研究計画書
ネットワークのトラフィック解析による集合知の獲得

慶應義塾大学 総合政策学部

自著：＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

希望プログラム:サイバーインフォマティクス（CI）

平成 21 年 5 月 28 日

概要

情報技術の発展に伴い，これまで無価値だった多様な情報を結びつけて新しい価値を創造できるようになった．インターネットの通信にも多様な情報が含まれており膨大なトラフィックを収集・解析することで，ユーザの利用形態やトレンドを調査できる．そこで，本研究は ISP や企業・組織などの様々なネットワークに流れるトラフィックに対して，データマイニングを利用して仮想的なユーザのプロファイルを作成し，パケットのヘッダ情報や位置情報といった情報を結びつけることで，ネットワークに潜在的に存在する集合知を得ることができる．また，プライバシへの配慮からユーザとの個人情報を直接結び付けないため，プライバシに配慮しつつユーザ全体の興味や活動の変化といったネットワークでの集合知をリアルタイムに得ることができる．これによって，インターネット出現に伴う様々な社会動向の変化が調査可能になり，多様な分野の研究に対して貢献が期待できる．

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．情報の伝達や集計が少ない労力や短時間でできるようになり，これまで知られていなかった事実を明らかにできる．例えば，走行中の自動車のワイパー情報を収集することで，雨が降っている地域をリアルタイムに把握する試みや，Where 2.0[1] というホスト情報と位置情報を取得することで，容易に詳細な地図で作成するプロジェクトがある．このように多くの情報を収集・分析することで生み出される新しい価値ある情報を，集合知と定義する．集合知を獲得する情報源の一つに通信ネットワークのトラフィックが挙げられる．トラフィックには，ホストの起動時間や，サービスの利用状況，位置情報など多くの情報が含まれており，それらを組み合わせることで非常に有益な情報の取得が期待できる．

ネットワークトラフィックの解析においては，仮想的にユーザを識別し，トラフィック情報とマッピングすることで，ネットワーク上の潜在的な情報を発見することや，集合知に新たな価値を付与することができる．例えば，ネットワーク上で利用されているアプリケーションの定着率を把握できる．どの地域の，どのようなユーザがどのアプリケーションに変更したのかについても把握できる．他にも，ユーザ情報と位置情報を組み合わせることで，より円滑なコミュニティの支援が可能となる．コミュニティのメンバの位置情報によって，最もメンバが多く集まる時間帯や場所に関する情報を取得したり，コミュニティ内での興味動向を取得できる．これらの例のように，仮想ユーザのプロファイルによって，ネットワーク上でのユーザに関する行動科学や社会学的な分析が可能となり，インターネットを活用した新しい研究アプローチとして期待できる．しかし，ユーザをプロファイルするためには，膨大なトラフィックからユーザを識別し，目的とする情報を探し出す必要がある．そこで，ネットワーク上で収集したユーザ情報を，データマイニング技術を用いて組み合わせることで解決する．

本システムは大小様々な規模のネットワークにおいて送受信されるトラフィックを収集し，データマイニング手法を用いて仮想ユーザをプロファイルすることでネットワークに関する集合知の取得を目的とする．また，本システムの想定する利用者は，該当するネットワークの管理者，もしくは管理者から許可されたユーザである．しかし，仮想ユーザのプロファイルに利用する情報は，プライバシの問題と密接な関係があり，個人が特定されることによる被害が増加 [2] する近年において，プライバシの配慮は必要不可欠である．そこで，本システムはパケットのヘッダ情報や位置情報などを利用し，ペイロードを解析せずに仮想ユーザを作成することで，プライバシに配慮する．仮想ユーザを作成し，トラフィック情報とマッピングすることで，ネットワーク上で新たな集合知を獲得できる．そして，利用者が目的のデータを容易に収集・解析が可能であり，かつプライバシに配慮したシステムの設計，実装する．

2　現状のネットワークトラフィックによる集合知獲得に関する問題点

ネットワークトラフィックから集合知を得る場合，無差別に情報を収集するのに比べ，ユーザを区別することで，より有意義な集合知を得ることができる．集合知獲得におけるユーザの区別は，IP アドレスを用いる場合が多い．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，ユーザの区別には適さない．例えば，ネットワークの運用手法によって IP アドレスの付与方法は変化する．IP アドレスが付け変わると，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．また，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や，NAT が導入されるネットワーク環境下において，正確に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手法としてユーザが登録したサービスの情報を利用する手法が挙げられる．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[3]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，このような手法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 問題解決手法の必要要件

