様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　　２　日現在

機関番号：　２４４０２
研究種目：　挑戦的萌芽研究
研究期間：　2013～2015
課題番号：　２５５４００３４

研究課題名（和文）ストリームマイニングによるユーザビヘイビア同定とその応用に向けた基礎的検証

研究課題名（英文）Research on Identification of User Behavior by Stream Mining and Its Application

研究代表者
阿多　信吾（Ata, Shingo）

大阪市立大学・大学院工学研究科・教授

研究者番号：３０３２６２５１

交付決定額（研究期間全体）：（直接経費）　　2,700,000円

研究成果の概要（和文）：本研究では、ユーザの挙動によって生成された通信トラヒックから、その主要因となるユーザ挙動を推定する、ビヘイビア識別技術の提案を行う。提案手法ではある時間におけるユーザの利用アプリケーションやサービスといったユーザの挙動をユーザビヘイビアと定義する。ユーザビヘイビアごとにトラヒックを収集し、それらの統計的性質を特徴値とした学習データを元に、機械学習によってユーザビヘイビアを識別する手法を新たに提案する。実トラヒックを用いた精度評価を行い、9 種類のアプリケーション単位では 91%、43 種類のビヘイビア単位では 81% の識別精度を実現できることを明らかにした。

研究成果の概要（英文）：Recently, network traffic is becoming strongly biased by user's action taken in an application. In this research, we propose a method to infer such action (we define as "user behavior") from the monitored traffic. The proposed method firstly composes a set of traffic features (statistical features of measured traffic flows) and then applies a Supervised Machine Learning (ML) algorithm to identify the user behavior from the statistical features. Through experimental results by using actual traffic, we show that the proposed method achieves around 91% accuracy of identification for 9 major applications, and around 81% accuracy of identification for 43 user behaviors.

研究分野：　情報ネットワーク工学

キーワード：　ユーザビヘイビア　アプリケーション識別　トラヒック計測　機械学習　ネットワーク　フロー

１．研究開始当初の背景

近年インターネットではセキュリティが非常に大きな問題であり、ウィルスや悪意のあるユーザによる攻撃のみならず、SPAM などの不要トラヒックの送信、なりすましや乗っ取りなどの問題、プライバシを含む個人情報の漏洩、ネットワークの信頼性などが非常に重要視されている。

特になりすましなどの問題は、本人以外が何らかの方法により本人の端末に不正に侵入し、それを踏み台にして別の端末やユーザへの攻撃を行うなど、非常に深刻な社会問題となりつつある。一般的になりすましによる攻撃は、本人自身の操作であるか乗っ取りにより生成されたものかを外部から容易に判別できない。また、異常トラヒックの検知などでは、たとえ被害ノードでの検出が行われたとしてもその探知は容易ではない。特に、詐称および隠匿されたヘッダ情報から攻撃者を特定することはほぼ困難であり、抜本的な解決法が望まれる。本研究では、このようなりすましを早期に検出するため、ユーザの挙動（ビヘイビア）に着目する。ユーザ端末が送受信するトラヒックパターンは、使用するユーザのビヘイビアに強く依存することから、逆にトラヒックパターンを分類することができれば、ユーザのビヘイビアを推測できる可能性が高くなる。ビヘイビアを推測することで、一連のフロー群に対する統一的な制御が可能になるほか、主要因となるフローの決定にも大きく寄与できるものと考える。

２．研究の目的

本研究では、ユーザビヘイビアの分類とその同定によるネットワーク制御を最終目的とし、ユーザビヘイビア同定の精度に関する検討、およびビヘイビア同定にもとづくネットワーク制御の有効性について明らかにする。

