Predict the Popularity of YouTube Videos Using Early View Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

The goal of the project is to use machine learning techniques to predict the popularity of YouTube videos based on the views in the preceding days. The problem is formulated as the calculation of minimum of the mean relative squared error between the predicted view count and the actual view count. Four methods, namely univariate linear model, multivariate linear model, radial basis functions and a preliminary classification method, are realized in the project. The experiment results show the multivariate linear model combined with category classification achieve the best result, with the mean relative squared error of 0.2014.

1 Introduction

YouTube[1] is a widely used website for uploading, watching and sharing videos. It is claimed that over 4 billion hours of video are viewed each month on YouTube [1]. In the single year of 2011, YouTube experienced more than total 1 trillion views or about 140 views from each person on Earth.

Predicting how much attention of the video will receive on YouTube is of great significance the design and service management. For example, predicting the future video popularity is useful in planning advertisements, and the earnings and costs estimated are relevant to the views on YouTube as well. Acquiring an approximation of the popularity ahead of time enables market strategy adjustments and taking measures to change the popularity. It is also possible to reveal user behaviors and dynamic properties on social system [5].

Recent works have been done in understanding the characteristics of social systems and predicting the popularity using various methods. Szabo and Huberman [2] studied two social systems, and one of them was YouTube. They observed an approximate linearity between future popularity and early view data after log transformation on the data. The property makes the relation independent of the video themselves, and lays the foundation of linear regression. The algorithm is named univariate linear (UL) model in our paper since there is only one major feature. Based on the characteristic analysis in [4], Figueiredo et al incorporated the daily views in the history with different weights into the UL model, and derived a multivariate linear (ML) model [6]. Then radial basis functions were added to the ML model to achieve a further but limited improvement. These techniques have advantages on the simplicity and the accuracy, while lack the considerations on the characteristics of the video. Crane et al [5] build a model on the dynamics of video viewing on YouTube and inspect the endogenous and exogenous bursts of view number. Furthermore, they proposed four popularity evolution patterns, grouped most of the videos based on the entire view distributions. In the project, we applied the regression to the UL model and ML model, implemented the algorithms, and evaluated the prediction results. Finally, the characteristics of the YouTube video system was combined with the ML model to analyze the prediction performance.

[1]http://www.youtube.com

The rest of the paper is organized as follows: Section 2 describes the video view counts and popularity distribution, and formulate the prediction problem. In Section 3, we introduced the existing algorithms and newly proposed models to solve the problem. Experiment results are presented in section 4, followed by conclusions in section 5.

2 Popularity Prediction Problem

YouTube provides the statistical information along with the video on the web, and we focus on the video popularity associated with time. Figure1 shows an example of the view counts of *Gangnam Style M/V*[2], which is the most viewed YouTube video in the category "music". The cumulative distribution of total view number is in Figure1-a. Around 1.5 trillion views have cumulated in the 273 days from July 15, 2012 (the video was uploaded) to April 13, 2013 (the data was sampled). The corresponding daily view count distribution is depicted in Figure1-b.

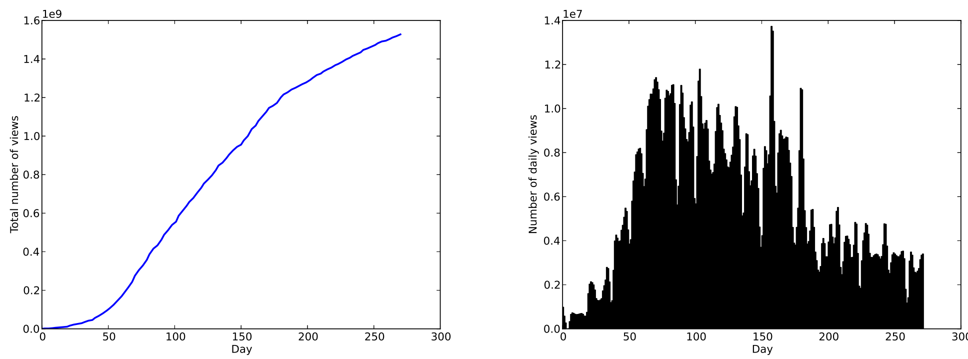

Figure 1: Total views and daily views of the video *Gangnam Style M/V*

To predict the relations between video popularity and the time in a dataset, the description of the data is exhibited. Linearity is always the first option to be examined to describe the property of data and model choice in machine learning. Among a large number of videos, there is no strongly linearity in days and view counts, or between daily view counts and total view counts. An intuition in the popularity growth is that the view patterns are similar for the majority of videos. An example of the total views at day 30 and day 7 are shown in Figure 2. The dataset contains 5652 YouTube videos.

