Abuse Detection and Prevention Systems at a Large Scale Video Sharing Website [Extended Abstract]

Yu-To Chen, Pierre Grinspan, Blake Livingston, Palash Nandy, Brian Palmer
YouTube, Inc.

1. Overview

Abuse on a large-scale website such as YouTube comes in various forms — scraping, email spam, hate speech, or "black-hat" search engine optimization to name a few — and must be fought accordingly. The detection of abusive behaviors uses supervised (machine learning) as well as unsupervised algorithms to mine billions of requests and predict the likelihood of abuse of each. We describe here an unsupervised system scoring incoming traffic for spam, more specifically bot activity, through some simple statistical anomaly detection rules we found to be quite effective.

Each incoming record includes a number of characteristics or *entities*, such as user id or IP address, used to group the traffic into *slices*; for each entity type we have a number of metrics known to correlate with anomalous behavior in that large values, observed over large slices, reliably indicate anomalous behaviour. Each metric, for each slice, produces a score $\beta \in [0, 1]$ which represents a conservative estimate (0 for most slices), based on that metric alone, of the proportion of undesirable (automated) traffic.

We map metric values to scores using the key fact that each metric is a log-ratio; thus the most standard outlier detection method — flagging as suspicious any value x of metric X more than α standard deviations σ above the mean μ_X — naturally extends into a scoring method by yielding a *goodness* score $\gamma = \exp(-\max(0, x - \mu_X - \alpha\sigma))$, then a *badness* (or spam) score $\beta = 1 - \gamma$. The quantity $\sigma = \sqrt{\sigma_X^2 + \sigma_x^2}$ combines two very different sources of noise : the (weighted) standard deviation σ_X of the metric across entities, and the sampling error σ_x on its measurement x.

2. Metric types

These metrics are implemented using Sawzall [3].

2.1. Ratio and Unique metrics

Assume we have a binary feature such that given a slice of N records, the ratio $q = C/N$ of records with that feature on is cause for suspicion if well below its average p; the associated *ratio* metric is $R_C = \log(p/\tilde{q})$, where the sample estimate $\tilde{q}$ of q and its standard error σ_{r_C} can be obtained by standard methods [2]. The resulting spam score is quite intuitive : say that ratio turns out 10 times smaller than it should be, even taking into account the error margin, i.e. $r_C = \mu_{R_C} + \alpha\sigma + \log(10)$; this will result in $\beta = 0.9$.

If we break this slice's traffic by another entity type or *dimension* D (e.g. a user's comments broken down by target video), an interesting extension using the "novelty" feature ("Is this a new video within this slice?") replaces the counter C with the number U of unique items along this dimension, producing the associated *unique* metric U_D.

2.2. Divergence and Concentration metrics

Assume the distribution $\mathcal{P} = \{p_d\}_{d \in D}$ along a certain dimension is largely stationary (at least across its most common values); an extension of the ratio metric idea measures, given the distribution $\mathcal{Q} = \{q_d = \frac{n_d}{N}\}$ within a slice, the divergence from that reference distribution : $KL_D^{(1)} = \sum_{p_d > q_d} p_d \log \frac{p_d}{q_d}$. Of course one can also look for the predominance of usually rare items, leading to $KL_D^{(2)} = \sum_{q_d > p_d} q_d \log \frac{q_d}{p_d}$. In cases (think of the "user" dimension) of an extremely spread out reference distribution ($p_d \ll 1$) we can, in practice, dispense with it entirely and use absolute, rather than relative, entropy, leading to the *concentration* metric $C_D = \frac{\sum_k n_k \log(n_k)}{N} = \log N - H(D)$, for which [1] for example provides useful variance estimates.

References

[1] A. Antos and I. Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19(3-4):163–193, 2001.

[2] L. D. Brown, T. T. Cai, and A. DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, 16(2):pp. 101–117, 2001.

[3] R. Pike, S. Dorward, R. Griesemer, and S. Quinlan. Interpreting the data: Parallel analysis with sawzall. *Sci. Program.*, 13:277–298, October 2005.