３．研究の方法

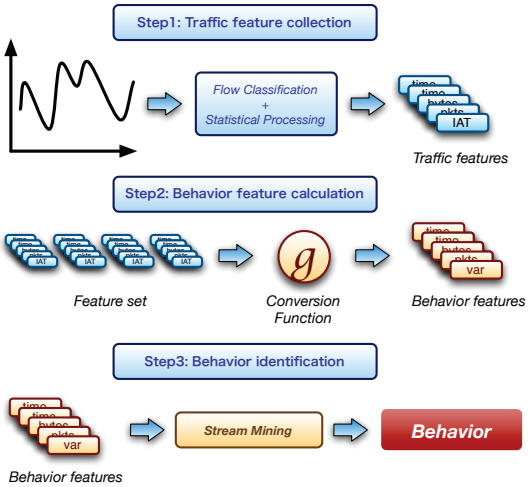


図１　ビヘイビア同定手順

提案するビヘイビア同定手順を図 1 に示す。まず計測トラヒックをフローごとに分類して統計処理を施すことで特徴量（traffic feature）を抽出する。これを複数のフローあるいは時間インターバルで収集し、特徴セット（feature set）を作成する。次に特徴セットに対して変換関数 g を与え、ビヘイビア特徴量（behavior feature）を導出する。その後ストリームマイニングによりビヘイビア特徴量が最も合致するビヘイビアを選択する。ストリームマイニングでは過去のビヘイビアと特徴量の関係を学習データとして保持し、ビヘイビア分類に使用される。

本研究では課題を 4 つ設定する。課題 1「ユーザビヘイビアを特徴付けるトラヒック種別の特定」ではユーザビヘイビアを分類するために必要十分となるトラヒック特徴値が何かを明らかにする。課題 2「ユーザビヘイビア同定のためのストリームマイニング技術」については、ユーザビヘイビアを高速かつリアルタイムに同定するためのストリームマイニング技術について検討する。課題 3「ユーザビヘイビアパターンの収集」では、実際に被験者を利用した長期間の実験により、ユーザの行動とそれによって生成されたトラヒックパターンを長期にわたり記録し、ユーザビヘイビアとそれに関連するトラヒック特徴のマッピングのための情報を収集する。課題 4「ユーザビヘイビア同定による応用としてのテストベッド構築」ではユーザビヘイビア同定による信頼性制御によるルーティングを対象としたテストベッド環境の構築を対象として実現する。

課題 3 のトラヒックパターンの収集は本研究における最も重要な課題項目である。具体的な実験方法を示す。まず、被験者に対して 1 日程度自由に計算機を操作してもらい、それらのトラヒックをすべて記録する。一方で、被験者がどのような行動を取ったのかを別途（時刻、行動のペアで）記録してもらう。両者の記録について時間をもとに対応付けを行えば、トラヒックの変化と行動の変化の関連性が明らかとなる。課題 3 については 2 年間の計画で、より多くの被験者からの長期にわたるデータを収集することを目指す。一方課題 1 および 2 については、これまで研究代表者のグループで行ってきたアプリケーション識別に関する研究を拡張し、トラヒックの特徴を示す統計値の数を増やし、さらにマイニングアルゴリズムについても既存の学習アルゴリズムの比較によりビヘイビア分類に適したアルゴリズムを選択することを目指す。テストベッド環境については、ビヘイビア分類としては最も単純な人的操作および機械操作のトラヒックパターンの分類を達成すべく学習アルゴリズムおよびトラヒック統計値の選択を行う。

４．研究成果

(1) アプリケーション識別のリアルタイム性向上と特徴量決定

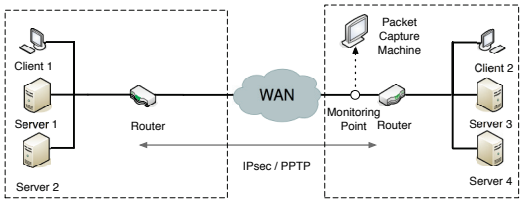


図 2　ビヘイビア計測環境

本項目では、ビヘイビア分類の基礎検討として、アプリケーション識別をリアルタイムに実現するための特徴量の決定、および識別精度向上のための多段階識別アルゴリズムについて提案し、その効果を述べる。

本研究では平文トラヒックと暗号化トラヒック(IPsec，PPTP）を計測対象のトラヒックとする。計測環境を図 2 に示す。

次に、識別対象のアプリケーションを表 1 に示す。アプリケーションは大分類および小分類によって区分されている。また、各アプリケーションで取得したフロー数についても示している。

表 1　計測対象アプリケーション

大分類	小分類	取得データ数		
		平文	IPsec	PPTP
BULK	FTP	9549	4067	2000
	HTTP	6008	3587	10178
INTERACTIVE	FTP	4601	1906	2533
	SSH	2000	4416	7828
P2P	BitTorrent	5990	1463	1476
STREAMING	HTTP (YouTube)	15097	3770	2140
	HTTP (Hulu)	43824	2512	2211
WEB	HTTP (Dynamic)	6196	5513	6100
	HTTP (Static)	20210	4616	8864

