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

藤井 達也[1,a)]　渡名喜 瑞稀[1]　利光 能直[2]　柴田 怜[2]　北條 大和[2]　齋藤 孝道[1]

概要：ブラウザ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利用し，ブラウザごとの情報の組み合わせの差異により，個々のブラウザを識別する技術である．一部の事業者は，ユーザに対して効果的な広告を配信するために，ブラウザフィンガープリンティングを利用し，ユーザの推定を行っている．本論文では，会員制サイトへのログイン認証前のアクセス時に採取可能なブラウザフィンガープリントを用いて，深層学習により会員 ID の推定を行った．会員 ID の推定は，会員 ID と会員のアクセス時に採取されたブラウザフィンガープリントの組を複数用意し，推定したい会員のアクセス時に採取されたブラウザフィンガープリントが，用意したどのブラウザフィンガープリントに紐づくかを推定することで実現した．結果として，推定した会員 ID のうち 87%を正しく推定することができた．

キーワード：ブラウザフィンガープリンティング，深層学習

Proposal and Implementation of An ID Estimation Using Deep Learning for PCs and Mobile Devices

Tatsuya Fujii[1,a)]　Mizuki Tonaki[1]　Yoshinao Toshimitsu[2]　Satoshi Shibata[2]　Yamato Hojyo[2]
Takamichi Saito[1]

Abstract: Browser fingerprinting is a technology that uses multiple pieces of information collected from a browser to identify an individual browser by the differences in the combination of browser's features. Some company uses browser fingerprinting to estimate users in order to provide effective advertisements to the target web viewers. In this paper, we apply browser fingerprinting to estimate the member's ID by deep learning at the time of access before login authentication. We estimate the member IDs by estimating which browser fingerprints were linked to multiple pairs of known browser fingerprints that we stored before. As a result, we correctly estimated 87% of the estimated member IDs.

Keywords: Browser Fingerprinting, Deep Neural Network

1. はじめに

Web 閲覧の際，Web サーバ側で採取可能な情報，User-Agent 文字列などを用いてブラウザを識別するブラウザフィンガープリンティングと呼ばれる技術がある．

本論文では，ブラウザフィンガープリンティング技術と，深層学習 (Deep Neural Network;DNN) を用いた会員 ID 推定を試みた（図 1）．なお，ブラウザフィンガープリントとしては，今回，タイムスタンプ，IP アドレス，User-Agent 文字列の情報を利用した．いわゆるパッシブフィンガープリンティングである．また，会員 ID 推定とは，会員サイトに会員登録し，会員サイト側に ID の登録のある利用者が会員サイトにログイン認証する前の状態で対象サイトを閲覧した際，当該閲覧のアクセスのみで，会員 ID を推定することを言う．

モデル作成と会員 ID 推定方法の概要を説明する．会員

[1] 明治大学
Meiji University
[2] 明治大学大学院
Graduate School of Meiji University
[a)] ee177052@meiji.ac.jp

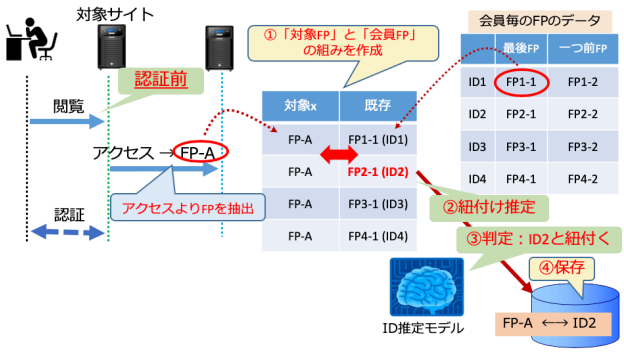

図 1　会員 ID 推定の概念

制サイトへのアクセス時に採取されたアクセスデータ約 190 万を保存し，これらをデータセットとする．採取された日付を基準とし，データセットを過去のアクセスデータと新規のアクセスデータに分割する．過去のアクセスデータを学習し，2 件のアクセスデータが持つ会員 ID の一致の有無を推定するモデルを作成する．会員 ID 推定は，新規のアクセスデータが持つ会員 ID が，どの過去のアクセスデータが持つ会員 ID に一致するのかを，作成したモデルにより推定することで実現する．

結果として，推定した会員 ID のうち 87%を正しく ID 推定することができた．また，過去のアクセスデータに推定を行うアクセスデータの会員 ID が存在するかどうかの推定も行い，その結果，F_1 値は 0.569 であった．

