L-022

クライアントのフロー挙動に着目したP2Pファイル共有ソフト識別手法の提案

近藤　泰彦†　　中村　康弘†
†防衛大学校情報工学科

1　はじめに

Winny 等の P2P ファイル共有ソフトを介したウィルス感染や情報漏洩事案が多発しており，企業や官公庁を問わず問題となっている．しかし，これらのソフトではポート番号の変更や通信の暗号化等を行うためフローの識別が困難であった．

ペイロード部のマッチングを行わなずに，プロトコルやアプリケーションを識別する研究 [1] があるが，同一プロトコル上を流れるコンテンツ及びアプリケーションの推定や，P2P のような複雑な挙動を示すアプリケーションの識別においては充分ではない．

そこで，LAN 内のクライアントの送受するフロー挙動に着目する．アプリケーション毎のフローの特性に基づき学習データを作成し，これとの類似度を求めることで識別する手法を提案する．HTTP 上を流れる動画コンテンツの識別については既に検討が行われており、同一プロトコル上でのコンテンツ識別が可能であることが示されている [2]．本研究では，さらに P2P ファイル共有ソフトの複雑なトラフィックを識別する手法を提案し，その有効性を確認した．

2　アプリケーション識別手法

パケット長やパケット到着間隔等の物理的な情報を用いて，アプリケーションを識別する手法がある [1]．

図 1 は SSH 及び HTTPS のフローのパケット長を時系列上にプロットしたものである．また，パケット長と到着間隔の対応を図 2 に示した．

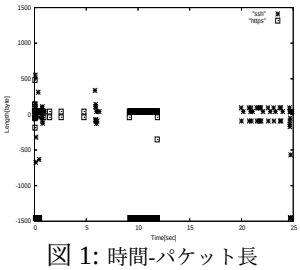

図 1: 時間-パケット長

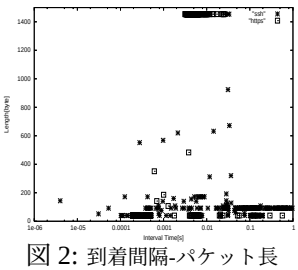

図 2: 到着間隔-パケット長

この例では，パケット長の分布だけを用いても，全トラフィックがこの二種のみであれば識別は可能である．

しかし，従来手法では LAN 外のサーバの応答も挙動に含むため，接続先や途中経路上の性能，負荷によって挙動が異なり，識別が困難になる．

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

3　提案手法

提案手法は LAN 内にあるクライアント側でのフローの挙動に着目したパケット長，到着時間間隔の収集を行う．

具体的な挙動の収集及び識別の手順を以下に示す．

1. 監視対象のクライアントのトラフィックをキャプチャできるように監視装置を設置する
2. 送受信される全トラフィックを IP 及び TCP ヘッダを用いてフローに分割する
3. フロー毎にパケットの向き (クライアントからの送信は ＋，受信は － で表す)，長さ及びパケット到着時刻を記録する
4. 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との時間間隔を計測する
5. (直前のパケット長 (*prelen*)，現在のパケット長 (*cltlen*)，時間間隔) の組をフロー毎に集計する
6. 同様の手法で予め集計されたアプリケーション毎の学習パターンと比較し識別する

例として SSH のフロー挙動の一例を図 3 に示す．

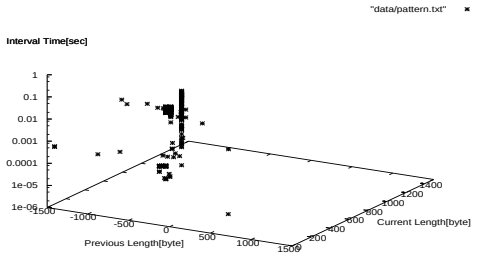

図 3: SSH のフローの挙動

識別には単純類似度法を用いた．つまり，図 3 のような各アプリケーション毎のプロット図をある時間間隔（本研究では対数スケールに変換し 0.1）毎の濃度分布に変換する．それを未知パターンとして同様の手法で得られる学習パターンとの単純類似度を求め，識別の評価尺度とする．単純類似度の値が最も大きい学習パターンのアプリケーションを，未知パターンのアプリケーションと識別する．