After log transformation, an approximately linear relationship comes out between the total view numbers at two different days [3]. Similar patterns are observed in the view counts of other days. Next, the description of the problem is formulated after introducing the background.

2.1 Problem Formulation

For a YouTube video v, denote the total view count at the date t by $N(v, t)$. Given the daily view counts of days earlier than the reference date t_r (inclusively), the goal is to predict the total view number of the video at the target date t_t in the future. Let $N(v, t_t)$ be the total view number at day t_t, and $\hat{N}(v, t_r, t_t)$ be the predicted total view number. Therefore, $\hat{N}(v, t_r, t_t)$ is a function of the view counts in the first t_r days:

$$\hat{N}(v, t_r, t_t) = \Phi(N(v, 1), \cdots, N(v, t_r)) \quad (1)$$

where $t_r \geq 1$, and $t_t > t_r$.

For different services with regard to the popularity prediction, relative errors are likely to be more important than the absolute errors in the various contents [4]. The relative squared error (RSE) is

[2] http://www.youtube.com/watch?v=9bZkp7q19f0

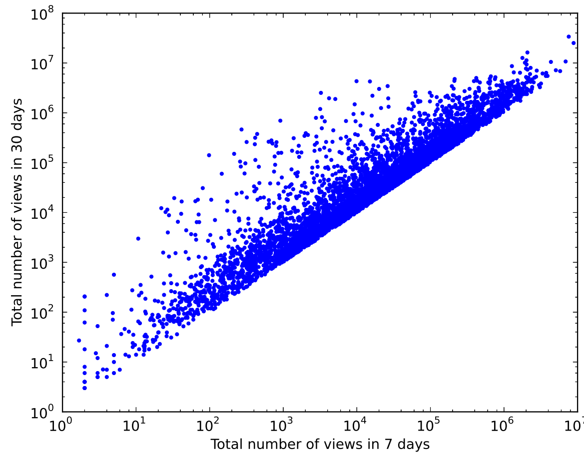

Figure 2: Total views at day 30 versus total views at day 7

used to evaluate the prediction of the total view number:

$$RSE = (\frac{\hat{N}(v, t_r, t_t)}{N(v, t_t)} - 1)^2 \tag{2}$$

In a set of videos V, the mean of the relative squared error (MRSE) of all videos is the criterion of the performance:

$$MRSE = \frac{1}{|V|} \cdot \sum_{v \in V} (\frac{\hat{N}(v, t_r, t_t)}{N(v, t_t)} - 1)^2 \tag{3}$$

3 Methods

In these section, the four methods of video popularity prediction are presented.

3.1 Univariate Linear (UL) Model

The popularity in a future date is strongly correlated with the total view count at the reference date via a log transformation, described by Szabo et al [3]. In the model, only $N(v, t_r)$ is used instead of the total view count from the beginning of the sequence in 1. Then the relation in 1 is:

$$\hat{N}(v, t_r, t_t) = r_{t_r, t_r} \cdot N(v, t_r) \tag{4}$$

Considering a uniformed model for all videos, r_{t_r, t_r} is expected to be independent of the video, and only decided by t_r and t_t. Adopting linear regression in equation 3, the optimal coefficient is:

$$r_{t_r, t_r} = \frac{\sum_{v \in V} (\frac{N(v, t_r)}{N(v, t_t)})}{\sum_{v \in V} (\frac{N(v, t_r)}{N(v, t_t)})^2} \tag{5}$$

