Modeling the Evolution of User-generated Content on a Large Video Sharing Platform

Rishabh Mehrotra
Department of Computer Science
University College London
r.mehrotra@cs.ucl.ac.uk

Prasanta Bhattacharya
Department of Information Systems
National University of Singapore
prasanta@comp.nus.edu.sg

ABSTRACT

Video sharing and entertainment websites have rapidly grown in popularity and now constitute some of the most visited websites on the Internet. Despite the high usage and user engagement, most of recent research on online media platforms have restricted themselves to networking based social media sites like Facebook or Twitter. The current study is among the first to perform a large-scale empirical study using longitudinal video upload data from one of the largest online video sites. Unlike previous studies in the online media space that have focussed exclusively on demand-side research questions, we model the supply-side of the crowd-contributed video ecosystem on this platform. The modeling and subsequent prediction of video uploads is made complicated by the heterogeneity of video types (e.g. popular vs. niche video genres), and the inherent time trend effects. We identify distinct genre-clusters from our dataset and employ a self-exciting Hawkes point-process model on each of these clusters to fully specify and estimate the video upload process. Our findings show that using a relatively parsimonious point-process model, we are able to achieve higher model fit, and predict video uploads to the platform with a higher accuracy than competing models.

Categories and Subject Descriptors

I.2.6 [**Artitifical Intelligence**]: Learning; J.4 [**Computer Applications**]: Social and Behavioral Sciences

Keywords

Hawkes Process; Video Uploads; Genre Clusters; Popularity Effect; Self-reinforcing Effect

1. INTRODUCTION

In the current study, we depart from existing work by explicitly modeling the supply side of these video distribution and sharing platforms. Moreover, instead of focusing on large general-purpose video sites like Youtube and Vimeo, we instead choose to focus on interest-based video sites like 9Gag.TV and XVideos.com which offer a unique context for us to study both content-level heterogeneity as well as temporal variations in upload numbers. For our empirical analysis, we use data from a large-scale adult entertainment website which is ranked among the top two most frequently visited adult entertainment sites on the internet. Our dataset comprises information about the uploaded video (e.g. title, descriptions etc.) as well as user-generated tags associated with each upload. In order to uncover genre-level patterns from the upload data, we first perform a clustering analysis, forming association clusters based on tag co-occurrence. Thus, videos which have a common set of associated tags are inducted into the same video cluster, implying that each video cluster qualitatively represents a specific genre or taste category. We find evidence for distinct upload patterns in each of the 37 clusters that we uncover from our analysis, supporting our hypothesis that video uploads demonstrate significant rate-heterogeneity depending on the specific video genre. Drawing on this insight, we perform predictive modeling on each of the clusters individually to generate insights for the platform owners. For our model estimations, we employ a parametric self-exciting process model, also termed as a Hawkes process model in literature [3]. Such models provide an elegant and parsimonious extension to the popular Poisson model, by incorporating the history of events into consideration. Hawkes models have been popularly used in recent studies to model natural phenomena like wildfire assessments [7], civilian deaths in conflicts [4], financial settings[2] and even online check-ins on a social media site[1].

Such self-exciting models are ideal candidates for fitting multi-spell events with bursty traffic where there are infrequent spikes in frequency followed by a period of mean reversal when the frequency retreats to its mean value. In our current study, we apply the Hawkes model to each of our 37 identified clusters and obtain parameter estimates that we later use to make predictions. We show that our model fits the data better than comparable variants of the Poisson model that have been used in recent research. Moreover, our model provides lowest average prediction error spanning different splits of the training and test data, as compared to other baseline models. We contend that these findings will increase our understanding of video-based UGC production on online platforms, and will also aid platform owners in provisioning of costly platform resources.

2. MODELING VIDEO UPLOADS

User generated content has become the de-facto form of media publishing on some of the most popular Internet platforms. Our aim in this research is to model the video upload process in crowd-contributed video ecosystems so as to be able to predict future upload volumes. We first explain the process of extracting clusters of videos based on their tag

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2745945.

associativity, following which, we exploit Hawkes Process to model the upload process and predict future uploads.

Graph-based Tag Cluster Formation:
The way a popular category (e.g. funny cat videos) gets flooded with user generated content varies drastically from the content generation process in niche categories (e.g. black hole videos). While videos do not have explicit category/cluster labels, we propose to make use of the associated tag information to form such video clusters. Given a set of videos along with their tag associations, we build a complete graph $G_V = (T, E, w)$, whose nodes T are the set of all tags associated with the set of videos V, and whose E edges are weighted by the tag-tag affinities. The weighting function w is a tag affinity function $w : E \rightarrow \mathbb{I}$ where $\mathbb{I}$ is the set of integers. For each pair of tags, the edge weight is defined as $w(t_1, t_2) = |\{v\}|$ such that $t_1 \in tags(v)$ & $t_2 \in tags(v)$, i.e., the total number of videos in which these two tags co-occur. We define the video clusters as the set of vertex-partitions induced by the connected components of the graph G_V. The rationale is to drop weak edges and to build clusters on the basis of the strong edges which identify the genre-clusters of related videos.

