Web Video Popularity Prediction using Sentiment and Content Visual Features

Giulia Fontanini[*] , Marco Bertini[†] , Alberto Del Bimbo
MICC, Università degli Studi di Firenze
Viale Morgagni 65 - 50134 Firenze, Italy
{*name.surname*}@unifi.it

ABSTRACT

Hundreds of hours of videos are uploaded every minute on YouTube and other video sharing sites: some will be viewed by millions of people and other will go unnoticed by all but the uploader. In this paper we propose to use visual sentiment and content features to predict the popularity of web videos. The proposed approach outperforms current state-of-the-art methods on two publicly available datasets.

Keywords

Video popularity; social networks; visual sentiment; affective computing

1. INTRODUCTION

Video accounts for the largest percentage of internet traffic and this figure is going to increase in the next few years, from 64% in 2014 up to 80% in 2019 [1]. Among US digital video viewers the two most popular sites are YouTube and Facebook [2]; YouTube reports to have 1 billion users and every minute hundreds of hours of videos are uploaded. However not all these videos receive the same attention from the viewers: some will be viewed billion of times, getting tens of millions of views in the very first days[1], while the vast majority will go unnoticed by all but the uploader.

In this context the ability to predict video popularity, i.e. the number of views of a video, is important to guide the design of technical services for video streaming and distribution, and to support economical decision-making processes. Advertisers, online social networks, content providers and content delivery networks are interested in predicting how many views an individual video may obtain, since this impacts on their business. For advertising, the popularity count is associated to the effectiveness of an advertising campaign; for social networks it may impact the algorithm that creates a personalized timeline of the users, as for Facebook; for content providers it is associated with ad revenues, and for content distribution networks, the prediction of access patterns improves the consumption of resources needed to distribute video content [10, 20]. Studying video content popularity can be also useful to help content producers to design new popular content. For these reasons there have been several studies on popularity prediction, typically considering earlier views of a video as predictor of its future views [3, 6, 11, 12, 16, 19], or analyzing characteristics of social networks that lead to video propagation and discovery [15, 18]. However, all these works do not consider the visual content of videos.

In this paper we propose to take into account both content and visual sentiment of videos to predict their popularity; we also propose to improve models based on the number of past views by taking into account different types of view patterns. The proposed approach has been tested on two standard datasets and improves over previous state-of-the-art methods in predicting the popularity of a video, based on the early days view patterns.

Figure 1: Frames of popular videos, from left to right: movie trailer, children performance in a talent show, cartoon.

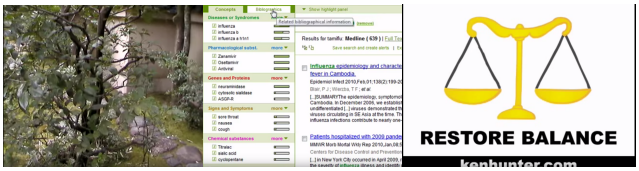

Figure 2: Frames of unpopular videos, from left to right: shooting of nature, presentation of a search engine, political propaganda.

2. PREVIOUS WORK

Crane and Sornette [9, 17] have analyzed the patterns of popularity evolution of YouTube videos, identifying four main classes: the majority of videos, that have no peak in popularity, belong to a "memoryless" class, while those that experience a peak of popularity are classified in "junk"

[*]giulia.fonta@gmail.com
[†]Corresponding author

[1]`http://youtube-trends.blogspot.it/2015/10/adeles-new-single-played-over-1m.html`

ICMR'16, June 06-09, 2016, New York, NY, USA

DOI: http://dx.doi.org/10.1145/2911996.2912053

(i.e. those that have a burst but do not spread through the social network), "viral" (i.e. slow increase of popularity, a burst due to word-of-mouth and then slow decrease) and "quality" (i.e. those characterized by a very sudden peak, due to external events such as being featured in the first page of YouTube, followed by slow decay as the video is shared by users).

Szabo and Huberman [19] have proposed three models to predict popularity based on the number of previous views of YouTube videos, observing that log-transformed long term popularity is strongly correlated to early popularity. The best performing model simply states that future popularity, in terms of video views, at target time t_t is related to the number of views at a reference time t_r (with $t_r < t_t$) by a constant factor α that depends only on target and reference dates, and can be fitted by linear regression.