3.2 Multivariate Linear (ML) Model

Contrary to the UL model only using the data at the day t_r, the multivariate linear model in [6] thinks that daily views are not equally important in contributing to the views at the day t_t, and gives different priorities to the views in days earlier than t_r. For simplicity, the total view number at the target day is a linear combination of the daily view counts multiplied by weights. Denote the view number in the *i*th day by x_i, and

$$x_i(v) = N(v, i) - N(v, i - 1) \tag{6}$$

The predicted popularity at the target date $\hat{N}(v, t_r, t_t)$ is expressed as:

$$\hat{N}(v, t_r, t_t) = \Theta_{t_r,t_t} \cdot X_{t_r}(v) \tag{7}$$

where $\Theta_{t_r,t_t} = (\theta_1, \cdots, \theta_{t_r})$ and $X_{t_r}(v) = (x_1(v), \cdots, x_{t_r}(v))^T$ are the parameter vector and feature vector respectively. Then the MRSE in equation 3 becomes:

$$MRSE = argmin\frac{1}{|V|} \cdot \sum_{v \in V}(\frac{\Theta_{t_r,t_t} \cdot X_{t_r}(v)}{N(v, t_t)} - 1)^2 \tag{8}$$

The approximate linearity of $\hat{N}(v, t_r, t_t)$ is shown in the UL model, and $N(v, t_t)$ is a scalar. Let $X^*(v) = \frac{X_{t_r}(v)}{N(v,t_t)}$, the criterion is:

$$MRSE = argmin\frac{1}{|V|} \cdot \sum_{v \in V}(\Theta_{t_r,t_t} \cdot X^*(v) - 1)^2 \tag{9}$$

This linear least squares problem is solved via a single value decomposition of the matrix composed of $X^*(v)$ [6].

3.3 RBF Model

Neither UL nor ML model tackles with the variance in the dataset. Only a single set of parameters is used for all videos, and it is weak to follow the different popularity growth patterns. Henrique et al [4] use ML model and radial basis functions (RBFs) together to depict the approximately linear functions, which is formally defined as:

$$\hat{N}(v, t_r, t_t) = \Theta_{t_r,t_t} \cdot X_{t_r}(v) + \sum_{c \in C} \omega_c \cdot RBF_c(v) \tag{10}$$

where C is the set of examples chosen as centers for the kernel and ω_c is the corresponding weight. The Gaussian kernel of the RBF is aimed at capture the variance in behaviors, and the RBF is:

$$RBF_c(v) = e^{(-\frac{||X(v)-X(c)||^2}{2 \cdot \delta^2})} \tag{11}$$

Different from the original, consider ML and RBF features together and then find optimums. To avoid large derivation from ML model results, the residuals of the prediction and real values are the inputs as actual object values. It is a degraded model used in the project compared with the original RBF algorithm. We compare the results with the ML, and only accept when the new error is better than ML model. Another difference is the way of finding RBF parameters. There are no best prior parameters. To solve the problem, they use grid search to find global optimal values. We choose centers of feature vector rather than directly using part of the dataset acting as centers.

3.4 Evolution Model

In this method [5], the theory is based on an one-peak assumption in the daily view distribution. Details are not repeated here. We classify the fraction of peak views into different categories (memoryless, viral, quality and junk), and train models on each category. Theoretical depictions of the four categories are shown in Figure 3. The colors yellow, green, blue and red represent memoryless, viral, quality and junk videos, respectively. The values of views per day in the figure is lees important than the trends of curves. A burst is assumed to happen at the day 50. The obvious difference in patterns could possibly increase the performance of the ML model in the prediction.

4 Evaluation

4.1 Datasets

YouTube API[3] contains the module YouTube Insight, a tool to retrieving statistics data from the video. The API provides video reports including view demographics and view report showing the

[3]https://developers.google.com/youtube/2.0/developers_guide_protocol

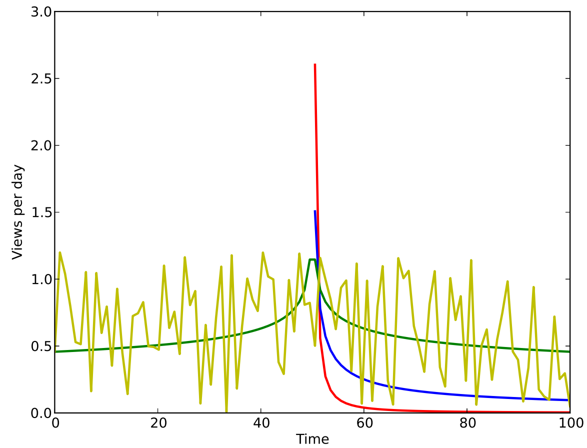

Figure 3: Four patterns in the evolution model

way that viewers interacted with the video. More information is available in the forms such as viewer locations, referrers, comments, favorites and ratings. However, it requires the authentication to access the statistical data of the objective videos or channels.

