様式 C-19

科学研究費補助金研究成果報告書

平成　２１　年　３　月　２７　日現在

研究種目：基盤研究（C）
研究期間：2007～2008
課題番号：19500120
研究課題名（和文）
任意長文字列統計処理を利用したネットワークログ解析
研究課題名（英文）
Analysis Methods for Network Traffic Log using Substring Statistics
研究代表者
梅村　恭司　Kyoji Umemura
豊橋技術科学大学・工学部・教授
研究者番号：80273324

研究成果の概要：
P2P トラフィックの検出方法を検討するにあたり、実データのトラフィックではどれが P2P トラフィックなのか判別が難しいこと、Winny のみのデータを分析することで P2P の特徴を捉えやすくなると考えたため実験環境を製作した。今回の実験では P2P ソフトは Winny とした。Winny のノードが定期的にキー情報を交換する様子を観測しようと考え、実験環境として実際の環境に近いネットワークの構築を行った。
この実験環境のデータを用いて P2P トラフィックの特定プログラムを作成した。これはパケットキャプチャのファイルを入力して、画面にどのアドレス間で通信が行われているか表示するプログラムである

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	1,500,000	450,000	1,950,000
２００８年度	1,500,000	450,000	1,950,000
年度			
年度			
年度			
総　計	3,000,000	900,000	3,900,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：知識発見とデータマイニング

１．研究開始当初の背景

インターネットの発展に伴いネットワークトラフィックが増加傾向になる。その中にあるのが P2P 技術を利用した新たなアプリケーションである。
年著作権問題やウィルスによる機密情報流失問題から P2P の特定方法が必要とされている。
P2P は特定のリソースがサーバに集中しないネットワークであるが、ファイル共有などで著作権の侵害の温床になっているという現実がある。P2P ネットワークで最も一般的に共有されているファイルは人気のある音楽の mp3 ファイルや映画の DivX コーデックを使った AVI ファイルである。またウィルスがＰ２Ｐ通信を通じて蔓延し、機密情報や個人情報が漏れるという事件も起きていて重大な社会問題にもなった。このようなことから P2P トラフィックを特定したいという欲求

がある。
問題を起こした P2P ソフトの中で有名なものに Winny がある。Winny の大きな問題は情報漏えいが頻発することが第一に挙げられる[1]。これは Winny ネットワークに Antinny というウィルスが蔓延しており、このウィルスは HDD 上の任意のファイルを勝手にアップフォルダにコピーするものである。コピーされたファイルは知らない間にアップロードされ Winny のネットワークに蔓延する。また一度アップロードされたファイルの完全削除が難しいためプライバシー情報や機密情報などが流れてしまうといつまでも被害を受け続けるという事件もある。
P2P トラフィックの特定方法に関しては様々な研究がされている。ペイロード部を解読するものも商品化もされているが処理負荷や個人情報保護の観点から問題が多い。

２．研究の目的
P2P トラフィックを特定することで注意を促し、事故を未然に防ぐ手助けが出来るツールがあればいいのではないかと考え、P2P トラフィック表示ツールの作成を行った。
そのために、実験環境の作成を行った。これは実環境に近いネットワークで P2P トラフィックのピュアなデータを得たいと思ったためである。
実験環境でのピュアな P2P トラフィックデータを得て、P2P トラフィックの可視化ツールを作成した。P2P トラフィックの可視化ツールの主な機能は時間の指定と P2P トラフィックの特定である。P2P トラフィックの検出方法に関しては他の研究でも行われている SYN パケットフラグの呼応を用いた。しかしパケットフラグは例えば暗号化されていたりすればパケットを見ることは出来なくなる。そこでパケットフラグを見ずに P2P トラフィックを検出する方法を二つ提案した。ひとつはパケットの呼応のタイミングを見ることで P2P トラフィックを検出する方法、もう一つは短いパケットの呼応を確認して P2P トラフィックを検出する方法である。