Figuereido *et al.* [12] have analyzed how the popularity of individual videos evolves since the video's upload time, characterizing the types of the referrers that most often attracted users to each video (e.g. searching or external linking). Analysis has been performed on three datasets of YouTube videos, that have been made publicly available. In [11] Figuereido has proposed the use of K-spectral clustering applied to the time series of video views to predict the popularity patterns of YouTube videos, according to the four classes of [9, 17].

Pinto *et al.* [16] have proposed two models for video popularity prediction: a Multivariate Linear (ML) model that extends the Szabo-Huberman (S-H) model sampling the number of views at regular intervals up to t_r; the ML model is then expanded considering video similarity evaluated using RBFs computed on the same feature vectors of ML (MRBF model). Experiments performed on the [12] datasets show that MRBF outperforms both S-H and ML.

Borghol *et al.* [3] have analyzed the impact of content-agnostic characteristics, finding that for "young" videos user characteristics and keywords become relatively important w.r.t. previous views pattern.

Li *et al.* [15] have addressed the problem of popularity prediction in social networks, analyzing the propagation of videos shared by users. Brodersen *et al.* [5] have shown that social sharing broadens the geographic reach of YouTube videos. Roy *et al.* [18] have proposed a method that uses knowledge from social streams to better predict videos with sudden bursts of popularity.

All these methods do not consider visual data of videos. Conversely, Khosla *et al.* [14] have shown that visual content is a useful feature to better predict the popularity of social image. Gelli *et al.* [13] have shown that adding visual sentiment analysis further improves popularity prediction of social images.

3. THE PROPOSED METHOD

In the proposed method specific prediction models for different popularity trend models are learned, differently from previous approaches [16, 19] that use only one model. Another novelty is the inclusion of visual content in the popularity prediction model. The goal is to forecast the number of views at target time t_t using the information available at reference time t_r, that may be the first day of the video upload, or one of the following days. In particular the goal of the method is to predict the early pattern of popularity, i.e. during the first days of video upload, in which the information related to the number of past views is still very limited or not available.

3.1 Popularity trends model

In the first step of the method is learned a classifier for the 4 different classes of video popularity trends $P_i \in \{P_1, \ldots, P_4\}$ identified in [9,17], using a variation of the method proposed in [11]. Firstly, a feature vector s_v that contains the number of views over time is used to cluster the videos of the training set, using the K-Spectral Clustering algorithm [21]. For each reference day t_r a model for the prediction of the popularity trend, based on Extremely Randomized Trees is learned, using video metadata (e.g. video category, upload date, etc.) and view features (e.g. # of views and # of comments, etc.); differently from [11] no referrer features are used since it is assumed that at the moment of video uploading these are not yet available.

3.2 The popularity prediction model

For each of the reference day t_r and each class of popularity trends $P_i \in \{P_1, \ldots, P_4\}$, a popularity prediction model is learned, combining both video content and early view information.

Visual content representation. Videos are subsampled extracting a frame every two seconds. Each frame is processed to obtain both content and sentiment features using CNNs. Regarding the former we have used the VGG-M-128 network [7], pre-trained on ImageNet dataset, using the 128-D features of the FC7 layer. For the latter descriptor we have used the DeepSentiBank network [8], using the 2089-D vector of the probabilities of visual sentiments expressed as couples of adjective-noun pairs, that are part of the Visual-Sentiment Ontology (VSO) developed in [4]; PCA is applied to this vector to reduce it to 131-D. Finally, the two feature vectors are combined together to represent the frame.

To represent the whole video, the frame feature vector are combined together using Fisher Vector encoding, and performing then a step of dimensionality reduction with PCA. This allows to evaluate visual similarity of videos of different duration and with different order of similar scenes.

Number of previous visualizations. Similarly to [16] we represent the number of previous views of a video at time t_r as a vector of the number of views for each day up to t_r; let $x_i(v)$ be the number of views of video v at day i, the feature vector is $X_{t_r} = (x_1(v), x_2(v), \ldots, x_{t_r}(v))$.

