科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　　５　日現在

機関番号：　１１３０１
研究種目：　基盤研究(C)
研究期間：　2011～2013
課題番号：　２３５００００７５

研究課題名（和文）非観測情報の統計的推定によるＷｅｂアプリケーション識別

研究課題名（英文）Web Application Identification Based on Statistical Estimation of Unobserved Information

研究代表者
和泉　勇治（Waizumi, Yuji）

東北大学・情報科学研究科・准教授

研究者番号：９０３３３８７２

交付決定額（研究期間全体）：（直接経費）　　3,800,000円、（間接経費）　　1,140,000円

研究成果の概要（和文）：ネットワークのトラヒックを観測することにより，一つのWebサイトにより提供される複数のWebアプリケーションの種別を識別する方式を構築した．Webの利用形態からIPアドレスやポート番号を利用することが出来ないため，ネットワークの流通するパケットの特徴を適切に捕捉し，それから直接観測することの出来ない利用者のWeb利用の挙動を間接的に推定することにより，約90%の正確度でWebアプリケーションの識別が可能な方式を構築することが出来た．

研究成果の概要（英文）：We developed an identification method for Web application provided from a Web site using network traffic only. Identifying method cannot use port number or IP address because all traffic from a Web site are blocked if packets which have the IP address of the Web site. So we proposed a new identification method which uses only packet sizes and the number of TCP connection. The identification accuracy achieves 90%.

研究分野：　総合領域

科研費の分科・細目：　情報学・計算機システム・ネットワーク

キーワード：　ネットワーク・セキュリティ技術　ネットワークアプリケーション識別　Webアプリケーション

１．研究開始当初の背景

クラウドの言葉に代表されるように，近年様々なウェブサービスを提供するウェブサイトが増加しており，それらのサイトで提供されるサービスはウェブブラウザ上で動作するウェブアプリケーションによって実現されている．そのため，Google や Yahoo! のように複数のウェブアプリケーションが利用できるウェブサイトでは，一つのウェブサイトにアクセスすることで多様なサービスを利用することが可能となっている．

そのようなウェブサイトは，営利企業なども含め様々な組織で利用されるようになってきており，ネットワーク資源を大量に消費する動画共有サービスなどの管理対象のネットワークを利用する上で不適切なサービスも存在すると考えられる．つまり，ユーザの利用するウェブアプリケーションを把握することはネットワーク管理において必須事項であると言える．

利用中のウェブアプリケーション種別の把握は，IP アドレスやポート番号を利用するのみでは不可能である．http の通信は 80 番ポート，または，443 番ポートを利用しているため，これらのポート番号を遮断すると全ての http での通信が不可能となってしまう．また，複数種類のウェブアプリケーションを提供しているあるサイト上の一部のウェブアプリケーションの利用を規制したい場合，該当サイトの IP アドレスに基づいて通信の遮断等を行うと必要なウェブアプリケーションも利用できなくなってしまう問題がある．つまり，IP アドレスやポート番号に依存しない Web アプリケーションの識別方式の開発が必要であると言える．

２．研究の目的

本研究は，パケットヘッダに記録されている IP アドレスやポート番号を利用せず，通信特性の基づいた識別方式を提案することを目的とする．特に，Web ブラウザ上で実行される Web アプリケーションの識別を対象とし，HTTP 上での動画再生，メール送受信，オフィスソフトの利用など Web アプリケーション毎に異なると予想される処理時間などの非観測情報を確率モデルにより統計的に推定し，不正なアプリケーションの利用抑制やより高度なネットワークマネージメントを実現する基礎技術の開発を行う．

３．研究の方法

ある同一の Web サイト上で提供される複数のアプリケーションを利用し，その通信を観測することで，Web アプリケーション種別毎のトラヒックデータベースを構築する．
構築したデータベースから Web アプリケーション毎の特性を複数の数値化手法を適用することにより，定量的に明らかにする．特に，クラスタ解析などを利用し，Web アプリケーション種別毎のトラヒックが生成するクラスタの分離性や分離に不足している情報などを検討する．この検討を基に数値化の特徴量の抽出方法を確定する．
確定した特徴抽出方法により抽出された特徴量に学習機械などを適用し，識別の可能性とその性能を検討する．

