ウェブ閲覧URL列へのTransformer Encoderの適用による広告閲覧ユーザの属性推定

崔　洙瑚† 橋口　友哉† 木村　塁†† 大島　裕明†

† 兵庫県立大学 応用情報科学研究科　〒650–0047 兵庫県神戸市中央区港島南町 7–1–28
†† 株式会社 KDDI 総合研究所　〒102–8460 東京都千代田区飯田橋 3–10–10
E-mail: †{aa19e505,aa19j508,ohshima}@ai.u-hyogo.ac.jp, ††ui-kimura@kddi-research.jp

あらまし　本研究では，ウェブ閲覧 URL 列を入力として，Transformer Encoder を用いてユーザの属性を推定する手法を提案する．多くのウェブページでは広告を配信しており，ユーザが広告が設置されているウェブページを閲覧した場合，どのウェブページからきたかという情報や他のウェブページを閲覧した履歴の情報からどのような広告を配信するかが決定される．ウェブ閲覧 URL 列はどのような広告を配信するか決めるために蓄積されたユーザが閲覧したページの URL の履歴である．ユーザが閲覧したウェブページからどのような広告を配信するか決められるが，広告を配信されたユーザのウェブ閲覧 URL 列ではユーザが誰であるか特定することはできない．広告を閲覧したユーザのウェブ閲覧 URL 列からユーザ属性を推定することができれば，ユーザ属性によって広告を配信することでよりウェブ広告の効率を上げることが可能になると考えられる．本研究では，ユーザのウェブ閲覧 URL 列からユーザの属性を推定する問題に取り組んだ．本研究は，ウェブ閲覧 URL 列に適した前処理を考え，前処理を行った URL 列で Transformer Encoder の事前学習モデルを作成した．さらに，ユーザ属性の推定する実験を行うために事前学習モデルにファインチューニングを行った．また，比較手法には既存研究の手法を用いた．

キーワード　ユーザ属性推定，ウェブ広告，ウェブ閲覧履歴

1 はじめに

近年，飛躍的なインターネットの発展とともに広告業界の中でもウェブ広告の成長が注目を浴びている．パソコンからスマートフォンなど，デジタルデバイスの普及によって従来のマス広告からウェブ広告へ注目が上がり，ウェブ上でユーザに合わせて広告を自動に配信する技術も発展してきた．例えば，検索キーワードや広告枠が設置されたウェブページの内容と，広告の内容の関連度を高いものを配信する広告など，ユーザに対して表示する広告は入札要求（RTB: Real-Time Bidding）で行われている．広告を自動に配信するために広告プロバイダはユーザのウェブ上の行動であるウェブ閲覧履歴を収集しているが，広告プロバイダが配信しているウェブページ以外のウェブ閲覧履歴は入手できない場合もあり，入手できるユーザのウェブ閲覧履歴には限度がある．図 1 は，ユーザが広告プロバイダが配信しているウェブページを閲覧した場合，ユーザ毎に取得されるウェブ広告閲覧履歴を表している．ウェブ広告閲覧履歴はすべてのウェブページの閲覧履歴が取得されるわけではなく，ユーザが閲覧したウェブページのなかで，広告プロバイダによる広告配信の入札要求があったウェブページの閲覧履歴である．

ユーザのウェブ広告閲覧履歴では広告掲載ウェブページのテキスト情報などは含まれず，あるユーザがどの URL に閲覧したかといった情報に限定されている．それらをもとに Demand Side Platform（DSP）を入札を行う必要があり，ユーザにより適したウェブ広告を発信するためには，限定された情報から

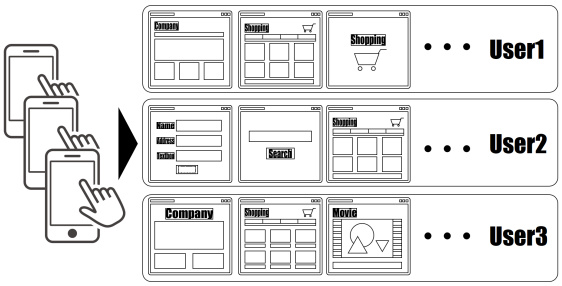

図 1　ユーザ毎に取得されるウェブ広告閲覧履歴

ユーザの属性を推定する必要がある．しかし，あるユーザがどの URL に閲覧したかといった情報は膨大であり，ユーザが誰であるかを識別するのは難しい．そのため，本研究では広告プロバイダが配信している広告のウェブページを閲覧した履歴であるウェブ広告閲覧履歴からユーザのウェブ閲覧 URL 列を抽出してユーザ属性を推定する問題に取り組む．

本研究の提案手法としてはウェブ広告閲覧履歴から抽出したウェブ閲覧 URL 列への Transformer Encoder を利用する手法を提案する．ウェブ閲覧 URL 列は広告プロバイダが広告を配信しているウェブページを閲覧した場合に取得できるウェブページの URL を閲覧した順に並べている列である．ユーザ属性が単なれば，趣味，関心，ライフステージ等が単なるため，閲覧するウェブページの傾向にも違いがあると考えられる．そこで，どのようなウェブページどのような順序で閲覧したかを知るこ

とができるウェブ閲覧 URL 列から Transformer Encoder を用いたユーザ属性を推定するモデルを作成する．また，ユーザ属性を推定することに最適なモデルを作成するためにウェブ閲覧 URL 列からどのような前処理を行うかについて説明する．

以下，2 節では関連研究について述べる． 3 節では，本研究で用いたデータと問題定義を行う． 4 節では，本研究で利用する Transformer Encoder 事前学習モデルの作成，そのモデルに対して行うユーザ属性推定を行うファインチューニングの手法について説明する． 5 節では上記の手法を用いた実験の概要について説明した後，実際に作成した分類モデルの精度の比較を行う． 6 節ではまとめと今後の課題について説明する．