2. ブラウザ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（以降，フィンガープリンティングと呼ぶ）は，ブラウザから採取した情報の組み合わせによって，端末上のブラウザを識別する技術である．また，フィンガープリンティングのためにブラウザから採取する情報を特徴点と呼び，特徴点の値や，特徴点の組み合わせをブラウザフィンガープリント（以降，フィンガープリントと呼ぶ）という．

2.1 フィンガープリンティングの種類

フィンガープリンティングはその手法に基づき，2 種類に分類される．2 種類の分類を以下に示す．

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

アクティブフィンガープリンティングとは，JavaScript や CSS を端末上のブラウザで実行し，メソッドの実行結果やプロパティの値から取得できる特徴点を用いて行うフィンガープリンティングである．取得する情報としては画面解像度やフォントリストなどがある．

パッシブフィンガープリンティングとは，ブラウザから Web サーバへの通信の際に送信される HTTP ヘッダや TCP ヘッダなどの受動的に取得できる情報を用いて行うフィンガープリンティングである．情報を取得するためにスクリプトの実行を必要としないため，アクティブフィンガープリントに比べ情報の取得が比較的高速である利点がある．一方で，アクティブフィンガープリンティングに比べ取得可能な特徴点の種類が少なく識別が困難であるとされる．

2.2 関連研究

Eckersley ら [1] は，フィンガープリントを採取するサイトを構築し，94.2%のフィンガープリントがユニークであることを示した．

Laperdrix ら [2] はフランスの Web サイトにおいて，2,067,942 件のフィンガープリントを採取し，大規模なデータにおけるフィンガープリントのユニーク性や，最新の Web 技術におけるフィンガープリンティングの有用性について調べた．結果として，デスクトップやラップトップマシンのフィンガープリントは 35.7%，モバイル端末のフィンガープリントは 18.5%がユニークであった．大規模なデータにおけるフィンガープリントは有用ではなく，フィンガープリンティングによるトラッキングの危険性は低いと論じた．

田邉ら [3] は，フィンガープリンティングにおいて，特徴点の最良の組み合わせを分析した．最良とされた組み合わせで使用される特徴点の多くが，JavaScript から採取可能な特徴点であったことを示した．

高橋ら [4] は，JavaScript や CSS を利用せず，HTTP ヘッダのみから採取可能な特徴点のみを利用し，パッシブフィンガープリンティングの実験を行った．結果として，特徴点の中でも特にグローバル IP アドレスと User-Agent 文字列の情報が識別において有用であることを示した．

北條ら [5] は，パッシブフィンガープリンティングで採取可能な情報のうち，タイムスタンプ，User-Agent 文字列，IP アドレスのみを用いて，深層学習によりモバイル端末の識別を行った．結果として，F1 値が 0.99 以上という精度でモバイル端末の識別が可能であることを示した．

3. データセット

本論文では，複数の PC とモバイル端末から，38 日間の内に採取した 2,809,300 件のアクセスデータ（以降，データセットと呼ぶ）を実験に使用した．また，アクセス元を識別するために，アクセスデータは会員ごとに異なる固有の会員 ID を持つ．会員 ID は全部で 213,727 種類だった．会員 ID は，深層学習と会員 ID 推定を行う際の正解ラベルの作成にのみ使用した．

なお，実験の際には，後述する 3.2 節の方法によりアクセスデータが持つ特徴点をもとに，**表 4** に示す特徴点のうち，ISP 名，都市名，OS のバージョン，Web ブラウザのバージョン，OS のメジャーバージョンのいずれかを生成できなかったアクセスデータを取り除き実験を行った．取

り除いた結果，1,896,864 件のアクセスデータ，155,744 種類の会員 ID であった．会員ごとのアクセス数（会員 ID の出現回数）を集計し，最大値，最小値，中央値，平均値について**表 1** に示す．

表 1　会員 ID の出現回数に関する情報

種類	最大値	最小値	中央値	平均値
155,744	389	1	10.0	12.18

3.1　アクセスデータが持つ特徴点の詳細

データセット中に含まれるアクセスデータはそれぞれ，**表 2** に示す特徴点と会員 ID を持つ．本節では，アクセスデータが持つ特徴点ついてデータの分布を記す．

表 2　データセットのアクセスデータが持つ特徴点の例

特徴点	例
タイムスタンプ	2019-07-10 07:11:14
User-Agent 文字列	Mozilla/5.0 (Android 5.1.1; Tablet; rv:68.0) Gecko/68.0 Firefox/68.0
IP アドレス	192.168.100.1

