ツイートが運ぶ個人属性の分析

余岳* 笹原和俊†

概要

近年，ソーシャルメディアの投稿内容からユーザーの個人属性を推測する研究が盛んである．本研究は，Twitter のデータを用いて，テキストの内容から深層学習を含む異なる機械学習のアルゴリズムを用いて，性別，職業，年齢層の 3 つの個人属性の予測に関する実験を行う．word2vec によって単語の分散表現を作り，複数のツイートを 1 つのブロックにまとめてツイートをベクトル化し，このツイートのブロックのベクトルに基づいて予測精度を調べた．その結果，適切なパラメータを選んだ場合，3 つの個人属性は 60 〜 70% の精度でアルゴリズム的に予測できることが明らかになった．さらに，個人属性を予測できたことの理由を明らかにするため，性別を例とし，男女の共通単語のコサイン距離と使用頻度の予測精度への影響を調べた．性別予測する場合，男女が共通して使用する単語が予測精度に大きく影響することが分かった．

1 はじめに

ソーシャルメディアは，情報のハブかつ意見交換のプラットフォームであり，研究者にとっては人間行動に関するビッグデータを収集できる場所でもある．このような時代背景から，計算社会科学 (Computational Social Science) と呼ばれる新しい学際科学が誕生し，現在盛んに研究が行われている [1]．例えば，Twitter では「ゆるいつながり」による多様なコミュケーションを観測することができ，ソーシャルデータの入手が容易なため，計算社会科学の研究で頻繁に利用されている [2, 3].

近年，ソーシャルデータを用いた研究で，特に活発に行われているのは個人属性の推定に関する研究である．ソーシャルデータに含まれる大量の言語表現には，本人が意図しなくても，パーソナリティーを反映した情報が含まれていると考えられる [4]．Schwartz らは，Facebook の投稿に含まれる単語とトピックの使用頻度に基づいて辞書を作成し，性格，性別，年齢を判別できることを示した [5]．Kosinski らは，Facebook の「いいね !」（気にいったことを示すシグナル）の頻度に基づいて，線形回帰などの比較的簡単な方法でも個人属性を推定できることを示した [6]．Liu たちは，中国のソーシャルメディア Weibo の投稿に基づいて，autoencoder を学習させ，投稿者のビッグ・ファイブと呼ばれる人間の性格因子（神経症傾向，外向性，経験への開放性，協調性，誠実性）を推定できることを示した [7]．これらのように，ソーシャルデータに散在する個人情報を数値化・推定する手法は，現在，多くの研究者が取り組んでいるものの，まだ確立した一般性の高い方法は存在しない．

本研究の目的は，ソーシャルデータのテキストのみから個人属性をどの程度推定できるかに関するベースラインの知見を獲得し、その理由を調査することである．そこで，Twitter のデータを用いて，機械学習のアルゴリズムによって性別，職業，年齢層の 3 つの個人情報を予測する実験を行った．本論文ではその予備的な結果について報告する．

* 名古屋大学大学院情報学研究科, yotake1987@nagoya-u.jp
† 名古屋大学大学院情報学研究科, JST さきがけ, sasahara@nagoya-u.jp

2 方法

2.1 データ収集と処理

本研究では予測モデルを構築するために，Twitter の個人プロフィールが明記されていたり，Wikipedia などの Web の記事によって性別と職業が確認でき，投稿されたツイート数が 3000 以上のアクティブアカウントを 120 人特定した．また，これらのアカウントの年齢の情報も Web 検索で可能な限り特定した（年齢は非公開な場合が少なくない）．これらのアカウントのうち，100 人をトレーニング用，残り 20 人をテスト用にした．

これら 120 人のアカウントの特徴は，次のようにまとめられる．男性と女性のアカウント数は等しく，偏りがない．職業は 10 種類を選択し，それぞれの職業に属するアカウント数は等しく，偏りがない．年齢層は，1980 年以後の生まれ（デジタルネイティブ）か，それ以前の生まれ（デジタル移民）の 2 種類にした．

次に，これらのアカウントから Twitter の公式 API を用いて，リツイートや返信も含むユーザータイムラインを可能な限り収集した．その結果，トレーニング用のアカウントからは 314382 個，テスト用のアカウントからは 64027 個のツイートが収集された．