2 関 連 研 究

ユーザはウェブ上から日々様々なサービスを利用しており，同時に様々なユーザの行動履歴が存在している．ユーザの行動履歴を対象として，ユーザ属性の推定に関する研究が行われている．

2.1 ユーザ属性推定に関する研究

ウェブ上のユーザの行動履歴からユーザ属性を推定する研究は広く行われている．Hu ら [1] は，ユーザのウェブ閲覧履歴からウェブページ内のテキスト情報とユーザが閲覧してウェブページのカテゴリ情報を用いてウェブページの特徴量を生成して性別と年齢といったユーザ属性の推定する手法を提案した．Weber ら [2] は，ウェブ検索エンジンの Yahoo!のクエリログデータからユーザのウェブ検索キーワードを抽出してユーザ属性を推定する手法を提案した．Weber らはユーザが検索したキーワードから特徴ベクトルを作成し，k-mean 法を用いてクラスタリングを行うことでユーザ属性推定を行った．Bi ら [3] は Facebook から配信されているユーザ間のコンテンツの興味を表した履歴を用いて，検索エンジンサイトによる検索キーワードからユーザ属性を推定する手法を提案した．Bi らは属性が分かるユーザが興味を表したコンテンツのタイトルを用いてモデルを学習して検索エンジンサイトに検索した検索キーワードを学習したモデルに転移学習することでユーザ属性の推定を行った．Peersman ら [4] や Alekseev ら [5] は，ソシャルネットワークのサービスを利用しているユーザのプロフィールや作成したコンテンツのテキスト情報を用いてユーザ属性の推定を行った．Malmi ら [6] や Kalimeri ら [7] はユーザのスマートフォンに設置されている多数のアプリケーションからユーザ属性を推定することは可能であることを示している．Dong ら [8] や Ying ら [9] はスマートフォンによる Gowalla，Foursquare，Facebook のようなウェブサービスのログデータを用いてユーザ属性の推定を行った．このようなウェブサービスはユーザから属性情報や位置情報を他のユーザと共有している．ユーザの位置情報やウェブサービス利用時間のウェブログデータを用いて，ユーザ属性を推定を行った．このようにスマートフォンによるログデータを用いてユーザ属性の推定する事が多い [10] [11] [12]．Ito ら [13] は，Twitter ユーザのユーザ属性を推定する手法を提案している．他にも Twitter を利用して，Twitter ユーザのユーザ属性を推定する研究は近藤ら [14] の位置情報付きツイートを用いて，ユーザの地域クラスタリングを行う研究，Miller ら [15] や Burger ら [16] によるユーザの性別推定の研究，Rao ら [17] による属性分類，Pandya ら [18] による年齢推定において URL とハッシュタグの利用に関する研究がある．また，榊ら [19] は Twitter からユーザの職業の推定を行っている．E コマースウェブサイトの購入履歴からユーザ属性推定に関する研究としては Lu ら [20]，Jiang ら [21] の研究がある．

2.2 文書検索において用いられる手法の活用に関する研究

崔ら [22] は，本研究と同様にウェブ広告閲覧履歴を用いて，ユーザ属性の推定を行う手法を提案している．ウェブ広告閲覧履歴から URL 列を抽出して閲覧した URL の頻度や Word2Vec，Doc2Vec，fastText といった文書検索において用いられる手法を用いてユーザ特徴ベクトルを作成し，ユーザ属性の推定を行った．また，星ら [23] も本研究と同様にウェブ広告閲覧履歴を用いて，ユーザのライフイベントの予測を行う手法を提案している．ライフイベントとは，結婚や出産といった生活上の大きなイベントのことである．ユーザはライフイベントの前には，そのイベントに関連した情報を収集すると考えられる．そこで，星らは，ウェブ広告閲覧履歴において，URL を語とみなして Word2Vec を適用することで，ある URL を固定長のベクトルで表現した．あるユーザのウェブ広告閲覧履歴に含まれる URL を学習した Word2Vec のモデルを用いてベクトル化し，閲覧履歴の平均値をとることでユーザ特徴ベクトルを生成した．そして，生成されたユーザ特徴ベクトルを用いて，ユーザのライフイベントの予測を行った．Tagami ら [24] は，ユーザのウェブ閲覧履歴から，広告をクリックするかどうかや，広告主のウェブページを閲覧するかどうかを予測する手法を提案した．ウェブ広告の配信において顧客になりそうなユーザを絞り込むことで広告の効率向上が期待される．Tagami らはウェブ閲覧履歴からウェブページの URL 列を抽出し，Doc2Vec と Word2Vec の手法を用いて URL をベクトル化する手法を提案した．ウェブ閲覧履歴から学習された Doc2Vec と Word2Vec を用いて，ユーザ特徴ベクトルを生成し，ユーザ特徴ベクトルから広告をクリックするユーザなのかを予測した．Hu ら [25] は，ユーザの検索ログを用いてユーザ特徴ベクトルを生成する手法を提案した．あるユーザが検索エンジンなどで検索したキーワードである検索ログデータから TF-IDF を用いて，ユーザ特徴ベクトルを生成し，さらに Word2Vec によって計算された単語の重みを付けることでユーザ特徴ベクトルを生成した．Kanagasabai ら [26] は，通信業者が取得することができるユーザのウェブ閲覧履歴を用いて，ユーザの興味を推定する手法を提案している．ウェブ閲覧履歴における URL を Word2Vec によってベクトル化し，それを用いてユーザの特徴ベクトルを生成した．ユーザの興味が推定されると，広告の効率向上などに利用することができると考えられる．