3.1.1　タイムスタンプ

タイムスタンプの期間は連続した 38 日間で，全て YYYY-MM-DD HH:MM:SS 形式で保存されている．日付ごとの PC とモバイル端末のアクセス数を調べた結果を**図 2** に示す．図 2 の通り，38 日間のすべての日付において PC とモバイル端末から一定のアクセスがあったことがわかる．

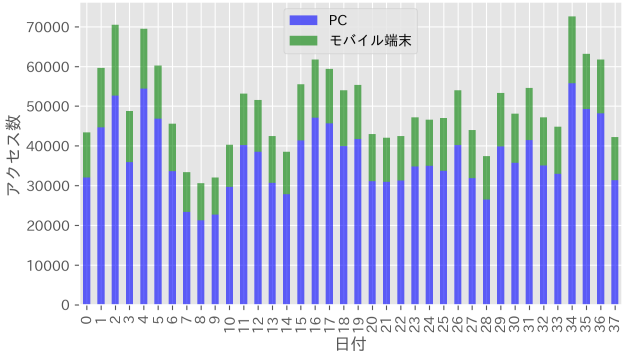

図 2　日付毎のアクセス数

3.1.2　User-Agent 文字列

データセット内に User-Agent 文字列（以降，UA 文字列と呼ぶ）は 15,432 種類存在した．PC の UA 文字列が 1,861 種類，モバイル端末では 13,571 種類であった．**表 3** に，UA 文字列の出現回数に関する情報を示す．

表 3　UA 文字列の出現回数に関する情報

端末	種類	最大値	最小値	中央値	平均値
PC	1,861	246,487	1	8	760.38
モバイル端末	13,571	114,254	1	4	35.50

3.1.3　IP アドレス

データセット内の IP アドレスは全てグローバル IP アドレスであり，354,722 種類存在した．

3.2　表 2 で示した特徴点から新たな特徴点を生成

表 2 で示した特徴点の値を用いて新たな特徴点を生成した．表 4 に，生成した特徴点を示す．

OS 名，OS バージョン，Web ブラウザ，Web ブラウザのバージョンは，user-agents2.0[6] を用いて，UA 文字列から抽出した．ISP 名は，pyisp[7] を用いて，IP アドレスから取得した．国名，都市名，市区町村，緯度，経度は GeoIP2[8] を用いて，IP アドレスから取得した．GeoIP2 は MaxMind 社の GeoLite2 のデータベースを使用するライブラリであり，GeoLite2 に保存されている位置情報は過去のある時点での情報である．

本節では，表 4 で示される生成した特徴点のいくつかについてデータの分布を記す．

表 4　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
IP アドレス	ISP 名
IP アドレス	国名，都市名，市区町村名，緯度，経度
UA 文字列	OS 名，OS のバージョン，Web ブラウザ名，Web ブラウザのバージョン，機種名，機種のブランド名，OS のメジャー・マイナー及びメンテナンスバージョン Web ブラウザのメジャー・マイナー及びメンテナンスバージョン

3.2.1　OS 名

データセット中の OS は 7 種類存在した．**表 5** に，データセット中の OS の種類とその割合を示す．

表 5　OS の種類とその割合

端末	OS の種類	全体の割合
PC	Windows	0.719602
	Mac OS X	0.025367
	Chrome OS	0.001016
	Ubuntu	0.000016
モバイル端末	iOS	0.132612
	Android	0.121376
	Windows Phone	0.000012

3.2.2　Web ブラウザ

データセット中の Web ブラウザは 31 種類存在した．**表 6** に，PC とモバイル端末ごとに上位 3 つの Web ブラウザの種類とその割合を示す．

表 6　Web ブラウザの種類とその割合（上位 3 つ）

端末	Web ブラウザの種類	全体の割合
PC	Chrome	0.266558
	Edge	0.237755
	IE	0.178860
モバイル端末	Mobile Safari	0.130249
	Chrome Mobile	0.081065
	Chrome Mobile WebView	0.019771

3.2.3　ISP 名

データセット中の ISP 名は 511 種類存在した．ISP 名に関する集計結果の内，上位 5 つを**表 7** に示す．多くが国内からのアクセスであり，ISP 名の出現回数の上位 5 つの内，3 つが日本国内の大手キャリアで占められていることがわかる．