その後，次の手順でデータの処理を行った．収集したツイートを，日本語形態素解析ツール MeCab と日本語辞書 NEologd を用いて分かち書きの処理をした．分かち書きしたツイートから，長さが 4 未満の情報量が少ないツイートを削除した．クリーニング後に残ったツイートは，トレーニング用が 312169 個，テスト用が 63454 個である．トレーニング用の全ツイートをコーパスとして（全単語数は 11308535 個，異なり語は 395491 個），word2vec [8] を用いて（window size は 5，イテレーション回数は 20），全ての異なり語を単語ベクトルの辞書に変換した．

各ツイートのベクトルは，そに含まれる単語のベクトルの平均値を使用した．単体のツイートに含まれる情報量は少ないため，複数のツイートを束にして学習させる方が精度が向上する可能性があると考えられる．そこで，複数のツイートを束にしたものをツイートブロックと呼び，そこに含まれる各ツイートをベクトル的に平均化したものツイートブロックのベクトルとして扱った．

2.2 個人属性の予測

以上の処理から得られたツイートブロックのベクトル値を入力として，機械学習を用いてモデルを構築し，個人属性の予測精度を調べる．使用した学習アルゴリズムは，Linear Support Vector Classification（Linear SVC），K-Neighbors，AdaBoost，Random Forest の 4 つと深層学習である．

2.3 個人属性の分析

個人属性の予測が成功する原因を調べるために，次のようなデータ分析を行う．ツイート・コーパスに存在している，男女が共通して使用する単語は，性別の予測するときに重要な役割を果たすと推測され，事情は他の個人属性も同様だと考えられる．ここでは，性別を例としてデータを分析し，この考えを検証する．トレーニング用のツイート・コーパスを性別ごとに分け，前節と同様の処理で単語の分散表現を作り，男女ごとに 2 つの単語ベクトルの辞書を作成した．男性の単語ベクトル辞書（単語数 237141）と女性の単語ベクトル辞書（単語数 241645）に基づき，共通単語の集合（単語数 83295）を抽出した．そして，男女の共通単語において，式 1のようにコサイン距離を測った．

$$S = 1 - \cos(\theta) = 1 - \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \left\|\vec{b}\right\|} = 1 - \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} b_i^2}} \tag{1}$$

ここで，S は共通単語のコサイン距離であり，a_i と b_i はそれぞれ，男性と女性の辞書に存在している単語ベクトルである．i は単語ベクトルの要素のインデックスであり，n は次元数である．

個人属性の予測には，男女で共通する単語の意味の違い（コサイン距離 S）だけでなく，トレーニング用のツイート・コーパスにおけるその単語の頻度 n にも関係しているため，式 2のように共通単語の重み W を求めた．

$$W = S \cdot n \tag{2}$$

予測精度への影響を調べるため，トレーニング用のコーパスから（単語数が 376577）から，重み W の大きい順に従って単語を取り除き，新しい単語ベクトル辞書を作成した．この辞書を用い，トレーニング用の 100 人とテスト用の 20 人のすべてのツイートをベクトル化した．また，非共通単語をランダムに取り除き，同様の処理をして，比較対象とした．

3　結果

3.1　個人属性の予測

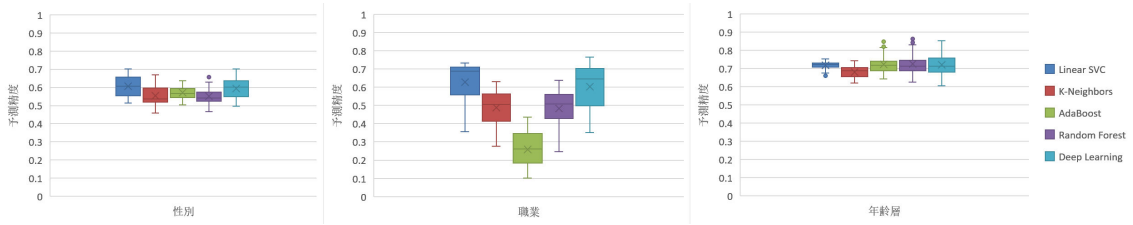

図1　各課題における予測精度の分布

図 1は，word2vec の埋め込み次元 N とツイートブロックのサイズ L の全ての組み合わせにおいて，3 つの個人属性の推定における精度の分布を示したものである．年齢層は，どの学習アルゴリズムにおいても平均して 70% 程度の予測精度を示していることから，ツイートのテキストのみから年齢層を推定することは，他の 2 つと比べて容易であることがわかる．また今回の実験では，Linear SVC と深層学習は他のアルゴリズムと比べて，安定して高い精度が得られることがわかった．