３．研究の方法
Ｐ２Ｐトラフィックを得るために実験環境（図１）の作成を行った。Ｐ２Ｐトラフィックのターゲット以外も稼動しているような環境ではどのトラフィックがターゲットなのか分かり辛い。そこで、Ｐ２Ｐ以外のトラフィックが極力少ない、ターゲットのふるまいを正確に分析できるトラフィックを得られる環境が必要であると考えたからである。３台のサーバマシンと３台のクライアントマシンを使い、仮想マシンを用いて仮想空間を作りネットワークを構築した。その中でＰ２Ｐソフトを動作させパケットキャプチャを行った。


図１　実験環境の外観

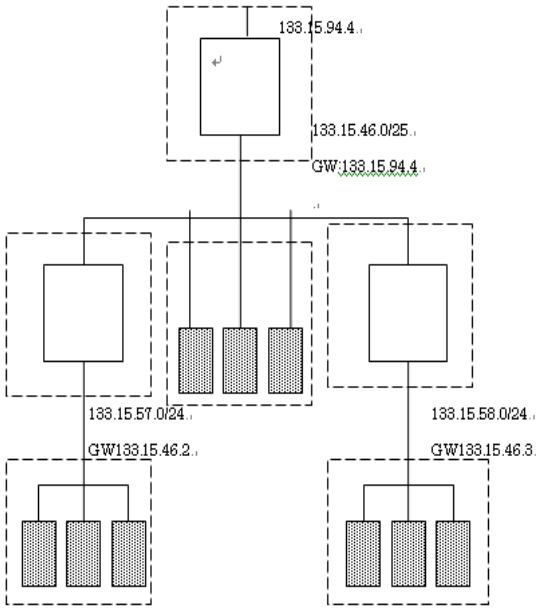
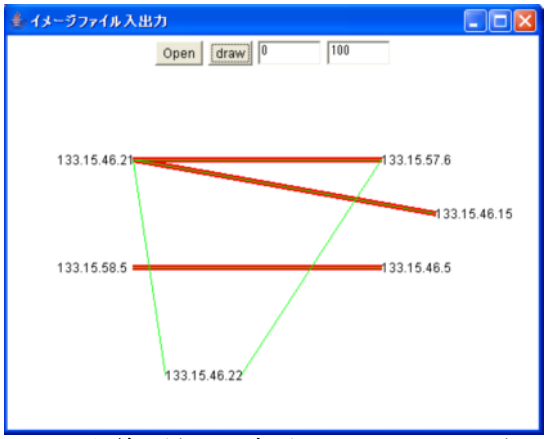


図２：実験環境のネットワーク論理構成

実験環境を構築する際に、Ｐ２Ｐソフトは Ｗｉｎｎｙ、仮想マシンはVMware を用いた[。 パケットキャプチャソフトにはＷｉｒｅＳｈａｒｋを用いた。作成したプログラムはＷｉｒｅＳｈａｒｋのパケットダンプフ



ァイルを読み込んで表示する。パケットダンプのファイルは時間データとパケットが時
図３：プログラムの画面

系列順に並べられている。パケットダンプファイルを読み込み、画面上に通信が行われた区間の表示を行えるものとした。その上で表示する時間の指定が行えるもの、Ｐ２Ｐトラフィックを特定できるものをとる。このプログラムにおいて，P2P トラフィックのの特定のアルゴリズムを組み込み，それを表示に反映させる。

４．研究成果

(1) 最初の成果は P2P トラフィック検出ぷろぐらむである。そのユーザインタフェースを図３に示す。

最初の機能は，表示する時間の指定である。テキストボックスが二つあり、左に開始時刻、右に終了時刻を入力し、指定することが出来る。開始時刻、終了時刻共にファイルの最初から計測時刻である。指定することで開始時刻から何秒後に何処で通信が行われていたかを調べることができる。テキストボックスに数字の入力をせずに描画を行った場合はファイルのトラフィックが最初から最後まで描画されることになる。

次の機能は，Ｐ２Ｐトラフィックの特定である。トラフィックの中で、特にＰ２Ｐ通信と考えられるものは他の通信とは異なる表示をした。今回の方法として、ＳＹＮパケットの呼応が確認された場合に赤い線で表示されるようになっている。線の太さでも区別がつくように赤線は平常の線より太くなっている。一定時間の閾値は、ＳＹＮパケットの呼応が行われている例 100 件を調べ、100 件の呼応時間の平均をとり 0.022 秒とした。