表 7　ISP 名の分布（上位 5 つ）

ISP 名	割合
KDDI KDDI CORPORATION, JP	0.184679
OCN NTT Communications Corporation, JP	0.174106
GIGAINFRA Softbank BB Corp., JP	0.132062
OPTAGE OPTAGE Inc., JP	0.051845
JTCL-JP-AS Jupiter Telecommunication Co. Ltd, JP	0.051443

3.2.4　都市名

データセット中のアクセス元の都市名は，1,874 種類存在した．**表 8** に都市名の分布を示す．表 8 から，全アクセスの約 14%が都市 T からのアクセスであるのに対し，それ以外の都市からのアクセスは広く分布していた．

表 8　アクセス元の都市名の分布（上位 5 つ）

都市名	アクセス数	割合
都市 T	264065	0.139211
都市 O	75440	0.039770
都市 Y	74512	0.039282
都市 N	45412	0.023941
都市 K	36376	0.019177

3.3　統計量を計算した特徴点を生成

統計量を以下の手順で計算し，新たな特徴点とする．

(1) 文字列の特徴点を連結することで，新たな特徴点（以降，連結特徴点と呼ぶ）を生成（**表 9**）
(2) 連結特徴点の値が同じであるアクセスデータの集合（以降，連結特徴点同一集合と呼ぶ）を作成（**表 10**）
(3) 連結特徴点同一集合において，**表 11** に示す特徴点 X と特徴点 Y の**表 12** に示す統計量を計算

表 12 で示されている，平均値からどの程度離れているのかを表す特徴は，計算対象の特徴点の値が [平均値 − 標準偏差, 平均値 + 標準偏差] および [平均値 − 2 × 標準偏差, 平均値 + 2 × 標準偏差] の範囲に入っているかどうかをそれぞれ True または False で表す特徴点である．

表 9　生成した連結特徴点

名前	連結する特徴点
連結特徴点 A	UA 文字列，都市名，ISP 名
連結特徴点 B	UA 文字列，ISP 名
連結特徴点 C	UA 文字列，都市名

表 10　作成した連結特徴点同一集合

名前	対象の連結特徴点	種類
連結特徴点 A 同一集合	連結特徴点 A	185,764
連結特徴点 B 同一集合	連結特徴点 B	49,669
連結特徴点 C 同一集合	連結特徴点 C	116,048

表 11　連結特徴点同一集合において統計量を計算する特徴点

連結特徴点同一集合	特徴点 X	特徴点 Y
連結特徴点 A 同一集合	時刻 (hour)，曜日	なし
連結特徴点 B 同一集合	時刻 (hour)，曜日，緯度，経度	都市名
連結特徴点 C 同一集合	時刻 (hour)，曜日	ISP 名

表 12　計算する統計量およびその対象の特徴点

計算する統計量	計算対象
連結特徴点同一集合の要素数	特徴点 X，特徴点 Y
平均値，標準偏差，四分位数	特徴点 X
平均値および中央値との差	特徴点 X
平均値からどの程度離れているのかを表す特徴	特徴点 X
最頻値	特徴点 Y
最頻値の出現回数	特徴点 Y
最頻値と値が一致するかどうか	特徴点 Y
計算対象の特徴点の値の種類	特徴点 Y

3.4　特徴点の変化について

端末の特徴点が，どの程度変化したかを示す．特に変化しやすいと考えられる 6 種類の特徴点に関して，会員 ID ごとに特徴点の変化回数を計算した．

表 13 と**表 14** において，mean は平均値を，std は標準偏差を，25%，50%，75%は，それぞれ四分位数を表す．また，UA は UA 文字列，IP は IP アドレス，OS v は OS のバージョン，Web v は Web ブラウザのバージョンを示す．

各期間における端末の特徴点の変化の詳細を表 13 と表 14 に示す．表 14 より，6 種類の特徴点の中央値は全て 1.0 である．また標準偏差も小さく，28 日から 37 日において多くの端末は特徴点がほとんど変化しなかったといえる．また，表 13 と表 14 より，0 日から 37 日と 28 日から 37 日の期間を比べると，前者の期間における端末の特徴点の変化回数のほうが多かったと言える．

