7X-05

クロスデバイスフィンガープリンティング手法の提案と実装

礒貝　竜真†　　利光　能直‡　　北條　大和‡　　菊田　翼‡　　高山　眞樹‡
藤井　達也†　　齋藤　孝道†
明治大学†　　明治大学大学院‡

1　はじめに

スマートフォンの普及に伴い，複数の端末を利用するユーザが増加した．2016年のPew Research Centerの調査によると，90%以上の米国の家庭に3台以上の端末があるとされている[1]．そのため，Webトラッキングにおいて，同一端末からだけでなく，異なる端末からのアクセスデータの紐づけも求められる．しかし，従来のブラウザフィンガープリンティング手法では，特に大規模なデータセットにおいて，異なる端末からのアクセスデータの紐づけが可能であるかの事例はあまりなかった．そこで本論文では，クロスデバイスフィンガープリンティング手法の提案と，1,617,292件のアクセスデータを用いた検証を行った．検証では，認証を伴う会員制のWebサイトにおいて，認証後のアクセスデータをもとに，認証前のアクセスデータの会員IDを推定することを想定した実験を行った．本実験の会員ID推定の概念図を図1に示す．PCおよびモバイル端末からのアクセスデータをそれぞれ20,000件，合計40,000件のアクセスデータに対してクロスデバイスにおける会員ID推定を行った．

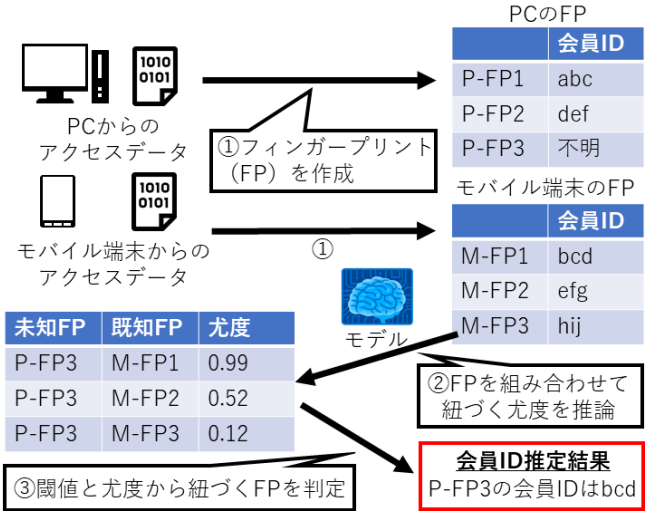

図1: 会員ID推定の概念図

Proposal and implementation of cross-device fingerprinting techniques
†Tatsuma ISOGAI　‡Yoshinao TOSHIMITSU　‡Yamato HOJYO
‡Tsubasa KIKUTA　‡Masaki TAKAYAMA　†Tatsuya FUJII
†Takamichi SAITO
†Meiji University
‡Graduate School of Meiji University

2　クロスデバイスフィンガープリンティング

ブラウザフィンガープリンティング（以下，フィンガープリンティング）とは，ブラウザがWebサーバにアクセスすることによって，Webサーバが取得できる情報をもとに，同一ブラウザまたは同一端末からのアクセスデータを紐づける手法である．また，Webサーバが取得できる情報を特徴点と呼び，取得した特徴点の組み合わせをブラウザフィンガープリント（以下，フィンガープリント）と呼ぶ．フィンガープリンティングの研究事例として，藤井ら[2]は，非クロスデバイスにおける会員IDの推定を行った．推定の精度は，4章で定義する会員ID推定の精度で0.776であった．

クロスデバイスフィンガープリンティングとは，フィンガープリンティングなどを用いて，同一ユーザの使用する複数の端末からのアクセスデータを紐づける手法である．クロスデバイスフィンガープリンティングの研究事例として，Zimmeckら[3]は，PCとモバイル端末をF_1値で0.91の精度で紐づけた．しかし実験参加者は126名であり，サンプル数が少ない．

3　実験の方法

本実験における既知とは会員IDが判明している状態を指し，未知とは会員IDが判明していない状態を指す．

3.1　データセット

本実験では，国内ウェブサイトにおいて28日間で収集した1,617,292件のアクセスデータ（以下，データセットOと呼ぶ）を用いた．内訳はPCから996,011件，モバイル端末から621,281件，また，アクセスデータには固有の会員ID（178,599種類）が付与されていた．

3.1.1　実験に用いた特徴点

本実験では，データセットO内のアクセスデータが持つ特徴点（表1）に加え，これらをもとに生成した特徴点（表2）を用いた．

表1: アクセスデータが持つ特徴点

特徴点	例
タイムスタンプ	yyyy-mm-dd hh:mm:ss
UA文字列	Mozilla/5.0 (Linux; Android ...
IPアドレス	aaa.bbb.ccc.ddd

3.1.2　データセットの分割

データセットOを表3のように分割した．データセットAおよびCは既知のデータセット，データセットB

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.