４．研究成果

未学習の Web アプリケーションのトラヒックに対し 89.09%の精度で識別可能な識別方式を提案した．
トラヒックの数値化方法としては，TCP のコネクション毎にパケットサイズの逆数の遷移パターンをベクトル化した特徴量と，識別対処のコネクションが生成された時刻から一定時間過去に生成されたコネクション数を併用したものが最も高い精度を達成することが明らかとなった．
識別においては，パケットサイズの逆数のベクトル間距離にコネクション数の差分の重み付き和を利用する手法を提案した．

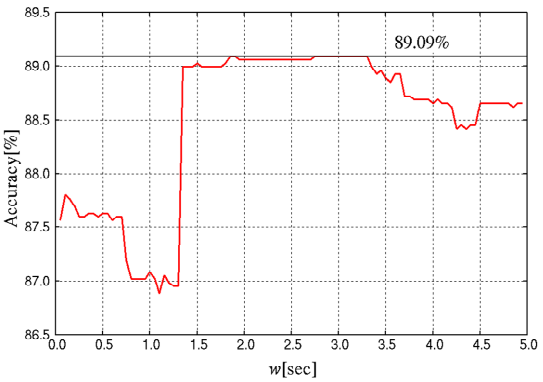

図１：重みによる識別精度の変化

図１は重み付き和の重みαと識別精度の関係である．α=0.011 のとき最高の 98.09%の識別精度を達成している．

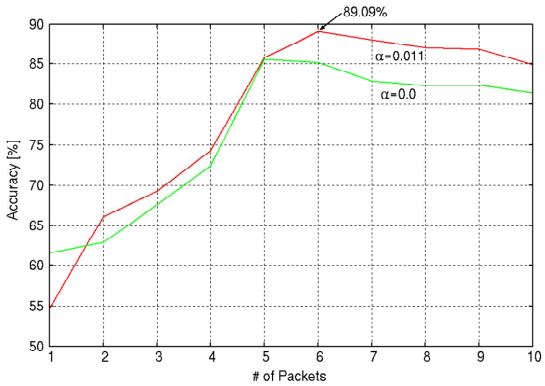

図２：パケットサイズとその逆数を利用した識別

図２は，パケットサイズとその逆数を利用した識別精度を示している．図中の赤線が逆数を利用した識別結果である．横軸は，識別に利用したパケット数で，パケット数が６の時

に最も高い精度を得ている．パケット数が１以外でパケットサイズの逆数を利用した識別の方が高い精度を得られており，Web アプリケーションの識別においては，パケットサイズの逆数を利用することが有効であると言える．このことは，画像や動画など，大きなサイズのコンテンツをダウンロードする場合，個々のパケットが MTU に達していることが多くなり，サイズの大きなパケットを類似性を評価する距離の算出に利用するとアプリケーションの違いが現れなくなるからであると考えられる．
図３は，コネクションが同時に発生したと判断する時間間隔の変化に対する識別精度の変化を示している．識別対象のコネクションが発生する２～３秒前に発生したコネクション数をパケットサイズの逆数のベクトル間距離に付加することで識別精度が向上することが明らかとなった．

