5ZD-08

埋め込みモデルを用いたFingerprintのベクトル化による端末推定の試み

山本 美桜 † 市野 雅暉 ‡ 升田 尚幸 § 加藤 志門 ¶ 齋藤 孝道 ‖
明治大学 † 明治大学大学院 ‡ 明治大学大学院 § 明治大学大学院 ¶ 明治大学 ‖

1 はじめに

近年，Webサイト上でのなりすましをはじめとする不正行為が増加している．その対策の1つとして，ブラウザフィンガープリンティングを用いた不正端末の検知がある．先行研究 [1] では，推定対象と推定候補のフィンガープリント組を作成し，機械学習モデルを用いて推定した．しかし，端末候補の増加に応じて組数が増加し，推定時間が爆発的に増加する課題がある．課題解決には，推定候補数の増加に影響を受けない手法が必要である．埋め込みモデルは複数の特徴点を比較が容易な形式に圧縮するため，推定時間の短縮が期待できる．よって本論文では，埋め込みモデルを用いたブラウザフィンガープリントのベクトル化による端末推定手法の提案を行う．

2 関連知識

2.1 ブラウザフィンガープリンティング

Webブラウザが Webサーバにアクセスした際に取得できる情報の組み合わせをブラウザフィンガープリントと呼ぶ．ブラウザフィンガープリンティングとは，ブラウザフィンガープリントを用いて端末を推定する技術である [2]．

An Attempt at Device Estimation through Vectorization of Fingerprints Using Embedded Models
† Mio Yamamoto
‡ Masaki Ichino
§ Naoyuki Masuda
¶ Shimon Kato
‖ Takamichi Saito

2.2 Two-Tower モデル

Two-Tower モデルとは，類似したオブジェクトを同じベクトル空間にペアリングし，関連する候補を近接させることで主に推薦を行う機械学習モデルである．主に情報検索において推薦システムとして使用される．Two-Tower モデルで使用されるモデルについて以下に示す．

Query Tower
: 推薦対象と推薦候補の関係性を導くモデル

Candidate Tower
: 推薦候補の埋め込みを行うためのモデル

3 提案手法

Two-Tower モデルの学習について記述する．使用する特徴点として，端末識別子，タイムスタンプ，User-Agent 文字列，IP アドレスを使用する．端末識別子は端末に一意に付けられた値であり，推薦候補となる．以下に具体的な学習手順を示す．

1. IP アドレスから地理情報，ISP を算出し，関数 *user-agents* を用いて User-Agent 文字列から OS，ブラウザ，デバイス機種に関する情報を抽出する [3]
2. 先に示した特徴点と 1 で算出した値を使用して Query Tower を作成する
3. 端末識別子を使用して Candidate Tower を作成する
4. 2 つのモデルを利用し，最も適する上位 n 件の端末を推薦する Two-Tower モデルを作成する

端末推定の方法について記述する．学習手順の 1 と同様に新たな特徴点を算出し，端末識別子を除いた特徴点を用いて Two-Tower モデルに入力して推定を行う．

4 実験

提案手法を用いて，端末推定の精度および処理時間を求める．会員制 Web サイトから取得した 1 日分のアクセスデータ 120,000 件をデータセットとした．このうち，学習データには 100,000 件，テストデータには 20,000 件を使用して，Two-Tower モデルを用いた推薦を行う．

推薦結果の上位 n 件 ($n = 1,\ 5,\ 10,\ 50,\ 100$) に正解の端末が含まれていた場合正しく推定されたと判断し，その割合を正解率として算出する．また，推定にかかった時間を計測し，テスト件数で割ることにより 1 件あたりの推定時間を求める．

5 結果

実験結果を表 1 に示す．上位 1 件 ($n = 1$) の正解率は 0.71，上位 10 件以降 ($n \geq 10$) での正解率は 0.80 以上という結果が得られた．また，1 件あたりの推定時間は 0.0064 秒であった．

表 1　n に対する正解率の推移

上位 n 件	1	5	10	50	100
正解率	0.71	0.77	0.80	0.87	0.89

6 考察

先行研究 [1] における端末推定の正解率は 0.875 であった．先行研究の結果と比較しても，正解率に大幅な低下はなく推定が可能であると判断できる．さらに，組を作成する処理がないため，端末候補の増加による影響を受けにくい．よって，提案手法は有用であると考える．

提案手法は過去手法とは異なり，複数の候補を列挙できるという特徴がある．膨大な候補数を削減する方法としての利用も考えられ，推定時間の更なる削減が期待できる．

7 研究倫理

実験を行う際，個人識別はせずプライバシーを尊重した．論文中では，統計的処理によりオリジナルデータについての推察をされることがないようにした．また研究に使用されたデータセットは，学術的な目的にのみ使用し，我々の管理下で厳重に保管されており，他者への提供をしない．

8 まとめ

本論文では，埋め込みモデルである Two-Tower モデルを用いたブラウザフィンガープリントのベクトル化による端末推定手法について提案し，その有用性を確認するための実験を行った．結果として，先行研究と比べ大幅な精度の低下はなく，端末の推定が可能と判断できる．さらに，提案手法は端末候補数の増加による影響を受けにくいため，推定時間の短縮にも期待ができる．

参考文献

[1] 藤井達也, 渡名喜瑞稀, 利光能直, 柴田怜, 北條大和, 齋藤孝道. PC とモバイル端末における深層学習を用いた ID の推定手法の提案と実装. コンピュータセキュリティシンポジウム 2020 論文集, pp. 50–57, oct 2020.

[2] P.Eckersley. “how unique is your web browser?”. *in Proc. of the 10th international conference on Privacy enhancing technologies (PETS’ 10)*, 2010.

[3] 北條大和, 齋藤祐太, 齋藤孝道. 深層学習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. コンピュータセキュリティシンポジウム 2019 論文集, 第 2019 巻, pp. 252–259, oct 2019.