計測地点において取得したパケットはヘッダ情報の 5-tuple（送信元 IP アドレス、宛先 IP アドレス、送信元ポート番号、宛先ポート番号、プロトコル番号）にもとづいてフローに分類し、フローごとに統計値の算出を行う。フロー統計情報はパケットサイズの平均値や中央値、パケット到着間隔の最大値など 29 種類の統計情報を初期特徴値として用いる。また、これらの統計情報についてはクライアントからサーバ方向、サーバからクライアント方向および両方向の集約値を用いる。

図 3 に、大分類および小分類によるアプリケーション種別の識別結果を示す。図の横軸は観測対象パケット数（フロー先頭からのパケット数)、縦軸は全体の識別精度を表している。この図から、アプリケーションの分類が細分化されることで識別精度が大きく低下していること、また、よりパケット数が少ない領域では精度が大幅に低下していることが分かる。すなわち、リアルタイム識別など少ないパケットでアプリケーション識別を行う場合、精度向上が必要であることが示されている。

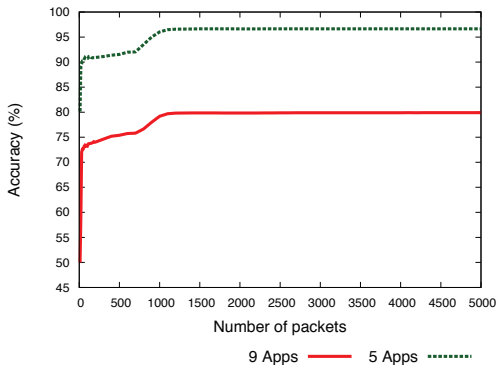


図 3　識別対象アプリケーション数による影響

この問題を解決し、より少ない特徴量を用いてリアルタイムかつ高精度に識別可能なアプリケーション識別手法を提案する。本研究では、アプリケーション識別における機械学習の識別器を複数用いる多段階型アプリケーション識別を行う。

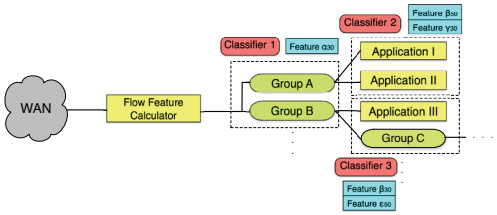
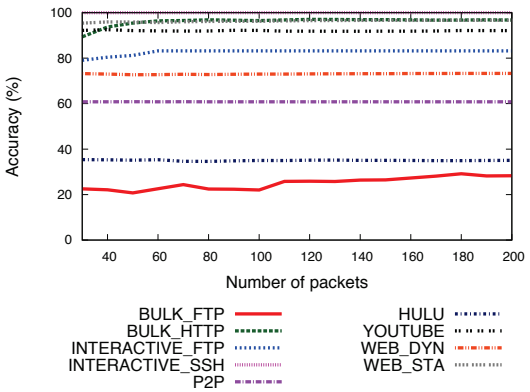


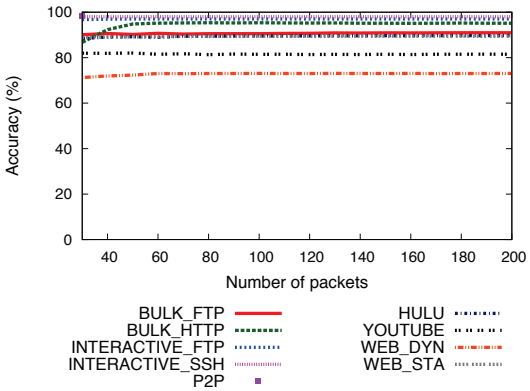
図 4　多段階型アプリケーション識別

多段階型アプリケーション識別モデルを図 4 に示す。まず、計測したトラヒックをフローに集約し、トラヒッ ク特徴量を算出する。そして、算出した特徴量をもとに、アプリケーションのグループ分類を行う。グループ分類に関しては、グループ分類アルゴリズムに従う。次に、分類されたそれぞれのグループでアプリケーション識別を行う。アプリケーションのグループ分類とグループ内アプリケーション識別では異なる特徴量を用いる。識別に必要なパケット数は特徴量によって異なることから、それぞれの識別器で異なる特徴量を用いることで、早期にアプリケーション識別を行うことが可能となる。