4　P2P ファイル共有ソフトウェアの識別

P2P ファイル共有ソフトウェアが扱うネットワークは多様なノードが接続し，ファイル転送や検索クエリ

等の送受信及び中継等を行い挙動が複雑である．そこで，提案手法を次のように改善する．

多様な接続ノードに対する挙動取得

P2P ネットワークに接続するノードは，その端末やインフラ等によって様々な MTU（Maximum Transmission Unit）値が設定されている．しかし，最大パケット長の送受信は，MTU 値が違っても同じ処理をすることが普通である．そこで，フロー毎に MSS（Maximum Segment Size）を取得し，最大パケット長を送受信する挙動を同一の挙動とみなすようにする．

可変長データに対する挙動取得

本手法では，検索クエリやキー情報等を含む可変長データが多く送受信される P2P のようなプロトコルでは挙動が安定しない．そこで，アプリケーション毎にパケット長を任意の長さのグリッドに分割し，学習パターンを再編する．異なったパケット長の同一の処理を，同じ挙動として認識できるようにする．

多様なサーバントの通信に対する挙動取得

フロー毎にサーバントの役割が異なるアプリケーションの場合，主に送受信されるコマンドやデータが異なる．そこで同一アプリケーションで複数パターンを生成する．学習パターン生成時に類似度を測りながら分類した．

実験は，主要な P2P ファイル共有ソフトに対する識別率を評価する．

実験環境は 4 のような一般的なインターネット環境を用意した．

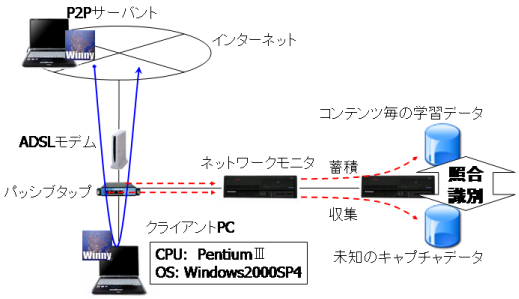

図 4: 実験環境

学習パターンの作成に約 50 フロー，識別実験に 50 フロー用いた．但し，著作権やウィルス感染等の問題を配慮して，全てのアプリケーションにおいて検索機能のみを利用し，ダウンロードは行わない．また，ファイル転送の中継も行わないよう注意した．

学習パターンは表 1 のように生成した．また，識別結果は表 2 の通りである．

表 1: P2P ファイル共有ソフト学習パターン

Application	Grid Width of Prelen	Grid Width of Cltlen	Number of Pcaket	Number of Session
Winny	1	1	37151	45
	1	1	3729	5
Limewire	1	1	15482	20
	10	1	21800	29
WinMX	42	1	46303	51
Cabos	24	1	40403	50
Share	20	32	35110	43
	1	32	9172	13
PerfectDark	1	1	37664	50

表 2: P2P ファイル共有ソフト識別実験結果

Application	Aver.Similarity	Accuracy
Winny	0.822	92%
Limewire	0.758	86%
WinMX	0.614	98%
Cabos	0.881	96%
Share	0.678	86%
PerfectDark	0.758	78%

概ね高い識別率を得た．LimeWire と Cabos は同一のプロトコル Gnutella を実装しているが，高い識別率が得られている．

可変長データを多く含む挙動のアプリケーションでは，平均類似度が低い．また，PerfectDark の識別率が低いのは Winny と誤識別したためで，両者の挙動が酷似しているためである．対して，Winny は挙動が安定しているため，識別率が高い．

5 まとめ

本研究では，クライアント側のフローに着目した挙動収集及び識別手法を用いて，P2P ファイル共有ソフトウェアの識別を可能にした．また，実環境において P2P ファイル共有ソフトウェアの識別実験を行い，高い識別率を得た．本手法は，部内ネットワークにおいて，匿名性の高い P2P ファイル共有ソフトウェアのような不正な通信の識別に有効である．

参考文献

[1] 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 ”．信学技報, Vol. 105, No. 470, pp. 13–16, 12 2005.

[2] 近藤泰彦, 中村康弘．”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 ”．第 70 回情処学全大, Vol. 3, No. 2Y-4, pp. 93–94, 3 2008.