Deriving Latent Social Impulses to Determine Longevous Videos

Qingbo Hu
qhu5@uic.edu

Guan Wang
gwang26@uic.edu

Philip S. Yu
psyu@uic.edu

Department of Computer Science
University of Illinois at Chicago
Chicago, Illinois, U.S.A

ABSTRACT

Online video websites receive huge amount of videos daily from users all around the world. How to provide valuable recommendation of videos to viewers is important for video websites. Previous studies focus on analyzing the view count of a video, which measures the video's value in terms of popularity. However, the long-lasting value of an online video, namely *longevity*, is hidden behind the history that a video accumulates its "popularity" through time. Generally speaking, a longevous video tends to constantly draw society's attention. With a focus on Youtube, this paper proposes a scoring mechanism to quantify the longevity of videos. We introduce the concept of *latent social impulses* and use them to assess a video's longevity. In order to derive latent social impulses, we view the video website as a digital signal filter and formulate the task as a convex minimization problem. The proposed longevity computation is based on the derived social impulses. Unfortunately, the required information to derive social impulses is not always public, which disallows a third party to directly evaluate the longevity of all videos. Thus, we formulate a semi-supervised learning task by using videos of which the longevity scores are known to predict the unknown ones. We develop a *Gaussian Random Markov Field* model with *Loopy Belief Propagation* to solve it. The experiments on Youtube demonstrate that the proposed method significantly improves the prediction results comparing to two baseline models.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: On-line Information ServicesWeb-based services; G.3 [**Probability and Statistics**]: Time series analysis

Keywords

Longevity Evaluation, Semi-supervised Learning, Social Media

1. INTRODUCTION

Contemporary Internet world has witnessed the rise of numerous online video websites. Investors favor a video website with a large number of registered users and daily visits due to the potential advertising value. Therefore, in order to attract viewers, video websites may solicit third parties like search engines to help with video recommendations. Without access to a user's view history, it is difficult for a search engine to recommend high quality videos. In this case, how to evaluate a video's value based on its public information becomes particularly important. Previously, the view count, a.k.a. "popularity", is widely used to assess a video's value. In recent years, video sharing websites such as Youtube start to offer statistical data of how a video accumulates its view count through time. This provides us an opportunity to analyze another perspective of a video's value, which is the long-lasting value, a.k.a. *longevity*. Longevous videos have view count growth patterns that suggest promising long-term value, i.e. the ability to continuously attract society's attention. Once a trend triggers society's attention to a video, people's latency usually differ a lot. Therefore, even though we may observe continuous increase of view count in history, there may be only a few number of trends stimulating such increase. We name such trends as *latent social impulses* and introduce a method to derive them from the video's view count history. Moreover, we also propose a method to quantify a video's longevity based on the derived social impulses and a model to predict it when it is unobtainable. The full version of this paper can be found in [2].

In this paper, we have collected a dataset that contains 65,016 videos and 117,370 edges connecting them (each edge indicates that the two videos are connected by the "related video" relationship assigned by Youtube). Afterwards, we use Youtube API 2.0 to collect 6 public features of all videos: ***accumulated view count, favorite count, average rating, length of the video, "like" count, and "dislike" count***. Moreover, we also apply the same method introduced in [1] to retrieve the view count histories. For each video, the attained result consists of N points denoting the view count increments in different time intervals.

2. SOCIAL IMPULSES AND LONGEVITY QUANTIFICATION

To derive the latent social impulses, we deem the video website as a digital signal filter, which takes the social impulses (denoted by vector **x** = $(x_0, x_1, ...x_N)$) as input and outputs a view count history (denoted by vector **y** = $(y_0, y_1, ...y_N)$). We further define the *Impulse Response Function (IRF)*, denoted by h_*, as the following exponential form: $h_i = e^{-\gamma \cdot i}$, for $i \in [0, N]$, where γ is a tunable parameter. According to the basic knowledge of digital signal filter, the following discrete convolution holds: $y_{i(i=\{0,1,...,N\})} = \sum_{j=0}^{i} x_j \cdot h_{i-j}$. This means that the effect of a previous input signal has a exponentially diminishing effect on the output, and γ controls how fast the effect decreases. Therefore, the problem of

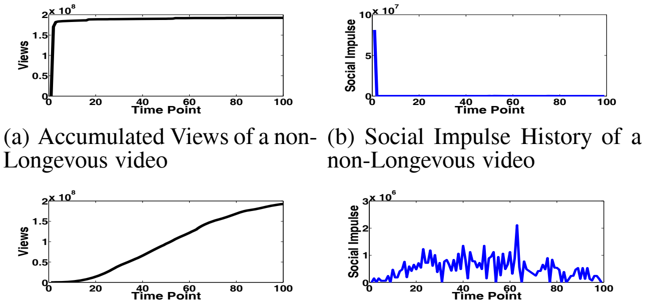
WWW' 14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577287.

deriving latent social impulses is to infer the system's input signals according to the outputs, which can be further formulated as the following least square problem:

$$min_{\boldsymbol{x}} : \sum_{i=0}^{N} \left[y_i^* - \sum_{j=0}^{i} x_j \cdot h_{i-j} \right]^2$$
$$subject\ to: \ x_j \geq 0, \forall j$$

where $\mathbf{y}^* = (y_0^*, y_1^*, ..., y_N^*)$ denotes the observed data. Methods such as gradient descent can easily solve this convex problem.