今回使用したトラフィックはＰ２Ｐ以外のトラフィックはほぼ無いが、画面（図２）を見るとＰ２Ｐ通信ではないと判定されている通信が存在する。これは誤判定である。誤判定が起きた原因は、今回の実験に用いた

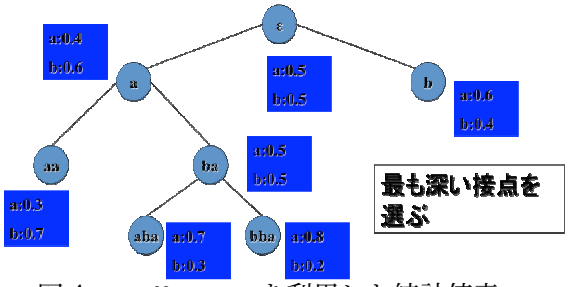


図４　Suffix Tree を利用した統計値表

他の通信は接続と切断を繰り返しているため検出が出来たがこのアドレス間ではデータのやりとりを行っていたため検出が出来なかった。これは検出方式の特性を示しているものといえる。

(2) 文字列の統計値のアルゴリズムが次の成果である。文字列処理のアルゴリズムの詳細は割愛するが，ほん研究での重要な成果であり，基本記述として，多くの研究に利用できた。最終的には，論文としてまとめることができ，この論文は，招待論文となっている。

この方法の特徴は，統計値を Suffix Tree というデータ構造で保持し，すべての文字列に対して，統計値を効率よく保持することを実装したものである。Suffix Tree を利用する方法は図４に示すが，文字列の先頭から，文字が増えるごとに枝をつくっていき，そこごとに統計値を保持する。保持するだけでなく，その統計値を木の性質をつかって，枝から談判する方法で求めるというところが，アルゴリズムの一番と特徴である。

(3) 文字列特徴量を利用した P2P トラフィックの判定アルゴリズムが次の成果である。これを利用して，トラフィックの流量を文字列で表現して，そこに対して統計処理を施し，異常なトラフィックの場所を特定する。

その結果を図５に示す。横軸はある一日の記録であり，そのなかで，P2P の通信が行われていると推定した場所を太字（濃い色）で

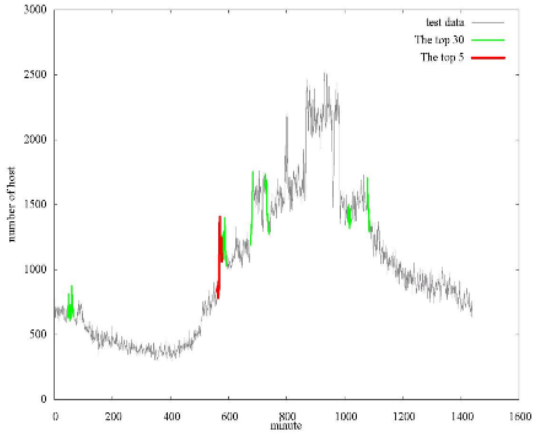


図５　文字列処理による P2P 検出

示したものである。これは実際に P2P トラフィックであることが確認できた。

(4) アルゴリズムの評価
テスト環境で，P2P トラフィックを流したときに, 検出したトラフィックが P2P である割合を示した。なお，検出の数は，あらかじめそろえておいた。表１に結果を示す。この結果，タイミング方式で検出するのがよいという結果がえられた。ただし，この結果はまだ不十分である。それは，検出の数の総数がまだ, 実際に想定去る P2P のトラフィックの存在数よりも小さく，取りこぼしが問題となるからである。研究期間内に，取りこぼしの問題を解決できるアルゴリズムが発見できなかった。ここについては，引き続き，研究を継続したい。

表１：正解率（%）

流量方式	タイミング方式	長さ方式
5.869	80.50	38.18

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　１件）

Substring Statistics
平成21年3月　CICLing 2009, LNCS(Lecture Notes of Computer Science), vol.5449 Springer, pp. 53-71
著者：Kyoji Umemura, Kenneth Church