およびDは未知のデータセットを想定する．

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，UNIX時間形式のタイムスタンプ
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，ISP，国名，都市名，市区町村名，緯度，経度
UA文字列	OS，OSのバージョン，ブラウザ名，ブラウザのバージョン，機種名，機種のブランド名

表3: 分割したデータセット

データセットの種類	期間	デバイスの種類
データセットA	1〜21日	モバイル端末
データセットB	22〜28日	モバイル端末
データセットC	1〜21日	PC
データセットD	22〜28日	PC

3.2 モデル作成

同一会員IDを持つPCとモバイル端末からのアクセスデータを紐づけるモデルを作成した．モデルの学習および検証にはデータセットAおよびC，モデルのテストにはデータセットBおよびDを用いた．

3.3 会員ID推定

本論文では，未知のアクセスデータの会員IDを推定することを会員ID推定と呼ぶ．未知のデータセットBおよびDから古い順にそれぞれ20,000件のアクセスデータを抽出し，未知のデータセットXおよびYとし，合計40,000件のアクセスデータに対して会員ID推定を以下の流れで行った．まず未知のアクセスデータの会員IDが，既知のデータセット内に存在しないかを推定する（以下，会員ID無し推定と呼ぶ）．存在しないと推定した場合，会員ID推定を終了する．存在すると推定した場合，会員IDの値を推定する（以下，会員ID値推定と呼ぶ）．以下に具体的な手順を示す．

1. 未知のデータセットX内の各アクセスデータを，既知のデータセットC内の全てのアクセスデータと一対一で組み合わせ，ベクトルデータを作成
2. 3.2節で作成したモデルで推論し，作成した各ベクトルデータの会員IDが一致する尤度を求める
3. 最も大きい尤度の値が閾値以下である場合，会員ID推定を終了する．閾値より大きい場合，そのベクトルデータの会員IDを，会員ID値推定の推定結果とする

以上の処理を未知のデータセットYおよび既知のデータセットAにおいても同様に行った．

4 実験結果およびその考察

会員ID無し推定の精度の評価指標としてPrecision，Recall，Accuracy，F_1値を使用する．また，会員ID値推定および会員ID推定の精度を求める式を以下に示す．ただし，ID値推定正は会員ID値推定の正解数，ID値推定誤は会員ID値推定の不正解数，ID無し推定正は会員ID無し推定におけるTPである．なお，会員ID無し推定におけるTPとは，会員IDが存在しないと推定し，その推定が正しかった場合を示す．

$$\text{会員}\ ID\ \text{値推定の精度} = \frac{ID\ \text{値推定正}}{ID\ \text{値推定正} + ID\ \text{値推定誤}}$$

$$\text{会員}\ ID\ \text{推定の精度} = \frac{ID\ \text{無し推定正} + ID\ \text{値推定正}}{\text{推定したアクセスデータ数}}$$

実験の結果，会員ID無し推定の精度，会員ID値推定および会員ID推定の精度は表4，表5のようになった．

表4: 会員ID無し推定の精度（小数点第6位を四捨五入）

Precision	*Recall*	*Accuracy*	F_1
0.89554	0.94200	0.873	0.91818

表5: 会員ID値推定および会員ID推定の精度（小数点第5位を四捨五入）

会員ID値推定	会員ID推定
0.6985	0.8553

高精度に会員ID推定を行うことができた原因は，データセットO内の多くの会員が，特有の特徴点の値を持っていたからだと考えられる．例えばデータセットO内の約9割の会員が特有のIPアドレス，約6割の会員が特有のUA文字列の値を持っていた．

5 まとめ

本論文では，PCおよびモバイル端末からの1,617,292件のアクセスデータを用いて，クロスデバイスフィンガープリンティングを利用した会員ID推定を行った．結果として，40,000件のアクセスデータの会員ID推定を85%の精度で行うことができた．

参考文献

[1] A third of Americans live in a household with three or more smartphones, http://www.pewresearch.org/fact-tank/2017/05/25/a-third-of-americans-live-in-a-household-with-three-or-more-smartphones/(2017).

[2] 藤井 達也，渡名喜 瑞稀，利光 能直，柴田 怜，北條 大和，齋藤 孝道，2020，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（CSS2020），pp.50-57

[3] Zimmeck, S., Li, J. S., Kim, H., Bellovin, S. M. and Jebara, T. A privacy analysis of cross-device tracking, 26th USENIX Security Symposium (USENIX Security 2017) (2017).

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.