Modeling Video Uploads as a Hawkes Process:
The *weighted connected components* of the tag affinity graph, as defined above, serve as the genre-clusters which represent the different types of video categories one usually observes on online video sharing sites. We treat each such genre-clusters as a separate process and employ a self-exciting Hawkes for each genre-cluster to model the cluster specific upload process. The Hawkes process is a specific class of self- or mutually-exciting point process models [3] and is well known for its self-exciting property, which refers to the phenomenon that the occurrence of one event in the past increases the probability of events happening in the future. Each genre-cluster is treated as a separate Hawkes Process while each video upload in each of these clusters is treated as an event. We model the intensity of video upload events involving a cluster c at time t as follows:

$$\lambda_c(t) = \mu_c + \sum_{p:t_p<t} g_c(t - t_p) \quad (1)$$

This intensity function can be interpreted as a rate at which video-uploads in a cluster occur. The summation in the second term is over all the events (i.e. uploads) that have happened up to time t. μ_c describes the background rate of event occurrence that is time-independent, whereas the second term describes the self-excitation part, so that the video upload in the past increase the probability of observing another upload in the (near) future. We will use a two-parameter family for the self-excitation term:

$$g_c(t - t_p) = \beta_c exp(-w_c(t - t_p)) \quad (2)$$

where β_c describes the weight of the self-excitation term (compared to the background rate), while w_c describes the decay rate of the excitation. Overall, each genre-cluster is defined by three parameters of the Hawkes Process viz. $< \mu, \beta, w >$, representing the upload process as characterized by a particular cluster. The estimates of these parameters were obtained by minimizing the negative of the log likelihood function [6].

Splits	Hawkes	Baseline 1	Baseline 2
15	**31.99**	32.00	33.50
30	**27.76**	31.39	30.96
45	**35.25**	40.60	38.16
60	**37.43**	37.58	38.63256
90	**39.30**	41.90	42.89

Table 1: Performance of the models in predicting total number of video uploads to the site within a future window of 2 weeks. Prediction error rates are presented.

3. EXPERIMENTAL EVALUATION

For our empirical analysis, we use data from a large-scale adult video sharing site[5]. Our dataset comprises an exhaustive collection of metadata from all videos published on a large scale adult video platform, since its creation in April 2007, up until February 2013, totaling over 800,000 videos. The associated metadata consists of an anonymized uploader identifier, video upload date and time, list of uploader contributed video-tags for the uploaded videos, and video popularity cues (e.g. number of views, number of comments etc). We evaluate the performance of our model on the task of video upload prediction. For all genre-clusters obtained experimentally (i.e. 37 in total), we segregate the data into two components, training set and testing set and fit a separate Hawkes Process on the training data and perform MLE to obtain estimates of the model parameters which are used to predict the number of videos that would be uploaded in a given future time frame. The number of events (video uploads) between time interval t and $t+\delta t$ can be computed using the counting process as below ($\delta t > 0$):

$$N(t + \delta t) - N(t) = \int_t^{t+\delta t} \lambda(\tau) d\tau \quad (3)$$

We run two comparable baseline models viz. a piecewise constant nonhomogenous Poisson processes (NHPP), and NHPP with drifting[1]. Our results as illustrated in Table 1 show that the Hawkes model outperforms both of the Poisson based baseline models across all sample sizes. We find that our model is able to predict video uploads to the site with prediction error rates lowest among comparable models used in recent studies.

While prior studies on UGC in general, and videos in particular have focused on modeling demand-side of this ecosystem, we contend that this is among the first studies to analyze the user-generated supply side nature of these video distribution platforms.

4. REFERENCES

[1] Y.-S. Cho, G. Ver Steeg, and A. Galstyan. Where and why users check in. 2014.

[2] V. Filimonov and D. Sornette. Apparent criticality and calibration issues in the hawkes self-excited point process model: application to high-frequency financial data. *arXiv preprint arXiv:1308.6756*, 2013.

[3] A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

[4] E. Lewis, G. Mohler, P. J. Brantingham, and A. L. Bertozzi. Self-exciting point process models of civilian deaths in iraq. *Security Journal*, 25(3):244–264, 2012.

[5] A. Mazières, M. Trachman, J.-P. Cointet, B. Coulmont, and C. Prieur. Deep tags: toward a quantitative analysis of online pornography. *Porn Studies*, 1(1-2):80–95, 2014.

[6] R. D. Peng. Multi-dimensional point process models in r. *Department of Statistics, UCLA*, 2002.

[7] R. D. Peng. *Applications of multi-dimensional point process methodology to wildfire hazard assessment*. PhD thesis, University of California, Los Angeles, 2003.