2.3 Attention メカニズムである Transformer を活用した研究

本研究では，提案手法として双方向 Transformer [27] の Encoder を用いる．言語の分野では，既存の RNN といった時系列データに対する手法よりも Transformer による手法が精度が高くなると報告されている．本研究では 2.2 節の研究から，URL を言語と考え，Transformer を用いることで，精度が向上するのではないかと考えた．

Transformer は言語以外にもの応用されており，既存の手法より精度が高くなっていることが報告されている．Huang ら [28] は，Transformer を用いて音楽を生成する手法を提案した．音楽のメロディーのデータセットから音をトークンにして Transformer のモデルを学習することでメロディーを表現する音楽の生成を行った．Dosovitskiy ら [29] は，Transformer を用いて画像を認識する手法を提案した．画像を認識する Transformer のモデルを作成するために画像をパッチに分けて各パッチを単語のように扱うことで画像を識別する Transformer のモデル作成を行った．Hernandez ら [30] は，Transformer の事前学習において学習データ数は多ければ多いほど精度を向上させることができると主張している．本研究では大量のウェブ閲覧 URL 履歴が存在しており，それらを用いて Transformer の事前学習モデルを作成することでユーザ属性推定の精度を向上を期待してる．

3 データと問題定義

本節では，本研究で利用するユーザのウェブ広告閲覧履歴データとウェブ閲覧 URL 列について述べ，研究に使用するデータセットの説明を行う．その後，本研究における問題定義を行う．

3.1 ウェブ広告閲覧履歴データ

本研究でのウェブ広告閲覧履歴とは，あるユーザが広告枠のあるウェブサイトを閲覧した際に，ユーザに配信する広告を決めるために DSP に送信された広告配信要求ログから抽出したものである．ユーザはウェブブラウザを用いて自由にウェブ閲覧を行っており，DSP は広告主からのターゲティング条件と広告配信要求とマッチさせて，入札する広告を決める．本研究でのウェブ広告閲覧履歴は Supership 株式会社から DSP 基盤により収集された広告配信要求ログから抽出した．ウェブ広告閲覧履歴は，以下の 3 つ組のデータである．

- ユーザ識別子
- 閲覧日時
- 閲覧 URL

本研究で利用するウェブ広告閲覧履歴は，2015 年 1 月から 2016 年 8 月までに得られたこのような 3 つ組のデータの集合である．ユーザ識別子は，ウェブページを閲覧したユーザに割り当てられ，他のウェブページも閲覧したかどうかを識別することが可能である．しかし，利用するデバイスが同一であっても，利用するブラウザが異なる場合には，異なるユーザ識別子が割り当てられる．そのため，ユーザ識別子が異なっても，同一のユーザである場合が存在する．このような場合において，ユーザを判別することは原則としてできないため，上記期間に収集されたデータ全てを用いるものとする．

3.2 ウェブ閲覧 URL 列

ウェブ閲覧 URL 列とは，ウェブ広告閲覧履歴データから抽出したものであり，ユーザが閲覧したウェブページの URL で構成されている．本研究ではユーザ識別子ごとにウェブ広告閲覧履歴の期間中のすべての閲覧 URL を閲覧日時順に並べて作成した．

3.3 ユーザ属性データ

Transformer Encoder を利用してユーザ属性推定モデルを作成し，モデルを評価するためにはユーザ属性が分かる正解ラベルのデータが必要である．あるユーザ識別子を持つユーザに対するユーザ属性データにおいて，本研究ではユーザのユーザ属性の推定を行う．

- **性別（2 種類）**

男性, 女性

- **結婚の有無（2 種類）**

未婚, 既婚

- **年齢カテゴリ（10 種類）**

12 才～19 才, 20 才～24 才, 25 才～29 才, 30 才～34 才, 35 才～39 才, 40 才～44 才, 45 才～49 才, 50 才～54 才, 55 才～59 才, 60 才以上

以下では，ユーザ属性について説明を行う．本研究では，ユーザ属性データを取得するために，2016 年 9 月に実施したアンケート調査の結果を用いる．本調査は，一般的なウェブアンケート調査であり，そこでは，ユーザ属性データを含む様々な質問に対する回答が行われた．

本研究で用いるユーザのウェブ閲覧 URL 列に対する対象のアンケート調査においてユーザ属性データを一意に取得できた 2,132 名のデータが，本研究で用いるユーザ属性データである．そのうち，閲覧した URL の件数が少ないユーザは，適切なユーザ属性推定を行うことが難しいと考えられる．そこで，ユーザが閲覧した URL の件数が 100 件未満のユーザをフィルタリングして，本研究の対象から除外することとした．その結果，属性データが得られた 2,132 名のユーザのうち 1,991 名のユーザを，本研究で対象とするユーザとした．それらのユーザによる本研究で用いる URL の総数は，90,106,600 件であり，また，URL の種類数は，4,819,890 件である．

3.4 問題定義

本研究の問題定義を行う．本研究で取り組むのは，ユーザのウェブ閲覧 URL 列があるときに，そのウェブ閲覧 URL 列からユーザ属性を推定することである．その入力と出力は以下の通りである．

入力　ユーザのウェブ閲覧 URL 列

出力　男性 or 女性

従って，本研究で取り組む問題は入力されたユーザのウェブ閲覧 URL 列からユーザ属性を分類する分類問題である．性別

表 1　ウェブ閲覧 FQDN 列に対する前処理の手順

1. オリジナル	[a.com, a.com, b.com, a.com, c.com a.com]
2. 連続部の集約	[a.com, b.com, a.com, c.com a.com]
3. 順序を保持したサンプリング	[a.com, b.com, c.com]

