G-33

DNS パケットを基に IP アドレスを利用者単位に分類する手法の精度向上

小川 考輝[†]　　川橋 裕[‡]

Koki Ogawa　Yutaka Kawahashi

1. はじめに

近年，インターネット利用率上昇に伴い，情報セキュリティに関するインシデントが多発，多様化している．インシデントが発生した場合，早急に原因の究明と対処を行う必要がある．

同一ネットワーク内での機器の判別を行う情報として，MAC アドレスが挙げられる．しかし，近年ではプライバシーの観点などから，自身の MAC アドレスをランダム化する機器が増加している．現在では，原因となっている機器の MAC アドレスを調査することで，インシデントに関係した IP アドレスを確認することが可能である．しかし，MAC アドレスのランダム化が一般的になることで，MAC アドレスを用いた方法でインシデントに関係する IP アドレスを発見することが困難になると考えられる．

以上のことを踏まえ，MAC アドレス以外の情報を用いて，IP アドレスを利用者単位に分類し，管理者が IP アドレスの一覧を確認することができる手法が必要であると考えた．

先行研究[1]では，ネットワーク上を通信する際に発生するパケットを取得し，パケットから取得できるアクセス先のドメイン名から送信元 IP アドレスを利用者単位に分類した．IP アドレスごとのドメインのアクセス数をベクトルで表現し，2 つのベクトルの類似度を計算することで，類似度の計算に用いられた 2 つの IP アドレスが同一の利用者であるのかを判断した．これにより，利用者が使用した IP アドレスの一覧を確認できるようにすることを可能とした．しかし，分類精度に課題があり，精度向上が求められていた．

本研究では，パケットから取得したアクセス先のドメイン名のうち，利用者単位に分類する際に不要なドメイン情報の選定および削除を行うことで，分類精度の向上ができると考えた．

2. 関連研究

関連研究である「Why We Still Can't Browse in Peace: On the Uniqueness and Reidentifiability of Web Browsing Histories」[2]は，ウェブブラウザの Firefox を開発している Mozilla 社が行った研究である．Firefox ユーザ 52000 人を対象に，1 週間分の閲覧履歴と，翌週 1 週間分の閲覧履歴を照合した．研究にて収集した約 66 万個のドメインのうち，50 個のドメインに対する閲覧履歴を使用したデータセットを用いた場合，約 50%の精度で個人を特定できたことがわかった．

さらに，150 個のドメインに対する閲覧履歴を使用したデータセットを用いた場合，識別率は約 80%以上に達したと述べている．以上のことからアクセス先ドメインによって個人の特徴を見出すことが可能であることが考えられる．

3. 提案手法

図 1 に提案手法の概略図を示す．

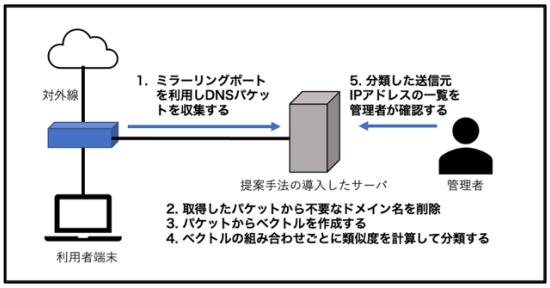

図 1．提案手法概略図

3.1 不要なドメイン情報の削除

収集の対象となるパケットは，Web アクセスを行った際に利用者端末から送信される DNS パケットである．その中に，利用者の特徴を見出すのに不要なドメイン情報も含まれている．取得したドメイン情報から，利用者の特定に不要なドメイン情報を削除することで，IP アドレスの分類精度がより向上すると考えられる．

本研究では，Web サイトに表示された広告のドメインを削除対象とした．今回収集したドメイン情報の内，広告に用いられるドメイン名の一部を図 2 に示す．

googleads4.g.doubleclick.net.
ad.doubleclick.net.
securepubads.g.doubleclick.net.
c.amazon-adsystem.com.
s.amazon-adsystem.com.

図 2．広告に用いられる DNS の一部

† 和歌山大学大学院 システム工学研究科

‡ 和歌山大学 学術情報センター

3.2 コサイン類似度の計算

DNS パケットから，IP アドレスとドメインのアクセス履歴を取得し，IP アドレスごとにベクトルを作成する．作成したベクトルの一部を図 3 に示す．その後，すべてのベクトルの組み合わせでコサイン類似度を計算する．2 つのベクトルがなす角のコサイン値により，類似性を判断する．値が 1 に近いほど，ベクトルの類似度が高いことを示す．コサイン類似度の値をもとに，2 つのベクトルの類似度を求める．図 4 に計算結果の例を示す．