3.2　個人属性の項目ごとの予測

性別，職業，年齢層の 3 つの個人属性には，予測しやすい（しにくい）項目があると考えられる．例えば，政治家やタレントは他の職業よりも予測がしやすいかもしれない．それを調べたのが図 2で，深層学習（全ての N と L の組み合わせ）によって予測が成功した項目の内訳を示している．性別の場合は，男性より女性の方が予測しやすい．職業の場合は，経営者とアスリートは予測しやすく，作家とミュージシャンは予測しにくい．年齢層の場合は，デジタルネイティブよりデジタル移民の方が予測しやすいことがわかった．

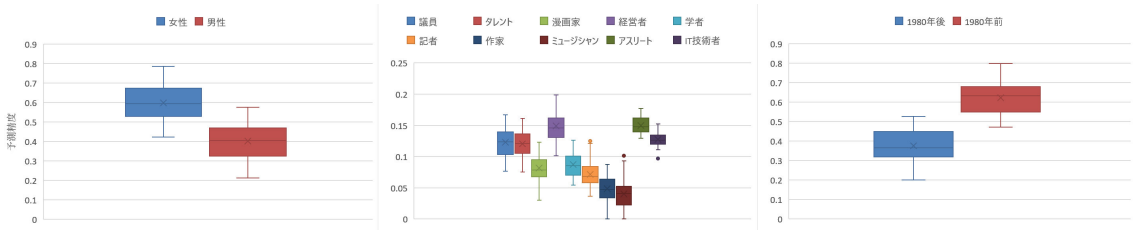

図2　個人属性における各項目の予測のしやすさ

3.3　男女の共通単語の性別予測に対する影響

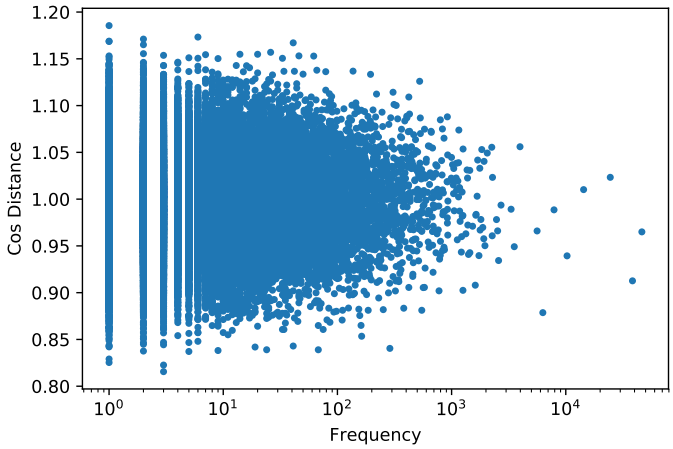

図3　男女の共通単語のコサイン距離と頻度の分布

図 3は男女で共通する単語のコサイン距離と頻度の分布図である．この分布から，これらの変数間には弱い負の相関が認められた（$R = -0.0097, p < 0.0221$）．

トレーニング用のツイート・コーパスから，重み W の大きい順に共通単語を 1 つずつ取り除いて，性別の予測精度を調べた．同様に，非共通単語をランダムに 1 つずつ取り除いて予測精度（10 回の施行の平均値）を比較した．その結果が図 4である．取り除かれた共通単語の個数を増やすにつれて，予測精度は最初の 70% から 55% に大幅に下がった．一方，比較対象の非共通単語の場合は，取り除かれた単語数を増やしても予測精度がずっと 68% 程度に安定している．