の場合，男性または女性の 2 つのカテゴリに分類する二値分類問題となる．同様に，結婚の有無についても二値分類問題となる．年齢カテゴリは 10 カテゴリの多クラス分類問題となる．

また，推定の対象となるユーザは収集したウェブ閲覧 URL 列において，ユーザ属性ラベルが付与されている 1,991 名を用いる．評価手法については，正解率，適合率，再現率，F 値などで行うものとする．

4　ユーザ属性の推定手法

本節では，入力されるウェブ閲覧 URL 列からユーザ属性を推定する手法について説明する．まず，本研究で使用する Transformer Encoder による事前学習モデルについて説明し，その後，モデルに対するファインチューニングの手法について記述する．

4.1　URL の前処理手法

本研究では，ユーザのウェブ閲覧 URL 列内の URL を考慮した分散表現を用いて分類を行うため，Transformer Encoder 事前学習モデルを作成した．本研究の対象である 1,991 名のユーザが閲覧した URL の種類数は，4,819,890 件であり，モデルの学習において膨大な種類数となる．また，全く同じ URL を複数のユーザが閲覧していることは非常に稀である．

そこで本研究では URL を FQDN（Fully Qualified Domain Name）に変換する．FQDN とは URL のホスト名とドメインだけが含まれた部分である．例えば，「`https://www.example.co.jp/top.html`」という URL から FQDN 部分に変換すると「`www.example.co.jp`」となる．このように変換することで，ある FQDN を複数のユーザが閲覧することになり，ユーザの特徴を得られるのではないかと考えた．本研究ではウェブ閲覧 URL 列からウェブ閲覧 FQDN 列に変換する前処理を行った．

4.2　ウェブ閲覧 FQDN 列の前処理手法

URL を FQDN に変換する前処理を行って分析すると，同じ FQDN が何度も連続して出現することが分かった．これは，同じウェブページにおいて，複数のページを次々と閲覧していた場合であると考えられる．同じウェブサイトのいくつかのページを閲覧するということは一般的であり，頻繁に行われると考えられる．

表 1 は本研究で行った前処理の手順を示している．**オリジナル**はユーザが閲覧した URL 列を FQDN 列に変換したものである．**連続部の集約**は**オリジナル**から 2 回以上連続して出現する同一の FQDN を 1 回の FQDN として集約したものである．**順序を保持したサンプリング**は FQDN の出現した順番を保持したまま，重複をなくしたものである．

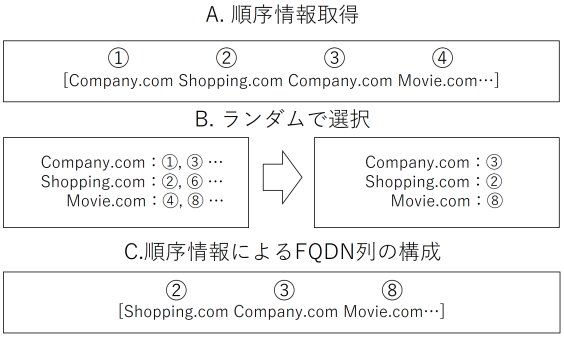

図 2　順序を保持した FQDN 列の作成

連続部の集約は，ある FQDN から，次にどの FQDN へアクセスしたかといった，FQDN 単位の移動の情報を維持する．そのため，ある FQDN[a.com] から他の FQDN[b.com] にアクセスした後に，また [a.com] にアクセスしていることを確認できる．しかし，**連続部の集約**を行っても特定の FQDN が交互に繰り返される場合，何度も同じ FQDN が交互に出現する．Transformer Encoder 事前学習モデルの作成において，同じ FQDN が重複して現れるような**オリジナル**と**連続部の集約**はユーザ属性推定で結果が悪くなることが予備実験を通して分かった．そのため，順序を保持した状態で重複をなくす適切な前処理を行う．

本研究では，単に集合化を行うのではなく，**順序を保持したサンプリング**を行うことで，FQDN が出現した順番を保持したまま，FQDN 列から重複をなくす．**順序を保持したサンプリング**の前処理の詳細な手順を図 2 に示す．**順序を保持したサンプリング**は**連続部の集約**が行われた FQDN 列に対して，行われる．**順序を保持したサンプリング**は FQDN の種類ごとで順序の情報をランダムで選択し，FQDN 列を再構成する．このような前処理を行うことで，Transformer Encoder モデルの学習に用いるウェブ閲覧 FQDN 列の頻度数が多い特定の FQDN が学習に与える影響を可能な限り軽減できると考えた．

前処理を行うことであるウェブ閲覧 FQDN 列に出現する FQDN の種類において FQDN 間の類似性が上手く計算され，ユーザ属性推定モデルに適用できると考えた．図 3 はユーザ毎に取得されるウェブ閲覧 FQDN 列の例を表している．あるユーザの全ての閲覧 FQDN 列を閲覧日時順に並べて作成したウェブ閲覧 FQDN 列から前処理として**連続部の集約**を行い，126 個ごとで FQDN 列を分割する．126 個にした理由は，Transformer Encoder 事前学習モデルを作成する際に，最大入力長を決める必要があるためである．本研究では，Transformer Encoder モデルの最大入力長を 128 とし，分割された FQDN 列の前後に開始と終了を意味する特殊記号を挿入し，学習を行う．126 個に分割されたウェブ閲覧 FQDN 列から**順序を保持したサンプリング**を行うことで，Transformer Encoder モデルに入力する重複がない FQDN 列を作成する．