The model. We extend the MRBF model by learning for a specific popularity trend P_i, determined at the previous step of Sect. 3.1, the following model for popularity prediction:

$$\hat{N}_{P_i}(v, t_r, t_t) = \Theta_{(t_r,t_t)} \cdot X_{t_r}(v) + \sum_{v_c \in C} \omega_{v_c} \cdot RBF_{v_c}(v)$$

where $RBF_{v_c}(v)$ is a Gaussian radial basis function that captures the similarity of video v from some training video v_c of the video collection C: $RBF_{v_c} = e^{\left(-\frac{\|\hat{X}(v)-\hat{X}(v_c)\|^2}{2\cdot\sigma^2}\right)}$. $\Theta_{(t_r,t_t)}$ and ω_{v_c} are the parameters to be learned. The first part of the model extends the Szabo-Huberman (S-H) model, where the popularity of a video v at time t_t based on the number of views at time t_r (i.e. $N(v, t_r)$) is $\hat{N}(v, t_r, t_t) = \alpha_{t_r,t_t} \cdot N(v, t_r)$, by evaluating the differences

of number of views over different days up to t_r. The second part exploits the similarity of the video to be analyzed based on a number of "training videos". In our model $\hat{X}$ is Fisher Vector representing visual content, as described above, and v_c videos are selected through k-means clustering, to select visually representative videos from the collection C. In [16] instead $\hat{X} = X_{t_r}$, i.e. the number of previous views, without considering visual content, and v_c videos are randomly selected. The model can be rewritten as:

$$\hat{N}_{P_i}(v, t_r, t_t) = \Theta^*_{(t_r,t_t)} \cdot X^*_{t_r}(v)$$

where Θ^* is the concatenation of Θ with ω_i and X^* is the concatenation of X_{t_r} with the RBF features. The model can be trained with linear or ridge regression. In the following we will refer to this model as PMRBFV, where P is related to the first step of popularity trend prediction and V is related to the use of visual features.

4. EXPERIMENTAL RESULTS

Dataset. For the evaluation, we have used the publicly available Top and Random datasets [12], that contain videos from all world-wide top lists provided by YouTube and random sample of YouTube videos, respectively. For each dataset, we removed videos no longer available in the platform, as well as videos with missing or inconsistent statistical information. We also discarded videos that were in the system for less than 30 days, since we have followed the experimental setup of [16,19], that requires to evaluate popularity prediction on the 30^{th} day. After this skimming, our Top and Random datasets consist of 4,840 and 13,144 videos, respectively.

Experimental setup. The results were obtained using 4-fold cross validation: each dataset was randomly partitioned into 4 complementary and equal-sized subsets. The analysis was then performed using 3 folds as training set and the other one as test set. To reduce variability, the process was repeated 4 times, using a different fold as test set in each round and averaging the results over the rounds.

We used grid search to optimize the four main parameters that are defined for each model:

- the number of Gaussians $\in [16, 32, 64]$, used in the Fisher Vector encoding;
- the number of centroids $\in [10, 50, 100, 200, 500]$, used in the Fisher Vector clustering, i.e. the number of v_c videos used in the RBF features computation;
- the σ values $\in [1, 10, 50, 100, 1000]$, used in RBF features computation;
- the type of regression $\in [linear, ridge]$, used when fitting the models.

Popularity trends prediction. In this experiment we evaluate the performance of the classifier used to predict the pattern of popularity trend, used in the first step of the proposed method.

Table 1 shows average F_1 score performances, for different reference days and for each dataset, resulting in 0.70 and 0.55 F_1 scores, at 7^{th} day, for Top and Random dataset, respectively. It is possible to notice that performances on Random dataset are slightly lower, probably due to the fact that Random videos have an higher variability of patterns, compared to Top videos one, resulting in a more difficult classification.

	Day						
Dataset	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Top	0.57	0.60	0.63	0.66	0.69	0.70	0.70
Random	0.38	0.47	0.50	0.51	0.53	0.54	0.55

Table 1: F_1 score performances, in average on the four patterns, of popularity trends prediction task.

