3Q1-9in1

# Entity Linkingを用いたユーザのサイト回遊におけるデモグラフィック推定の検討

Demographic Estimation from Users' Page Views Based on Entity Linking

原 淳史 Atsushi Hara　馬場 惇 Jun Baba　岩崎 祐貴 Yuki Iwazaki　田中 駿 Shun Tanaka

株式会社サイバーエージェント
CyberAgent，Inc.

As demands for branding products by web advertise increases, it is more important to estimate demographics of user who want to appeal by advertisement. Existing approaches estimate age or gender mainly by using simple page information, but those are not enough for estimating some demographics like occupation or annual income. In this paper, we propose an approach using entity and those category in pages which user wandered to estimate demographic of user and compare with existing approaches, and verify the effect using entity.

## 1. はじめに

近年，インターネット広告配信において，企業のブランドイメージの向上 (ブランドリフト) を目的とした広告配信が行われるようになっている．ブランドリフトを目的とした広告配信は，従来の広告配信における商品購入や資料請求などを目標としておらず，商品やサービスを認知してもらいたいユーザ像に広告を配信することを主な目標にしている．そのため，想定ユーザ層に広告配信をするために，ユーザのデモグラフィックやサイコグラフィックを推定する必要がある．

ユーザのデモグラフィックを推定するための 1 つの方法として，ユーザが回遊したウェブページの情報を利用することが考えられる．先行研究としてウェブサイトの本文や画像などのコンテンツに着目し，デモグラフィックを推定する研究は多々提案されている．しかし，セキュリティの発達により，最近では同一セッション内，もしくは，ログインしていなければ，ユーザが閲覧していたページと同一のコンテンツにアクセスできないウェブサービスが増えてきている．また，コンテンツを取得できたしたとしても，文書量の増加や 1 ページあたりの情報量の増加から，アルゴリズムの複雑さによる特徴語の抽出やデモグラフィックの推定にかかる処理時間やシステムコストの増大などの課題がある．

一般的に，ウェブサービスのドメイン URL はトップページに接続できることが多く，メタデータを取得することが容易で，そのメタデータには検索エンジン対策としてそのウェブサイトのテーマや特徴を表現している単語や文章を記述しているウェブサービスが多い．Entity Linking は名詞，特に固有表現に対して Knowledge Base と紐付け，文章理解を助ける手法である．そのため，本稿では比較的入手しやすいドメイン URL のメタデータのタイトルやディスクリプション，キーワードを元に，Entity Linking を用いてウェブサイトの特徴語を抽出し，ユーザのデモグラフィックを推定する手法を提案する．

連絡先: 原 淳史，株式会社サイバーエージェント，Email: hara_atsushi@cyberagent.co.jp
連絡先: 馬場 惇，株式会社サイバーエージェント，Email: baba_jun@cyberagent.co.jp

## 2. 関連研究

デモグラフィック推定の取り組みとして， Hu らのページ毎にデモグラフィックの分布の推定を行った後に閲覧履歴からユーザのデモグラフィックの推定を行う手法 [Hu 07] を提案した．また， Kabbur らのページ内コンテンツやリンクのテキスト情報を用いてユーザのデモグラフィックデータを推定する手法 [Kabbur 10] を提案した．これらのような研究から，ユーザが閲覧したウェブページ内の文章や画像を用いる手段が有効であるといえる．

Entity Linking は文章中の名詞，特に固有表現の名詞と Knowledge Base を紐付ける手法であり，近年多く研究されている．これにより，例えば「ワールドカップ」という表記に対して，サッカー，もしくはラグビーのワールドカップを結びつけることで文章の理解を助けることが可能となり，紐付いた Entity を特徴語としてみなすことで文章の意味を表現することが可能になる．しかし，現在の多くの Entity Linking の研究は文章中の Entity になりうる Entity Mention の推定や Entity Mention がどの Entity と紐づくかを推定する研究例 [Wei 13] が多く， Entity Linking を用いたデモグラフィックを推定する実験例は少ない．

## 3. 提案手法

本章では， Entity Linking を行ったページのメタデータを用いたデモグラフィック類推手法について述べる．はじめに，本提案手法で用いたデータについて説明し，そのデータを用いた Entity Linking について説明する．次に， Entity Linking から得られた結果を用いたデモグラフィックの類推手法について説明する．Entity Linking では Wikipedia や Freebase などの Knowledge Base を参照するが，本稿では日本語 Wikipedia を利用する．