Longevous videos refer to the videos which can continuously attract society's interest. In this paper, we develop a longevity scoring method by considering two important factors: (i) longevity refers to the videos which have more active social impulses. Thus, instead of incorporating the strength of social impulses in the longevity function, we should use the occurrence of social impulses. (ii) Social impulses occur later are more valuable, because they imply that even after a long time since a video is published, it can still attract viewer's attention. Therefore, we first propose r_i to evaluate the longevity based on the derived $\mathbf{x}$: $r_i = \sum_{i=0}^{N}[\mathbb{I}_{\{z|z\geq\epsilon\}}(x_i) \cdot (1 + log(1+i))]$, where $\mathbb{I}_{\{z|z\geq\epsilon\}}(x_i)$ is an indicator function determining the occurrence of a social impulse at the time i, and ϵ is a small positive number used to filter out noises. $(1 + log(1 + i))$ is a weighting term providing larger values to later social impulses. We use a log function to avoid the weight dominating r_i. In order to simplify the computed scores, we scale r_i to [0,100] only taking integer values: we compute the percentage number of each r_i to the largest r_i and further round the result to the closest integer, a.k.a. $Round(\frac{r_i}{max_i\{r_i\}} * 100)$. Under this definition, the video having a larger score will be more longevous. Figure 1 shows a case study comparing a non-longevous video to a longevous one (when γ = 0.3). The non-longevous video has a score of 13, while the longevous one has a score of 90. Obviously, the view count of the longevous continuously increases and it has more social impulses.



(a) Accumulated Views of a non-Longevous video

(b) Social Impulse History of a non-Longevous video

(c) Accumulated Views of a Longevous video

(d) Social Impulse History of a Longevous video

Figure 1: Case Study of Longevous and non-Longevous Videos

3. PREDICTION OF LONGEVITY SCORES

In real life, the view count history of a video may not be visible for everyone because of the privacy policy. As a result, some videos' longevity cannot be directly quantified by a third party. However, there are two types of data that are always available: public features we have crawled through API and the connections among videos. Therefore, given some videos' longevity scores and all the public information, we can formulate the problem of inferring the unknown scores as a semi-supervised learning task. Inspired by the framework in [3], we propose a *Gaussian Markov Random Field (GMRF)* model to solve it.

Each node (video) is represented as a vector $\mathbf{v_i} = (v_{i1}, ..., v_{im})$ (m is the total number of features), where v_{ik} is the value of the i^{th} node's k^{th} feature. For every edge connecting two nodes, say $\mathbf{v_i}$ and $\mathbf{v_j}$, we assign the following edge weight: $w_{i,j} = \exp\big(-\sum_{k=1}^{m}(v_{ik} - v_{jk})^2/\varsigma_k^2\big)$, where ς_k^2 is the variance of the k^{th} feature. As one can see, the more similar the feature values of two adjacent nodes are, the larger the weight their edge has. Furthermore, let s_i be the longevity score of the i^{th} node, we use the following Gaussian formed node potential function: $\psi_i(s_i) = \exp\big(-\frac{1}{2} \cdot \frac{(s_i-\mu)^2}{\sigma^2}\big)$, where μ and σ^2 are constants that can be estimated by fitting the distribution of longevity scores to Gaussian distribution. The edge potential function assigned to two connected nodes also has a Gaussian form: $\psi_{ij}(s_i, s_j) = \exp\big(-\frac{1}{2} \cdot w_{i,j}(s_i - s_j)^2\big)$. Finally, we have $p(\mathbf{s}|\mu, \sigma) = \frac{1}{Z}\prod_i \psi_i(s_i) \prod_{i,j} \psi_{ij}(s_i, s_j)$, where Z is the partition function. Since there is no parameter needs to be learned in the proposed GMRF, we can focus on the inference step to obtain the unknown longevity scores. Because of the isolated points, the matrix formed solution in [3] will contain singular points that cannot be estimated. To overcome this, we adopt *Loopy Belief Propagation (LBP)* to perform the inference.

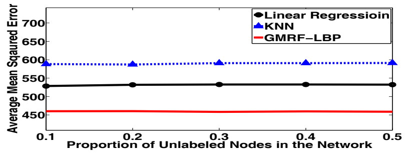


Figure 2: Average MSEs of Predicted Scores

In order to examine the performance of the proposed model, we compare it with two straightforward baseline methods: (1) a linear regression model which uses nodes' features to predict the longevity scores; (2) a K-nearest neighbor (KNN) model to estimate an unknown score by using the known scores of the node's nearest neighbors. We randomly pick unlabeled nodes in the network to predict their scores and further change the proportion of unlabeled nodes to test the robustness of different models. For each model, we run 50 times and report the average mean squared errors (MSEs) of the predicted scores in Figure 2. The proposed GMRF model significantly outperforms the other two baselines.

4. CONCLUSION

In this paper, we have introduced the method of scoring an online video's long-lasting value, a.k.a. longevity. A longevous video tends to keep attracting society's attention. We further introduce the method of deriving latent social impulses and quantifying a video's longevity. Since a third party cannot assess the longevity of all videos due to the privacy policy, in the second part of this paper, we propose a GMRF model with LBP to predict unknown longevity scores. Experimental results show that the proposed model significantly outperforms other baselines.

5. REFERENCES

[1] F. Figueiredo, F. Benevenuto, and J. M. Almeida. The tube over time: characterizing popularity growth of youtube videos. In *Proceedings of ACM WSDM '11*, pages 745–754, 2011.

[2] Q. Hu, G. Wang, and P. S. Yu. Deriving latent social impulses to determine longevous videos. *CoRR*, arXiv:1312.7036, 2013.

[3] X. Zhu, Z. Ghahramani, J. Lafferty, et al. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, volume 3, pages 912–919, 2003.