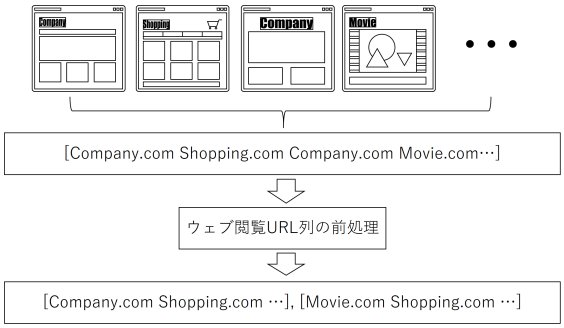

図 3　ユーザ毎に取得されるウェブ閲覧 FQDN 列の前処理

4.3　Transformer Encoder の事前学習

4.3.1　Transformer Encoder の事前学習に用いる学習データ

本研究の使用するデータには，ユーザ属性のラベルが付与されたユーザのウェブ広告閲覧履歴データ以外にも，大量のラベルなしユーザのウェブ広告閲覧履歴データが存在している．

本研究では，この大量のラベルなしデータを用いて，Transformer Encoder を事前学習する．用いたデータは 2016 年 4 月から 2016 年 7 月のウェブ閲覧 FQDN 列である．**連続部の集約**を行ったデータの FQDN 総数は，14,290,248 件で，FQDN 種類数は，990,279 件である．

ユーザの FQDN の履歴に対して， 4.2 節で説明した**順序を保持したサンプリング**を行うことで学習データを擬似的に増やすことが可能である．本研究では，学習の都合上，サンプリング回数を 1 回にしている．また，10 種類以下の FQDN で構成された FQDN 列はノイズになると考え，除くことにした．サンプリングを行い，作成された事前学習に用いる学習データは 2,823,748 件である．

また，Transformer Encoder の事前学習に用いる学習データと推定するユーザ属性ラベル付きのデータでデータリークが生じる可能性はない．

4.3.2　事 前 学 習

本研究における Transformer Encoder は既存の Transformer Encoder 事前学習モデルである RoBERTa [31] のモデルを参考に作成した．本研究の事前学習の目的は，事前学習された学習済みのモデルを用いて，比較的少量のラベル付きデータで教師あり学習を行い，ユーザ属性推定問題の精度を向上させることである．

事前学習に用いるデータは 4.3.1 で説明した 2,823,748 件のサンプリングされた FQDN 列である．事前学習の手法は既存の Transformer Encoder 事前学習モデルである BERT [32] の Mask language を参考にした．本研究の Transformer Encoder の事前学習を図 4 に示す．Transformer Encoder に入力された FQDN の系列から [MASK] に当てはまる FQDN を推定する．このように当てはまる FQDN を推定するには，推定候補となる FQDN をある程度絞る必要がある．本研究では，推定候補の FQDN を 32,000 件とする．推定候補の FQDN は，**連続部の集約**を行った 4.3.1 節の FQDN から出現回数上位 32,000 件の FQDN を対象とした．対象とした FQDN はユーザ属性ラベル付きウェブ閲覧 FQDN 列におけるカバー率が 92 %であったため，事前学習に使用したモデルがユーザ属性を推定する際に問題がないと考えられる．事前学習のハイパーパラメータを表 2 に示す．また，オプティマイザには Adam を用いた．

表 2　事前学習に用いたハイパーパラメータ

ハイパーパラメータ	数値
バッチサイズ	32
学習率（最終層）	2.0×10^{-5}
学習率（他の層）	1.0×10^{-3}
学習バッチ数	847,680

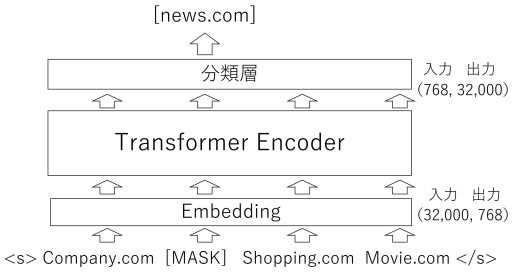

図 4　Transformer Encoder の事前学習

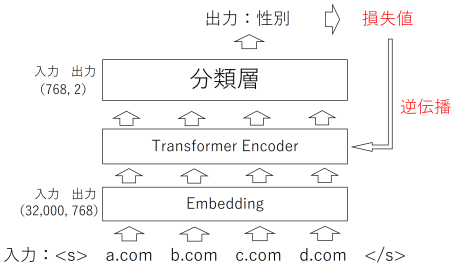

図 5　ユーザ属性分類のためのファインチューニング

4.4　ユーザ属性推定のファインチューニング

4.3.2 節で事前学習を行った Transformer Encoder モデルをユーザ属性推定に適用するためのファインチューニングを行う．本研究では，ユーザ属性推定を行うラベル付きデータは 3.3 で説明したユーザ 1,991 名のデータであり，性別，結婚の有無，年齢を推定する．1,991 名のデータをユーザ単位で訓練とテストに 8 : 2 に分割し，さらに，訓練を検証に分けるため，9 : 1 で分割する． 4.2 節で説明した**順序を保持したサンプリング**を行い，事前学習同様，10 種類以下の FQDN 列を除いた結果，訓練と検証は合わせて 100,536 件であった．

ユーザの属性推定のファインチューニングの概要を図 5 に示す． 4.3.2 節で事前学習を行った Transformer Encoder モデルにユーザ属性ラベルを推定する分類層を追加する．ファインチューニングはユーザのウェブ閲覧 FQDN 列を入力し，予測結果のクロスエントロピーロスで行われる．