多段階型アプリケーションの識別結果を図 5 に示す。その結果、グループ分類にパケットサイズの最大値（サーバからクライアント方向）の分散値を用い、グループ内アプリケーション識別に 20 種類の特徴量を用いることで、開始から 30 パケット時点でのアプリケーション識別精度が約 71%から約 88%へと大幅に向上することが可能であることが明らかとなった。



(a) 非多段階型識別



(b) 多段階型識別

図 5　多段階識別結果

(2) ユーザビヘイビア識別への応用

次に本研究の主目的であるユーザビヘイビアの識別手法について述べる。対象とするユーザビヘイビアは、一般的な利用者が主として行う動作をアンケートにより調査したものを対象とする。表 2 に対象とするビヘイビアを示す。表が示すとおり、代表的なアプリケーション 9 種類に対し、さらにそのアプリケーション内で主として行われる動作をビヘイビアとして決定した。ビヘイビアの合計は 43 種類である。

表 2　対象ビヘイビア

アプリケーション	ビヘイビア
Amazon	トップページ表示，ログイン，商品ページ，カート閲覧，検索，購入処理
Dropbox	起動，同期，アップロード，削除，名前変更，フォルダ作成，移動
Facebook	ログイン，投稿，画像投稿，タイムライン，プロフィール，トップ
Gmail	受信ボックス，メール送受信，メールを開く
Google	検索，画像検索，トップページ表示
Skype	起動，通話，ビデオ通話，メッセージ送受信，ファイル送受信
Twitter	タイムライン，ツイート，画像投稿，トップページ表示
Yahoo! JAPAN	トップページ表示，検索，ニュース，動画付きニュース
YouTube	動画視聴，動画検索，マイチャネル，ログイン，トップページ表示

これらのビヘイビアを識別するための手法を提案する。ビヘイビアによって生成されるフローは複数存在することから、単一フローに対するアプリケーション識別ではビヘイビアを決定することはできない。したがって本研究では、アプリケーション識別で用いるフロー特徴量をさらに複数フローに対して集約し、それらの統計量を識別の特徴量とした機械学習を行うことで、ビヘイビア識別を行う。図 6 にビヘイビア識別手法の概要を示す。

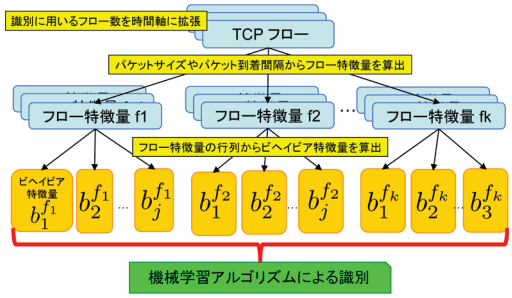


図 6 ビヘイビア識別手法概要

表 3 および表 4 に提案手法によるアプリケーション識別精度、およびビヘイビア単位の識別精度を示す。

表 3　アプリケーション識別精度

Amazon	90%	Skype	88%
Dropbox	92%	Twitter	84%
Facebook	93%	Yahoo! JAPAN	97%
Gmail	90%	YouTube	96%
Google	92%		