4.　実験について

本節では実験の詳細，データセットのアクセスデータを

表 13　0 日から 37 日における端末の特徴点の変化回数

	UA	IP	ISP 名	都市名	OS v	Web v
mean	2.0844	2.7944	1.1690	1.6996	1.2609	1.5520
std	1.4705	5.4720	0.4369	1.6757	0.5622	0.7428
min	1.0	1.0	1.0	1.0	1.0	1.0
25%	1.0	1.0	1.0	1.0	1.0	1.0
50%	2.0	1.0	1.0	1.0	1.0	1.0
75%	3.0	2.0	1.0	2.0	1.0	2.0
max	84.0	270.0	12.0	41.0	7.0	26.0

表 14　28 日から 37 日における端末の特徴点の変化回数

	UA	IP	ISP 名	都市名	OS v	Web v
mean	1.3142	1.5566	1.0978	1.2717	1.1112	1.2598
std	0.6097	1.7224	0.3223	0.7834	0.3419	0.5063
min	1.0	1.0	1.0	1.0	1.0	1.0
25%	1.0	1.0	1.0	1.0	1.0	1.0
50%	1.0	1.0	1.0	1.0	1.0	1.0
75%	2.0	1.0	1.0	1.0	1.0	1.0
max	23.0	83.0	6.0	24.0	5.0	8.0

ベクトルデータに変換する方法，および会員 ID 推定の方法を示す．

4.1　実験概要

本論文では，過去のアクセスデータを深層学習で学習することで，新規のアクセスデータの会員 ID 推定が可能かを検証した．

以下の手順で会員 ID 推定を行った．

(1) 過去のアクセスデータを学習し，2 件のアクセスデータが持つ会員 ID の一致の有無を推定するモデルを作成
(2) 作成したモデルをテスト
(3) 作成したモデルを用いて，新規のアクセスデータの会員 ID が，どの過去のアクセスデータの会員 ID に一致するかを推定

過去のアクセスデータと新規のアクセスデータを区別するために，データセットを 0 日から 34 日と 35 日から 37 日の期間で分割した．また，過去のアクセスデータのうち，28 日から 34 日のアクセスデータをデータセット A，過去のアクセスデータ全体をデータセット B，新規のアクセスデータ全体をデータセット C と呼ぶ．各データセットの用途を**表 15** に示す．

実験では，次に 2 点について検証した．

- 過去のアクセスデータを学習することで，新規のアクセスデータの会員 ID 推定が可能か
- 会員 ID 推定時に用いる過去のアクセスデータの期間と会員 ID 推定の精度との相関があるか

4.2　ベクトルデータの作成

アクセスデータをベクトルデータに変換する手順を説明する．本実験では，任意の 2 件のアクセスデータを組み合わせ，一次元のベクトルデータを作成した．

表 15　各データセットの用途

名前	期間	用途
データセット A	28 日から 34 日	深層学習の学習用アクセスデータ，会員 ID 推定時に用いる過去のアクセスデータ
データセット B	0 日から 34 日	会員 ID 推定時に用いる過去のアクセスデータ
データセット C	35 日から 37 日	深層学習のテスト用アクセスデータ，会員 ID 推定を行う新規のアクセスデータ

実験に使用する一次元のベクトルデータは以下の手順で作成した．

(1) 2 件のアクセスデータを結合した組を作成
(2) 2 件のアクセスデータが持つそれぞれの特徴点の値を比較した情報を表す特徴点を新たに付加
(3) 4.2.1 節に示す方法で，正解ラベルを付加

特徴点の値を比較した情報を表す特徴点を**表 16** に示す．

表 16　比較した情報を表す特徴点

比較した特徴点	比較した情報を表す特徴点
値が文字列の特徴点，値が真偽値の特徴点	値の一致有無
値が数値の特徴点	値の差
緯度および経度	2 件のアクセスデータの位置情報の直線距離
直線距離およびタイムスタンプ	2 件のアクセスデータの距離を移動する場合のタイムスタンプの差から計算する速度
OS バージョンおよびタイムスタンプ	OS のバージョンとアクセス時刻の前後関係に矛盾があるかどうか

2 件のアクセスデータを結合した組に比較した情報を持つ特徴点を付与した後，特徴点を数値化した．数値化の詳細を**表 17** に示す．

表 17　ベクトルデータの数値化

数値化する方法	対象の特徴点
ハッシュ化することで数値化する	値が文字列の特徴点
値を 0 または 1 で表す	値が真偽値の特徴点
値をそのまま使用する	値が数値の特徴点