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．

- 管理の容易さ
  ネットワーク管理者が既存手法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，リアルタイムで把握することが望ましい．このことから，ユーザの識別は即時に行われる必要がある．

- 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．

- ユーザの秘匿性
  第 1 章で述べた通り，近年ではユーザが特定されることで，個人が侵害されるプライバシ問題が注目されている．そのため，ユーザの集合知を取得する際に，ユーザを直接特定できないようにするなどプライバシに配慮する必要がある．

4 関連研究，手法

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手法とベイズ統計とよばれる統計手法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」[4] という研究がある．この手法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，この手法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手法がある．このアプリケーションによる手法は，「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」[5] 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」[6] 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．

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として Google Ad Planner[7] 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

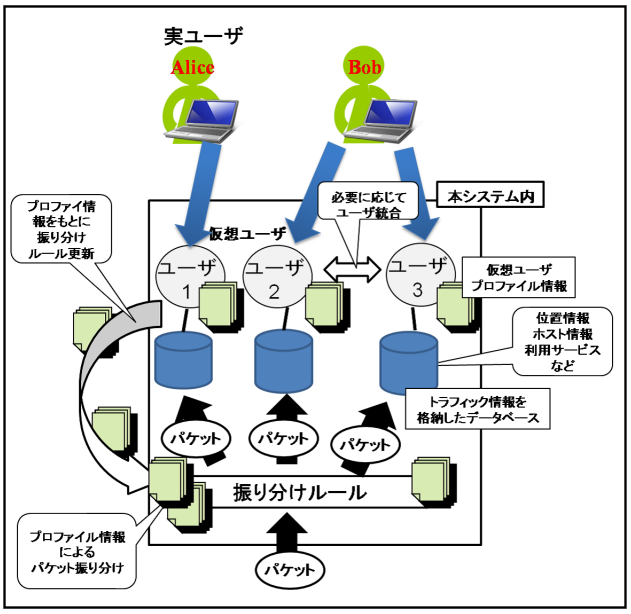

図 1: 本研究動作概要

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インタレストマッチ [8] をはじめとして多数存在する．しかし，この手法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

本研究では仮想ユーザのプロファイルを作成するにあたり，データマイニング手法を用いることで，膨大な量の情報からユーザプロファイルを作成する．ここで述べる仮想ユーザとは，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プロファイル情報を元に，実ユーザを推測したものである．

5.1 仮想ユーザをプロファイルする手法

ネットワークのトラフィックから各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，仮想ユーザのプロファイルを作成する．ユーザに関する情報は，ネットワークの中継地点に本手法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ユーザである．本システムで用いるデータマイニング手法は，サポートベクターマシンと呼ばれる複数のユーザ情報と，線形入力素子を利用し，2 クラスのパターン識別器を構成する手法や，ベイズ統計という事象発生の確率を求めるなど数多くの手法が存在する．各データマイニング手法を検証し，仮想ユーザのプロファイルがより正確にできる適切なデータマイニング手法を考案する必要がある．

システムの概要を図 1 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．本システムは，仮想ユーザ間のトラフィックデータベースから関連性を抽出することによって，集合知を獲得する．しかし，プロファイル情報は，ユーザの特徴やふるまいを元に作成されるため，Bob のように複数の仮想ユーザを推測できる実ユーザを想定する．そのような場合は，定期的に，仮想ユーザ間同士で共通事項を探し，発見された場合に該当仮想ユーザを統合する．

本システムで扱うユーザ情報は，プライバシの問題と密接な関係がある．電気通信事業法や個人情報保護法があり，ユーザプライバシに配慮しなければならない．そのため，ユーザの秘匿性に関して法的側面の問題について十分に考慮する必要がある．本システム利用者は，ユーザプロファイルの際に，ユーザが許容できる情報を規定する必要がある．例えば，データ自体はどこまで収集し，ユーザ識別に利用可能かという問題が挙げられる．ヘッダ情報は閲覧可能なのか，位置情報やホスト情報だけなのかなどのレベルはネットワーク，もしくはネットワーク管理者ごとに異なる．その問題に対して，本システムの取得するデータやその取扱い，ユーザプロファイル結果の使用方法についてのガイドラインを作成し，本システムを適用するネットワークユーザに同意を得ることで，プライバシ問題を考慮する．同時に，本研究では可能な範囲であらゆる情報を利用するが，どの段階においてても仮想ユーザをプロファイルできるようにシステムを設計する．