### 3.1 利用データ

ユーザが来訪したメディアのページは，セッションやログインしていなければ取得することができないページが多い．そのため，本稿では，ユーザのサイト回遊のデータとして，ユーザ訪問したメディアページのドメインページを使用する．特にドメインページに含まれるメディアの性質を表す Title，Keyword，Description の 3 つのメタデータをドメインページから抽出し，利用した．

Keyword と Description は似た内容が設定されることが多く，意味的に重複することが考えられるため，これらの 3 つのメタデータを Title と Keyword，Title と Description をつなげた 2 種類のデータを作成し，形態素解析を行い，得られた名詞を素性データとした．形態素解析器には MeCab[*1] を利用した．

### 3.2 Entity Linking

3.1 で得られた素性データに対して， Entity Linking を行う．本稿では， 鈴木正敏ら [Suzuki 16] によって提案された日本語 Wikipedia エンティティベクトル [*2] を利用し，素性データを Entity に置換する．素性データを Entity に置換する際，候補 Entity に対し，Wikipedia 記事ベクトル間の cosine 距離を足し上げ，ラティスのエッジスコアの総計で Entity を決定し，置換する．

$$\hat{y} = \arg\max Score(x, y) \tag{1}$$

ここで,$\hat{y}$は最適経路, $Score(x, y)$ は経路 $y$ の最順位付けを行うスコア関数とする．また，Entity として置換される名詞を素性として利用し，Entity として置換されない名詞は特徴のない名詞と判断し，素性データから除外する．

### 3.3 デモグラフィック推定

3.2 で得られた Entity に置換された素性データを用いて，ユーザごとの特徴ベクトルを作成する．ユーザが訪問したメディアの Entity 群には 1 を，訪問していない Entity 群には 0 を特徴ベクトルの値として与える．したがって，ユーザが訪問した Entity $UE$ を

$$U_iE_j = \begin{cases} 1 & (if\ user\ access) \\ 0 & (otherwise) \end{cases}$$

としたとき，User-Entity 行列 UE が得られる．ここで, $i$ , $j$ はそれぞれユーザと $Entity$ のインデックスを示し, $(i = 1$ , 2,..., $N)$, $(j = 1, 2, ..., M)$ の値域を取る．$N$ はユーザ数，$M$ は $Entity$ 数とする.. デモグラフィック属性の学習には，Freedman が提案した Gradient Boosting Decision Tree(GBDT)[Freedman 01] に用いる．

## 4. 評価実験

### 4.1 実験概要

本実験では 2017 年 2 月のある 1 週間の広告接触ログと，それに紐づく第三者機関のデモグラフィック属性データを利用して，ユーザデモグラフィック属性の予測タスクを行った．利用するデモグラフィック属性は表 1 に示す通り，性別，年代，未既婚，子供有無の属性である．

表 2 に示すように, 学習対象のユーザについては，広告接触したユニークドメイン数が 5 以上あるユーザに制限し，71,196 ユニークユーザを対象とした．一方，学習対象のドメインについては，接触したユニークユーザが 10 以上あるドメインに制限し，1276 ドメインを実験対象のドメインとした．

本実験では，以下の特徴量で分類タスクの精度比較を行い，Entity-Linking の有効性について評価する．以下で作成される bag-of 特徴では，節 3.3 で示したように，特徴量の値は接触回数ではなく接触したか否かの 0/1 で構成される．

*1 https://sourceforge.net/projects/mecab/
*2 http://www.cl.ecei.tohoku.ac.jp/ m-suzuki/jawiki_vector

| デモグラフィック属性 | 内容 |
|---|---|
| 性別 (gender) | 1: 男性, 2: 女性 |
| 年代 (age) | 20: 20-29 歳，30: 30-39 歳，40: 40-49 歳，50: 50-59 歳，60: 60-69 歳 |
| 未既婚 (marriage) | 1: 未婚，2: 既婚 |
| 子供有無 (child) | 1: 無，2: 有 |

表 1: デモグラフィック属性の種類

| 項目 | 統計量 |
|---|---|
| 学習対象ユーザ数 | 71,196 UU |
| 学習ユーザへの広告接触数 | 72,947,263 UU |
| 広告接触ログに含まれるドメイン数 | 4,482 ドメイン |
| 名詞・Entity 抽出可能ドメイン数 | 1,276 ドメイン |
| 取得した名詞数 | 4,431 語 |
| 取得した Entity 数（title&keywords） | 2,049 語 |
| 取得した Entity 数（title&description） | 2,459 語 |

