

Personalized PageRank を利用した網羅的 Twitter ユーザ属性推定

上里 和也[†] 浅井 洋樹[†] 山名 早人^{†‡}

[†] 早稲田大学大学院 基幹理工学研究科 〒169-8555 東京都新宿区大久保 3-4-1

[‡] 国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2

E-mail: {k.uesato, asai, yamana}@yama.info.waseda.ac.jp

あらまし Twitter をはじめとするソーシャル・ネットワーキング・サービス (SNS) のユーザ属性情報は、各ユーザの嗜好や特徴を示す情報として、ターゲティング広告などの分野で有用である。しかし、自身のプロフィールとして属性情報を登録しないユーザや、誤った情報を登録するユーザが存在するため、ユーザプロフィールから直接得られる正確な属性情報は限定的である。このような背景から、正しいユーザ属性を推定する研究が行われているが、既存研究では性別や国籍といった特定の種類の属性を対象としたパターン抽出により推定を行うため、推定可能な属性の種類が限定される。そこで本研究では、Personalized PageRank を利用し、推定対象ユーザと関連の深いユーザのプロフィール文から複数の単語を抽出することで、推定する属性の種類を限定することなく、多様な属性を推定する手法を提案する。12 名の学生を対象とした属性推定実験を行った結果、1 ユーザあたり最大 7.25 個の異なる属性を推定し、最高で 0.802 の MAP@30 で推定に成功しており、推定属性の多様性及び推定の正確性において、推定対象ユーザのツイートを利用するベースライン手法を上回ることに成功した。

キーワード Twitter, 属性推定, Personalized PageRank

1. はじめに

近年インターネット上では、ユーザごとに興味を引くやすい広告を選択し表示するターゲティング広告や、e-コマースサイトの商品推薦システムなど、各ユーザの嗜好に合わせたコンテンツを提供するシステムが広く利用されている。これらのシステムでは、インターネット上の様々なデータから対象ユーザの特徴を抽出し、その特徴を利用してユーザの嗜好に最も適合すると考えられるコンテンツを選択する。ここで利用されるユーザの特徴を表すデータとして、大規模なソーシャル・ネットワーキング・サービス (SNS) のユーザ属性情報は有用である。大規模な SNS の 1 つである Twitter^aは、2015 年 11 月の時点で 3 億を超えるアクティブユーザを有している^b。また、多くの SNS ではユーザが自身の属性情報をプロフィールとして登録することができ、これらの属性情報は億単位の大規模なユーザの特徴を直接的に示していると考えられる。

しかし、多くの SNS において属性情報の登録は必須ではないため、ユーザ本人のプロフィールから直接得られる属性情報は限られている。また、誤った属性情報を登録するユーザも存在し、居住地域の情報を公開している Twitter ユーザのうち、34%のユーザが誤った

情報を登録しているという報告もされている[1]。

以上のような背景から、SNS ユーザの正確な属性情報を推定する手法が研究されている。近年の研究では、推定対象ユーザの投稿内容などの情報に加えて、SNS 上で推定対象ユーザと交流を持つユーザの情報を利用して、推定に利用する情報を拡充することで、対象ユーザのみの情報を用いた時と比べてより高い精度での推定に成功している[2][3][5][6][7][8][9][10][11]。しかしこれらの手法では、位置情報[4][5][7][8][11]や国籍[3]など、予め定めた特定の種類の属性を対象としたパターン抽出によって推定を行っている。そのため、推定可能な属性の種類は予め想定されたものに限定されている。

そこで本研究では、Twitter ユーザを対象とした幅広い種類の属性の推定を目的として、推定対象属性の種類を限定せず、多様なユーザ属性を推定する手法を提案する。具体的には、フォローやメンションによって Twitter 上で交流を行うユーザ間では、一部で同一の属性を持つと仮定する。この仮定のもとで、ユーザをノード、ユーザ同士の相互交流関係を無向エッジとしてユーザグラフを構成する。ここで構成したユーザグラフに対して、ランダムウォークをベースとする Personalized PageRank[12]を適用することで、推定対象ユーザとその他のユーザの関連度を算出する。この関連度を利用して、他のユーザのプロフィール文から推定対象ユーザと関連の深い単語を複数抽出し、これを推定結果として出力する。ユーザ自身によって様々な属性情報について記述されているプロフィール文を利用することで、多様な属性に関連する単語を抽出する

