動画トラフィックの抽出に関する研究

大倉 圭介 †, 岡村 耕二 ††

† 九州大学 システム情報科学府, 〒 812-8581 福岡県福岡市西区元岡 744
†† 九州大学 情報基盤研究開発センター, 〒 812-8581 福岡県福岡市東区箱崎 6-10-1

A Study on measurement of Video Traffic

Keisuke OKURA † Koji OKAMURA ††

† Graduate School of Information Science and Electrical Engineering(ISEE),
Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 092-802-3600 Japan
†† Research Institute for Information Technology,
Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581, Japan

概要

現在, ブロードバンド環境の普及等により, インターネットにおける動画の視聴によって発生するトラフィック (以下, ビデオトラフィックと呼ぶ) は全トラフィック中でも大きな割合を占めるようになっている. そのため, ビデオトラフィックを把握することは重要である. ビデオトラフィックの把握のために, 全トラフィック中よりビデオトラフィックを抜き出す技術が必要となる. 動画配信サーバの IP アドレスを全て調べるのは困難であるため, 本論文では, IP アドレスの代わりに, ルータの BGP 経路情報を用いて動画配信サーバの IP アドレスから AS 番号を調べ, その AS 番号を用いてビデオトラフィックを抽出する. さらに, その抽出したトラフィック中からノイズを取り除き, さらに精度の高いビデオトラフィックの抽出を行うため, ファイルサイズによるフィルタリング処理を行う手法を提案する. 本論文では, 本手法を九州大学の対外ルータを通過したフローデータに対して適用し, 評価, 考察を行っている. 今回は Youtube からのトラフィックについて解析した.

1 はじめに

現在, ブロードバンド環境の普及等により, インターネットにおける動画の視聴が人気となっている. 動画の視聴によって発生するトラフィックは通常の Web トラフィックよりも大きなトラフィックである. 以下, この動画の視聴によって発生するトラフィックのことをビデオトラフィックと呼ぶこととする. 例えば, Youtube[11] は世界中の視聴者より一日に数百万件を超えるビデオクリップのリクエストを受けている. 北米では Youtube のトラフィックが 2007 年には総トラフィック量の 10 パーセントを占めていた [7]. このように, ビデオトラフィックは総トラフィック中でも大きな割合を占めるようになっている. 現在, ビデオトラフィックがどのような挙動をしているのかを明らかにし, また, 今後どのように変化していくかを予測することによって, 回線の増強等に利用するために, このトラフィックを把握することは重要である.

ビデオトラフィックのトラフィック量は一般的な web トラフィックのそれよりも大きい. その原因は, ビデオクリップファイルのダウンロードである. ビデオクリップファイルのサイズは html ファイルや画像ファイルよりも大きいからである. よって, このダウンロードによるトラフィックを解析することがビデオトラフィックを解析する上で重要となる. しかしながら, ビデオクリップファイルのダウンロードはポート番号８０番を用いた HTTP 通信であるため, パケットのヘッダ情報や, トラフィックパターンを見て HTTP トラフィック中よりビデオクリップファイルのダウンロードによるトラフィックを抜き出すことは困難である. そのため, パケットのペイロード部より HTTP ヘッダ情報を参照することによってビデオクリップファイルを抜き出す手法が用いられている. しかし, ペイロード部の

情報まで保存していくと保存データ量が増加し, 長期間にわたる調査, また, データの保存にはコストが多くかかってしまう. そのため, 本研究ではペイロード部のデータを用いずにビデオトラフィックの抽出を行う.

全トラフィック中より特定のトラフィックを抜き出す研究は, トラフィックパターン等の解析によって行われている [3],[5]. しかし, トラフィックパターンは時々刻々と変化する回線の使用可能帯域等の影響を受けるため, 抽出の精度を高めることが難しい. そこで, 本研究ではビデオクリップファイルのトラフィックの抽出の精度を高めるため, トラフィックパターンではなく, ビデオ提供側の情報を用いる. しかし, 多数ある動画配信サーバの IP アドレスを全て把握することは難しい. そこで, IP アドレスの代わりに AS 番号を用いて判別を行う. そのためには, IP アドレスより AS 番号を調べる仕組みが必要となる. そこで, 本研究ではルータの BGP 経路情報を用いて IP アドレスから AS 番号を調べる. さらに, ビデオクリップファイルのサイズが通常の Web トラフィックのサイズよりも大きいという特性 [2] を用いて, ファイルサイズによるフィルタリング処理を行い, 精度を高めたトラフィック抽出の手法を提案する. 本手法を九州大学の対外ルータを通過したフローデータに対して適用し, 評価, 考察を行う. 今回手法の適用の対象としたのは, Youtube である.