表 2: 実験データの統計量

**bag-of-domain(BoD):**
ドメイン URL を利用して bag-of 特徴量を作成する．

**bag-of-noun(BoN):**
ドメインのメタ情報からタイトルを抜き出し，その名詞のみを利用して bag-of 特徴量を作成する．

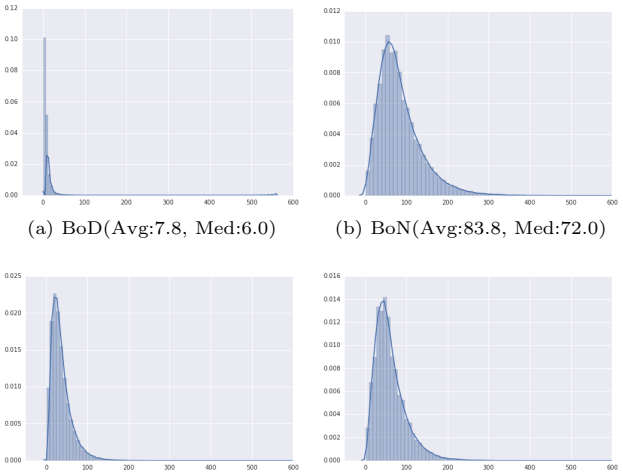
**bag-of-Title-Keyword-Entity(BoTKE):**
ドメインのメタ情報からタイトルとキーワードを抜き出し，その名詞から Entity を抽出する．その Entity から bag-of 特徴量を作成する．

**bag-of-Title-Description-Entity(BoTDE):**
ドメインのメタ情報からタイトルとデスクリプションを抜き出し，その名詞から Entity を抽出する．その Entity から bag-of 特徴量を作成する．

各デモグラフィック属性をいくつかのクラスに分け，1 対多の二値分類を全クラスで行い，評価尺度は Area Under ROC (AUROC) を利用した．本実験の教師クラスは，性別と年代の組み合わせ，未既婚と子供有無の組み合わせ，の 2 種類を用意する．また，予測器の GBDT のパラメータは，木の数を 200, 木の最大の深さを 7，学習率を 0.10，サブサンプルを 0.80 とし，全ての比較手法と属性クラスで同一とした．

### 4.2 比較手法での特徴次元

図 1 は，各比較手法ごとに作成した特徴量における，1 ユーザあたりの非ゼロ特徴数の分布を表す．1 ユーザの行動を平均して何次元の特徴で表現しているかがわかる．図 1a が示すように，BoD が 1 ユーザあたり平均で 7.8 個の素性を持ち，比較手法内で最も少ない．一方,BoN が最も多く，平均 83.8 個の素性を持つ（図 1b）．Entity で特徴量を作成する BoTKE と BoTDE では，それぞれ平均 35.4 個と 59.7 個であったことから，BoN の素性数より少ない次元でユーザの行動を表現しようとすることが分かる（図 1c/1d）．特徴空間を表す基底という観点から考えると，最も冗長な基底を持つのが BoN であり，そこから BoTDE，BoTKE，BoD の順番でより少ない基底で特徴空間を表していることになる．



(a) BoD(Avg:7.8, Med:6.0) (b) BoN(Avg:83.8, Med:72.0)

(c) BoTKE(Avg:35.4, Med:30.0) (d) BoTDE(Avg:59.7, Med:52.0)

図 1: 1 ユーザ当たりの非ゼロ特徴数の分布

### 4.3 実験結果

| | BoD | BoN | BoTKE | BoTDE |
|---|---|---|---|---|
| class | auroc | auroc | auroc | auroc |
| gender1_age20 | 0.8448 | 0.8402 | **0.8483** | 0.8473 |
| gender1_age30 | 0.7432 | 0.7551 | **0.7591** | 0.7535 |
| gender1_age40 | 0.7480 | 0.7673 | **0.7707** | 0.7684 |
| gender1_age50 | 0.8290 | 0.8361 | **0.8383** | 0.8356 |
| gender1_age60 | 0.8660 | 0.8750 | 0.8747 | **0.8790** |
| gender2_age20 | 0.8152 | **0.8252** | 0.8109 | 0.8246 |
| gender2_age30 | 0.7384 | **0.7391** | 0.7350 | 0.7350 |
| gender2_age40 | 0.6625 | **0.6581** | 0.6535 | 0.6539 |
| gender2_age50 | **0.7253** | 0.7191 | 0.7211 | 0.7118 |
| gender2_age60 | 0.7816 | **0.8010** | 0.7957 | 0.7989 |
| marriage1_child1 | 0.7174 | 0.7237 | **0.7244** | **0.7244** |
| marriage1_child2 | 0.5808 | 0.5816 | 0.5863 | **0.6006** |
| marriage2_child1 | 0.6018 | 0.6025 | 0.6037 | **0.6068** |
| marriage2_child2 | 0.6796 | **0.6898** | 0.6869 | 0.6873 |