^a Twitter, <https://twitter.com/> (2015 年 12 月 19 日アクセス)

^b Leading social networks worldwide as of November 2015, ranked by number of active users (in millions), Statista – The statistics portal, <http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/> (2015 年 12 月 19 日アクセス)

ことが可能である。

また、我々は本稿の提案手法をこれまでに一度提案しているが、5名の被験者を対象とした簡易的な推定実験の結果を示すにとどまっており、推定結果に対するより詳細な評価が課題となっていた[13]。これに対して、本稿では推定対象ユーザのツイートを利用するベースライン手法との推定結果の比較や、Personalized PageRankのパラメータと推定結果との関係の調査を通して、より詳細な評価・考察を行う。

本稿は次の構成をとる。まず2節では、これまでに提案されてきたSNSユーザの属性推定手法について説明し、3節では、我々が提案する網羅的SNSユーザ属性推定手法について説明する。続く4節で12人の被験者に対して行った属性推定実験の方法や結果について説明し、最後に5節で本稿をまとめる。

2. 既存の SNS ユーザ属性推定手法

本節では、これまでに提案されてきたSNSを対象としたユーザ属性推定手法について説明する。推定に利用する情報の種類に基づいて、1) 推定対象ユーザの情報のみを利用する手法、2) 他のユーザの情報を利用する手法の2つに既存手法を大別し、それぞれの手法について説明する。

2.1. 推定対象ユーザの情報のみを利用する手法

推定対象ユーザの情報のみを利用する手法[1][14][15][16][17]では、推定対象ユーザ自身が投稿したツイートの本文やプロフィール情報を利用し、推定対象ユーザの属性を推定する。例えばBurgerらは、Twitterユーザを対象とし、推定対象ユーザ自身の情報から性別の分類を行う手法を提案している[15]。それぞれのユーザが投稿したツイート及び、スクリーンネーム、フルネーム、プロフィール文から取得した文字n-gram、単語n-gramを特徴量として、線形分類器であるWinnowによって2値分類を行うことで、91.8%の正答率を得ている。

しかしこれらの手法では、推定対象ユーザがツイートをほとんど投稿しない場合や、プロフィール情報を公開しない場合には、推定に利用できる情報が不足し、正確な推定が困難であるという問題点が存在する。実際にGongらは、あるTwitterユーザのコミュニティでは、毎週平均24.4%のユーザがツイートを全く投稿していないことを報告している[2]。またLiらは、プロフィール情報として居住地域を公開しているTwitterユーザは全体の20%程度であると報告している[7]。さらに、34%のユーザが誤った位置情報をプロフィール情報として記載していることが報告されており[1]、推定対象ユーザのプロフィール情報を利用する手法では、このような誤った属性情報が推定精度低下の原因とな

り得る。

2.2. 他のユーザの情報を利用する手法

他のユーザの情報を利用する手法では、SNSにおける記事の購読機能やメッセージ機能によって推定対象ユーザと交流を行うユーザは、一部の属性を推定対象ユーザと共有していると仮定し、主に推定対象ユーザと交流を持つユーザの情報を利用して推定を行う[2][3][4][5][6][7][8][9][10][11]。

他のユーザの情報を利用する手法では、推定に利用できる情報を拡充することで、推定対象ユーザのみの情報を使う手法と比べ、より高精度な推定に成功している。特に、記事の投稿頻度が低いユーザや属性情報を公開しないユーザなどに対しても、他のユーザの情報によって推定に利用する情報量の不足を補うことが可能となり、推定精度の低下を軽減することができる。

