パケットヘッダ情報を用いた仮想ユーザ識別システムの提案

上原 雄貴† 水谷 正慶‡ 武田 圭史†† 村井 純††

†慶応義塾大学 総合政策学部 ‡慶応義塾大学大学院 政策・メディア研究科
†† 慶応義塾大学 環境情報学部
252-8520 藤沢市遠藤 5322
{nakajima,mizutani,keiji,jun}@sfc.wide.ad.jp

あらまし 情報技術の発展に伴い，一人のユーザが複数台のホストを所有する場合や，単一のホストを複数人で使用するなど，ユーザの利用形態は多岐に渡った．しかし，ユーザの行動や興味動向の調査を行う際，そのようなユーザが識別できないため，ユーザを基準としたネットワークの実態調査が困難である．そこで，パケットのヘッダ情報のみを用いて，ユーザを識別するシステムを提案する．本システムはトラフィックからヘッダ情報のみを利用するため，汎用性やユーザの網羅性に優れている．これにより，ネットワークにおけるユーザに焦点を当てた実態調査が可能となり，ネットワークでの人間の行動学調査や新しいサービスの創発，犯罪捜査などに応用可能である．

Proposal of virtual user identification system with packet header analysis

Yuki Uehara† Masayoshi Mizutani‡ Takeda Keiji†† Jun Murai††

†Faculty of Policy Management, Keio University
‡Graduate School of Media and Governance, Keio University
††Faculty of Environment and Information Studies, Keio University
252-8520, 533 Endo ,Fujisawa-shi,Kanagawa,Japan
{nakajima,mizutani,keiji,jun}@sfc.wide.ad.jp

Abstract With the advancement of information technology, use of computers has changed. For example, one person can have many hosts or multiple users may share one host. Therefore, identifying network user became difficult. In this paper, we propose virtual user identification system with packet header analysis. Since the system monitors only packet header information to preserve use's privacy, it is considered acceptable and deployable in various environments. We can apply the system to survey user behavioral and the crime investigation on computer network.

1 はじめに

情報技術の発展によって，複雑な情報を収集・解析し，情報に新しい価値を生み出すことができるようになった．これによって，ネットワークにおけるユーザの利用形態が多様化する傾向にある．特に，複数のホストを所持するユーザの増加は著しい．他にも，家庭におけるホストの普及によって，一つのホストを複数人が利用するケースも存在する．

しかし，そのような調査にあたり，一人で複数のホストを保有するユーザや共有ホストを利

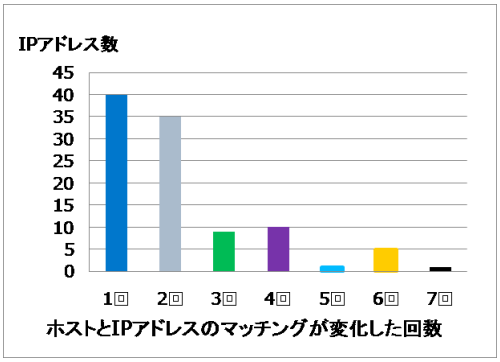

図 1: ホストの IP アドレス遷移回数

用するユーザを識別することができないといった問題点や，ネットワークごとに取得できる情報が異なるという問題が存在する．そのため，ユーザを基準としたネットワークの実態調査が困難である．
また，ユーザの識別を行う場合，ペイロード解析をする場合が多い．しかし，ペイロード解析はプライバシと密接に関係するため，ネットワークによってあ困難である．そこで，本研究はパケットのヘッダ報情のみを用いてネットワーク上におけるユーザを識別するシステムを提案する．これにより，ネットワークにおける人間の行動科学・自然科学の調査や，調査によって得られた統計情報を公開することによる新しいサービスの創発，犯罪捜査などに利用することが期待できる．

2 現状のユーザ識別に関する問題点

既存の技術によってネットワーク上でのユーザを把握する場合，ログ解析や認証技術の導入，アカウント登録やトラフィック解析などが挙げられる．しかし，この研究はホストの識別は可能だが，一人で複数台ホストを持つユーザや，共有ホストを利用するユーザを識別することができない．また，ユーザの識別に利用可能な情報はネットワークポリシや構成によって変化し，ユーザの同意が必要な場合が多く，汎用性に乏しい．

ネットワークトラフィックからユーザを識別する際には，IP アドレスを基準とする場合が多い．しかし，IP アドレスはパケットを送受信する際に，インターネット上におけるホストを区別するための識別子であるため，ユーザの識別には適さないという問題点が挙げられる．

