The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

新聞記事のアクセスログを用いたユーザ属性の逐次推定

Sequential User Profiling from Newspapers' Access Log

工藤 航　　鳥海 不二夫
Wataru Kudo　　Fujio Toriumi

東京大学
The University of Tokyo

With the spread of social media and e-commerce websites, the technology of user profiling from users' action history has attracted a lot of interest. Users' action history can be acquired both in the passive way and in the active (interactive) way, and previous studies have found out how to presume the users' profile from their action history which is acquired passively. The purpose of this study was to find out the best way to interact with users for efficient user profiling. First, we constructed a CNN (Convolutional Neural Network) model which presumes user's profile. Next, we proposed three ways to interact with users for efficient user profiling. Two ways of the three were discovered to be effective to streamline the user profiling.

1. はじめに

近年，ソーシャルメディアや EC サイトなどの普及に伴い，蓄積された膨大なデータを用いた研究が盛んに行われている．その一例として，各種のサービスにおけるユーザの行動履歴からユーザの性別や年齢といった属性を推定する手法の研究が挙げられる．属性未知のユーザに対する属性推定は，社会科学などの学術的な分野，および企業のマーケティング戦略のような商業的な分野に大きく貢献することが期待できる．

属性推定に用いる行動履歴の取得方法として，ユーザとの対話的なやり取りの中でユーザの反応を記録する方法が考えられる．EC サイトやマルチメディアにおけるコンテンツのリコメンドに対するユーザの反応を取得することも，対話的な情報取得の一例と言える．

本研究では，対話的な情報取得が行える状況の下で，ユーザから引き出す情報を属性推定という目的に応じて能動的に選択することによって，受動的に情報が蓄積される場合に比べて効率的に推定を行うことができるという可能性に着目する．本研究では，まずユーザが読んだ新聞記事のタイトルからユーザの属性を推定できるモデルを構築する．次に，構築したモデルを用いて記事の入力に対し逐次的な推定を行うことで推定の効率を評価し，効率的な推定を行うためのユーザとの対話戦略について明らかにする．

2. データ概要

本研究で用いるデータの概要を表 1 に示す．本研究では，ユーザが読んだ記事のタイトルからユーザの性別・年齢・職業を推定するモデルを構築する．性別は { 女性，男性 } の 2 種類，年齢は {20 代以下，30 代，40 代，50 代，60 代以上 } の 5 種類，職業は { 会社員，主婦，学生，無職，自営業 } の 5 種類とした．

表 1: データ概要

データ提供	日本経済新聞社
対象ユーザ	日経電子版に登録されたユーザ
対象期間	2017/5/21 から 2017/8/18 まで
ユーザ数	1,606,192
記事数	104,896
行動履歴	ユーザが読んだ記事のタイトル
推定する属性	性別，年齢，職業

3. 属性推定モデルの構築

3.1 手法

本研究では，テキストの意味的な特徴からユーザ属性を推定するモデルを構築するべく，Kim[Kim 14] が提案した畳み込みニューラルネットワーク (CNN) を参考にしたモデルを用いた．まず，ユーザが読んだ記事のタイトルというテキスト情報に対し，工藤ら [Kudo 05] が開発した形態素解析エンジン MeCab を用いて単語に分ける．一つ一つの単語に対し，Mikolov ら [Mikolov 13] が提案した手法を用いて鈴木ら [鈴木 16] が日本語版 Wikipedia の本文全体から学習を行った word2vec モデルを用いて単語の埋め込みを行う．これらの処理を経て行列化された個々のユーザに関するテキスト情報を CNN の入力とする．

比較対象としては，ランダムに推定を行った場合の期待される結果のほかに，Lilleberg[Lilleberg 15] が提案したサポートベクターマシーン (SVM) を参考にした手法を採用した．表 2 に，各手法の特徴を示す．