表 4　ビヘイビア識別精度

1	AMAZON 購入処理	91%	23	GOOGLE 画像検索	87%
2	AMAZON カート閲覧	80%	24	GOOGLE 検索	55%
3	AMAZON 商品ページ	91%	25	GOOGLE トップページ	72%
4	AMAZON ログイン	97%	26	SKYPE ログイン	98%
5	AMAZON 検索	81%	27	SKYPE メッセージ	98%
6	AMAZON トップページ	92%	28	SKYPE ファイル送受信	83%
7	DROPBOX 起動	93%	29	SKYPE ビデオ通話	64%
8	DROPBOX アップロード	90%	30	SKYPE 通話	97%
9	DROPBOX 同期	73%	31	TWITTER ログイン	97%
10	DROPBOX 名前変更	41%	32	TWITTER ツイート	80%
11	DROPBOX フォルダ作成	34%	33	TWITTER 読み込み	88%
12	DROPBOX フォルダ移動	45%	34	TWITTER 画像投稿	81%
13	DROPBOX 削除	33%	35	YAHOOJP ニュース	99%
14	FACEBOOK 読み込み	82%	36	YAHOOJP 検索	92%
15	FACEBOOK ログイン	98%	37	YAHOOJP トップページ	91%
16	FACEBOOK 画像投稿	82%	38	YAHOOJP 動画付きニュース	95%
17	FACEBOOK 投稿	74%	39	YOUTUBE ログイン	88%
18	FACEBOOK プロフィール閲覧	84%	40	YOUTUBE マイページ	72%
19	FACEBOOK トップページ	97%	41	YOUTUBE 検索	89%
20	GMAIL メール開封	88%	42	YOUTUBE トップページ	97%
21	GMAIL 受信ボックス	98%	43	YOUTUBE 動画	97%
22	GMAIL メール送受信	97%			

これらの結果より、提案手法によるアプリケーション識別精度は約 91%で、単純なアプリケーション識別より多少劣化するものの、約 81%の精度でアプリケーション内の挙動判別も行えることが示されており、非常に高い精度でビヘイビア識別が実現できていることが示された。

さらに、ビヘイビア識別に必要な特徴量と識別精度の関係を図 7 に示す。この図は、ビヘイビア識別に用いた特徴量を 672 種類から順に減少させた場合の全体の識別精度の変化を示したものである。その結果、ビヘイビア識別精度を維持した状態で特徴量を 143 種類まで減少できることが明らかとなった。

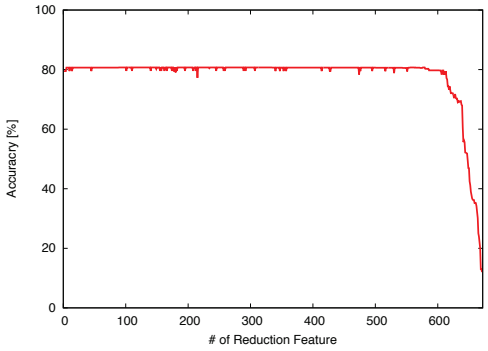


図７　ビヘイビア識別精度に必要な特徴量の種類数

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３件）

① Yuichi Kumano, Shingo Ata, Nobuyuki Nakamura, Yoshihiro Nakahira, Ikuo Oka, “Enhancing Immediacy of Identification with Multi-Stage Application Identification,” Proc. 7th International Conference on New Technologies, Mobility and Security (NTMS 2015), July 2015, doi:10.1109/NTMS.2015.7266492（査読有）

② Yusuke Iemura, Shingo Ata, Ikuo Oka, “Identification of User Behavior based on Time Variation of Traffic Statistics,” Proc. 16th Asia-Pacific Network Operations and Management Symposium (APNOMS 2014), September 2014.（査読有）

③ Yuichi Kumano, Shingo Ata, Nobuyuki Nakamura, Yoshihiro Nakahira, Ikuo Oka, “Towards Real-time Processing for Application Identification of Encrypted Traffic,” Proc. 2014 International Conference on Computing, Networking, and Communications (ICNC 2014), pp.136-140, February 2014, doi: 10.1109/ICCNC.2014.6785319（査読有）

〔学会発表〕（計５件）

① 津室雄志、阿多信吾、中村信之、岡育生、「モバイル端末におけるアプリケーショントラヒックの統計的分析に関する研究」電子情報通信学会総合大会、2016年３月

② 家村勇輔、熊野由一、中村信之、阿多信吾、岡育生、「複数フローに着目したユーザビヘイビア分類手法のためのフロー特徴分析」、電子情報通信学会技術研究報告、2015年４月

③ 熊野由一、阿多信吾、中村信之、中平佳裕、岡育生、「識別性能を向上させる多段階型アプリケーション識別手法」、電子情報通信学会技術研究報告、2015年３月

〔その他〕
ホームページ等
http://www.c.info.eng.osaka-cu.ac.jp/
に研究成果の概要を公表

６．研究組織
(1)研究代表者
阿多　信吾（ATA SHINGO）
大阪市立大学・大学院工学研究科・教授
研究者番号：　30326251