4.2.1　正解ラベルの作成

正解ラベルを作成する方法を説明する．3 節で述べたとおり，データセットのアクセスデータは，会員ごとに異なる固有の会員 ID を持つ．正解ラベルは，ベクトルデータを構成する 2 件のアクセスデータが持つ会員 ID の一致の有無に応じて作成した．正解ラベルは，組み合わせた 2 件のアクセスデータの持つ会員 ID が同一であるかどうかを表しており，一致していればラベルの値を 1，一致していなければラベルの値を 0 とした．

4.3 ニューラルネットワークの構造

本論文で使用したニューラルネットワークの構造を**表 18** に示す．また，各層で BatchNormalization を行い，損失関数は binarycrossentropy，最適化関数には Adagrad（パラメータはデフォルトの値）を用いた．すべての実験でバッチサイズは 2,000 を用いた．

表 18 ニューラルネットワークの構造

層	ユニット数	活性化関数	DropOut
入力層	特徴点の数	relu	無し
中間層 1	500	relu	0.3
中間層 2	500	relu	0.3
中間層 3	500	relu	0.3
中間層 4	50	relu	0.3
出力層	2	sigmoid function	0.3

4.4 モデルの作成

推定モデルの作成について説明する．作成は以下の手順で行った．

(1) データセット A からアクセスデータをランダムに 2 件抽出する
(2) (1) で取得したアクセスデータを組み合わせ，ベクトルデータを約 2000 万件作成する
(3) (2) で作成したベクトルデータで教師あり学習を行い，推定モデルを作成する

作成したモデルは，ベクトルデータを構成する 2 件のアクセスデータが持つ会員 ID が同一である尤度を出力する．モデルの出力がしきい値以上の場合は，2 件のアクセスデータが持つ会員 ID が同一であると推定する．

4.5 モデルのテスト

モデルのテストついて説明する．データセット A とデータセット C よりアクセスデータをランダムに 1 件ずつ抽出し，それらを組み合わせ，約 100 万のベクトルデータを作成した．作成したベクトルデータでモデルの精度を検証した．

4.6 会員 ID 推定

会員 ID 推定について説明する．3 節で述べたとおり，データセットのアクセスデータは，会員ごとに異なる固有の会員 ID を持つ．会員 ID 推定は，4.4 節で作成したモデルを用いて，新規のアクセスデータの会員 ID が，どの過去のアクセスデータの会員 ID に一致するかを推定することで実現した．

会員 ID 推定の方法を以下に示す．

(1) 新規のアクセスデータから推定を行うアクセスデータを 1 件取得する
(2) 推定を行うアクセスデータを過去のアクセスデータ全てと組み合わせ，複数のベクトルデータを作成する
(3) モデルを用いて，(2) で作成した各ベクトルデータを構成する 2 件のアクセスデータの会員 ID が同一である尤度を取得する
(4) 後述する会員 ID なし推定を行い，推定を行うアクセスデータの会員 ID が過去のアクセスデータに存在しないと推定された場合には (a) を，存在すると推定された場合には (b) を行う
    (a) 会員 ID の推定を行わない
    (b) 取得した尤度の中で，値が高かった上位 3 つのベクトルデータから過去のアクセスデータの会員 ID を取得し，最も出現回数の多い会員 ID を推定値とする

会員 ID なし推定について説明する．会員 ID なし推定とは，推定を行うアクセスデータの会員 ID が，会員 ID 推定時に用いる過去のアクセスデータに存在するかどうかを推定することを言う．会員 ID なし推定は，会員 ID 推定の (3) で取得した全ての尤度がしきい値以下である場合は，過去のアクセスデータには推定を行うアクセスデータの会員 ID が存在しないとすることで実現した．

会員 ID 推定時に用いる過去のアクセスデータに，推定を行うアクセスデータの会員 ID が存在しない場合には，正しく会員 ID 推定を行うことができない．そこで，会員 ID なし推定を行い，会員 ID 推定時に用いる過去のアクセスデータに推定を行うアクセスデータの会員 ID が存在すると推定された場合のみ，会員 ID 推定を行うこととした．

4.7 実験の詳細

4.7.1 実験 1

実験 1 では，過去のアクセスデータを学習することで，新規のアクセスデータの会員 ID 推定が可能かを検証した．

表 19 に実験 1 における会員 ID 推定時に用いたアクセスデータを示す．

表 19 実験 1 における会員 ID 推定時に用いたアクセスデータ

用途	アクセスデータ
会員 ID 推定時に用いる過去のアクセスデータ	データセット A 全体
会員 ID 推定を行う新規のアクセスデータ	データセット C のうち 20,000 件