表 2: 各手法の特徴

	特徴抽出方法	意味	時系列性
CNN	畳み込み	○	○
SVM	ベクトル加重和	○	×

連絡先: 工藤航，東京大学大学院システム創成学専攻 鳥海研究室，東京都文京区根津 1-24-5，09030938349，kudo@torilab.net

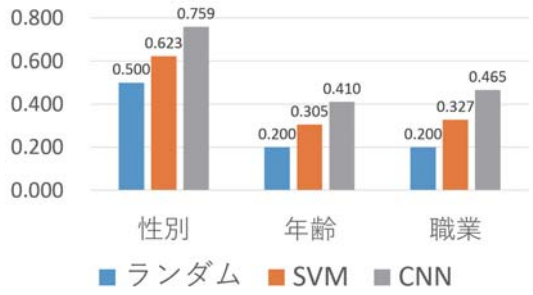

図 1: 正解率の比較

3.2 結果

各手法によって性別・年齢・職業という 3 種類の属性に対し推定を行ったときの正解率を図 1 に示す．機械学習を用いることで正解率を高められること，及び CNN を用いることで最も精度良く推定を行えることがわかった．

3.3 考察

図 1 に示したとおり，全ての属性において，推定の正解率は CNN(畳み込みニューラルネットワーク) > SVM(サポートベクターマシーン) > ランダム という結果になった．このことから，ユーザの読む記事のタイトルの意味的な特徴からユーザの属性を推定することが可能であると示された．また，CNN のモデルが最高の性能を示したことから，本研究の属性推定問題に対して記事のタイトル中の単語の時系列性を考慮すること，および単語の意味的な特徴を抽出する方法として畳み込みを用いることの有効性が示された．

4. 効率的な属性推定

4.1 記事の優先度を測る 3 つの指標

ユーザから対話的に情報を取得して効率的な属性推定を行うためには，推定の手がかりになりそうな記事を優先的に選択し，ユーザの反応を取得していく必要がある．そこで，属性推定という目的に応じて各記事の優先度を比較するための数値的な指標を 3 つ提案した．

指標 1 (被アクセスエントロピー)

読者の属性分布に偏りがある記事ほど属性推定の手がかりになりやすいと考えられる．そこで，各記事の読者の属性分布のエントロピーを計算し，指標 1 (被アクセスエントロピー) と定義する．分布の偏りが大きいほどエントロピーの値が小さくなるため，被アクセスエントロピーの値が小さい記事から優先的に選択していくことによって推定の効率化を図る．

指標 2 (推定出力エントロピー)

各記事のタイトルをモデルに入力することで，属性の各分類ラベルへの所属確率を表す推定出力を得ることができる．この推定出力の偏りが大きい記事ほど属性推定の手がかりになりやすいと考えられる．そこで，推定出力の確率分布のエントロピーを計算し，指標 2 (推定出力エントロピー) と定義する．推定出力エントロピーの値が小さい記事から優先的に選択していくことにより，推定の効率化を図る．

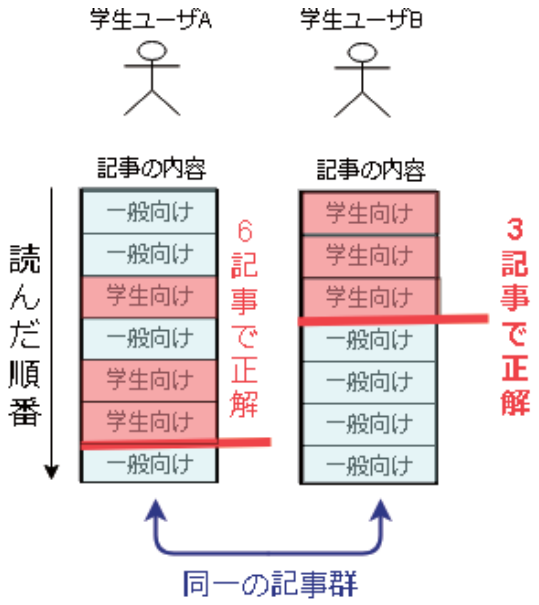

図 2: 同一の記事を読んだ 2 人の学生ユーザに対する推定効率の違い

指標 3 (正解ラベル確率上昇度)

ユーザが読んだ記事から属性推定を行う場合，記事の入力に対して逐次的に推定を行うことによって，一つ一つの記事がもたらした推定への影響を測ることができる．ここで，属性が既知であるユーザに対して逐次推定を行った際，各ユーザの正解ラベルの確率を大きく上昇させた記事ほど推定への貢献度が高いと考えることができる．