Video popularity and comparison. The proposed method has been compared to the method of Szabo and Huberman (S-H) [19], and the ML and MRBF methods of Pinto *et al.* [16]. Our method has three principal variants, based on the type of visual feature that is used for Fisher Vector encoding: Content and Sentiment models take into account just the corresponding visual content feature type, while Mixed model uses as video content feature the concatenation of the both. Performance is evaluated using mean Relative Squared Error (mRSE), as in [16,19], computed as the arithmetic mean of RSE values for all videos of a collection, where:

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

with $N(v, t_t)$ is the total number of views video v receives up to day t_t ($N(v, 0) = 0$), and $\hat{N}(v, t_r, t_t)$ is the total number of views predicted for v at target date t_t based on data from the first t_r days. Following the experimental setup of [16,19] $t_t = 30$, while $t_r \in \{1, \ldots, 7\}$.

Table 2 shows performances, in terms of mRSE, for the MRBF model [16] and our PMRBFV visual models, for Top and Random dataset respectively. The proposed visual models always outperform the MRBF baseline model: in particular, for Top dataset, visual sentiment features lead to significant decreases in error, w.r.t. the baseline MRBF model, reaching up to **4.33**% reduction on average on the first 7 days, while, for Random dataset, Mixed model decreases in error on average of **7.79**%.

Day	**MRBF**	**Content**	**Sentiment**	**Mixed**
1	0.5071	0.3980	**0.3965**	0.3998
2	0.3831	0.3139	**0.3133**	**0.3133**
3	0.2985	0.2587	**0.2570**	0.2604
4	0.2411	0.2153	0.2143	**0.2126**
5	0.2052	0.1825	**0.1821**	**0.1821**
6	0.1810	**0.1620**	0.1641	0.1635
7	0.1599	0.1453	0.1454	**0.1452**
Mean	*0.2823*	*0.2394*	***0.2390***	*0.2396*

Day	**MRBF**	**Content**	**Sentiment**	**Mixed**
1	0.4329	0.2854	0.2846	**0.2845**
2	0.3606	0.2454	**0.2439**	0.2442
3	0.2963	0.2157	0.2161	**0.2151**
4	0.2461	0.1808	**0.1796**	0.1808
5	0.2093	0.1570	0.1571	**0.1564**
6	0.1847	0.1405	0.1407	**0.1400**
7	0.1614	0.1256	0.1250	**0.1249**
Mean	*0.2702*	*0.1929*	*0.1924*	***0.1923***

Table 2: Mean Relative Squared Error performances of visual predictive models with respect to the baseline MRBF model [16], for the first 7 days: *top)* Top dataset, *bottom)* Random dataset. The lower the figure, the better the performance.

Full comparisons to all the other models are reported in Figure 3 and 4, for Top and Random datasets respectively. For further information, we have performed additional comparisons with two new models, namely PML and PMRBF, which refer to the implementation of the ML and the MRBF models trained for each of the trends types, predicted using the first step of our method. In all the cases the proposed method outperforms the other state-of-the-art ML and MRBF methods.

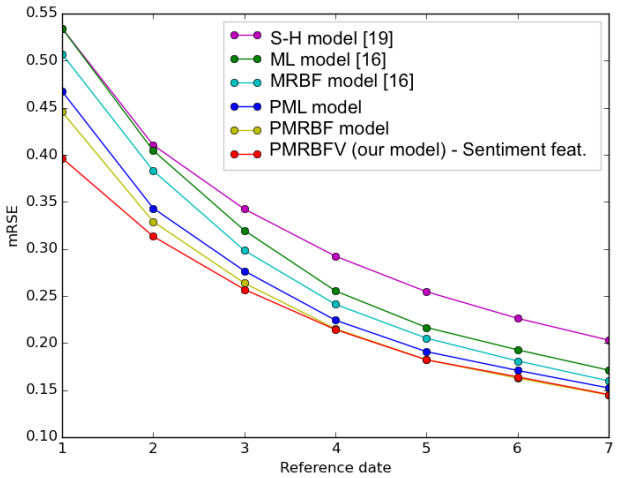

Figure 3: Comparison of all methods performances, in terms of mRSE, for Top dataset.

