4ZA-05

DNNを用いたパッシブフィンガープリンティング手法の提案と実装

北條　大和*　　細谷　竜平†　　齋藤　祐太†　　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

ブラウザフィンガープリンティング（以降，フィンガープリンティングと呼ぶ）には，JavaScriptやCSSを用いてブラウザから情報を採取するアクティブフィンガープリンティング，および，ブラウザから送信されるHTTPリクエストのヘッダ情報のみを使用するパッシブフィンガープリンティングの2種類がある．

パッシブフィンガープリンティングは，採取可能な情報が少なく識別精度の向上が見込めないとされていた．また，PCに対するフィンガープリンティングは高精度で識別できる手法が知られているが，モバイル端末の識別は困難であるとされている[1]．

本論文では，パッシブフィンガープリンティングのみで採取した情報を，深層学習（Deep Neural Network: DNN）で作成した予測モデルでモバイル端末の識別を行った．実験は，データ作成，学習，精度検証の3つの手順を取る．結果として，モバイル端末を99 %の精度で識別可能だとわかった．実験の流れを図1に示す．

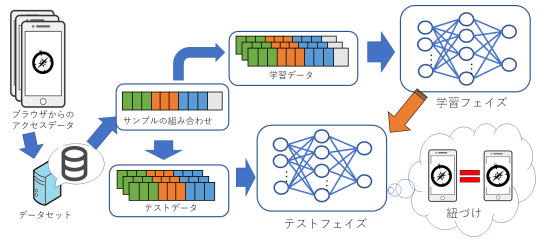

図1: 実験概要

2　フィンガープリンティング

フィンガープリンティングは，ブラウザから採取可能な情報を複数組み合わせ，ブラウザや端末の識別を行う技術である．採取された情報の種類を特徴点と呼び，特徴点の組み合わせや，特徴点の値をフィンガープリントと呼ぶ．この技術の2種類の分類を以下に示す．

1. パッシブフィンガープリンティング
2. アクティブフィンガープリンティング

パッシブフィンガープリンティングは，ブラウザからサーバへの通信の際に，ブラウザから送信されるHTTPヘッダや，IPアドレス，User Agent文字列の情報のみ利用し，端末やブラウザの識別を行う技術である．

アクティブフィンガープリンティングは，ブラウザ上でJavaScriptやCSSを実行することで得た情報を使用し，端末やブラウザの識別を行う技術である．

3　実験に使用する特徴点について

本論文におけるデータセットのサンプル数は約2,200万である．これらは，連続した15日間でモバイル端末のブラウザから採取した．サンプル内で実験に使用した特徴点を表1に示す．当該データセットにおけるOSは Android と iOS のみである．それぞれiOSが約1,200万個，Androidが約900万個のサンプルが存在する．

表1: 実験に使用する特徴点

特徴点	例
タイムスタンプ	1488898808
IP アドレス	192.168.100.1
OS 名	iOS
OS のバージョン	10.0.2
機種名	SOV32

実験では，任意の2つのサンプルを結合した組を複数用意し，使用する．実験にあたり，結合したデータに特徴点2種類を新たに生成した．

1つ目は，採取したフィンガープリントから導出できる情報（以降，生成した特徴点と呼ぶ）である．表2に，生成した特徴点を示す．ISP名は，pyisp[2]を使用して取得し，国や緯度などはGeoIP2[3]を使用して取得した．その他は元の特徴点を別々に分けて使用した．

表2: 生成した特徴点

元の特徴点	生成した特徴点
タイムスタンプ	年, 月, 日, 時, 分, 秒, 曜日
IP アドレス	第1オクテット，第2オクテット 第3オクテット，第4オクテット
IP アドレス	ISP 名
IP アドレス	国，県，市区町村，緯度，経度
OS バージョン	Major，Minor，Maintenance

2つ目は，2つのサンプルを比較した際の一致の有無を情報として持つ．比較した情報を持つ特徴点は以下である．

1. フィンガープリントが一致したかを示すラベル
2. 緯度と経度から計算した直線距離を示すラベル

4　実験概要

実験では採取したデータを，採取期間の前半7日間と，後半8日間に分けて使用し，前半のサンプルを初回

Proposal and Implementation of Passive Fingerprinting using Deep Neural Network
* Yamato HOJYO †Ryohei HOSOYA †Yuta SAITO *‡Takamichi SAITO
* Meiji University
†Graduate School of Meiji University
‡Rangeforce, Inc.

