

YouTube の動画の反復視聴行動に関する分析

Analysis of Repeated Viewing Behavior of YouTube Videos

芦原 大和

上山 憲昭

Takakazu Ashihara

Noriaki Kamiyama

福岡大学大学院 工学研究科 電子情報工学専攻

Graduate School of Engineering, Fukuoka University

1. はじめに

Web 閲覧や動画配信サービスを快適に行うため、ネットワークの様々な場所にキャッシュサーバを設置しユーザの近くに存在するキャッシュサーバからデータを配信するキャッシュ配信が広く行われている。しかし悪意を持ったユーザが意図的に低人気のコンテンツに多数の要求を行うことでキャッシュの効果を低下させる Cache pollution attack (CPA) の問題が指摘されている [2]。CPA 発生時にはできるだけ迅速に攻撃ホストを特定し、キャッシュを保護することが重要である。

そこで筆者らは、少ないメモリ量とメモリアクセス回数で高精度に攻撃ホストを特定する、ホスト ID とコンテンツ ID の組をキーとする Bloom Filter (BF) を用いた CPA ホストの検知法を提案した [1]。また正常ホストの誤検知を抑制するため BF による検知回数に閾値を設けること、また継続的な CPA ホスト検知を可能とするため BF を 2 つ用意して交互に運用することを提案した [3]。しかし本方式には、検出回数が時間とともに増加していき、いずれほとんどのユーザが誤検出される。本問題に対しては検出回数を周期的に減少させる方策が有効と思われるが、検出閾値や減少レートの適切な設定のためには、本方式は同一ホストが同一のコンテンツを要求することで検知を行うため、正常ユーザが同一コンテンツを繰り返し要求する時間間隔の分析が不可欠である。そこで本稿では、YouTube の視聴履歴の分析を行い、ユーザの反復視聴の時間間隔が指数分布より分散が大きな分布に従うことなどを明らかにする。

2. 分析条件

YouTube は最も人気のある動画コンテンツの一つであり、日々多くの動画コンテンツが視聴および配信されており、モバイルデータ使用量の 38 パーセントを占めている [4]。そこで YouTube の視聴履歴を分析する。筆者らの研究室のメンバーや友達、家族の合計 21 人から Google アカウントに残る YouTube の視聴履歴を収集した。JSON ファイルで出力された YouTube の視聴履歴を個人情報保護の観点から Python プログラムにてコンテンツ名、チャンネル名をハッシュ値に変換し日時を取得した。YouTube の広告はデータから除外し、最長 3 年間の視聴履歴を分析した。

3. 分析結果

全ユーザのコンテンツの要求回数を降順に並べたグラフを図 1(a) に示す。コンテンツの要求数はおおそ、パラメタ $\theta = 0.492$ の Zipf 分布に従うことが確認できる。YouTube の要求数分布はパラメタ $\theta = 0.8$ の Zipf 分布に従うという報告がなされているが [5]、今回取得したデータの総量が少ないことが要因と考えられる。

図 1(b) は各人が 2 回以上視聴した各コンテンツに対し、その人のそのコンテンツの平均視聴間隔に対し、その人の視聴回数を、1 つのグラフに散布図としてプロットする。2 回以上視聴した各コンテンツの平均視聴間隔は広く分布しており、また視聴回数の多いコンテンツの中にも、平均視聴間隔の短いものあれば、平均視聴間隔の長いものもある。すなわち反復視聴現象は、短期間に集中的に行われる場合も、長期間にわたって多数の視聴が行われる場合もある。

同じユーザが 2 回以上視聴した各コンテンツの平均視聴間隔の累積分布を図 2(a) に示す。また平均間隔を一致させた指数分布も合わせて示す。同一コンテンツの同一ユーザによる視聴間

隔の累積分布は指数分布と類似の形状を示すが、その分散は指数分布と比較して大きい。2 回以上視聴した各コンテンツの平均視聴間隔の平均値は 1441.677 時間、最小値は 0.033 時間、最大値は 64347.117 時間であった。

ユーザごとの 2 回以上視聴したコンテンツの要求時間間隔の最小値の累積分布を図 2(b) に示す。多くのユーザの同一コンテンツの要求時間間隔の最小値は 24 時間以内である。

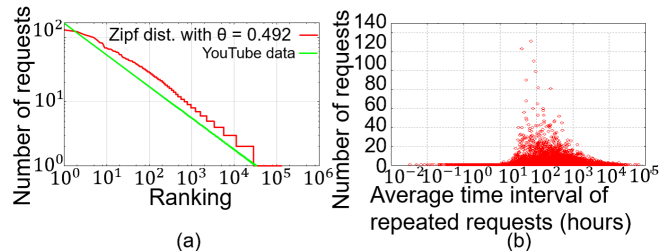


図 1: (a) 動画の要求数分布, (b) 平均視聴間隔と要求数の散布図

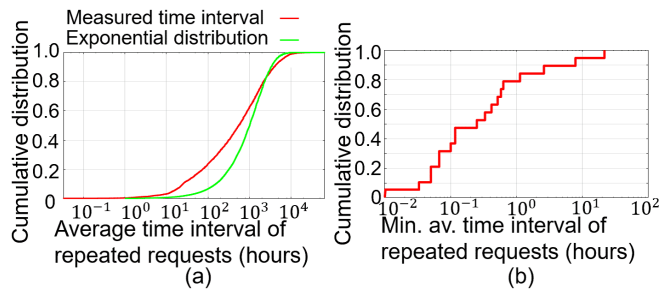


図 2: (a) 平均視聴間隔の累積補分布, (b) 平均視聴間隔のユーザごとの最小値の累積分布

4. まとめ

本稿では、YouTube の視聴履歴の分析を行い、ユーザの同一コンテンツの視聴間隔は指数分布と同様の形状であるが分散がより大きな分布に従うことを明らかにした。

謝辞 本研究成果は JSPS 科研費 18K11283 の助成を受けたものである。ここに記して謝意を表す。

参考文献

- [1] 芦原大和, 上山憲昭, "MEC キャッシュにおけるキャッシュポリューション攻撃の検知", 信学会 NS 研究会, NS2019-108, 2019 年 10 月.
- [2] S. Paul, A. Seetharam, A. Mukherjee, and M. K. Naskar, Investigating the Impact of Cache Pollution Attacks in Heterogeneous Cellular Networks, IEEE ICNP 2017.
- [3] 芦原大和, 上山憲昭, "Bloom Filter を用いたキャッシュポリューション攻撃の検知", ICM 研究会, 2020 年 10 月
- [4] S. Lall, and R. Sivakumar, A YouTube Dataset with User-level Usage Data: Baseline Characteristics and Key Insights, ICC 2020
- [5] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon, Analyzing the Video Popularity Characteristics of Large-Scale User Generated Content Systems, IEEE/ACM ToN, Vol.17, NO.5, pp.1357-1370, Oct. 2009.