フローの応答特性に着目したWEBコンテンツ識別手法の提案

近藤 泰彦† 中村 康弘†
†防衛大学校情報工学科

1 はじめに

WEBを介したストリーミングメディアの配信やSSLによる暗号化通信が一般化するに伴い，組織ネットワークのトラフィック管理が複雑，困難化してきている．このため，通信パケットのペイロード部の分析を行わず，そのパケット長と到着間隔等の情報を用いてアプリケーションプロトコルを識別する手法が提案されている．

しかしながら，HTTPのように同一プロトコル上で様々な種類のコンテンツを転送するような場合には，その識別は難しい．

そこで本研究では，通信パケットの送受信応答特性に着目することにより，同一プロトコル上のコンテンツ識別を行う手法を提案する．また，実際のHTTP通信に適用してその識別能力を評価する．

2 関連研究

通信の暗号化やポート番号の偽装への対策として，フローのペイロードの分析を行わず，パケット長やパケット到着時間に着目して通信内容を解析する研究がある．

例えば，予め収集したサイト毎の送受信されるデータサイズやパケット到着時刻を用いて，接続先サイトを推定する研究 [1] や，パケット長，到着間隔を用いてアプリケーションプロトコルを識別する研究 [2] 等がある．

いずれの研究においても，同一プロトコルを使用するアプリケーションや転送されるコンテンツの識別までは至っていない．しかし，ファイアウォールのポリシー設定で許可されることが多いHTTPは，YouTube[3]に代表されるストリーミングメディア等の利用が活発で，トラフィック量の増大が顕著である．また，SSLを用いた通信内容の暗号化も多用される傾向にある．

このようなトラフィック管理の困難化及び複雑化に対して，ペイロードを分析せずにコンテンツを識別する手法が求められていると考える．

†Yasuhiko Kondo †Yasuhiro Nakamura
†Department of Computer Science, National Defense Academy

3 提案方式

3.1 フローの応答特性

一般にWEBブラウザにおいてコンテンツを描画する際は，そのコンテンツのMIMEタイプに応じたコーデック，プラグイン及びアプリケーションが用いられる．つまり，WEBコンテンツを表示するクライアント側において，コンテンツ毎の描画処理は異なり，リソースの消費や負荷も異なる．

本研究では，その差異をより明確にするために，クライアント側でのフローの応答特性に着目した．ここで応答特性とは，ある長さを持ったパケットを受信または送信した後に，次にパケットを送信するまでの時間間隔およびそのパケット長のことを表す．本手法はサーバ側での負荷や，途中経路上のネットワークの混雑等を反映しにくいため，より正確な識別が期待できる．

3.2 識別手法

フローの応答特性に着目するフロー情報収集及び分析手法の概要を図1に示す．

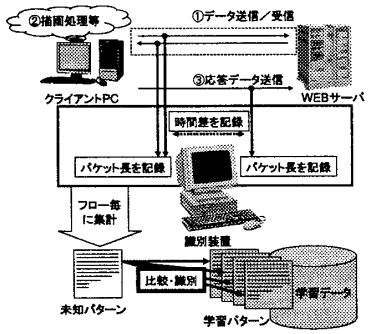

図 1: フロー情報収集及び分析図

具体的な手順を以下に示す．

1. 送受信される全トラフィックをフロー毎にパケットの向き，長さ及びパケット到着時刻を記録する
2. クライアント側からの送信パケットを確認したら，そのフローにおける直前のパケット長 (送信は +，受信は − で表す) と時間間隔を計測する
3. (送信パケット長，直前のパケット長，時間間隔) の組をフロー毎に集計する