図 1 は筆者が所属する研究室のネットワークにおけるホストの IP アドレス遷移を示す．dhcpd のログを 2009 年 7 月 12 日 6 時から 7 月 18 日 6 時まで取得し，分析した．取得した期間はこの期間において、取得した IP アドレス数は 219 に対して取得した MAC アドレス数は 334 であった．そのうち，同一と見られるホストの IP アドレス遷移は 101 回あり，期間中に 7 回も IP アドレスがつけ変わるホストが存在した．このことから，IP アドレスのみを用いてユーザを識別することは困難である．

また，割り当てられている IP アドレスが変わると，新しく付与されたアドレスを新たなユーザとして誤って認識される場合がある．他にも，ユーザが別のホストを利用する場合や，ホストを共有している場合，ネットワークインターフェースカードを変更した場合に同一ユーザを識別することは困難である．

ユーザを区別するための研究としてユーザが登録したサービスの情報を利用する研究が挙げられる．この方法はユーザがサービスに登録したり特別なアプリケーションをインストールすることでユーザを識別する．例えば，Pathtraq[1] ではユーザのホストにアプリケーションやプラグインをインストールし，データ収集サーバに情報を発信することでユーザを常時識別する．しかし，このような研究はサービスに登録したユーザしか識別できない問題点が挙げられる．

また，機器をネットワークに接続する際の，ユーザ認証を利用する方法がある．例えば，ネットワーク管理者はネットワークに 802.1x 認証などの認証機構を導入することで，管理ネットワーク内のユーザを正確に識別できる．しかし，802.1x などの認証機構を導入する場合，当該ネットワークに接続するすべてのユーザの識別情報を登録・管理必要がある．そのため，ネットワーク管理者の導入・運用における負担が増加してしまう傾向がある．

3 システム要件

本システムの必要要件は以下の 5 点である．

- ユーザの識別結果の精度
  ユーザの識別結果の精度は重要である．プロファイルの結果が曖昧であった場合，データ自体の信憑性が薄れてしまう．利用者もプロファイル結果が信頼できないことを考慮しなければならず，負担が増える．本システムは仮想ユーザのプロファイルによって得られる様々な集合知を取

得することを目的としているため，プロファイルの結果は可能な限り高精度であることが求められる．

- 管理の容易性
  ネットワーク管理者が既存研究によってユーザのプロファイル作成を試みる場合，その負担は非常に大きなものとなる．例えば，前述した 802.1x 認証などを導入・維持するための負担が挙げられる．管理者の運用負担が増大すると，集合知を形成する動機が失われる可能性があるため，管理負担の低減が必要である．

- ユーザ識別の即時性
  目的とするユーザの識別は即時に行われる必要がある．ユーザ全体のトレンドは刻々と変化するため，可能な限り短時間で把握することが望ましい．

- ユーザの網羅性
  集合知は対象とするユーザの数によって集合知の価値が大きく変化するため，より多くのユーザを対象とするべきである．ユーザがシステムに登録する方法や特定のユーザのみから情報を収集する方法で獲得した集合知よりも，可能な限り多く取得した情報を分析することで，より多くの集合知の取得が期待できる．

4 関連研究

4.1 ベイズ統計を用いたユーザ嗜好の分析

事例ベース推論という研究とベイズ統計とよばれる統計研究を組み合わせることによってユーザの好みを検索する”Profiling Case-Based Reasoning and Bayesian Networks”[2] という研究がある．この研究はあらかじめデータベースに登録したデータを元にユーザの行動の頻度や傾向，他のユーザに対する影響度などを収集し分析することによってユーザを識別する．しかし，この研究は事前にユーザを登録する必要があり，取得する情報もデータベースが保有する情報しか利用できないため，網羅性に欠けていると言える．

4.2 クライアントエージェントを用いたユーザ情報収集

ユーザの使用するホストなどの端末にエージェントをインストールすることによって，ユーザの傾向や振る舞いを識別する研究がある．このアプリケーションによる研究は，” 高度なパーソナライズ実現のためのユーザプロファイル統合サービスエージェントの設計”[3] をはじめとして広く研究されている．これらの研究は，エージェントによってユーザを識別するが，すべてのクライアントにエージェントが導入されていなければならないため，ネットワークの管理が容易ではなく，ユーザの網羅性も欠けていると言える．