本論文では，第 2 章において，関連研究について述べる．第 3 章では，本研究の提案手法を述べる．第 4 章では，提案手法を Youtube トラフィックに対し適用し, 得られた結果を示し，本手法の評価, 考察を行う．そして第 5 章では, 本研究のまとめと今後の課題について述べる.

2 関連研究

Youtube トラフィックの解析に関する研究として HTTP ヘッダ情報を用いたビデオクリップファイルのトラフィックを抽出し解析する研究 [6] や Youtube の動画配信サーバの IP アドレスのリストを参照してトラフィックの抽出を行い, 解析を行う研究 [2] がある．[6] では www.youtube.com からのパケットをキャプチャし, そのパケットのペイロード部にある HTTP ヘッダのメッセージ情報を取得することにより, ビデオクリップファイルのダウンロード元の IP アドレスを調べ, それ以降に該当の IP アドレスから流れてくるパケットをビデオクリップファイルであると識別することにより, 抽出を行っている. 本論文では, パケットのペイロード部のデータを参照せず, ビデオクリップファイルの抽出を行う. また, その抽出したデータを用いて, ビデオクリップのリクエスト数やビデオクリップのファイルサイズ, ダウンロードにかかった時間等について解析を行っている. [2] では Youtube の動画配信サーバのリストを作成し, 該当 IP アドレスより流れてきたパケットをキャプチャし, そのデータをビデオクリップファイルとして扱う. そのデータを用いてファイルサイズや, 通信のビットレート等の解析を行っている. 本論文では, 動画配信サーバの IP アドレスではなく AS 番号に着目して抽出を行う. また, 特定のトラフィックを抜き出す研究として [3], [5] などがある. 特定のトラフィックを抜き出す研究には主に障害検知や, 不正アクセスの発見を目的としたものが多い. [3] はウェーブレット変換を, [5] は文字列解析を用いてトラフィックの抽出を行う. [3], [5] ではトラフィックパターンに着目しているが, 本論文では, トラフィックパターンに加えて経路情報を参照することにより, 抽出の精度を高めている.

3 AS 番号を用いたビデオトラフィック抽出法

3.1 AS 番号

AS(Autonomous System) とは，共通のポリシーや同じ管理下におかれているルータやネットワークの集合のことである．インターネットでは，AS 単位に分割して管理することでネットワークの管理を容易にしている．一つの AS には複数のホストが接続されており，AS 間と AS 内とでは別々に経路の制御がなされている．AS には，それぞれの AS を一意に識別するために AS 番号という 16bit の識別子が割り当てられている．AS や IP アドレスは ARIN[8] や，APNIC[9] など，RIR（Region Internet Registry）と呼ばれる組織が管理しており，これらの組織がそれぞれ持っている WHOIS データベースに，ユニークな AS 番号と共に所有組織名や国情報等が保持，運用されている．

3.2 BGP 経路情報

BGP(Border Gateway Protocol) は，AS 間の経路情報を交換するためのプロトコルである．BGP

の経路情報には，次にパケットを転送すべきルータのアドレスや，宛先に到達するまでに経由したAS 番号のリスト（AS パス）が含まれており，通常 AS パス長の短いものを最短ルートとして使用する．また，BGP にはパス属性と呼ばれるパスに関する情報が含まれており，パス属性を変更することで，より柔軟な経路制御を行うことができる．経路情報には，AS ごとに AS パスが割り当てられているのではなく，プレフィックスという単位ごとに AS パスが割り当てられている．プレフィックスとは，ネットワークアドレスとネットマスクから成り，一群のネットワークアドレスを指す．経路選択の際には，宛先ネットワークアドレスを含むプレフィックスを検索することで，それに対応する AS パスを知ることができる．また，BGP では到達可能な AS への経路情報を隣接ルータで交換しあうことで，到達可能な全てのプレフィックスへの経路情報を作成する．