Gongらは、ツイートの投稿頻度によってユーザ进行分类し、様々な投稿頻度のユーザの属性推定実験を行うことで、投稿頻度と属性推定精度の関係を調査している。具体的には、シンガポールのTwitterユーザを対象とし、結婚の有無、宗教、支持政党の3項目について、Naïve Bayesを利用した2値分類の推定実験を行っている。Naïve Bayesの入力テキストとしては、1) 推定対象ユーザ本人、2) 相互フォロー関係のユーザ、3) フォロワー、4) フォローしているユーザのツイートをそれぞれ利用し、推定精度を比較している。

推定実験を行った結果、投稿頻度に関わらず全ユーザを推定対象とした場合、どの項目の推定においても、他のユーザのツイートを利用する2)、3)、4)のいずれかの方法が、本人のツイートを利用する方法と比較して、高精度な推定が行えたと報告している。特に18週間に5件以下のツイートのみを投稿する非アクティブユーザを対象としたとき、本人のツイートのみを利用する方法における推定精度は、無作為に分類を行う場合と同程度にまで低下した。それに対して、他のユーザのツイートを利用する方法では、本人のツイートを利用した場合の平均2倍程度の精度での推定に成功している。

また、他のユーザの情報を利用する方法は、あるユーザの誤ったプロフィール情報を推定に利用してしまった場合にも、その他の複数ユーザの正しい情報を利用することができれば、推定精度への影響の軽減が期待できるという利点がある。

しかし、他のユーザの情報を利用する既存手法では、位置情報[4][5][7][8][11]や国籍[3]など、予め定めた特定の種類の属性を対象としたパターン抽出により、推定を行っている。そのため、推定可能な属性の種類が限定的であり、予め想定しない種類の属性についての推定は不可能であるという問題点がある。

3. 提案する網羅的ユーザ属性推定手法

本節では、本研究で提案する網羅的ユーザ属性推定手法について説明する。

3.1. 手法概要

提案手法では、高精度かつ非アクティブユーザや誤ったプロフィール情報に頑健な推定を行うため、2.2項で説明した手法と同様に、他のユーザの情報を利用して推定を行う。また、推定対象ユーザの様々な属性に関連する単語を他のユーザのプロフィール文から複数抽出するアプローチをとることで、推定対象属性の種類を予め定めることなく、多様な属性の推定を可能とし、既存研究における「推定可能な属性の種類が限定的である」という問題点を解決する。

提案手法の推定手順の概略を図1に示す。図1に示したように、提案手法の最終的な出力は、推定対象ユーザの属性情報に関連すると推定された単語のリストである。リスト中の単語は、推定対象ユーザとそれぞれの単語との関連度を示す「単語関連度スコア」によって順位付けされている。

我々が提案する単語関連度スコアでは、推定対象ユーザとより深い関係をもつと考えられるユーザのプロフィール文に頻繁に出現する単語に対して高いスコアが付与される。推定対象ユーザと他のユーザとの関係の深さは、ユーザをノード、ユーザ間の交流関係をエッジとしたユーザグラフに対して Personalized PageRank[12]を適用することによって算出する。Personalized PageRankの具体的な算出方法を3.2項、単語関連度スコアの算出方法を3.3項でそれぞれ説明する。