4.3 受動的にネットワーク上で情報を収集

“Passive Network Discovery for Real Time Situation Awareness”[4] は様々な Passive finger printing を利用することによって，ネットワークに負荷をかけることなく情報を取得し，ユーザを特定する研究である．この研究で用ている Passive finger printing によって取得できる情報は稼働中のホストや OS 情報，ホストの役割，提供サービス，プロトコル，ネットワークの IP アドレス設定である．しかし，この研究でユーザ識別に用いている情報は多いとは言えない．そして，収集したデータは統計解析していないため，ユーザ識別の精度を改良する余地がある．また，この研究の目的は，ネットワークトラフィックに着目することでセキュリティインシデントをリアルタイムで発見することであり，本研究の目的とは異なる．

次に”BLINC : Multilevel traffic classification in the dark”[5] はパケットのヘッダ情報を利用してネットワークに流れるプロトコルやアプリケーションを把握する研究がある．この研究はネットワークにおいてのトラフィックをアプリケーションごとに分類することが目的であり，本研究の目的と合致しない．

4.4 ユーザに着目した Web 統計解析サービス

ユーザの情報を統計解析することで，Web アクセスの統計情報を提示するサービスの例として”Google Ad Planner”[6] が挙げられる．統計情報として，Web サイトを閲覧したユーザの性別，年齢層，世帯収入，キーワードなどを表示できる．このサービスは，google アカウントを保持しているユーザの履歴や外部調査データを元に統計結果を算出している．このような Web 統計解析によるサービスは多数存在する．しかし，この研究は大規模検索サイトを保有している人や組織でしかこのサービスを扱うことがで

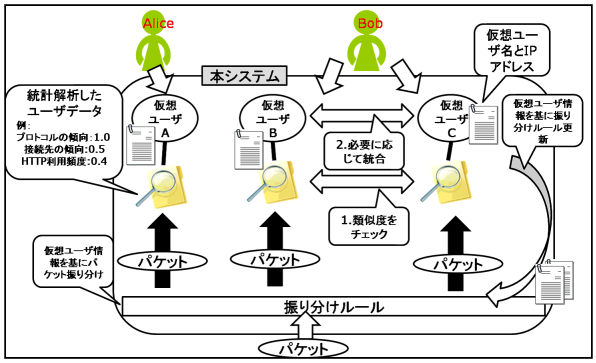

図 2: システム動作概要

きない．また，統計解析などのアルゴリズムなどは公開されていないため，精度を検証することは困難である．

5 アプローチ

3 節での必要要件を満たすシステムを提案する．本システムはネットワークの中継地点にトラフィック監視装置を設置し，定常的にパケットのヘッダ情報を収集することで，仮想ユーザを識別する．仮想ユーザとは，ホストの利用者である実ユーザの振る舞いや特徴などの集合を元にした識別子であり，実ユーザを現実に存在するユーザと定義する．

5.1 仮想ユーザに対する識別

ネットワークのトラフィックから各実ユーザの送受信するパケット情報と接続位置や利用時間など情報を分析することで，仮想ユーザを識別する．仮想ユーザに関する情報は，ネットワークの中継地点に本研究を用いたトラフィック監視装置を設置することで，定常的に収集する．そのため，本システムの利用者は対象ネットワークの通信を監視する権限保有者，もしくはネットワーク管理者から許可された実ユーザである．

システムの概要を図 2 の例で述べる．実ユーザの Alice は，本システムにおいて仮想ユーザ 1 として推定され，扱われる．そのため，仮想ユーザ 1 に関する情報から直接実ユーザの特定が困難となる．仮想ユーザは実ユーザの特徴など，推測できる情報によってプロファイル情報を作成する．このプロファイル情報を元に，仮想ユーザはトラフィックから位置情報やホスト情報など集合知の獲得に必要な情報を取得し，データベースに格納する．しかし，仮想ユーザ識別に利用する情報は，実ユーザの特徴やふるまいを元に作成されるため，Bob のような複数にまたがる仮想ユーザを想定する．その場合は，定期的に，仮想ユーザ間同士で共通事項を探し，発見された場合は該当する仮想ユーザを統合する．

5.2 システム設計

仮想ユーザを識別する本システムは，大きく 3 つの動作に分けられる．IP アドレスから仮想ユーザを判定する **IP 判定モジュール**，ホスト上のアプリケーションの動作からホストを分類する**パターン解析モジュール**，実ユーザの振る舞いを統計解析する**統計解析モジュール**の 3 つである．