5.2 本システム動作

本システムは，大きく 2 つの動作に分けられる．実ユーザのふるまいや特徴から仮想ユーザを作成し，実ユーザから仮想ユーザを推論するプロファイル動作とプロファイル情報を用いて，トラフィックから各仮想ユーザの情報を収集し，各データベースに格納する動作である．

- **仮想ユーザのプロファイル情報の作成**
  まず，本システムはパケットを取得時に，パケットの IP アドレスと，仮想ユーザのプロファイル情報を比較する．仮想ユーザのプロファイル情報に含まれる IP アドレスは，一定時間該当する IP アドレスから通信が発生しない場合削除される．図 2 で示すように，IP アドレスがマッチした場合，パケット情報は該当仮想ユーザのプロファイル情報に付与される．もし，IP アドレスがマッチしなかった場合に，ポート番号や位置情報などパケットの IP アドレス以外の情報を参照し，既存の仮想ユーザプロファイル情報と類似するかを調査する．類似していた場合は，同じくそのパケットの情報を該当する仮想プロファイル情報に追加・更新す

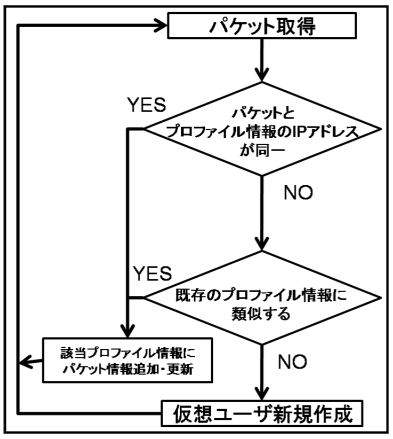

図 2: ユーザプロファイル動作

る．類似データが発見されなかった場合に，新しい仮想ユーザを作成する．これを繰り返すことによって，仮想ユーザやプロファイル情報を作成する．

- **仮想ユーザのデータベースに格納する動作**
  本システムは，各仮想ユーザのプロファイル情報を用いて，トラフィックから情報を収集する．仮想ユーザは保持しているプロファイル情報を用いてることで，トラフィック情報を分析する．トラフィック情報が該当する仮想ユーザのデータであると判断すると，データベースに格納する．これらを繰り返すことによって，仮想ユーザとトラフィック情報のマッピングを行う．利用者は仮想ユーザとマッピングした情報を利用することで，集合知を獲得できる．

5.3 本システムの必要要件に対する充足度

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集合知を得ることが可能になるため，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，データマイニング手法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，リアルタイムにユーザをプロファイルするため，即時性に富んでいる．また，本手法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．最後に，ユーザを抽象化することで直接特定できないように配慮し，秘匿性を保持している．

5.4 プロファイルの作成に用いる情報

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判明する．それによって，該当ユーザはどのホストと通信を多くする傾向があるかなどの情報を把握できる．この様に，ヘッダ情報を解析することによって多くの情報を得ることができる．

- **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個人のプロファイルを作成できる．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．また，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プロファイルを作成する材料となる．

- **接続位置**
  ネットワーク上の様々な一に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でき，プロファイルに利用することができる．

- **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．

- **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．

6 実現する世界

本研究では，ネットワーク上でユーザを識別することによって，様々な情報に新しい価値を付与することが期待できる．そして，ネットワーク上でユーザを識別し，様々な集合知を容易に取得することができる．以下に想定される利用例を挙げる．

- ユーザの興味動向の把握
  本システムは，仮想ユーザをプロファイルすることによって，ユーザの興味動向の把握が可能となる．. Web サイトの解析の際に，仮想ユーザからは，時間帯，地域情報などの情報が取得できる．そのため，話題や商品に関して，どの地域のユーザが興味を持っているかなどを把握することができる．これによって，広告効果の計測や，キーワードの注目度を把握することがで