そこで，属性既知の多数のユーザに対して逐次推定を行い，各記事がもたらした正解ラベル確率の上昇度を全て記録する．それらの上昇度について記事ごとに総和を計算したものを指標 3 (正解ラベル確率上昇度) と定義し，新たな別のユーザに対して推定を行う際に指標 3 の値が大きい記事から優先的に選択することによって推定の効率化を図る．

4.2 3 つの指標がもたらす推定効率の評価・比較

図 2 のように，2 人のユーザが同一の記事を読んだとしても，読む順番が異なれば正解に到達できるまでに必要な記事数は異なる．すなわち，推定の効率はユーザが記事を読む順番に依存すると考えられる．ここで，ユーザに記事タイトルを提示してユーザが読むかどうかを記録するという方法で行動履歴を取得する際，属性推定の手がかりになりそうな記事から優先的に選択して提示することによって推定の効率化が期待できる．

本研究では，4.1 節で述べた 3 つの指標に従った順番で記事を選択し，記事選択ごとに各ユーザの推定結果を更新していく逐次推定を行った．そして推定効率の評価基準として，記事選択のステップごとの属性推定の正解率の推移を採用した．図 3 で示すように，正解率がより速く上昇する推定の方が効率の良い推定であると評価できる．

そこで，推定効率の数値的な評価基準を得るために，図 4 のように正解率の推移曲線の積分値を計算し「効率面積」と定義する．効率面積が大きい推定の方が効率の良い推定であると評価できる．

提案した 3 つの指標がもたらす推定の効率を評価・比較す

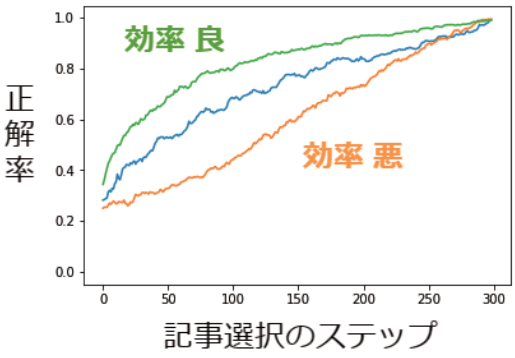

図 3: 正解率の推移

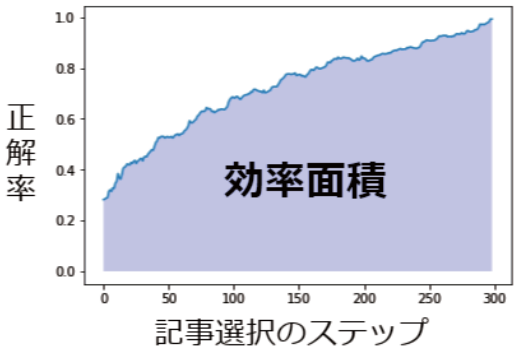

図 4: 効率面積の定義

るために，設定の異なる 2 つの実験を行った．1 つ目の実験と結果について 4.2.1 項，2 つ目の実験と結果について 4.2.2 項で述べる．

4.2.1 実験 1 (ユーザが読んだ記事の並べ替え)

1 つ目の実験は，各ユーザが読んだ記事を 3 つの指標の値によって並べ替えて入力を行う実験である．「指標によって並べ替えられた順番通りにユーザが記事を読んだと仮定したときに，推定効率が高くなるのはどの指標を用いた場合なのか」ということが明らかになる．この実験の結果から，個々の記事における「ユーザがその記事を読んだ際の推定への貢献度」を各指標が正しく予測できるかどうかという評価を得ることができる．

図 5 に，ユーザが読んだ順番で入力を行った場合，ユーザが読んだ記事をランダムに並べ替えて入力を行った場合，及び 3 つの指標によって並べ替えて入力を行った場合の推定の効率面積を示す．

図 5 より，読んだ順番に入力した場合とランダムに並べ替えた場合は同程度の効率を示すということ，そして提案した 3 つの指標によって効率的に推定を行えることが明らかになった．また，特に指標 2 (推定出力エントロピー)，指標 3 (正解ラベル確率上昇度) を用いることで効率の良い推定を行えることがわかった．

4.2.2 実験 2 (ユーザが読んでいない記事を含めた並べ替え)