図 6 はラベル付きユーザの FQDN 列の系列長を表している．図 6 からすべての FQDN をモデルに入力して学習することが

表 3　ファインチューニングに用いたハイパーパラメータ

ハイパーパラメータ	数値
epsilon	1.0×10^{-8}
バッチサイズ	64
学習率	2.0×10^{-6}
エポック数	10

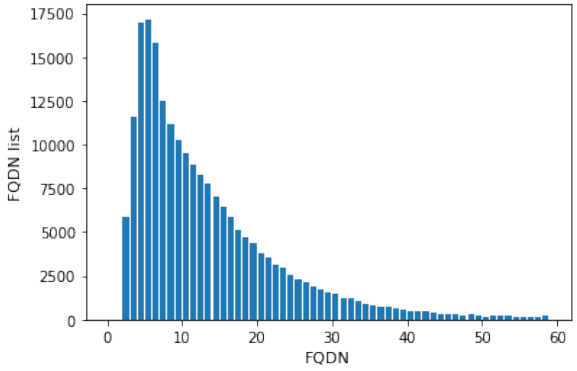

図 6　学習時に入力される FQDN 系列長のヒストグラム

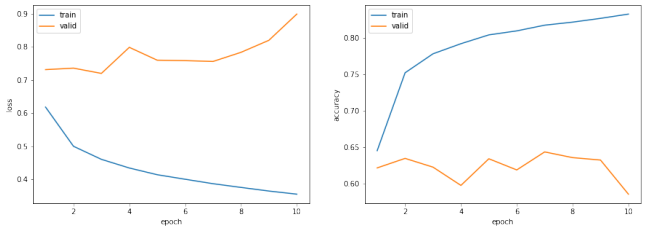

図 7　性別推定の loss と accuracy

出来ていることが分かる．ファインチューニングのハイパーパラメータを表 3 に示す．また，オプティマイザには AdamW を使用した．

図 7 は性別のファインチューニングの学習過程を表している．図 7 をみると，検証データの loss が上昇し，accuracy も低下している．他のユーザ属性のファインチューニングの学習過程においても同じくことが起きていることから過学習などによって，うまく学習が出来ていない可能性が考えられる．

5　評 価 実 験

5.1　実 験 設 定

本実験に使用するデータは広告プロバイダが入手できるウェブ広告閲覧履歴である．ウェブ広告閲覧からウェブ閲覧 URL 列を抽出し，URL を FQDN に変換することでウェブ閲覧 FQDN 列を作成する．ウェブ閲覧 FQDN 列において 2 回以上連続して出現する FQDN を 1 回の FQDN として集約する**連続部の集約**の前処理を行う．さらに，本研究では，順序を考慮したウェブ閲覧 FQDN 列を作成するために，**順序を保持したサンプリング**を用いた．

このように前処理を行ったラベルなしの大量のウェブ閲覧 FQDN 列を用いて，Transformer Encoder モデルで事前学習モデルを作成した．さらに作成した事前学習モデルに対して，

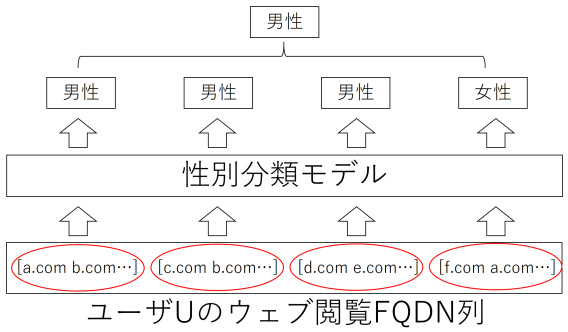

図 8　Transformer Encoder によるユーザ属性推定

表 4　Transformer Encoder モデルの性別推定の結果

	適合率	再現率	F 値
男性	0.71	0.76	0.73
女性	0.76	0.71	0.74

表 5　Transformer Encoder モデルの結婚の有無推定の結果

	適合率	再現率	F 値
既婚	0.33	0.41	0.37
未婚	0.80	0.74	0.76

性別推定に適用するファインチューニングを行った．

実験に先立って，1,991 名のユーザのデータを訓練データとテストデータを 8 : 2 に分割にする．ファインチューニングを行った Transformer Encoder モデルは訓練データを用いており，事前学習だけ行った Transformer Encoder モデルでは使用されてない．評価実験においてはテストデータとしての 399 名のユーザの属性を推定する．これらの学習の実装には，PyTorch（バージョン 1.5）を用いた．

図 8 では実験の Transformer Encoder による性別推定を表している．まず，一人のユーザに対するウェブ閲覧 FQDN 列に対して，**連続部の集約**を行い，126 件で分割した後，**順序を保持したサンプリング**を行った．サンプリングされた FQDN 列をそれぞれ Transformer Encoder モデルに入力し，性別推定を行う．本研究では，それぞれの FQDN 列の性別推定結果から最も多い性別の推定結果を予測結果とする．サンプリングされた FQDN 列が 300 個を超えた場合は 300 個までの FQDN 列の結果だけ取ることにした．テストユーザ 399 人のうち，FQDN 列が 300 個を超えたユーザは 39 人である．また，テストデータの性別の割合は男性 51%，女性は 49%であり，既婚の割合は 69%であり，未婚者は 31%である．

5.2　結　　果

表 4 は，ファインチューニングを行ったモデルの性別推定の適合率，再現率，F 値を示している．正解率は 0.73 となった．図 9 は，ファインチューニングを行ったモデルの性別推定の混同行列を表している．表 5 は結婚の有無の推定の適合率，再現率，F 値を示している．正解率は 0.66 となった．図 10 は，結婚の有無の推定の混同行列を表している．図 11 は，年齢カテゴリの推定の混同行列を表している．正解率は 0.24 となった．こ