3.3 ビデオトラフィック抽出法

本論文では，ビデオクリップファイルのダウンロードによるトラフィックを全トラフィック中よりAS 番号とビデオクリップのファイルサイズによるフィルタリング処理を行うことによって抽出を行う手法を提案する．

本手法ではフローデータを扱う．動画配信サーバのネットワークにはトップページと同じ AS 番号が割り当てられていると仮定する．そのため, フローの送信元 IP アドレスがその AS 番号に属していれば, ビデオトラフィックであると判別することができる．IP アドレスが AS に属しているかどうかは, 収集した BGP 経路情報のパス属性を調べることで判別することができる．BGP 経路情報のパス属性により, 対象の AS 番号に属するネットワークアドレスを調べることができる．そして, そのネットワークアドレスに属する IP アドレスのリストを作成する．調べたいフローの送信元 IP アドレスがそのリスト内にあればビデオトラフィックであることが判別できる．

しかし, AS 番号によって抽出したビデオトラフィック中には目的のビデオクリップファイルの他に html ファイルや, 画像ファイル等のダウンロードによるフローも存在する．これらその他のフローは動画の視聴ではなくても発生するトラフィックであり, ビデオトラフィック中の誤差であるため, これらを取り除きビデオクリップファイルのフローのみにする必要がある．本手法ではビデオクリップファイルがその他のファイルよりも大きいという特徴を利用し, ファイルサイズに閾値をもうけることで, 閾値以下のファイルサイズのフローを取り除くフィルタリング処理を行う．また, この閾値の決定を行うために, 閾値を昇順にとり, その際のビデオトラフィックのフロー数と総ファイルサイズの変化を調べる．この変化を調べることによって, 閾値を決定する．

本手法のフローチャートを図 1 に示す．

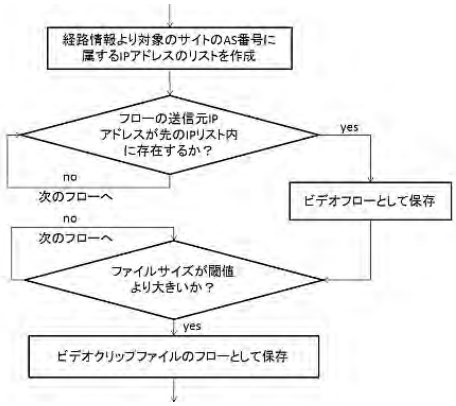

図 1: フローチャート

本手法では, トラフィックのパケットヘッダ情報を使用するため, 収集したトラフィックデータ統計情報はペイロード部の情報を用いるよりも小さくてすむ. そして, IP アドレスではなく AS 番号を用いることで, 動画配信サーバの IP アドレスの収集をする必要がない.

4 AS 番号を用いたビデオトラフィック抽出法の評価, 考察

本章では，AS 番号を用いたビデオトラフィック抽出法の評価, 考察を行う．

4.1 AS 番号を用いたビデオトラフィック抽出法の評価

本手法の評価には, サンプリングレート１０分の１で収集されている九州大学のフローデータを用いる．これは, AS 番号 2508 番（九州大学）のルータから抽出されたフローデータである．さらに, 収集されている九州大学のルータの BGP 経路情報も利用する．その中でも今回は 2007 年 10 月 14 日から 10 月 20 日までの一週間のデータを用いる．本

Date	全 HTTP フロー数	Youtube フロー数
10/14	16047844	12206
10/15	16691047	19805
10/16	17354265	21156
10/17	17778101	19567
10/18	19196725	21197
10/19	18147005	21818
10/20	17313213	12837

表 1: 1 週間のフロー数

論文では, Youtube での動画の視聴によって発生したトラフィックを抽出する．表１は, 14 日から 20 日までの全体の HTTP フロー数及び Youtube トラフィックのフロー数を示している．Youtube トラフィックの抽出には AS 番号によるフィルタリングを行っている．