本システムの設計を図 3 で示す．

1. **IP アドレス判定モジュール**
   IP アドレス判定モジュールは，パケットの IP アドレスを基に，仮想ユーザを推定する．ネットワークの DHCP リースタイム時間内に，該当する IP アドレスはすべて同じユーザであると判断する．しかし，DHCP のリースタイム時間内に該当する IP アドレスから通信がなかった場合は，そのユーザはネットワークから離れたものとする．

2. **パターン解析モジュール**
   ホスト上のアプリケーションの動作からホストを分類する．このモジュールは状況に応じて複数個作成する．また各モジュールごとにパターンを解析するアルゴリズムは変化する．例えば，メールサーバにアクセスする頻度や回数をユーザを分類するための識別子とする．

3. **統計解析モジュール**
   このモジュールは，ユーザの振る舞いをベースに統計解析を行い，類似する仮想ユーザを探す役割を持つ．例えば，ネットワークの接続時間，生活時間と接続先サーバなどの傾向から，関連性を見つけ出し，類似する仮想ユーザを発見する．このモジュールもパターン解析モジュールと同様に，用途に応じて複数作成する．

各モジュールで類似する仮想ユーザが発見した場合，ユーザが保有する情報は該当する仮想ユーザの情報に統合される．そして，これらのモジュールによって新規ユーザが既存の仮想ユーザとマッチしなかった場合にはじめて，新たな仮想ユーザが作成される．これら一覧の作業を繰り返すことによって，仮想ユーザを識別する．

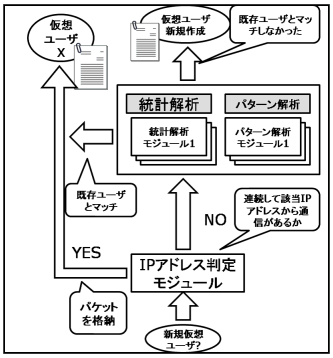

図 3: 仮想ユーザ識別のモデル

表 1: 送信先 IP アドレス上位リストの類似調査

	UserA	UserB	UserC
1 回	1	0	1
2 回	0	1	0

5.3 ユーザ識別に用いる情報

ユーザのプロファイル作成に用いる情報は，ユーザを推定できる情報である．この情報は，各個人が持っている独特の傾向のことを指す．同一ホストでもユーザの使用方法によって各個人は判別可能である．その推定できる情報を組み合わせることによって仮想ユーザのプロファイルを作成する．推定できる情報を以下に記述する．

- **パケットヘッダ情報**
  各ユーザはそれぞれの利用状況に応じた特徴的なパケットを送受信している．ヘッダ情報からは，送信先・発信元 IP アドレス，ポート番号から使用しているサービス，アプリケーションの使用頻度の情報が取得できる．そして，送信先 IP アドレスからはユーザの通信相手の情報が把握可能であるこれによって，該当ユーザはどのようなホストとの通信する傾向があるかを把握できる．そこで，本システムは各ユーザの送信先のリストを保持し，最も多くの通信をしている送信先ホストの上位を識別子とし，ユーザ識別を行う．

  同時に，各ユーザの送信先 IP アドレス上位はユーザの識別子となりえるのかの調査を行った．調査したネットワークは筆者が所属する研究室でのネットワークであり，取得期間は，2009 年 6 月 25 日 19 時 45 分から 22 時 20 分までと 7 月 12 日 0 時 46 分から 06 時 14 分までである．

  その結果は表 1 に示す．各期間における調査対象のユーザの上位 5 位のリストうち，UserA と UserC は同じ IP アドレス一つ，UserB に関しては二つ見られた．このことから，ユーザの送信先 IP アドレス上位 5 位のリストはユーザの識別子の一つになることが期待できる．

  また，ホストが利用する特徴的なプロトコルもユーザの識別子となる．ホストとユーザは密接な関係があるため，ホストの識別はユーザの把握の一因となり得る．そこで，本システムはパケットのヘッダ情報を利用してホストを識別する．例えば，ホストがファイル共有やネット BIOS，プリンタの検索などを行った場合，特徴的なパケットを出力する．この特徴を利用することによってホストを識別する．

- **起動時間・接続頻度**
  ユーザがホストをネットワークに接続した時間や接続時間帯の規則性を記録して，本人の生活習慣からもユーザの識別子となる．人間の生活習慣は多少のぶれが生じるが，傾向を把握することによって，パターンを取得できる可能性がある．そのため，ユーザがネットワークに接続する頻度やその接続時間はユーザの識別の材料となる．これに加え，本システムはホストを起動時に一番最初に利用するプロトコルや通信傾向も識別子とする．