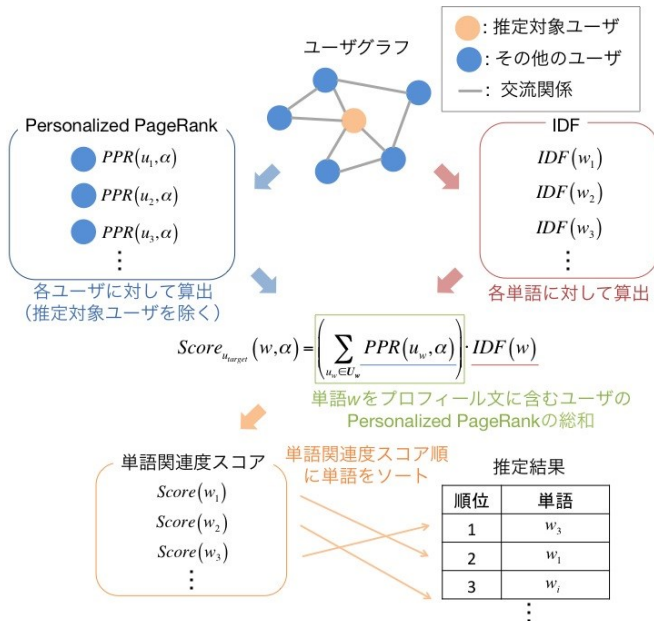


図1 提案手法概略

3.2. Personalized PageRank

Personalized PageRank[12]は、グラフデータを対象に、中心ノードとして定めた特定のノードとその他のノードとの関連度を算出するアルゴリズムである。具体的には、ランダムウォークによってエッジを辿ってノードを訪問し、それぞれのノードに対してその訪問回数に応じたスコアを付与する。中心ノードからランダムウォークを開始し、さらにランダムウォーク中に一定確率で中心ノードにジャンプすることで、中心ノードからより少ないホップ数で訪問可能なノードに対して高いスコアを付与する。提案手法では、推定対象ユーザとその他の各ユーザの関連度を定量化することを目的とし、ユーザグラフに対して Personalized PageRankを適用する。提案手法における各ユーザの Personalized PageRank 算出手順を以下に示す。

まず、ユーザ u をノード $u \in U$ 、ユーザ間の交流関係を無向エッジ $e \in E$ とするユーザグラフ $G = (U, E)$ を構成する。次に、構成したユーザグラフ G に対して、始点を推定対象ユーザ u_{target} としてランダムウォークを行い、それぞれのユーザの Personalized PageRank の値が収束するまで Personalized PageRank を更新していく。ただし、次の訪問ノードを決定する前に、 α の確率で推定対象ユーザ u_{target} にジャンプする（その他の場合はランダムウォークを続行する）。また、提案手法では推定対象ユーザ自身の Personalized PageRank を算出する必要がないため、ユーザ集合 $U' = U - \{u_{target}\}$ をランダムウォークの訪問対象とし、推定対象ユーザ u_{target} へジャンプする際は、訪問回数をカウントしない。

推定対象ユーザを u_{target} 、ランダムウォーク中の推定対象ユーザへのジャンプ確率を α 、総訪問回数が t であるときのユーザ $u' \in U'$ の訪問回数を $t_{u', \alpha}$ としたとき、ユーザ u' の Personalized PageRank $PPR_{u_{target}}(u', \alpha)$ は、式(1)で与えられる。

$$PPR_{u_{target}}(u', \alpha) = \frac{t_{u', \alpha}}{t} \quad (1)$$

また、総訪問回数が t の時点での各ユーザの Personalized PageRank を要素とするベクトルを $\mathbf{r}_t$ としたとき、 $\mathbf{r}_{t-1}$ との Personalized PageRank ベクトルの誤差 d_t は、式(2)にしたがって算出される。

$$d_t = \|\mathbf{r}_t - \mathbf{r}_{t-1}\|_1 \quad (2)$$

誤差の値が閾値 ϵ より小さくなったとき、Personalized PageRank が収束したと判断し、このときの Personalized PageRank を最終的な値として採用する。

3.3. 単語関連度スコア

単語関連度スコアは、推定対象ユーザの属性情報とその他のユーザのプロフィール文に含まれるそれぞれの単語との関連度を示す。具体的には、より Personalized PageRank が高いユーザのプロフィール文

に頻繁に出現し、かつ一般的に Twitter ユーザのプロフィール文には出現しにくい単語に対して高いスコアを付与する．単語 w をプロフィール文に含むユーザの集合を U_w ，ランダムウォーク中の推定対象ユーザ u_{target} へのジャンプ確率を α としたときのユーザ u の Personalized PageRank を $PPR_{u_{target}}(u, \alpha)$ ，ユーザグラフ G に含まれる全ユーザ U のプロフィール文を対象として算出した単語 w の IDF(Inverse Document Frequency) スコアを $IDF(w)$ としたとき，推定対象ユーザ u_{target} に対する単語 w の単語関連度スコア $Score_{u_{target}}(w, \alpha)$ は，式 (3) で与えられる．

$$Score_{u_{target}}(w, \alpha) = \sum_{u_w \in U_w} PPR_{u_{target}}(u_w, \alpha) \cdot IDF(w) \quad (3)$$