次に, Youtube トラフィック全体からビデオクリップファイルのフローのみを取り出すためのファイルサイズの閾値を第３章で述べた手法によって決定する．今回は閾値を１００バイト刻みで増加させ, フロー数, ファイルサイズの総計を見た．図２は閾値を変化させた際のフロー数の変化を, 図３はファイルサイズの総計の変化を表している．

図２において，０から１００キロバイトの閾値の領域でフロー数の変化がゆるやかになっている．その値域をより詳細に調べるため，１０バイト刻みで閾値を変化させた (図４)．図４より，１５キロバイト付近からフロー数の減少がゆるやかになっているのが分かる．よって閾値を１５キロバイト近辺にとることで, ビデオクリップファイル以外のフローは取り除かれると推定できる．

また, ファイルサイズの総計についても減少の幅が変化する点を詳細に調べるため，１０バイト刻みで閾値を変化させた (図５)．図５より, 閾値が 15 キロバイト付近でファイルサイズの総計の減少がゆるやかになっている．これは先ほどの図４から見てとれる結果と一致している．

よって, この手法におけるファイルサイズによるフィルタリングの閾値は 15 キロバイトであると推定できる．この閾値はビデオクリップファイル以外のファイルの最大値を示している．一般的に html ファイルよりも画像ファイルの方がサイズが大きいと考えることができるので, このファイルの最大

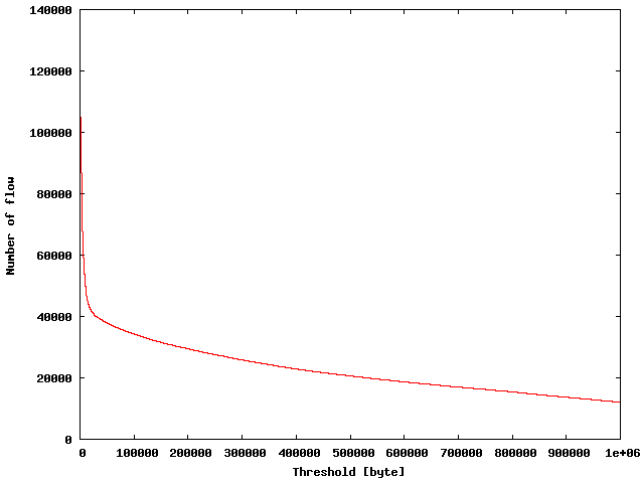

図 2: 閾値とフィルタリング後のフロー数

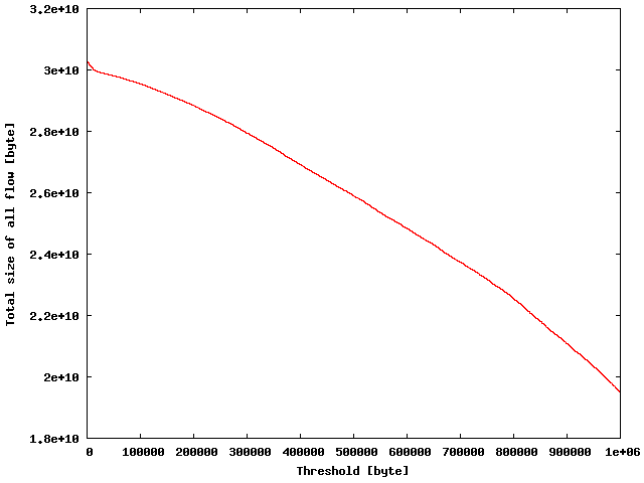

図 3: 閾値とフィルタリング後のファイルサイズの総計

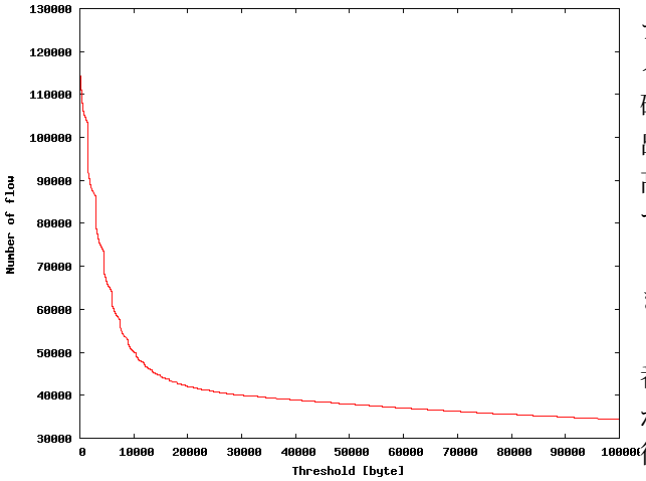

図 4: 閾値とフィルタリング後のフロー数 (詳細)