- **接続位置**
  ネットワーク上の様々な位置に本システムを設置することによって，ホストの接続位置を取得する．これによって，ユーザの行動範囲を把握することが可能となる．

- **ユーザが利用するサービス**
  ユーザが利用する送信先アドレスや，ポート番号から分かるサービスを識別子とする．例えば，Web サービスを使用する場合，利用している Web サイトによってユーザをプロファイリングできる．サービスの利用頻度や傾向はユーザごとに差異があるため，ユーザの識別子としても有効である．このため，Web アプリケーションや SNS を利用する頻度や時間帯を各ユーザごとに調べることによって類似するユーザを調査する．例えば，mixi[7] や Twitter[8] などにアクセスする時間帯や間隔を各ユーザごとに記録する．これらはユーザ特有の傾向であるためユーザ識別の識別子となりうる．また，ユーザの所

属するネットワークの Mail サーバ，Web サーバへのアクセス情報も重要である．そこで本研究は各ユーザのアプリケーションやサービスの利用頻度やアクセスする間隔に用いてユーザを識別する．

- **OS の種類**
  ユーザが主に使っているホスト OS も十分な識別子となる．IP ヘッダや TCP ヘッダを組み合わせて解析することによって，ユーザの OS 情報を知ることができる．Passive OS Fingerprinting を利用することによってユーザの所有しているホストを把握する．

6 今後の取り組み

6.1 実装及び評価

本システムは 3 章で述べた設計を C 言語で実装する．また，評価に関しては識別した仮想ユーザと実ユーザの合致数を比較する．評価のデータの取得方法は，筆者が所属する研究室のネットワーク環境で実験する．その際，許可を得てペイロード解析することで，仮想ユーザが実ユーザと合致しているのか確認する．評価の基準値は，本研究の目的によって変化するため，利用ケースに基づいて最適な値を模索する必要がある．

6.2 考察

本システムの要件に対する充足度について以下に述べる．本システムは，ネットワークの中継地点に設置するだけでユーザ識別が可能になるため，管理者による運用が容易である．次に，ユーザプロファイルの精度に関しては，多くのユーザに関する情報を集め，本人と推測できる情報を多数組み合わせることによって精度の向上を図る．その際に，データマイニング研究を用いることによって誤認識を防止する．そして，定常的に情報を収集し，リアルタイムに仮想ユーザを識別るため，即時性に富んでいる．また，本研究は大規模ネットワークの上流で利用するため，ネットワーク上すべての実ユーザを対象とすることから，ユーザの網羅性はあると言える．また，モジュールを増設が容易であることから拡張性にも優れている．

本システムはパターン解析モジュールや統計解析モジュールを追加することで，より高精度のユーザ識別が可能となる．そのため，ユーザやホストの特徴や傾向を把握し，相関があるデータを議論する必要がある．

7 まとめ

本稿ではパケットのヘッダ情報のみを用いて，ユーザを識別するシステムを提案した．ネットワークにおいて，実ユーザの実態調査を行う場合，一人で複数のホストを保有するユーザや共有ホストを利用するユーザを識別することができないといった問題点や，ネットワークごとに取得できる情報が異なるという問題が存在する．そこで，本システムはネットワークの中継地点に設置し，定常的にヘッダ情報のみを収集し，ユーザを識別する．ユーザ識別には，ホストの識別やデータマイニングを用いる．そのため，汎用性やユーザの網羅性，拡張性に優れている．今後の課題として，ユーザ識別に利用する情報や情報の組み合わせ手法について更に議論する必要がある．

参考文献

[1] CybozuLabos. pathtraq. `http://pathtraq.com`, 5 2009.

[2] Schiaffino Silvia N and Analia Amandi. User profiling case-based reasoning and bayesian networks. *7th Ibe-American Conference on Ai and Brazilian*, 2(1):19–22, 11 2000.

[3] 山崎賢児 and 勅使河原海. 高度なパーソナライズ実現のための統合サービスエージェントの設計. *IPSJ SIG Technical Report*, pages 105–110, 3 2005.

[4] Annie De Montigny-Leboeuf. Passive network discovery for real time situation awareness. 4 2004.

[5] T. KARAGIANNIS. Blinc : Multilevel traffic classification in the dark. *ACM Sigcomm, Philadelphia, PA, Aug. 2005*, 2005.

[6] Google. Google ad planner. `https://www.google.com/adplanner/planning/`, 5 2009.

[7] Inc mixi. [mixi]. `http://mixi.jp`, 9 2009.

[8] Twitter. twitter. `http://twitter.com`, 9 2009.