4　まとめ

本研究では，まず，ツイートのテキストのみから性別，職業，年齢層という 3 つの個人属性を予測する実験を機械学習を用いて行い、ベースラインの結果を得た．予測精度を比べると，年齢層（デジタルネイティブか否か）は，性別や職業よりもアルゴリズム的に推定しやすいことが示された．これは，言葉の選び方や使い方が年齢によって異なるからではないかと考えられる．テキストデータのみから，60 ～ 70% の精度で個人属性

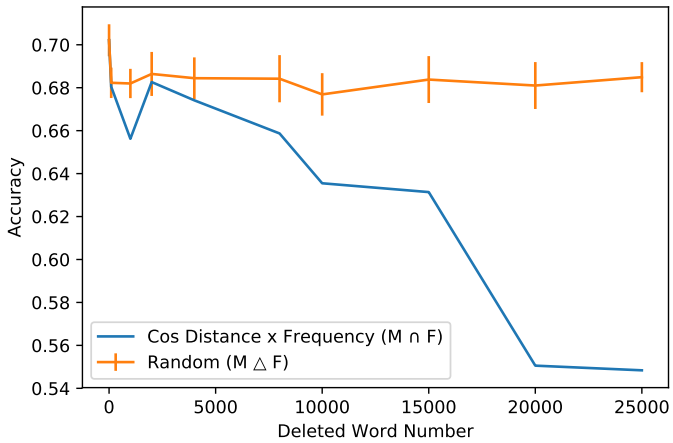

図4　男女の共通単語と非共通単語をの取り除いた際の予測精度への影響

を推定できるということは注目に値する．

ツイートから個人属性を予測する場合，単体のツイートよりも複数のツイートをまとめてブロックとして使用した方が精度が顕著に上がった．この結果は，ツイートの束をまとめてベクトル化したものが，個人属性の推定に必要な情報を比較的低次元で表現していることを示唆している．換言すると，テキストのみから個人属性を推定するためには，ある程度の量のデータが必要である．Twitter の場合は，60% 程度の予測精度を得るためには，50 個程度がその下限だということになる．

個人属性を予測できる理由を調べたところ，男女が共通して頻繁に使用する単語の中で意味が異なる単語は，性別の予測精度に大きい影響を与えることが示された．共通単語の場合は，取り除かれた単語数が増えていくにつれ，予測精度が著しく下がった．さらに，取り除かれた単語数は 15000 から 20000 の間にある時，最も予測精度が下がった．これは，一部の共通単語が他より予測精度に強い影響を与えていることを示唆している．非共通単語の場合は，取り除かれた単語数が増えても予測精度が大きく変動することがなかった．これらの結果によると，性別を予測する場合は，非共通単語と比べて共通単語が予測精度に大きく影響することが分かった．今後，性別以外の個人属性でも同様の調査をする予定である．

本研究の技術が確立すれば，ソーシャルデータから個人属性を精度良く推定することが可能になり，計算社会科学の分析やマーケティングへの応用など幅広い利用が期待できる．

謝辞

本研究は JSPS 科研費 (JP16K16112, JP15H03446, JP17H06383JST), JST さきがけ (JPMJPR16D6), JST CREST(JPMJCR17A4) の助成を受けたものです。

参考文献

[1] Lazer D, et al. (2009) Computational social science. *Science* 323(5915):721–723.

[2] Sasahara K, Hirata Y, Toyoda M, Kitsuregawa M, Aihara K (2013) Quantifying collective attention from tweet

stream. *PLoS ONE* 8(4):e61823.

[3] Takeichi Y, Sasahara K, Suzuki R, Arita T (2015) Concurrent Bursty Behavior of Social Sensors in Sporting Events. *PLoS ONE* 10(12):e0144646–13.

[4] Pennebaker JW, Chung CK, Ireland M, Gonzales A, Booth RJ (year?) The Development and Psychometric Properties of LIWC2007.

[5] Schwartz HA, et al. (2013) Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE* 8(9):e73791.

[6] Kosinski M, Stillwell D, Graepel T (2013) Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110(15):5802–5805.

[7] Liu X, Zhu T (2016) Deep learning for constructing microblog behavior representation to identify social media user' s personality. *PeerJ Computer Science* 2:e81.

[8] Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.