Google Chart API[4] is a tool that allows creating charts in web pages. For the view count data, YouTube requests the Google Chart API to plot the graph of the statistics using at most 100 points. The API call to generate graphs were intercepted and the view counts data were collected via chart wrappers. Videos with points less than 100 are accurate in the daily view counts, while videos longer than 100 days need interpolation to infer the view count at each day.

After the data collection and pre-processing, all videos are classified into two groups: Top videos and Random Videos [4]. Top Videos include top lists such as most viewed and dominant favorites in terms of region, category, period, etc. They have higher opportunities than other videos to appear in the recommendation list and thus possibly being viewed. 17127 videos exist in this category. While random videos do not have any preference in the dataset generation, and a total 18095 videos belong to the random category.

4.2 Experimental Results

Considering the variability of the data, we adopted cross validation which reduced the model's dependence on the data that were used in the training. The entire dataset was randomly divided into 10 folds of equal size. For each fold $k \in \{1, 2, \cdots, 10\}$, we trained the models on all folds but the kth, and tested on the kth fold. The same process was repeated for 10 times with a different fold as the validation data each time. After separated testing on different folds, the results of the ten tests were averaged as a whole.

In this section, the default reference day is 7, and the target day is 30. Actually the meaning of predicting the 30th data using the previous 29 data points is much less than predicting using only the first 7 days. Figure 4 depicts the predicted mRSE values using the UL and ML models in the top dataset.

Model equations for top dataset

- UL model: $\hat{N}(v, 7, 30) = 1.13857465 \cdot N(v, 7)$
- ML model: $\hat{N}(v, 7, 30) = 1.03362469 \cdot x_1(v) + 1.0388038 \cdot x_2(v) + 1.15533971 \cdot x_3(v) + 1.21999948 \cdot x_4(v) + 1.14799831 \cdot x_5(v) + 1.50634477 \cdot x_6(v) + 1.70576819 \cdot x_7(v)$

[4]https://developers.google.com/chart/

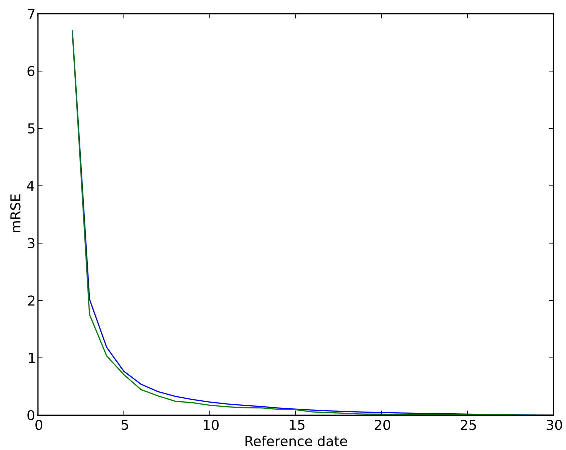

Figure 4: UL and ML Model Prediction Errors

Model equations for random dataset

- UL model: $\hat{N}(v, 7, 30) = 1.45587415 \cdot N(v, 7)$
- ML model: $\hat{N}(v, 7, 30) = 1.04951077 \cdot x_1(v) + 1.22277346 \cdot x_2(v) + 1.22278963 \cdot x_3(v) + 1.41108331 \cdot x_4(v) + 1.64858745 \cdot x_5(v) + 1.76170771 \cdot x_6(v) + 2.41083887 \cdot x_7(v)$

Generally, both models have the mean relative squared error decrease with the increasing day, namely more data is utilized. Compared with UL model, ML model has all daily view information in the days earlier than the reference day, providing the possible choices in the regression. At the left and right end, the two have similar results, but the reasons are not the same. At the first day, both UL and ML model predict based on only the first day data, thus generating the same result. When the data approaches the reference day, ML model attaches more importance to the recent data, which is the entire data the UL model could use. Through all reference dates, ML outperforms UL between the day 3 and day 11, with a 20% error deduction.