〔学会発表〕（計　14 件）
(1) 拡張固有表現獲得の精度向上
共著　平成 19 年 7 月,情報処理学会研究報告、徳島　pp.67-72
著者：塩入寛之，関根聡，梅村恭司

(2)文字列を特徴量とし反復度を用いたテキスト分類
共著　平成 19 年 7 月,情報処理学会研究報告、徳島　pp.121-126
著者：平田勝大，岡部正幸，梅村恭司

(3)テンプレートを構成する名詞の Katz モデルによる抽出の試み
共著　平成 19 年 7 月,情報処理学会研究報告、徳島　pp.145-149
著者：藤原大輔，高瀬暁央，梅村恭司

(4)語の出現文脈の一致判定における文脈出現頻度と異なり数の比較
共著　平成 20 年 3 月,言語処理学会第 14 回年次大会、東京　pp.749－752
著者：當間　雅，梅村　恭司

(5)前後に出現する長い共通文字列を用いる関連語判定法
共著　平成 20 年 3 月,言語処理学会第 14 回年次大会、東京　pp.757－760
著者：折原幸治，藤原大輔，梅村恭司

(6)相関障害への耐性の高い広域分散データ配置の検討
共著：平成 20 年 4 月，情報処理学会研究報告、徳島　pp.99-106
著者：阿部洋丈，梅村恭司

(7)スナップショットを用いたデバッグ環境の構築
共著：平成 20 年 4 月，情報処理学会研究報告、徳島　pp. 187-194
著者：菊池　誠，阿部洋丈，梅村恭司

(8)固有表現自動獲得に向けての固有表現とコンテキストの関連度
共著:平成 20 年 7 月，　電子情報通信学会、北海道函館　pp. 13-17
著者：塩入寛之，岡部正幸，阿部洋丈，梅村恭司

(9)二語の共通周辺文字列の長さに着目した語文脈類似判定
共著：平成 20 年 11 月，情報処理学会研究報告、福岡　pp. 99-104
著者：折原幸治，梅村恭司

(10)人間の動作に対するアノマリ型異常検知システムの実装
共著：平成 21 年１月，第 50 回プログラミング・シンポジウム、静岡箱根　pp. 181-184
著者：藤原大輔，菊池　誠，阿部洋丈，岡部正幸，梅村恭司

(11)情報量の最大化に基づく指向性背センサの方向制御
共著：平成 21 年１月，　第 50 回プログラミング・シンポジウム、静岡箱根　pp. 177-180
著者：我妻裕樹，阿部洋丈，岡部正幸，梅村恭司

(12)仮想ユビキタスセンサにおける測定値補完システムのプロトタイプ構築
共著：平成 21 年１月，　第 50 回プログラミング・シンポジウム、静岡箱根　pp. 173-176
著者：大堀達也，菊池　誠，齋藤義文，我妻裕樹，阿部洋丈，岡部正幸，梅村恭司

(13) "人間の動作に対するアノマリ型異常検知システムの実装"、情報処理学会７１回全国大会 3ZD（滋賀県草津市：2009 年３月１１日）
著者：藤原 大輔，菊地 誠，阿部 洋丈，岡部 正幸，梅村 恭司、

(14)SYN パケットの呼応に着目した P2P トラフィックの表示，
情報処理学会７１回全国大会 4V7（滋賀県草津市：2009 年３月１１日）
著者：三浦明日香，梅村恭司，阿部洋丈，岡部正幸

〔図書〕（計　0 件）

〔産業財産権〕
○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕

６．研究組織

(1)研究代表者
梅村　恭司（UMEMURA KYOJI）
豊橋技術科学大学　情報工学系・教授
80273324

(2)研究分担者
岡部　正幸（OKABE MASAYUKI)
豊橋技術科学大学　情報メディア基盤センター・助教
50362330

三輪　多恵子
豊橋創造大学　経営情報学部・准教授
60351178

(3)連携研究者

該当無し