2 つ目の実験は，1 つ目の実験に比べより現実的な状況を想定したものである．

ユーザが読んだ記事と読んでいない記事からなる候補記事群を 3 つの指標によって並べ替えて順に選択していく．選択された記事が実際にユーザが読んだものであった場合はその記事

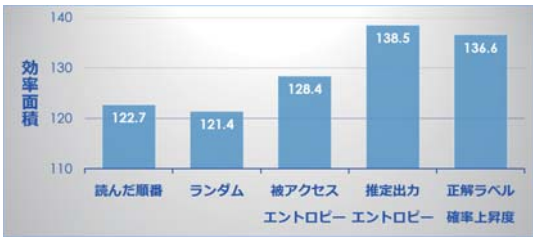

図 5: 効率面積の比較 (ユーザが読んだ記事を並べ替えた場合)

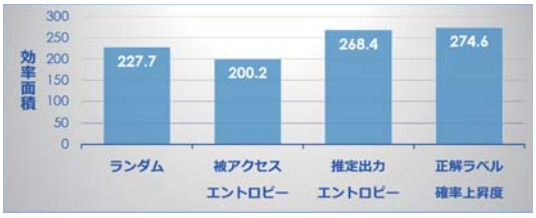

図 6: 効率面積の比較 (ユーザが読んだ記事と読んでいない記事からなる候補記事群を並べ替えた場合)

を入力し，各ラベルの所属確率の更新を行う．ユーザが読まなかった記事を選択してしまった場合は，確率に変化は起こらない．従って，効率の良い推定を行うためにはユーザの読む確率が高い記事を選択する必要がある．

この実験では，各指標によって推定への貢献度が高い記事を優先的に選択できるかということに加え，ユーザの読む確率が高い記事を優先的に選択できるかということも含めた総合的な推定効率化についての評価を得ることができる．

図 6 に，ランダムに並べ替えて選択を行った場合，及び提案した 3 つの指標によって並べ替えて選択を行った場合の推定の効率面積を示す．

4.2.1 項で行った実験とは異なり，指標 1 (被アクセスエントロピー) を用いた場合はランダムに並べ替えて選択を行った場合よりも推定効率が下回った．また，指標 3 (正解ラベル確率上昇度) を用いた場合の推定効率が指標 2 (推定出力エントロピー) を上回った．

4.3 考察

まず，実験 1 で示された結果について考察する．読んだ順番に入力した場合がランダムに並べ替えた場合と同程度の効率を示したことから，受動的に蓄積されたユーザの行動履歴を用いた属性推定は非効率であるということが示された．また，提案した 3 つの指標を用いることで推定を効率化できたことから，3 つの指標によって各記事の推定への貢献度を予測可能であるということが明らかになった．さらに，指標 2 (推定出力エントロピー) と指標 3 (正解ラベル確率上昇度) が指標 1 (被アクセスエントロピー) を上回ったことから，推定モデル由来の指標の方が統計情報をもとにした指標よりも優れているということが示された．

次に，実験 2 の結果について考察する．指標 1 (被アクセスエントロピー) を用いた場合がランダムに並べ替えた場合を下回ったこと，指標 3 (正解ラベル確率上昇度) が指標 2 (推定出力エントロピー) を上回ったことという 2 点において実験 1 と異なる結果となった．このことから，選択された記事をユーザが読む確率が指標ごとに異なり，そのことが推定の効率に影響

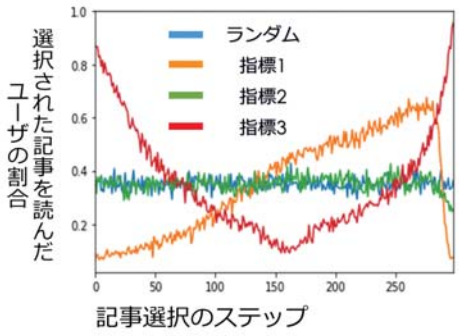

図 7: 選択された記事を読んだユーザの割合の推移

表 3: 各指標の評価 (◎:非常に有効，○:有効，△:効果なし，×:逆効果)

