4ZD-02

フィンガープリンティングによる Tor ブラウザと一般ブラウザとのアクセス紐付けの試み

伊藤 颯汰 †　福田 江梨子 †　木檜 圭祐 ‡　川越 響 ‡　渡名喜 瑞稀 ‡
高山 眞樹 ‡　利光 能直 ‡　齋藤 孝道 †§
明治大学 †　明治大学大学院 ‡　レンジフォース株式会社 §

1 はじめに

The Onion Router(Tor) を用いたアクセスを可能にする Tor ブラウザでは匿名性確保のために様々な対策を行っている．その一つに端末を識別する技術であるブラウザフィンガープリンティングの対策がある．Tor の追跡に関する研究として，利光ら [1] は，Tor ブラウザから実験用 Web サーバに接続した際のアクセスデータを基に，同じ Tor ブラウザからのアクセスの識別が可能であることを示した．本論文では，同一端末における Tor ブラウザと Tor 以外の Web ブラウザ（以下，一般ブラウザと呼ぶ）からの実験用 Web サーバへのアクセスにおいて，ブラウザフィンガープリンティングを用いてアクセスの紐付けを行い，端末の識別を試みた．実験の結果，F_1 値で 0.89 と高精度での識別ができた．

2 関連技術

2.1 The Onion Router(Tor)

Tor は接続元を秘匿するネットワークシステムである．Tor を使用した通信は Onion Router(OR) と呼ばれる中継用ノードを経由することで，接続先にオリジナルの IP アドレスを知られずに通信できる．

2.2 Character-level CNN(CLCNN)

CLCNN は，文字列を文字単位に分割し CNN を用いて学習，分類する手法である．CNN(Convolutional Neural Network) は深層学習の手法の一つであり，Convolution 層と pooling 層と呼ばれる層を組み合わせて構築する．RNNLM や Word2Vec などは単語毎に分割する分かち書きが必要である．CLCNN では文字単位で学習するため，分かち書きが必要なく，誤字や脱字による分類への影響が少ないという特徴がある．

3 データセット

3.1 アクセスデータの収集

実験を行うにあたり，どの端末からのアクセスかを識別するための識別子を入力する入力フォームが用意された実験用 Web サイトを用意した．

アクセスデータの収集は 2 回実施した．1 回目は，2020 年 11 月 21 日から 2020 年 12 月 27 の間，29 端末を対象に実施し，Tor ブラウザから 1,192 件，一般ブラウザから 949 件のアクセスデータを収集した．2 回目は，2021 年 10 月 24 日から 2021 年 11 月 18 日の間，70 端末を対象に実施し，Tor ブラウザから 1,419 件，一般ブラウザから 1,356 件のアクセスデータを収集した．

全体として，99 端末における Tor ブラウザからのアクセスデータ 2,611 件と一般ブラウザからのアクセスデータ 2,305 件を収集した．

3.2 ベクトルデータの作成

収集したアクセスデータのうち，HTTP POST リクエスト部分をベクトルデータに変換する．全てのアクセスデータに共通する文字列を削除し，Tor ブラウザからのアクセスデータの先頭と一般ブラウザからのアクセスデータの末尾を連結した．本論文では，この連結した文字列を組み合わせデータと呼ぶ．CLCNN に入力するため，Unicode 値に変換し，Keras のライブラリである Embedding 関数により固定次元の分散表現に変換した．

3.3 教師データの作成

教師データは，アクセスデータの連結において識別子の一致有無に応じて作成した．同一端末によるアクセスであれば正解ラベルとして 1 を，異なる端末によるアクセスであれば不正解ラベルとして 0 を付与した．以下，正解ラベルの組み合わせデータを正解データ，不正解ラベルの組み合わせデータを不正解データと呼ぶ．

4 実験

実験 1 では，正解データに対する不正解データの比率（以下，TF 比率と呼ぶ）を収集したアクセスデータの組み合わせの TF 比率と同じにして学習を行った．全てのアクセスデータから全組み合わせ 6,018,355 通りの組

An Attempt to Link Web Access between Tor Browser and Normal Browsers with Fingerprinting
†Sota ITO †Eriko FUKUDA ‡Keisuke KOGURE ‡Hibiki KAWAGOE ‡Mizuki TONAKI ‡Masaki TAKAYAMA ‡Yoshinao TOSHIMITSU †§Takamichi SAITO
†Meiji University
‡Graduate School of Meiji University
§Rangeforce, Inc.

