the power-law fitting on our data set results in an exponent of 1.54, which falls into the expected range 1.5∼2.5 [8]. Furthermore, we examine the category distribution of videos in our data set. Fig. 3(b) shows that videos in the data set fall into a wide range of 14 categories, with a preference on popular categories such as *People& Blogs*, *Gaming*, *Entertainment*, *Music*, and *Film& Animation*. Based on these two aspects, we conclude that our data set is sufficiently representative.

3. INITIAL POPULARITY FORECAST

In this section, we introduce the methodology to forecast the beginning-day popularity of YouTube videos at the beginning of their uploading. Specifically, we first discuss how to evaluate the features and select a subset of them as predictors. We then train a neural network model based on these predictors and evaluate the performance of the model.

Table 1. SU values of features

len	*cat*	N_f	N_{ac}	N_{am}	N_v	N_{rc}	N_r	*rand*
0.09	0.07	0.30	0.31	0.04	0.17	0.35	0.15	0.01

3.1. Predictor Evaluation

3.1.1. Preprocessing

We first preprocess the features as follows: 1) compute the average view counts N_{ac} and the average comments N_{am} from N_{tc}, N_{tm}, and N_v ($N_{ac} = N_{tc}/N_v$, $N_{am} = N_{tm}/N_v$) for each channel in the data set, in order to reduce correlations between features; 2) compute the average view counts N_{rc} of the related videos in $\vec{r}_i$ from the corresponding $\vec{N}_{rc}$ and replace $\vec{N}_{rc}$ with it, for ease of evaluation; 3) convert the categorical variable cat to a numerical variable by dummy coding; 4) normalize all features to make them comparable.

3.1.2. Correlation Measurement

We proceed to measure the correlations between the features and the target, i.e., the beginning-day view counts N_c. Since it is not safe to assume simple linear correlation between features and the target, we perform correlation measurement by adopting *symmetrical uncertainty* (SU), which is based on the entropy and is able to capture both linear and non-linear correlations [9] . Suppose $H(X)$ and $H(Y)$ are the entropies of X and Y respectively, and $H(X|Y)$ is the entropy of X after observing Y, then the additional information about X provided by Y can be reflected by the information gain as defined in Eqn. (1). The SU value between X and Y is then defined as in Eqn. (2). SU has the range [0,1] with the value 1 indicating that knowledge of the value of X completely predicts the value of Y and the value 0 indicating that X and Y are independent.

$$IG(X|Y) = H(X) - H(X|Y) \tag{1}$$

$$SU(X,Y) = \frac{2IG(X|Y)}{H(X) + H(Y)} \tag{2}$$

We hence compute the SU value for each feature by treating Y as N_c, and X as the feature. For comparison, we also compute the SU value for a random variable. The result is shown in Table 1. It indicates that the most relevant features are the average view counts of related videos N_{rc}, the average view counts of all videos uploaded by the corresponding channel N_{ac}, and the number of followers N_f of the channel. The number of videos that the channel uploaded N_v and the number of related videos N_r have relatively lower correlations with N_c. The length of the video len, the video category cat, and the average number of comments N_{am} that the channel received only slightly outperform the random variable. The predictors for initial popularity forecast can then be selected by a user-defined threshold of SU value.

3.2. Model Construction

Our popularity forecast task can be formulated into a regression problem. We propose to solve this problem by training a neural network model, since we are dealing with a large data set and unstructured features, which can be well handled by neural networks [4].

Specifically, we train a two-layer feed-forward neural network. A training sample is $\{\vec{P}, V\}$, where $\vec{P}$ is a set of predictors and V is the target to forecast. By selecting the SU value threshold t as 0.3, 0.1, and 0.01, $\vec{P}$ becomes $[N_{rc}, N_{ac}, N_f]$, $[N_{rc}, N_{ac}, N_f, N_v, N_r]$, $[N_{rc}, N_{ac}, N_f, N_v, N_r, N_{am}, len, cat]$, respectively. In terms of the hidden layer, we choose the number of neurons from 10 to 30, and find that the best number of hidden neurons when t = 0.3, 0.1, and 0.01 is 18, 20, and 25, respectively. In addition, we divide our data set into the training set, the validation set, and the test set, and each set accounts 70%, 15%, and 15% of the overall videos, respectively.

3.3. Performance Evaluation

To evaluate the performance of the proposed regression model, we consider two online services where initial UGVs popularity is greatly valuable, i.e., in-video advertising and video caching. We aim to measure the general ability of our model in capturing the initially top popular videos. We do not evaluate the absolute video view count for individual videos since the view number of UGVs has extremely broad range and it is also unnecessary in practice to know the exact number of the view counts. We focus on the top 1%, top 2%, top 5%, and top 10% popular videos in the first day because these videos are able to capture the most significant portion ($\geq$ 90%) of the total views in the data set and hence is most meaningful for popularity forecast.