	推定への貢献度	読まれる確率	推定の効率化
ランダム	△	△	△
指標 1	○	×	×
指標 2	◎	△	○
指標 3	◎	○	◎

を与えたと考えられる．

そこで，選択された記事を読んだユーザの割合が記事選択のステップごとにどのように推移していくかを指標ごとに調べたところ，図 7 のようになった．

推定効率という観点でより重要だと考えられる序盤の記事選択に着目すると，指標 3 (正解ラベル確率上昇度) を用いることでユーザの読む確率が高い記事を優先的に選択できていることがわかる．一方で，指標 1 (被アクセスエントロピー) を用いるとユーザの読む確率が低い記事を優先的に選択してしまう傾向があることが明らかになった．以上のことから，各指標の評価をまとめると表 3 のようになる．推定への貢献度の予測，およびユーザが読む確率の予測という両面から考えると，指標 3 が最も推定を効率化できるということが明らかになった．

5. 結論

本稿では，日本経済新聞社が提供する Web サービス「日経電子版」に登録しているユーザを対象とし，テキスト情報からのユーザ属性の推定手法，属性推定の効率化という 2 つの連続したテーマで研究を行った．

第 3 章で述べたように，日経電子版に登録されたユーザの読んだ記事のタイトルから，ユーザの性別・年齢・職業が推定可能であるということがわかった．すなわち，読む記事には属性ごとに傾向があり，記事のタイトルのセマンティックな情報からその傾向を学習可能であるということが明らかになった．

また，第 4 章で述べたように，能動的な推定において効率的に推定を行う手法を発見することができた．提案した 3 つの指標のうち，2 つが推定の効率化に有効であるということが明らかになった．さらに，記事の推定への貢献度と記事が読まれる確率という 2 つの特性を反映した指標が有効であるということが示され，提案した 3 つの指標の中では指標 3 (正解ラベル確率上昇度) が最も有効であるということが明らかになった．

今後の課題としては，属性推定モデルの構築，推定の効率化という 2 つの観点から考えられる．

属性推定モデルの構築という観点からは，まず推定精度の向上という課題が挙げられる．本稿で畳み込みニューラルネットワークのモデル構築の際に参考にした Kim[Kim 14] の文献では，学習済み word2vec モデルのパラメタも更新の対象とすることにより推定精度を高められることが示されている．このように，タスクに特化するようにモデルのパラメタを更新するアプローチは性能を向上させることが期待できるため，推定精度を実用的な水準まで向上させるために今後取り組むべき課題であると言える．また，本稿では日経電子版に登録されたユーザに対する属性推定を行ったが，その他のデータセットを用いて本稿で得られた結果の再現性を検証することも課題である．

推定の効率化という観点からも，まずは複数のデータセットを用いた再現性の検証が挙げられる．また，本稿で提案した 3 つの指標はそれぞれ統計情報，モデルによる出力，過去に行われた逐次推定から得られる比較的単純な指標であり，推定モデルの内部構造に依存しない．したがって，異なる推定モデルを用いて本稿と同様の実験を行うことにより，本稿で提案した入力の優先順位の予測方法が汎用的であるかを検証することも課題の一つであると言える．逆に，推定モデルの内部構造を考慮した予測方法によって更に推定の効率を高められる可能性もあるため，その手法を検討することも課題の一つと考えられる．さらに，記事について「ユーザが読んだ際の推定への貢献度」と「ユーザが読む確率」の両方を反映した予測が行えるように，過去の逐次推定から学習する方法をより高度化していくことも課題として挙げられる．

参考文献

[Kim 14] Kim, Y.: Convolutional neural networks for sentence classification, *arXiv preprint arXiv:1408.5882* (2014)

[Kudo 05] Kudo, T.: Mecab: Yet another part-of-speech and morphological analyzer, *http://mecab. sourceforge. net/* (2005)

[Lilleberg 15] Lilleberg, J., Zhu, Y., and Zhang, Y.: Support vector machines and word2vec for text classification with semantic features, in *Cognitive Informatics & Cognitive Computing (ICCI* CC), 2015 IEEE 14th International Conference on*, pp. 136–140IEEE (2015)

[Mikolov 13] Mikolov, T., Chen, K., Corrado, G., and Dean, J.: Efficient Estimation of Word Representations in Vector Space, *CoRR*, Vol. abs/1301.3781, (2013)

[鈴木 16] 鈴木正敏，松田耕史，関根聡，岡崎直観，乾健太郎?FWikipedia 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 (2016)