The previous two models do not tackle with the variances in the dataset, since only a set of parameters are generated for the entire dataset. Then the radial basis function are introduced to reduce the variance. Compared with the original RBF features chosen in [6], our RBF model is degraded in the choices of centers. The radial basis function provided by scipy[5] is used. For convenience, log transformed view counts were used for calculation and fitting. The RBFs could fit the relative error, but also make a bad impact on the data points with no error by switching to the RBF value. Only 1 dimensional RBF is shown in Figure 5 for good visualbility.

The last method is to classify the video into one of the four categories. The criterion is the fraction of numbers around the peak. The mean of the peak average for the memoryless, viral, quality and junk are 0.04, 0.04, 0.08, and 0.28 respectively. Corresponding intervals are [0.01, 0.21], [0.01, 0.15], [0.02, 0.32], and [0.03, 0.62]. We switch the predicted model based on the category. For a value that lies in several categories, we choose the its difference from the mean as the weight when conducting a weighted sum as the model. Future direction can be K-means classification.

In general, the mean mRSE values for the four methods are 0.3452, 0.2547, 0.2443, and 0.2014. The combination of the ML and four categories classification, namely the 4th method has the best result.

[5]http://www.scipy.org/SciPy

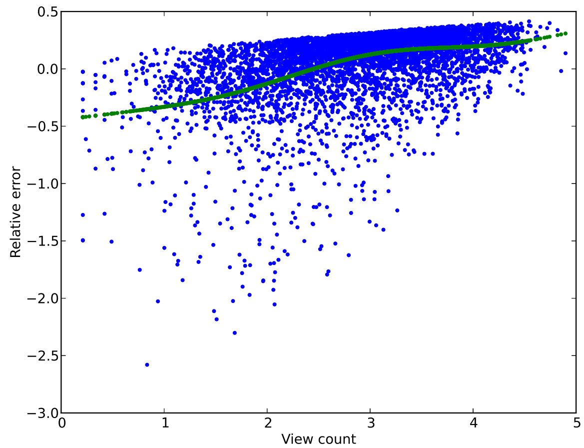

Figure 5: 1-d radial basis function on relative error

Table 1: Four methods performance

_	UL model	ML model	ML model and RBF	ML model and four categories
mRSE	0.3452	0.2547	0.2443	0.2014

4.3 More Considerations

Currently the dynamics are not clear. In most situations, more views lead to more comments, favorites, likes, dislikes. More likes cause more recommendations and views. It is hard to model without figure out relations. Then user behavior, internal mechanism of video recommendations are good areas to explore. Last but not least, no enough time to consider more options without enough data for the project. With more data in YouTube API, more specific questions like the study in terms of region is in [7].

5 Conclusions

In this article we have conducted four methods for predicting the long-term popularity of YouTube videos based on early measurements of view data. On a technical level, an approximate linear correlation exists between the logarithmically transformed video popularity at early and later times, with the residuals. The linearity is solved via linear regression while the residuals are dealt by different methods in the proposed models. ML has a big improvement over the UL model when the reference date is around 1/3 of the target date. Aimed at being more adaptive to the variance in the dataset, RBFs and classification are two methods. The better performance of the classification indicates that the dynamics of the social system plays an important role in the view number. In the future, we could possibly study why some videos are more popular than other videos, and their relations with referrals and recommendations.

References

[1] Statistics-YouTube. http://www.youtube.com/yt/press/statistics.html, 2013.

[2] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning: Data Mining, Inference, and Prediction*, Second Edition. Springer Series in Statistics, 2009.

[3] G. Szabo and B. Huberman. Predicting the popularity of online content. *Communic. of ACM*, 53(8), 2010.

[4] F. Figueiredo, F. Benevenuto, and J. Almeida. The tube over time: Characterizing popularity growth of youtube videos. In *Proc. Conference of Web Search and Data Mining*, 2011.

[5] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proc. National Academy of Sciences*, 105(41), 2008.

[6] H. Pinto, J. Almeida, and M. Gonalves. Using early view patterns to predict the popularity of youtube videos. In *Proceedings of the sixth ACM international conference on Web search and data mining*, 2013.

[7] A. Brodersen, S. Scellato, and M. Wattenhofer. YouTube around the world: geographic popularity of videos. In *Proceedings of the 21st international conference on World Wide Web*, 2012