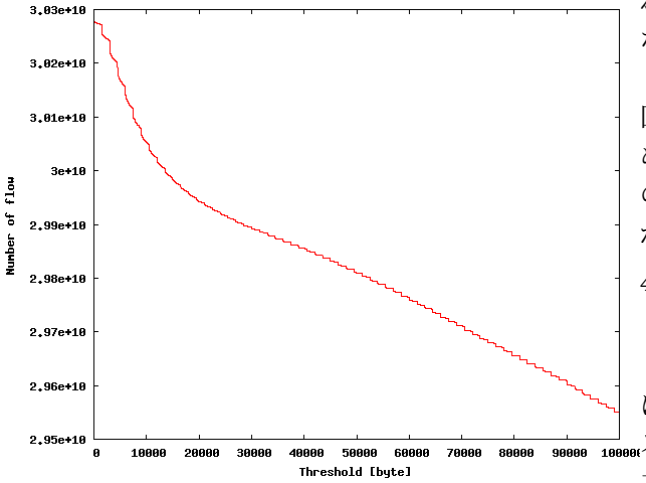

図 5: 閾値とフィルタリング後のファイルサイズの総計 (詳細)

値は画像ファイルの大きさであると推察できる. さらに, 15 キロバイトという大きさは Youtube の動画視聴ページの画像ファイルの大きさとしては妥当である. 15 キロバイトよりもサイズの小さなビデオクリップファイルや, 15 キロバイトよりもサイズの大きな画像ファイル等が存在しないという確証はないため, ビデオクリップファイルのみの抽出はできていないと考えられるが, この閾値はより高精度のビデオクリップファイルの抽出を実現していると推定できる.

閾値 15 キロバイトでフィルタリングを行ったとき, フィルタリング後のフロー数は約 45000 であり, 全フロー数が約 115000 であった. これは AS 番号のみで抽出した場合, フロー数に関しては誤差が大きいことを示している. また, フィルタリング後のファイルサイズの総計は約 30 ギガバイトであり, 全フローのファイルサイズの総計は約 30.03 ギガバイトであった. これは AS 番号のみで抽出した場合, ファイルサイズの総計に関しては誤差が少ないことを示している. よってファイルサイズによるフィルタリングはフロー数の解析の際には有効であると考えられる.

今回は Youtube からのトラフィックに対して解析を行ったが, 他の動画共有サービスによるトラフィックに対して本手法を適用した場合, 今回得られた閾値 15 キロバイトとは異なると考えられる. なぜならば, 動画視聴ページの仕様は異なっており, それによって発生する Web トラフィックも今回とは異なったサイズを持つことが考えられるためである. そのため、今後は別の動画トラフィックの仕組みを調べ、同様に手法を考案していく必要がある。

4.2 AS 番号を用いたビデオトラフィック抽出法の考察

本手法を適用できるかどうかは, 動画配信サーバにトップページと同じ AS 番号に属する IP アドレスを割り振っているかどうかに依存する. 表 3 に, 主要な動画共有サービスの動画配信サーバの AS 番号の一致不一致の一例を示している.

本手法は HTTP ヘッダ等のパケットのペイロード部の情報を用いないため, トラフィックデータの保存のための記憶容量は HTTP ヘッダ等のパケットのペイロード部の情報を用いる手法に比べて少くなる. 本論文で使用したデータでは, データサイ

動画共有サービス名	AS の一致不一致
ニコニコ動画	一致
Veoh	不一致
FC2 動画	一致
Google video	一致
My Space Video	不一致