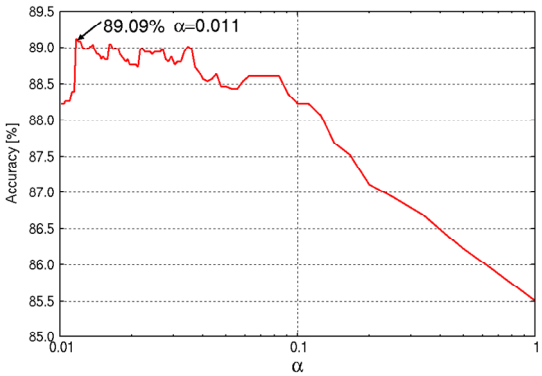

図３：コネクション同時発生数評価時間と識別精度の変化

コネクションの同時発生数は，Web ページを構成するコンテンツが複数のコネクションにより転送され，その発生のタイミングを定量的に評価する特徴量であると言える．この特徴量を加えることにより識別精度が向上することから，Web ページを構成する情報のダウンロードが Web アプリケーション種別毎に異なるタイミングで行われ，それを捕捉することが Web アプリケーション識別に有効であると判断出来る．このタイミングは，Web アプリケーションの利用方法がその種別ごとに異なり，追加の情報をダウンロードするユーザの挙動をコネクションの同時発生数により評価しているからと考えられる．この情報はネットワークトラヒックから直接観測することが出来無い情報であるが，コネクションの同時発生数に着目することにより，間接的にでもユーザの Web アプリケーション利用時の挙動を推定し，Web アプリケーション識別性能を向上させることに寄与する成果を得られたと言える．

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　４件）

1. Yuji WAIZUMI, Tsuyoshi SATO, Kazuyuki TANAKA ,“Network Application Identification Using Sequential Transition Patterns of Payload Length”,Interdisciplinary Information Sciences Vol. 18 No. 2, pp189-196, Dec. 2012,査読有, DOI 10.4036/iis.2012.189
2. Yuji Waizumi, Yohei Sato and Yoshiaki Nemoto, “A Network-based Anomaly Detection System Based on Three Different Network Traffic Characteristics”, Journal of Communication and Computer, Vol.7, pp.805-812, Jul. 2012, 査読有, http://www.davidpublishing.com/davidpublishing/Upfile/8/14/2012/2012081400385449.pdf
3. Yuji Waizumi, Hiroshi Tsunoda, Masashi Tsuji and Yoshiaki Nemoto , “A Multi-Stage Network Anomaly Detection Method for Improving Efficiency and Accuracy” Jouranl of Information Security , Vol.3 No.1, pp.18-24, Jan. 2012, 査読有, DOI:10.4236/jis.2012.31003
4. Hiroshi Tsunoda , Hidehisa Nakayama , Kohei Ohta , Akihiro Suzuki , Hiroki Nishiyama , Ryoichi Nagatomi , Kazuo Hashimoto , Yuji Waizumi , Glenn Mansfield and Yoshiaki Nemoto, “Development of a WLAN Based Monitoring System for Group Activity Measurement in Real-Time”, Journal of Communications and Networks, Vol. 13, No.2 13, pp.86-94, Apr. 2011 , 査読有,DOI: 10.1109/JCN.2011.6157407

〔学会発表〕（計　５件）

1. 和泉 勇治，田中 和之，“トラヒッ

ク解析に基づいたウェブアプリケーション識別”,電子情報通信学会通信方式研究会，2013 年 9 月 13 日，東北大学

2. 浅利 岳，安田 宗樹，和泉 勇治，田中 和之，“2 部グラフ型ボルツマンマシンに対する複合最尤法”電子情報通信学会ニューロコンピューティング研究会，2012 年 11 月 17 日，東北大学
3. 松尾 翔希，和泉 勇治，田中 和之，“通信特性に基づいたウェブアプリケーション識別に関する一考察”電子情報通信学会通信方式研究会，2012 年 9 月 21 日，東北大学
4. 高橋 洸，和泉 勇治，橋本 和夫，“能動学習における決定境界の安定性の検討”，電子情報通信学会情報論的学習理論と機械学習研究会，2012 年 3 月 13 日,統計数理研究所
5. 村山 竜太，安田 宗樹，和泉 勇治，田中 和之，“カラーチャンネル間の相関を考慮した Gaussian FoE モデル”，電子情報通信学会情報論的学習理論と機械学習研究会,2011 年 11 月 9 日，奈良女子大学

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
和泉 勇治（Waizumi Yuji）
東北大学．大学院情報科学研究科・准教授
研究者番号：90333872

(2)研究分担者
田中　和之　（Tanaka Kazuyuki）
東北大学．大学院情報科学研究科・教授
研究者番号： 80217017

(3)連携研究者
（　　）

研究者番号：