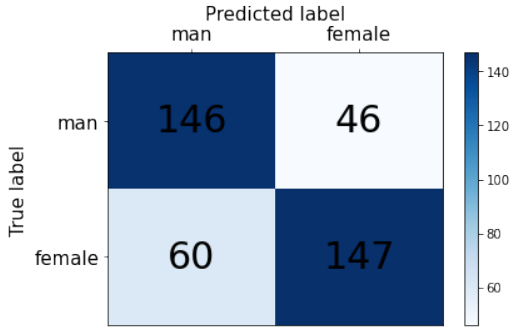

図 9　性別の推定結果の混同行列

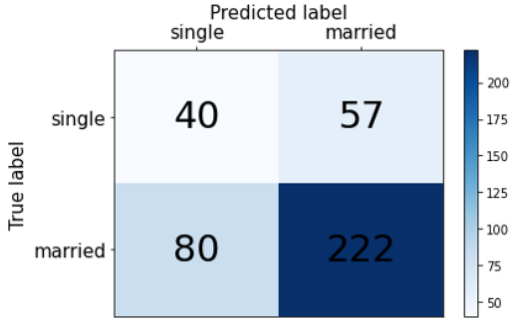

図 10　結婚の有無の推定結果の混同行列

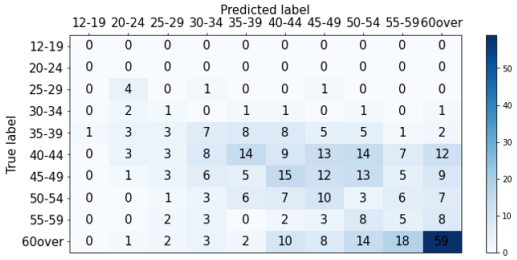

図 11　年齢カテゴリの推定結果の混同行列

れらの結果の比較対象として崔ら [22] のユーザ属性の推定結果が上げられる．Word2Vec によるユーザ特徴ベクトルを作成して性別の推定を行った結果の正解率は 0.85 であり，結婚の有無は 0.74，年齢カテゴリは 0.31 である．本研究での Transformer Encoder モデルによる推定手法は低い結果であった．

5.3　考　　察

大量のウェブ閲覧 FQDN 列から Transformer Encoder モデルの事前学習を行い，ファインチューニングを行ったモデルのユーザの性別推定の結果は正解率 0.73 であり，結婚の有無は 0.66，年齢カテゴリは 0.24 となった．これらの結果は既存手法の崔ら [22] のよりも低い結果であったが，ウェブ閲覧 FQDN 列を用いて，Transformer Encoder モデルを事前学習することによって，ユーザ属性を推定することが可能であることが分かった．正解率が低くなった原因としては，本実験の事前学習の時間が少ないことがあげられる．また，事前学習のデータ量を増やすことも考えられる．本実験では，サンプリングを行うことによって，事前学習のデータを増やすことは容易であるが本実験では行っていない．したがって，より学習を行った事前学習モデルを用いることで，ユーザ属性推定の精度を向上させることができると考えられる．

6　まとめと今後の課題

本研究では，広告プロバイダが入手できるウェブ広告閲覧履歴を用いて，あるユーザのウェブ閲覧 URL 列を抽出し，ユーザの属性を推定する問題に取り組んだ．ウェブ広告閲覧履歴には広告掲載ウェブページのテキスト情報などは含まれず，ユーザを特定できるユーザ識別子と，どの URL にアクセスして来ているといった情報に限定されており，それらをもとにユーザ属性を推定する手法を提案した．

本研究では，順序を考慮したウェブ閲覧履歴の前処理を考えた．まず，ウェブ閲覧 URL から FQDN と呼ばれる部分を変換する処理を行う．つぎに，あるユーザのウェブ閲覧 FQDN 列に対して，同一の FQDN が連続するときに一つにまとめる前処理を行う．その後，順序情報を用いて，FQDN 列から重複がない FQDN 列を作成する前処理手法を提案した．

本研究では，順序を考慮したウェブ閲覧 FQDN 列から Transformer Encoder モデルを事前学習した．その後，ラベル付きのユーザのデータからユーザ属性推定のファインチューニングを行った．ファインチューニングのモデルでユーザ属性の推定を行った結果，性別の正解率は 0.73 であり，結婚の有無は 0.66，年齢カテゴリは 0.24 となった．比較手法の性別の正解率は 0.85 であり，結婚の有無は 0.74，年齢カテゴリは 0.31 となった．

実験の結果，ウェブ閲覧 FQDN 列を用いて，Transformer Encoder モデルを事前学習することによって，ユーザ属性を推定することが可能であることが分かった．しかし，正解率はあまり高いとはいえない．原因としては，本実験の事前学習のエポック数が少ないことがあげられる．また，事前学習のデータ量を増やすことも考えられる．本実験では，サンプリングを行うことによって，事前学習のデータを増やすことは容易であるが本実験では行っていない．したがって，より学習を行った事前学習モデルを用いることで，精度は向上するのではないかと考えられる．

今後の課題としては，まずデータの量を増やして再度検証する必要がある．今回の実験においてはラベル付きのユーザデータ数が限られており，より多くのユーザで，長い期間のデータを用いることで結果を向上させたいと考えている．また，他のユーザ属性においてもユーザ属性の推定を行いたいと考えている．

謝　　辞

本研究の一部は JSPS 科学研究費助成事業 JP18H03243，JP17H00762，JP18H03244 による助成を受けたものです．ここに記して謝意を表します．

文　　献

[1] Jian Hu, Hua-Jun Zeng, Hua Li, Cheng Niu, and Zheng Chen. Demographic prediction based on user's browsing

behavior. In *Proceedings of the 2007 international conference on World Wide Web*, pp. 151–160, 2007.

[2] Ingmar Weber and Alejandro Jaimes. Who uses web search for what: And how. In *Proceedings of the 2011 ACM International Conference on Web Search and Data Mining*, p. 15–24, 2011.