表 2: トップページと動画配信サーバの AS 番号の一致不一致

ズの総計は 1 週間で約 30 ギガバイトであった．また, フローデータはサンプリングレート１０分の１でサンプリングされているため, 実際には 300 ギガバイト程度収集されていると推測される．パケットのペイロード部の情報を取得するためにフルパケットで 1 年間トラフィックデータを収集したとすると, 単純計算で約 15 テラバイト程度の記憶容量が必要となる．本手法で扱ったフローデータはペイロード部をキャプチャしていないため, 1 週間でのデータサイズは約 2 ギガバイトである．サンプリングをしなかったとすると, 約 20 ギガバイトであると推測される．また, 1 週間で収集されたルータの BGP 経路情報は約 20 ギガバイトである．よって, 1 週間あたり 40 ギガバイト, 1 年間では２テラバイト程度の記憶容量を必要とする．よって, 本手法ではトラフィックデータの収集, 保存にかかるコストを減少させることが可能である．

また, ルータの経路情報を収集をしていれば, 過去に収集されたフローデータに対して適用することができる．このため, 時系列でのフローデータの解析をすることが可能である．エンド側のネットワークにおいて時系列でのビデオトラフィックの増加の傾向を把握できれば, 対外接続の容量の増強などのネットワーク管理に利用することができる．

5 おわりに

今回, 本論文ではビデオトラフィック抽出のためのフローデータに対する AS 番号及びファイルサイズによるフィルタリング処理を行う手法を提案した．そして, 収集された九州大学のフローデータに対して本手法を適用することによって本手法の評価, 考察を行った．その結果, ファイルサイズのフィルタリングの際の閾値は 15 キロバイト付近をとるとビデオクリップファイルを取り出すことができると推定することができた．本手法の AS 番号によるフィルタリング処理のためには, 動画配信サーバがトップページの AS 番号に属する IP アドレスを持っていなければならないという条件が存在する．また, 本手法はルータの BGP 経路情報が保存されていれば, 過去のフローデータに対しても適用が可能である．

今後の課題として, 推定した閾値が適切かどうかを確かめるため, ビデオトラフィックを作成し, ビデオクリップがいくつあるかを確実にした状態での手法の適用を行い, その正誤率を調べる．また, ファイルサイズより精度のよいファイルタリングのための要素を発見することも課題である．そして, 過去のフローデータに対しても適用できるという利点を生かし, 蓄積されているフローデータに対し本手法を適用し, 抽出したフローデータを解析し, ビデオトラフィックの増加の傾向を調査する．

謝辞

論文執筆にあたりまして、御助言を下さった IC 査読委員、並びに QGPOP の皆様方に深く感謝致します．

参考文献

[1] Joachim Charzinski, “ HTTP/TCP connection and flow characteristics, ”Performance Evaluation, vol. 42, pp. 149-162, 2000.

[2] Phillipa Gill, Martin Arlitt, Zongpeng Li, Anirban Mahanti “ Youtube traffic characterization: a view from the edge ”Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, October 2007 .

[3] ウェーブレット解析を用いた周波数成分変化に基づくインターネット 脅威検出法, 石黒正揮, 鈴木裕信, 村瀬一郎, 暗号と情報セキュリティシンポジウム 2006, 2006.

[4] K. Lan, J. Heidemann, ”On the correlation of Internet flow characteristics.” Tech. Rep. ISI-TR-574, USC/Information Sciences Institute, July 2003.

[5] 岡部正幸，三輪多恵子，梅村恭二, “ 文字列解析に基づくネットワークトラフィックデータか

らの異常発見 ”，インターネットカンファレンス，pp.67-74 October 2006.

[6] Michael Zink, Kyoungwon Suh, Yu Gu and Jim Kurose, ”Watch Global Cache Local: YouTube Network Traffic at a Campus Network - Measurement and Implications”, Proceedings of the SPIE, Volume 6818, pp. 681805-681805-13 (2008).

[7] Ellacoya Networks. ”Ellacoya Data Shows Web Traffic Overtakes Peer-to-Peer (P2P) As Largest Percentage of Bandwidth on the Network”, June 2007.

[8] American Registry for Internet Numbers (ARIN), http://www.arin.net/.

[9] Asia Pacific Network Information Centre (APNIC), http://www.apnic.net/.

[10] Resource IP Europeens Network Coordination Centre (RIPE-NCC), http://www.ripe.net/.

[11] http://www.youtube.com