4.7.2 実験 2

実験 2 では，会員 ID 推定時に用いる過去のアクセスデータの期間と会員 ID 推定の精度との相関があるのかを検証した．

表 20 に実験 2 における会員 ID 推定時に用いたアクセスデータを示す．会員 ID 推定時に用いる過去のアクセスデータとして，データセット A とデータセット B の 2 パターンを用意し，両者の会員 ID 推定の精度を比較した．

表 **20**　実験 2 における会員 ID 推定時に用いたアクセスデータ

用途	アクセスデータ
会員 ID 推定時に用いる過去のアクセスデータ	データセット A 全体，データセット B 全体
会員 ID 推定を行う新規のアクセスデータ	データセット C のうち 2,800 件

5. 実験結果

5.1 モデルの精度の算出に使った指標

モデルの推定値と正解ラベルに基づき，*Precision*，*Recall*，*Accuracy*，F_1 値をそれぞれ算出し使用する．以下に *Precision*，*Recall*，*Accuracy*，F_1 値を求める式を示す．

$$Precision = \frac{|TP|}{|TP + FP|}$$

$$Recall = \frac{|TP|}{|TP + FN|}$$

$$Accuracy = \frac{|TP + TN|}{|TP + FP + FN + TN|}$$

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

上記の識別精度の算出の際には，**表 21** に基づき，混同行列を算出する．

表 **21**　モデルにおける TP，TN，FP，FN の分類

		会員 ID	
		同一	異なる
モデルの推定	同一	TP	FP
	異なる	FN	TN

5.2 会員 ID なし推定の精度の算出に使った指標

会員 ID なし推定の推定値と正解ラベルに基づき，*Precision*，*Recall*，*Accuracy*，F_1 値をそれぞれ算出し使用する．*Precision*，*Recall*，*Accuracy*，F_1 値を求める式は，5.1 節で示したものと同じである．

上記の識別精度の算出の際には，**表 22** に基づき，混同行列を算出する．

表 **22**　会員 ID なし推定における TP，TN，FP，FN の分類

		ID の存在	
		ID がない	ID がある
会員 ID なし推定	ID がない	TP	FP
	ID がある	FN	TN

5.3 会員 ID 推定の精度の算出に使った指標

会員 ID 推定の推定値と正解ラベルに基づき正解率（狭義）を算出し使用する．また，会員 ID 推定の推定値，会員 ID なし推定の推定値，正解ラベルに基づき正解率（広義）を算出し使用する．正解率（狭義）とは，会員 ID 推定を行ったアクセスデータのうち，どれだけ正しく会員 ID の推定ができているかを表す指標である．正解率（広義）とは，会員 ID 推定と会員 ID なし推定を行った結果，両者がどれだけ正しく推定できているかを表す指標である．

以下に正解率（狭義）と正解率（広義）を求める式を示す．ただし，ID 推定正は会員 ID 推定の正解数，ID 推定誤は会員 ID 推定の不正解数である．また，ID なし推定正は会員 ID なし推定における TP，ID なし推定誤は会員 ID なし推定における FP である．

$$\text{正解率（狭義）} = \frac{\text{ID 推定正}}{\text{ID 推定正} + \text{ID 推定誤}}$$

$$\text{正解率（広義）} = \frac{\text{ID 推定正} + \text{ID なし推定正}}{\text{ID 推定正} + \text{ID 推定誤} + \text{ID なし推定正} + \text{ID なし推定誤}}$$

5.4 会員 ID 推定で用いたモデルの精度

会員 ID 推定で用いたモデルの混同行列を**表 23** に，精度を**表 24** に示す．

表 **23**　モデルの混同行列

TP	FP	FN	TN
287	3	121	999,588

表 **24**　モデルの精度

Precision	*Recall*	*Accuracy*	F_1
0.989	0.703	0.999	0.822

5.5 実験 1 について

実験 1 の結果を，**表 25**，**表 26**，**表 27**，**表 28** に示す．

表 28 より，正解率（狭義）は 8 割を，正解率（広義）は 7 割を超えており，概ね正しく会員 ID 推定を行うことができたと言える．

表 **25**　会員 ID なし推定の混同行列

TP	FP	FN	TN
1,916	2,526	375	15,183

表 **26**　会員 ID なし推定の精度

Precision	*Recall*	*Accuracy*	F_1
0.431	0.836	0.855	0.569

表 **27**　会員 ID 推定の正解数と不正解数

正解数	不正解数
13,613	1,945

表 28　会員 ID 推定の精度

正解率（狭義）	正解率（広義）
0.875	0.776

表 29　会員 ID なし推定の混同行列

期間	TP	FP	FN	TN
1 日から 38 日	28	140	42	2,590
29 日から 35 日	361	339	59	2,041

表 30　会員 ID なし推定の精度

期間	*Precision*	*Recall*	*Accuracy*	F_1
1 日から 38 日	0.167	0.400	0.935	0.235
29 日から 35 日	0.516	0.860	0.858	0.645