きる．同時に，そのキーワードが，どのような人に，いつ，どこに広がるのかを追跡することも可能であるため，マーケティングから世論調査まで幅広く利用できる．

- ネットワーク上におけるユーザ行動の把握
  ある時点において，ユーザがネットワーク上で同じ行動をした際の原因や法則の調査ができる．それによって，現象発生時のユーザ行動の関連性などを調べることで，ネットワークにおける潜在的な法則や関連性の発見が期待できる．

- 場におけるユーザ利用率の把握
  本手法で，ユーザと位置情報を関連付け，どの場所が一番人が多いのかを把握することができる．社内や教室などの場において，何人のユーザが利用しているかを把握するためにはセンサーなどを設置する必要があったが，本手法を用いることで，トラフィックからユーザ人数の把握が可能となり，人数に対して場の広さについて適切さを把握することができる．

- ユーザが最も多い時間帯の把握
  ユーザとホスト起動時間を組み合わせることによって，対象とするネットワークの利用時間帯を把握することができる．例えば，会社内であれば社員が一番多くいる時間帯や，最も人が少ない時間などを知ることができる．他にも，天候情報やユーザの興味動向に関する情報を組み合わせることで，サービスのアクセス数の予測をすることが可能となる．

7 これまでの活動

7.1 研究会活動

私は学部 2 年の秋に村井・徳田合同研究室に所属し，ネットワークとインターネットセキュリティを中心に学習してきた．研究室に所属した学部 2 年次にはパケットキャプチャを作成することで，ネットワーク上にどのようなトラフィックが送受信されているかを知ることができた．これによって，トラフィックの見方・解析の仕方の基礎を学んだ．また，研究と運用，実践の両立を目指し学習している．SFC Open Research Forum ではネットワーク構築のメンバーとして参加し，運用の上で多くの経験を積むことができた．他にも，少子高齢化社会に対応する新たな社会システムの創出を目的とする e-ケアプロジェクトのサーバ管理を行っている．それに加え，研究室のネットワークの管理，運用のグループにも参加している．

7.2 研究成果

学部 3 年次には，セキュリティインシデント発生時に被害を最小限に抑えるために，インシデント対応支援するシステムを実装し，村井研究室内のネットワークにて実験した．そして，私の研究を対外的に発表するため情報処理学会全国大会に参加し，以下の論文を執筆した．

上原雄貴，水谷正慶，武田圭史，村井純，セキュリティインシデント対応のためのユーザ特定支援システムの実装，情報処理学会第 71 回全国大会,March,2009.

8 志望理由

本研究は，ユーザに関する情報に対してデータマイニング手法を用いて，仮想ユーザにデータをマッピングし，類型化することによって，新たなデータや情報を発見することを目的とする．

本研究は村井・徳田研究室，ひいては慶應義塾大学大学院の一員として，インターネットの更なる発展に貢献できるものである．しかし，本研究の実現のためには，インターネットに関する知識だけでなく，法に関する知識，情報を分析する技術，社会学や心理学といった複数の専門分野を学ぶ必要がある．政策・メディア研究科では多くの分野を横断的に，かつ専門的に学ぶことができ，様々な専門分野や豊富な経験を持った指導者の方々が多数在籍しているため的確な指導を受けることができる．また，インターネットについて最先端の研究をしている村井・徳田研究室だけでなく，WIDE Project など，研究や実験をする環境が十分に整っている．このような実験できる大規模なネットワーク環境を保有している研究機関は非常に少ない．そのため，私は政策メディア研究科の進学を強く希望する．

参考文献

[1] O'Reilly. Where 2.0 conference 2008o'reilly conferences. http://en.oreilly.com/where2008/public/content/home, 5 2009.

[2] 法務省. 平成２０年における「人権侵犯事件」の状況について（概要）. http://www.moj.go.jp/PRESS/090327-2/090327-2.html, 5 2009.

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

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

[5] 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ユーザプロファイル統合サービスエージェントの設計. *IPSJ SIG Technical Report*, pages 105–110, 3 2005.

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

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

[8] Overture K.K. インタレストマッチ. http://ov.yahoo.co.jp/service/int/index.html, 5 2009.