4. 同様の手法で予め集計されたコンテンツ毎の学習パターンと比較し識別する

上記手法で収集される YouTube[3] のプロット図の例を以下に示す．ここで，時間間隔 *interval time* は対数スケールに変換している．

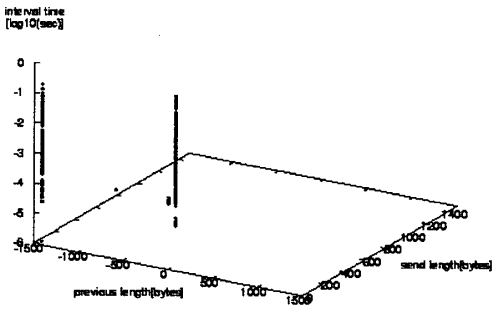

図 2: YouTube プロット図

識別には単純類似度法を用いた．つまり，図 2 のような各コンテンツ毎のプロット図をある時間間隔毎の濃度分布に変換し，それを学習パターンとして未知コンテンツとの類似度を求め，識別の評価尺度とする．この類似度が最も大きい学習パターンを，未知コンテンツのコンテンツと識別する．

4 評価と考察

評価には図 3 のような環境を用意した．

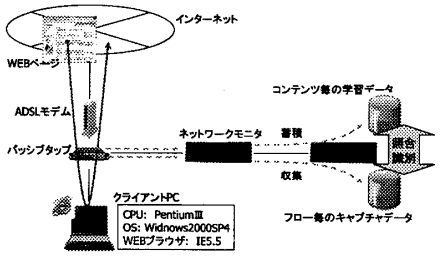

図 3: 評価環境

今回の評価には代表的なストリーミング動画サイトである，YouTube，ニコニコ動画 (RC2)[4]，Veoh[5]，Stage6[6] の各コンテンツ及びファイルのダウンロードを用いた．これらのコンテンツは，MSS(Maximum Segment Size) の長さを持つフラグメントされたパケットが連続して受信されるのみで，フローの挙動が酷似している．ここでは，それぞれ任意に 50 コンテンツ再生又はダウンロードを行い，学習パターンを得た．

続いて，それぞれのサイトにおいて任意のコンテンツを 50 回再生又はダウロードし，どのサイトにおけるコンテンツかを識別試験した．結果を表 1 に示す．

表 1: 識別試験結果

コンテンツ	識別率	平均類似度
YouTube	90%	0.991
ニコニコ動画 (RC2)	84%	0.983
Veoh	38%	0.986
Stage6	54%	0.967
ダウンロード	100%	0.992

YouTube 及びニコニコ動画 (RC2) は同じプラグインの Adobe Flash Player で再生されるが，主に再生されるビットレートの違いを反映し安定して識別している．

Veoh 及び Stage6 はそれぞれ専用のプラグインを必要とするため安定した識別が期待されたが，Veoh に関しては YouTube に学習パターンが酷似したため，Stage6 に関してはコンテンツ毎にビットレート及び動画サイズが異なるため，それぞれ高い平均類似度を得たものの識別率は低い．学習パターンの生成に更なる精度及び工夫が求められる．

また，ダウンロードコンテンツは動画コンテンツに誤識別することはなかった．

5 まとめ

本稿では，ペイロードの分析を行わずに，フローの応答特性のみを用いてコンテンツを識別する手法を提案し，実際の HTTP 通信を用いた実験により本手法が有効であることを確認した．本手法は

・コンテンツ毎の表示処理の差が明確

・LAN 外のトラフィック状況に左右されにくい

等の特徴を有しており，同一プロトコル上のアプリケーション及びコンテンツの識別を可能にしている．

参考文献

[1] G.Bissias, M.Liberatore, D.Jensen, and B.Levine. Privacy Vulnerabilities in Encrypted HTTP Streams, 2006.

[2] 北村強, 静野隆之, 岡部稔哉. フロー挙動分析技術に基づくアプリケーション識別手法, 2005.

[3] YouTube. http://www.youtube.com/.

[4] ニコニコ動画 (RC2). http://www.nicovideo.jp/.

[5] Veoh. http://www.veoh.com/.

[6] Stage6. http://www.stage6.com/.