4. 属性推定実験

4.1. 実験・評価方法

提案手法の評価を行うため，12 名の学生を被験者とし，属性推定実験を行った．実験では，まずそれぞれの被験者のアカウントに対して属性推定を行い，単語関連度スコア上位 30 の単語を出力した．次に，被験者本人へのインタビューを行い，被験者本人が推定結果として出力された 30 の単語に対して，1) それぞれの単語が自身の属性情報に関連するかどうかを判定し，2) ここで関連すると判定した単語うち，同一の属性に関連する単語をグルーピングする．その後，1) の出力単語の正誤判定結果をもとに，単語順位の Mean Average Precision@30(MAP@30) を算出し，2) の単語グルーピングの結果として得られたグループの数を推定成功属性数としてカウントした．ここで得られた MAP@30 を，単語の順位付け及び推定の正確性，推定成功属性数を，推定属性の多様性の評価指標とする．

4.2. 推定対象ユーザ本人のツイートを利用するベースライン手法

本実験では，提案手法との比較対象として，推定対象ユーザ本人のツイートから単語抽出を行う手法についても同様の推定実験を行った．

ベースライン手法では，提案手法における単語関連度スコアの代わりとして，それぞれの単語に対して「DF-IDF スコア」を算出し，DF-IDF スコアの値で単語の順位付けを行う．推定対象ユーザ u_{target} の全ツイートの集合を $T_{u_{target}}$ とし，そのうち単語 w を含むツイートの集合を $T_{w, u_{target}} \subseteq T_{u_{target}}$ としたとき，単語 w の DF (Document Frequency) スコア $DF_{u_{target}}(w)$ を式 (4) に従って算出する．

$$DF_{tweet, u_{target}}(w) = \frac{|T_{w, u_{target}}|}{|T_{u_{target}}|} \quad (4)$$

また，ユーザグラフ内の全ユーザのツイートの集合を T とし，そのうち単語 w を含むツイートの集合を $T_w \subseteq T$ としたとき，単語 w の IDF スコア $IDF_{tweet}(w)$ は式 (5) で与えられる．

$$IDF_{tweet}(w) = \log \frac{|T|}{|T_w|} \quad (5)$$

最終的な単語 w の DF-IDF スコア $DFIDF_{tweet, u_{target}}(w)$ は式 (6) で与えられる．

$$DFIDF_{u_{target}}(w) = DF_{u_{target}}(w) \cdot IDF_{tweet}(w) \quad (6)$$

4.3. データセット

本実験では，7,957,324 ユーザのプロフィール文，フォロー情報，最新の最大 2,000 ツイートのデータを収集し，利用している．また，それぞれのユーザのツイートを解析することでメンションの情報についても取得，利用している．収集したユーザはいずれも 2013 年 1 月～2013 年 12 月にツイートを投稿した日本語設定のユーザである．プロフィール文等のデータも，全て 2013 年当時のものであるため，実験では，2013 年時点の被験者の属性情報を推定しており，被験者へのインタビューにおいても，2013 年時点の被験者の属性情報をもとに評価を行っている．

実験で利用したデータの収集は，次のように行った．

- I. Twitter Streaming API^cを利用して最新ツイートを取得
- II. I で得られたツイートを投稿したユーザのデータを Twitter REST API^dを利用して取得
- III. II で取得したユーザの言語設定が日本語であれば，プロフィール文，フォロー情報，最大 2,000 ツイートをデータベースにストア (総ツイート数が 2,000 に満たないユーザは，全ツイートを取得)

4.4. 実験条件

本実験では，4.3 項の方法で収集した 7,957,324 ユーザをノードとして，エッジの種類が異なる二種類のユーザグラフを用意し，それぞれを利用して推定を行った．1 つ目は，相互フォローの関係をもつユーザ同士をエッジで接続する「フォロログラフ」，2 つ目は，相互にメンションを 1 件以上投稿しているユーザ間をエッジで接続する「メンショングラフ」である．どちらのグラフにおいてもエッジは無向であり，重み付けを行わない単純エッジである．本実験で利用したフォロログラフ，及びメンショングラフにおける 1 ユーザあたりの隣接ノード数の統計を表 1 に示す．