み合わせデータを作成した．その中からランダムに抽出した2割をテストデータ，残りの8割を学習データとしてモデルを作成した．学習データにおける正解データは93,216件，不正解データは4,721,468件である．

実験2では，TF比率を変えて学習を行い，実験1より高精度で識別可能なTF比率を調査した．実験1と同様に組み合わせデータを作成し，その中からランダムに抽出した2割をテストデータとした．これは実験1のテストデータと同じ組み合わせである．残りの8割の組み合わせデータからTF比率を変え，学習データとしてモデルを作成した．学習データにおける正解データは93,216件，不正解データは検証するTF比率に合わせてランダムに抽出した．

5 実験結果

5.1 識別精度の評価指標

識別精度の評価指標として，Accuracy，Precision，Recall，Specificity，F_1値，マシューズ相関係数(MCC)を使用する．また，表中の数値は，特に断りがない限り小数点第4位を四捨五入した数値である．

5.2 実験1

実験1の結果を表1にまとめる．表1から，RecallがAccuracyやPrecisionと比べて低く，同じ端末によるアクセスを異なる端末によるアクセスであると推定するケースが多いことが分かった．

表1: 実験1の結果

TF比率（正解：不正解）	Acc	Pre	Rec	Spe	F_1	MCC
1：51	0.994	0.955	0.703	0.999	0.810	0.817

5.3 実験2

実験2の結果として，TF比率を変化させた際の各評価指標の値を図1に示す．図1から，モデルの作成に使用する不正解データの比率を大きくするとPrecisionは高く，Recallは低くなることが分かった．

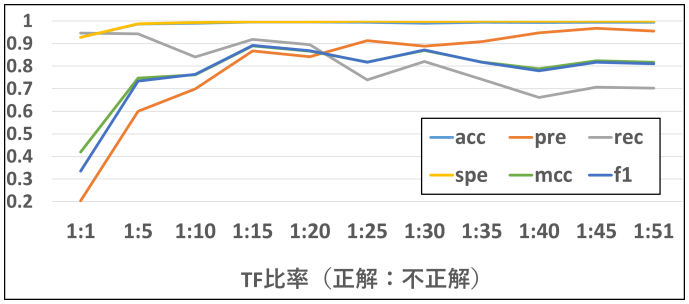

図1: TF比率に対する各評価指標の変化

Precisionと Recallの調和平均であるF_1値が最も高かった結果を表2にまとめる．表2から，TF比率を1：15に変えて検証した際，実験1の結果と比較してRecallが大きく向上していることが分かった．結果，F_1値が向上し，実験1よりも高精度の識別ができた．

表2: 実験2でF_1値が最も高い結果

TF比率（正解：不正解）	Acc	Pre	Rec	Spe	F_1	MCC
1：15	0.996	0.867	0.918	0.997	0.892	0.890

6 考察

実験1では，学習データとテストデータのTF比率を同じにして実験を行った．Recallが低いことから不正解データに過度に偏った学習であったと考えられる．実際にアクセスデータの識別を行う場合，学習データにおける不正解データの割合は大きくなり，実験1以上に偏りのある学習になると推測される．

実験2では，TF比率を調整することで不正解データに対する過度の偏りを軽減させた．図1からTF比率が1：1のとき，Recallは高い一方で，Precisionが低い．これは，不正解データに対する偏りが小さく，不正解データの比率が高いテストデータには不適であったと考えられる．結果として，TF比率が1：15のとき不正解データに対する過度な偏りが軽減され，高い精度で識別できたと考えられる．

7 まとめ

本論文では，Torブラウザと一般ブラウザからのアクセスデータを基に，フィンガープリンティングを用いてアクセスの紐づけを行い，CLCNNにより同一端末によるアクセスの識別を試みた．結果，F_1値で0.89と高精度で識別することができた．

謝辞

本研究の成果の一部は，JSPS科研費18K11305の助成を受けたものです．また，本研究はレンジフォース株式会社の支援により実施しています．

参考文献

[1] 利光 能直，齋藤 祐太，北條 大和，野田 隆文，齋藤 孝道，深層学習を用いたTor Browserアクセス識別の試み，2020暗号と情報セキュリティシンポジウム(2020).