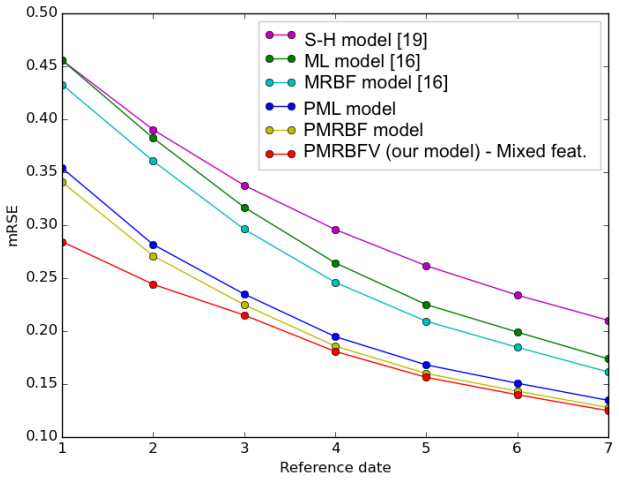

Figure 4: Comparison of all methods performances, in terms of mRSE, for Random dataset.

As the reference date t_r increases, both extensions of ML and MRBF models are improved by the additional steps of popularity trend prediction, as shown by the values obtained by PML and PMRBF. While the value of t_r increases the PMRBF tend to align with the visual PMRBFV models, for both datasets, showing the importance of the first step of our method. Despite this, the PMRBFV visual models remain the best in terms of performance, especially regarding the first few days after the uploading, in which visual features, as expected, are more important than the number of views, that can not provide yet much information.

Acknowledgments. This work is partially supported by the "Social Museum and Smart Tourism" MIUR project (CTN01_00034_231545).

5. REFERENCES

[1] Cisco visual networking index: Forecast and methodology, 2014–2019. Technical report, Cisco, 2015.

[2] Cross-platform video trends roundup. Technical report, eMarketer, 2015.

[3] Y. Borghol, S. Ardon, N. Carlsson, D. Eager, and A. Mahanti. The untold story of the clones: Content-agnostic factors that impact YouTube video popularity. In *Proc. of KDD*, 2012.

[4] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proc. of ACM MM*, 2013.

[5] A. Brodersen, S. Scellato, and M. Wattenhofer. YouTube around the world: geographic popularity of videos. In *Proc. of WWW*, 2012.

[6] 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 Transactions on Networking*, 17(5):1357–1370, 2009.

[7] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *Proc. of BMVC*, 2014.

[8] T. Chen, D. Borth, T. Darrell, and S.-F. Chang. DeepSentiBank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*, 2014.

[9] R. Crane and D. Sornette. Viral, quality, and junk videos on YouTube: Separating content from noise in an information-rich environment. In *Proc. of AAAI Spring Symposium: Social Information Processing*, 2008.

[10] J. Famaey, T. Wauters, and F. De Turck. On the merits of popularity prediction in multimedia content caching. In *Proc. of IEEE IM*, 2011.

[11] F. Figueiredo. On the prediction of popularity of trends and hits for user generated videos. In *Proc. of ACM WSDM*, 2013.

[12] F. Figueiredo, F. Benevenuto, and J. M. Almeida. The Tube over time: Characterizing popularity growth of YouTube videos. In *Proc. of ACM WSDM*, 2011.

[13] F. Gelli, T. Uricchio, M. Bertini, A. Del Bimbo, and S. F. Chang. Image popularity prediction in social media using sentiment and context features. In *Proc. of ACM MM*, 2015.

[14] A. Khosla, A. Das Sarma, and R. Hamid. What makes an image popular? In *Proc. of WWW*, 2014.

[15] H. Li, X. Ma, F. Wang, J. Liu, and K. Xu. On popularity prediction of videos shared in online social networks. In *Proc. of CIKM*, 2013.

[16] H. Pinto, J. M. Almeida, and M. A. Gonçalves. Using early view patterns to predict the popularity of YouTube videos. In *Proc. of ACM WSDM*, 2013.

[17] C. Riley and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. In *Proc. of the National Academy of Sciences*, 2008.

[18] S. D. Roy, T. Mei, W. Zeng, and S. Li. Towards cross-domain learning for social video popularity prediction. *IEEE Transactions on Multimedia*, 15(6):1255–1267, 2013.

[19] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

[20] Z. Wang, W. Zhu, X. Chen, L. Sun, J. Liu, M. Chen, P. Cui, and S. Yang. Propagation-based social-aware multimedia content distribution. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 9(1s):52:1–52:20, Oct. 2013.

[21] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proc. of WSDM*, 2011.