表 31　会員 ID 推定の正解数と不正解数

期間	正解数	不正解数
1 日から 38 日	2,120	512
29 日から 35 日	1,862	238

表 32　会員 ID 推定の精度

期間	正解率（狭義）	正解率（広義）
1 日から 38 日	0.805	0.767
29 日から 35 日	0.887	0.794

5.6　実験 2 について

実験 2 の結果を，**表 29**，**表 30**，**表 31**，**表 32** に示す．

表 32 より，会員 ID 推定時に用いる過去のアクセスデータの期間が長いと会員 ID 推定の精度が落ちると言える．

6.　考察

6.1　実験 1 について

実験 1 の結果から，概ね正しく会員 ID 推定を行うことができたと言える．この原因について考察する．

今回用いたデータセットについて以下のことがわかっている．

- 3.1.2 節の UA 文字列と 3.2.3 節の ISP 名の値を組み合わせると会員 ID の総数の 1/3 ほどのバリエーションが存在する
- 3.4 節で述べたように，28 日から 37 日において多くの端末は特徴点がほとんど変化していない
- 表 8 より，1,874 種類の都市の中で，都市 T からのアクセスが約 14%であったが，それ以外の都市からのアクセスは広く分布している

以上から，概ね正しく会員 ID 推定を行えるほどの多様性がアクセスデータに存在したことが推察される．

6.2　実験 2 について

会員 ID 推定時に用いる過去のアクセスデータの期間が長いと会員 ID 推定の精度が低くなる原因を考察する．

3.4 節で述べたように，アクセスデータの期間が長いと，端末の特徴点の変化回数が多くなる．よって，会員 ID 推定時に用いる過去のアクセスデータの期間が長いと，推定モデルが誤判定を起こしやすくなり，会員 ID 推定の精度が低くなったと考えられる．

7.　研究倫理

我々は，Menlo report[9] の精神に則り，倫理的配慮をして実験を行った．実験を行う際，個人識別はせず，プライバシーを遵守した．論文中では，オリジナルデータの統計的処理により，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，我々の研究室にて厳重に保管されており，他者への売却および提供をしない．

8.　まとめ

本論文では，タイムスタンプ，IP アドレス，UA 文字列から利用できる情報のみを深層学習で学習し，会員 ID 推定を行った．実験には，複数の PC とモバイル端末から，約 40 日間採取した約 190 万件のアクセスデータを用いた．

結果として，推定した会員 ID のうち 87%を正しく推定することができた．また，会員 ID なし推定を行った結果，F_1 値は 0.569 であった．

9.　謝辞

本研究の一部は JSPS 科研費 JP18K11305 の助成を受けたものです．

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] Gómez-Boix, Alejandro and Laperdrix, Pierre and Baudry, Benoit, Hiding in the Crowd: an Analysis of the Effectiveness of Browser Fingerprinting at Large Scale. WWW2018 - TheWebConf 2018 : 27th International World Wide Web Conference, Lyon, France, Apr 2018.

[3] 田邉一寿, 高橋和司, 安田昂樹, 種岡優幸, 細谷竜平, 小芝力 太, 齋藤祐太, 齋藤孝道, "Browser Fingerprinting における特徴の組み合わせに関する考察", コンピュータセキュリティシンポジウム 2017, 2017.

[4] 高橋和司, 安田昂樹, 種岡優幸, 田邉一寿, 細谷竜平, 野田隆 文, 齋藤祐太, 小芝力太, 齋藤孝道, "HTTP ヘッダのみを 用いた Browser Fingerprinting の考察", 暗号と情報セキュリティシンポジウム 2018，2018.

[5] 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, 2019.

[6] python-user-agents, https://github.com/selwin/python-user-agents

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

[8] GepIP2-python, https://github.com/maxmind/GeoIP2-python/

[9] Dittrich, D. and Kenneally, E. The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research. U.S. Department of Homeland Security, Aug 2012.