表 3: 各属性クラスの AUROC

実験結果を表 3 に示す．各行は，対象の属性クラスを正例，それ以外を負例として二値分類を行った時の，AUROC 値を比較している．太字は各属性クラスでの最大 AUROC 値を表す．

性別/年代クラスの推定精度においては，男性 (gender=1) のクラス全般で BoTKE，BoTDE が最大の AUROC を獲得しており，特にほぼ全ての男性クラスで BoTKE が効果的であるといえる．また，未既婚/子供有無クラスの推定タスクにおいては，提案手法が高い精度を獲得しており，特に BoTDE が 75%のクラスで最大の精度を持っている．しかし，女性（gender=2）のクラス全般では，提案手法ではなく BoN が高い AUROC を達成しており，次いで，BoD が高い精度を出している．

男性クラスと未既婚/子供有無クラスにおいて，提案手法が高い精度を達成できたのは，Entity Linking によりページコンテンツの内容を表記ゆれなどを吸収して表現することができているからであると推測できる．同じクラスのほぼ全てにおいて，BoN の方が BoD よりも高い推定精度を出していることから，ページコンテンツを利用することは有用であることが言える．そして，BoTKE と BoTDE が BoN よりも精度が高いこと，また，1 ユーザあたりの平均素性数を削減していることから，表記ゆれや同義語へのマッピングがうまく貢献したと考えられる．

しかし，女性クラスのほとんどで，BoN が Entity を利用した提案手法よりも高い精度を出せているのは，ドメインのわかりやすさにあると考えられる．女性が訪れやすい Web サイトは，女性に向けて作られていることが多く，ドメイン URL 単位で十分特徴を表現しやすいのではないかと推測できる．提案手法よりも BoD の方の推定効果が高いことからも，Entity による次元圧縮が効きづらいことがわかる．

## 5. おわりに

本稿では，ユーザが回遊した Web ページのメタデータ情報を Entity にマッピングし利用することで，ユーザのデモグラフィック属性を推定する手法を提案した．

今回の実験から，まず，各属性クラスで少なくともページコンテンツを利用した手法の精度が高く，デモグラフィック推定においてページコンテンツを利用することが有用であることを示した．また，提案手法は男性属性や未既婚/子供有無の属性に対して高い分類精度を示し, 推定するクラスによっては Entity の利用がうまく表記ゆれや同義語へのマッピングができ，効果的であることがわかった．

今後の研究として，本稿で利用した以外の属性クラスへ拡張いく予定である．また，よりページの主題を表現できる Topic モデルとの比較を行っていきたいと考えている．

## 参考文献


[Hu 07] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic Prediction Based on User ' s Browsing Behavior. WWW, 2007.

[Kabbur 10] S. Kabbur, E.-H. Han, and G. Karypis. Content-Based Methods for Predicting Web-Site Demographic Attributes. ICDM, 2010.

[Li 10] L. Li, T. Mei, X. Niu, and C.-W. Ngo, PageSense: style-wise web page advertising, WWW, 2010.

[He 15] X. He, W. Dai, G. Cao, R. Tang, M. Yuan, and Q. Yang. Mining target users for online marketing based on app store data. IEEE, 2015.

[Suzuki 16] 鈴木正敏, 松田耕史, 関根聡, 岡崎直観, 乾健太郎. Wikipedia 記事に対する拡張固有表現ラベルの多重付与. 言語処理学会第 22 回年次大会 (NLP2016), March 2016.

[Freedman 01] J.H. Freedman. Greedy function approximation: A gradient boosting machine. Ann. Statist., vol. 29 (2001), pp. 1189-1232.

[Jacob 10] Eisenstein, Jacob and O'Connor, Brendan and Smith, Noah A. and Xing, Eric P. A Latent Variable Model for Geographic Lexical Variation

[Wei 13] Shen, Wei and Wang, Jianyong and Luo, Ping and Wang, Min. Linking Named Entities in Tweets with Knowledge Base via User Interest Modeling. KDD, 2013.