We first investigate in-video advertising service, where a set of popular videos are required to determine in advance. The forecasting performance can then be evaluated by how much the predicted popular videos match the real popular videos. Therefore, we compute the percentage of the intersected videos between the predicted and real top 1% popular videos (i.e., the precision). We repeat the computation for top 2%, 5%, 10% popular videos and show the results in Fig. 4(a). We can see that the best performance of our model is achieved when the SU threshold is set as 0.1. Under such a setting, the predicted top 1%, 2%, 5%, and 10% popular videos are able to capture 49.8%, 55.7%, 60.9%, and 66.2% of the real top 1%, 2%, 5%, and 10% popular videos, respectively. In other words, given a set of just-uploaded videos, the proposed model is able to correctly forecast 50∼65% of the top popular videos without any accumulated popularity statistics at the video uploading moment.

Furthermore, we examine video replication for efficient storage or caching, where the total number of views captured by the replicated/cached contents is the essential factor for the performance. Therefore, it is critical to guarantee that the predicted top popular videos can capture as many views as the ground truth. The percentage of views contributed by different top videos for both ground truth and prediction is shown in Fig. 4(b). We observe that the best performance is also achieved when the SU threshold is 0.1. The predicted top 1%, 2%, 5%, and 10% popular videos can capture 50%, 63%, 78%, 87% of the total views of all videos in the data set. As a comparison, the real top 1%, 2%, 5%, and 10% popular videos can capture 73%, 84%, 92%, 97% of the total views. If we consider all videos in the data set as the whole online UGV pool and cache 1∼10% of all contents using the proposed forecasting model, then the cache hit ratio can reach as high as 70∼90% of the ideal cache hit ratio.

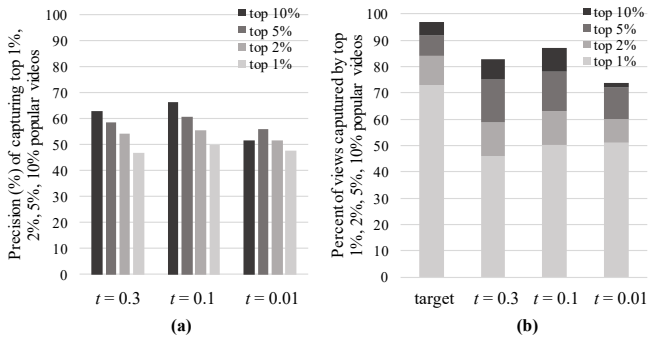

Fig. 4. Performance of the forecasting model

Additionally, the best model performance under SU value of 0.1 demonstrates that the predictors N_{rc}, N_{ac}, N_f, N_v, and N_r play the most essential role in initial popularity forecast. It is interesting to note that video length and video category features that have significant impacts on the long-term popularity patterns [1, 6] do not contribute sufficiently to the beginning-day popularity forecast.

4. CONCLUSION

In this paper, we present the first exploration of initial popularity forecast of user generated videos (UGVs). In particular, we have explored the possibility of using publicly observable features to forecast the beginning-day video popularity at the video uploading moment, based on a large and representative data set. Furthermore, we trained a neural network using the explored predictors to conduct the forecast. The results demonstrate that initial popularity forecast upon uploading moment is feasible and is promising for enriching services such as in-video advertising and video caching. Future work shall focus on investigating deeper into the predictors optimization and further improving the forecasting performance. For example, we can enhance the predictors associated with related videos by considering the impacts of video age.

5. REFERENCES

[1] 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, 2015.

[2] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li and C. Faloutsos, "Rise and fall patterns of information diffusion: model and implications," *ACM SIGKDD*, pp.6-14, 2012.

[3] H. Pinto, J. M. Almeida, and M. A. Goncalves, "Using early view patterns to predict the popularity of youtube videos," *ACM WSDM*, pp.365-374, 2013.

[4] Z. Wang, L. Sun, C. Wu, and S. Yang, "Guiding Internet-scale video service deployment using microblog-based prediction," *IEEE INFOCOM*, pp.2901-2905, 2012.

[5] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," *ACM WSDM*, pp.177-186, 2011.

[6] G. Chatzopoulou, C. Sheng, and M. Faloutsos, " A first step towards understanding popularity in YouTube," *IEEE INFOCOM*, pp.1-6, 2010.

[7] F. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," *ACM WWW*, pp.697-706, 2007.

[8] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon, "Analyzing the video popularity characteristics of large-scale user generated content systems," *IEEE/ACM Trans. Networking*, vol.17, no.5, pp.1357-1370, 2009.

[9] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowledge and Data Engineering*, vol.17, no.4, pp.491-502, 2005.