アクセス，後半のサンプルを次回アクセスと想定した．実験は，Android のサンプルのみ使用した場合，iOS のサンプルのみ使用した場合，両方の OS のサンプルを使用した場合の 3 パターンで行う．

実験により，予測モデルが次回アクセスのフィンガープリントから，端末やブラウザを識別可能かどうか検証する．実験を以下に定義する．

実験 1 前半と後半のサンプルを組み合わせた 2,000 万個のデータで，予測モデルの精度を検証する

実験 2 前半と後半のサンプルを組み合わせる際に，2 つのサンプルの日付の差を取り，差の絶対値に応じて 200 万個のデータを 14 つ作成する．作成したデータでそれぞれ精度を検証する

4.1 教師データ

教師データは端末固有の文字列である端末識別子から作成する．結合した組の端末識別子が一致していれば，教師データとして，正解ラベルを付与し，一致していなければ不正解ラベルを付与する．

4.2 ニューラルネットワークの構造

ニューラルネットワークの構造は，三層の中間層を持ち，各層は全結合層である．損失関数は交差エントロピー関数，最適化関数には Adam を使用した．また，本実験ではハイパーパラメータの最適化は行っていない．

4.3 予測モデルの作成

学習データとして，教師データの正解と不正解のラベルの数が均等になるように，前半のサンプルからランダムに抽出し，データを組み合わせた．Andorid のサンプルのみの場合，iOS のサンプルのみの場合，両 OS のサンプルを混合した場合それぞれについて 2,000 万個の学習データを作成し，3 つのモデルを学習する．

5 実験結果およびその考察

実験における精度の評価指標に，Precision，Recall，Accuracy，F_1 を使用する．この中で，F_1 を識別精度とし，予測モデルの性能の評価値とする．精度の検証には，iOS のみを学習したモデルであれば，iOS のテストデータを使用している．

表 3 と表 4 それぞれに，実験 1 と実験 2 の結果を示す．

表 3: 実験 1 の結果

各評価指標における数値	両 OS	Android	iOS
Precision	0.992	0.999	0.999
Recall	0.996	0.999	0.994
Accuracy	0.994	0.999	0.997
F_1	0.994	0.999	0.997

表 3 より，3 つのパターン全てにおいて，各数値が 0.99 以上という結果が得られた．この結果から，パッシブフィンガープリンティングのみでも，モバイル端末の識別が可能であると言える．

実験 2 では，次回アクセスまでの期間の長さに伴い，どの程度精度に変動があるかを検証した．表 4 に実験 2 の結果を示す．

表 4: 実験 2 の結果（数値は全て F_1 を使用する）

期間	両 OS	Android	iOS
1 日	0.975	0.964	0.973
2 日	0.978	0.964	0.899
3 日	0.985	0.938	0.893
4 日	0.984	0.963	0.897
5 日	0.978	0.930	0.821
6 日	0.982	0.958	0.843
7 日	0.959	0.970	0.775
8 日	0.958	0.965	0.808
9 日	0.918	0.938	0.770
10 日	0.914	0.959	0.727
11 日	0.930	0.959	0.725
12 日	0.975	0.930	0.685
13 日	0.947	0.940	0.794
14 日	0.961	0.965	0.855

表 4 から，iOS の場合次回アクセスまでの期間が長くなるにつれ，精度が低下する傾向がある事がわかった．

本論文のサンプルの機種名のフィンガープリントは，Android が 1,016 種，iOS が 16 種存在し，iOS のサンプルの 99 %が iPhone からのアクセスであった．Android は機種名のフィンガープリントが iOS に比べて種類が多く端末ごとに差が出やすいので，全体的に精度が高く識別がしやすかったと推察する．また，機種名や OS バージョンのフィンガープリントは変化することが少なく，機種名と OS バージョンの多様性が多いほど，長い期間も識別が可能であると推察する．

6 研究倫理

研究に使用したデータセットは提供者の同意を得て使用している．これらは研究のみに使用し，個人の識別はせず，他者への売却および提供をしない．

7 まとめ

本実験では，初回アクセス時のフィンガープリントを学習し，同一のモバイル端末を紐付けられるかを調査した．結果として，パッシブフィンガープリンティングのみでも，モバイル端末を 99 %の精度で識別可能だとわかった．

参考文献

[1] P. Eckersley，”How Unique is Your Web Browser?”, in Proc. of the 10th international conference on Privacy enhancing technologies (PETS’10)，2010

[2] pyisp，https://github.com/ActivisionGameScience/pyisp/

[3] GeoIP2，https://github.com/maxmind/GeoIP2-python/