[3] Bin Bi, Milad Shokouhi, Michal Kosinski, and Thore Graepel. Inferring the demographics of search users: Social data meets search queries. In *Proceedings of the 2013 International Conference on World Wide Web*, p. 131–140, 2013.

[4] Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. Predicting age and gender in online social networks. In *Proceedings of the 2011 international workshop on Search and mining user-generated contents*, pp. 37–44, 2011.

[5] Anton Alekseev and Sergey Nikolenko. Word embeddings for user profiling in online social networks. *Computación y Sistemas*, Vol. 21, No. 2, pp. 203–226, 2017.

[6] Eric Malmi and Ingmar Weber. You are what apps you use: Demographic prediction based on user's apps. *arXiv preprint arXiv:1603.00059*, 2016.

[7] Kyriaki Kalimeri, Mariano G Beiró, Matteo Delfino, Robert Raleigh, and Ciro Cattuto. Predicting demographics, moral foundations, and human values from digital behaviours. *Computers in Human Behavior*, Vol. 92, pp. 428–445, 2019.

[8] Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, and Nitesh V Chawla. Inferring user demographics and social strategies in mobile social networks. In *Proceedings of the 2014 ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 15–24, 2014.

[9] Josh Jia-Ching Ying, Yao-Jen Chang, Chi-Min Huang, and Vincent S Tseng. Demographic prediction based on users mobile behaviors. *Mobile Data Challenge*, Vol. 2012, pp. 1–4, 2012.

[10] L Podoynitsina, A Romanenko, K Kryzhanovskiy, and A Moiseenko. Demographic prediction based on mobile user data. *Electronic Imaging*, Vol. 2017, No. 6, pp. 44–47, 2017.

[11] Kajanan Sangaralingam, Nisha Verma, Aravind Ravi, Anindya Datta, and Varun Chugh. Predicting age & gender of mobile users at scale-a distributed machine learning approach. In *Proceedings of the 2018 IEEE International Conference on Big Data*, pp. 1817–1826. IEEE, 2018.

[12] Erheng Zhong, Ben Tan, Kaixiang Mo, and Qiang Yang. User demographics prediction based on mobile data. *Pervasive and mobile computing*, Vol. 9, No. 6, pp. 823–837, 2013.

[13] Jun Ito, Takahide Hoshide, Hiroyuki Toda, Tadasu Uchiyama, and Kyosuke Nishida. What is he/she like? estimating twitter user attributes from contents and social neighbors. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, p. 1448–1450, 2013.

[14] 近藤聖也, 吉田孝志, 和泉潔, 山田健太. 位置情報付きツイートを用いたユーザ属性推定と地域クラスタリング. 人工知能学会全国大会論文集 第 30 回全国大会, Vol. JSAI2016, pp. 1–3, 2016.

[15] Zachary Miller, Brian Dickinson, and Wei Hu. Gender prediction on twitter using stream algorithms with n-gram character features. *International Journal of Intelligence Science*, Vol. 2, No. 4, pp. 143–148, 2012.

[16] John D Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1301–1309, 2011.

[17] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2010 international workshop on Search and mining user-generated contents*, pp. 37–44, 2010.

[18] Abhinay Pandya, Mourad Oussalah, Paola Monachesi, Panos Kostakos, and Lauri Lovén. On the use of urls and hashtags in age prediction of twitter users. In *Proceedings of the 2018 IEEE International Conference on Information Reuse and Integration*, pp. 62–69. IEEE, 2018.

[19] 榊剛史, 松尾豊. ソーシャルメディアユーザの職業推定手法の提案. 日本知能情報ファジィ学会誌, Vol. 26, No. 4, pp. 773–780, 2014.

[20] Siyu Lu, Meng Zhao, Hui Zhang, Chen Zhang, Wei Wang, and Hao Wang. Genderpredictor: a method to predict gender of customers from e-commerce website. In *Proceedings of the 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 3, pp. 13–16. IEEE, 2015.

[21] Peng Jiang, Yadong Zhu, Yi Zhang, and Quan Yuan. Life-stage prediction for product recommendation in e-commerce. In *Proceedings of the 2015 ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1879–1888, 2015.

[22] 崔洙瑚, 木村塁, 南川敦宣, 黒柳茂, 申吉浩. ウェブ広告閲覧履歴を用いたユーザ属性の推定. 第 12 回データ工学と情報マネジメントに関するフォーラム DEIM2020, 2020.

[23] 星尚志, 秋山卓也, 木村塁, 黒柳茂, 南川敦宣. URL エンベディングを用いたライフイベント予測. 情報処理学会 研究報告データベースシステム, Vol. 167, No. 3, pp. 1–5, 2018.

[24] Yukihiro Tagami, Hayato Kobayashi, Shingo Ono, and Akira Tajima. Representation learning for users' web browsing sequences. *IEICE Transactions on Information and Systems*, Vol. E101.D, No. 7, pp. 1870–1879, 2018.

[25] Jianqiao Hu, Feng Jin, Guigang Zhang, Jian Wang, and Yi Yang. A user profile modeling method based on word2vec. In *Proceedings of the 2017 IEEE International Conference on Software Quality, Reliability and Security Companion*, pp. 410–414. IEEE, 2017.

[26] R. Kanagasabai, A. Veeramani, H. Shangfeng, K. Sangaralingam, and G. Manai. Classification of massive mobile web log urls for customer profiling analytics. In *Proceedings of the 2016 IEEE International Conference on Big Data*, pp. 1609–1614, 2016.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 2017 International Conference on Neural Information Processing Systems*, pp. 5998–6008, 2017.

[28] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer. *arXiv preprint arXiv:1809.04281*, 2018.

[29] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[30] Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4920–4928, 2019.