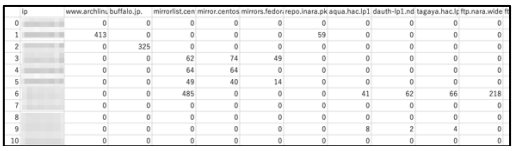

図 3. 作成したベクトルの一部

図 4. コサイン類似度の計算例

4. 実験・考察

本実験は，実験環境内の端末を自由に利用してもらい，環境内の端末から送信されたパケットを収集する．このとき，利用者には自身が使用した機器，時間，IP アドレスを記録してもらう．実験期間終了後に，送信元 IP アドレスとアクセスしたドメインの情報からベクトルを作成する．

コサイン類似度の計算を行った後，閾値を設定し，同一利用者であると考えられる IP アドレスの組み合わせを分類する．このとき，設定した閾値より低い類似度の IP アドレスの組み合わせを「類似なし」と判断する．1 回目の実験では閾値を 20%，2 回目の実験では閾値を 10%に設定した．

4.1. 実験結果

今回行った実験では，合計 1378 通りの IP アドレスの組み合わせがあった．その内，同一の利用者が使用した IP アドレスの組み合わせは 332 通りであることが，利用者の記録からわかった．

上記 332 通りのうち，先行研究および本研究の提案手法による判断の内訳を示す．その後，異なる利用者が使用した IP アドレスの組み合わせ 1046 通りの内訳を示す．

表 1. 同一の利用者が使用した IP アドレスの判断の内訳

	類似と判断	類似なしと判断	合計
先行研究 閾値20%	219	113	332
先行研究 閾値10%	293	39	332
本研究 閾値20%	220	112	332
本研究 閾値10%	291	41	332

表 2. 異なる利用者が使用した IP アドレスの判断の内訳

	類似と判断	類似なしと判断	合計
先行研究 閾値20%	102	944	1046
先行研究 閾値10%	227	819	1046
本研究 閾値20%	100	946	1046
本研究 閾値10%	227	819	1046

4.2. 考察

実験の結果，閾値が 20%の場合，正解数が 3 件増加し，閾値が 10%の場合，正解数が 2 件減少した．

図 5 の上から 2 行目の計算結果に注目すると，先行研究では類似度が 12%,本研究では約 2.4%であることがわかる．広告ドメインの削除によって類似度が低下していることから，同じ広告ドメインを使用していた可能性が考えられる．さらに精度が低下した原因として，不要なドメインを削除してしまったことにより，1 つの送信元 IP アドレスに対するドメインのアクセス数の合計が少なくなってしまった．その結果，利用者の特徴を見出すのに十分な情報量が取得できなかったことが考えられる．

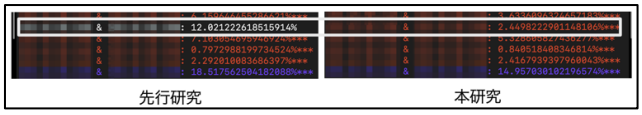

図 5. ある IP アドレスの組み合わせにおける計算結果

図 6 に，同一の利用者が別日にインターネットを利用した際のパケットの収集結果を示す．

図 6. 同一利用者の別日の記録

同じ利用者であるが，使用したドメインに大きな差が生じていることが判明した．これは，日によって利用者のインターネットの利用時間，利用目的が異なることが原因であると考えられる．先行研究では，1 日単位でパケットの収集を行っていたが，関連研究では 1 週間単位でインターネットの閲覧履歴を収集していたことから，本研究の提案手法を効果的に適用するには，パケットの収集期間を見直す必要がある．

5. おわりに

インターネット利用時に参照したドメイン情報に利用者ごとの特徴が現れることを利用して，ドメイン情報から IP アドレスの分類を行い，複数の IP アドレスを利用者単位で確認できるようにすることを目的とした．さらに，利用者単位に分類する際に不要なドメイン情報の選定および削除を行うことで，分類精度の向上ができると考えた．実験をおこなった結果，閾値が 20%の場合 1%の分類精度向上に成功した．今後は，さらなる分類精度の向上のため，パケット収集方法の見直し，アクセスポイントへの接続情報などの新たな情報を利用した手法を取り入れたいと考えている．

参考文献

[1] 堀内 亮汰 "IP アドレスをアクセス先ドメイン名に基づいて利用者単位に分類する手法の提案",2021 年度修士論文 和歌山大学大学院システム工学研究科

[2] Sarah Bird, Ilana Segall, Martin Lopatka, "Replication: Why We Still Can't Browse in Peace: On the Uniqueness and Reidentifiability of Web Browsing Histories", Mozilla, 2020-08-10, https://www.usenix.org/system/files/soups2020-bird.pdf , (参照 2021-10-20)