^c Twitter Streaming API,

<https://dev.twitter.com/streaming/overview> (2016 年 1 月 4 日アクセス)

^d Twitter REST APIs,

<https://dev.twitter.com/rest/public> (2016 年 1 月 4 日アクセス)

表 1 隣接ノード数の統計

グラフの種類	隣接ノード数		
	平均値	中央値	標準偏差
フォローグラフ	119.5	46.0	655.3
メンショングラフ	22.4	12.0	29.4

Personalized PageRank におけるランダムウォーク中の推定対象ユーザへのジャンプ確率 α は、0.05, 0.20, 0.50, 0.80, 1.00 の 5 つの値を利用してそれぞれ推定を行った。パラメータ α が大きい値である場合には、ランダムウォーク中に頻繁に推定対象ユーザにジャンプする。そのため、ランダムウォークが推定対象ユーザからより少ないホップ数で到達可能なユーザのみに訪問することとなる。

提案手法では、ランダムウォークが一度以上訪問したユーザのプロフィール文が単語抽出の対象となるため、この場合、より推定対象ユーザに近いユーザのプロフィール文のみから単語を抽出することになる。逆に、パラメータ α が小さな値であれば、ランダムウォークがより広範囲のユーザに到達可能となり、より多くのユーザのプロフィール文から単語抽出を行うことになる。なお、Personalized PageRank の収束判定基準となる ε の値は、実験的に 10^{-4} と定めた。

次に、単語抽出の際に行う前処理や形態素解析ツールについて説明する。本実験では、形態素解析ツールとして lucene-gosen 4.10.0^eを利用している。形態素解析時に利用している辞書は、naist-jdic 0.4.3^fである。また本実験では、上記の辞書において「名詞-一般」、「名詞-固有名詞」の 2 種類の品詞が付与されている単語のみを利用している。

また、提案手法において IDF スコアを算出する際は、全ユーザから 100,000 ユーザをランダムサンプリングし、これらのユーザのプロフィール文を対象に計算を行っている。同様に、ベースライン手法では、10,000 人のツイートを対象とし、IDF スコアを算出している。

4.5. 実験結果・評価

提案手法及びベースライン手法を適用したときの平均推定成功属性数を表 2 に、MAP@30 を表 3 に示す。提案手法においては、Personalized PageRank のパラメータ α の値を異なる 5 つの値に変化させ、推定を行った際の結果をそれぞれ示している。

^e GitHub – lucene-gosen,
<https://github.com/lucene-gosen/lucene-gosen> (2016 年 1 月 4 日アクセス)

^f OSDN – NAIST Japanese Dictionary,
<https://osdn.jp/projects/naist-jdic/releases/31880> (2016 年 1 月 4 日アクセス)

表 2 平均推定成功属性数

α		0.05	0.20	0.50	0.80	1.00
提案手法	フォローグラフ	7.08	6.75	6.00	6.25	6.25
	メンショングラフ	7.25	6.00	5.75	5.67	5.58
ベースライン		4.83				

表 3 MAP@30

α		0.05	0.20	0.50	0.80	1.00
提案手法	フォローグラフ	0.599	0.769	0.802	0.794	0.794
	メンショングラフ	0.705	0.723	0.671	0.674	0.697
ベースライン		0.488				

表 4 出力単語リスト及び評価結果の例

出力単語リスト		評価結果	
順位	単語	正誤	属性の種類
1	情報	○	専門分野・大学
2	音楽	○	趣味
3	相互	×	—
4	バンド	○	趣味
5	アカウント	×	—
6	アニメ	×	—
7	趣味	×	—
8	ギター	○	趣味
9	東京	○	生活地域
10	ライブ	○	趣味

また、推定結果として出力された単語リスト及び被験者インタビューの結果の例を表 4 に示す。

ベースライン手法との比較

提案手法では、いずれのユーザグラフを利用した場合にも、 α の値に関わらず推定成功属性数及び MAP@30 がベースライン手法を上回っている。プロフィール文及びツイート本文の例をそれぞれ図 2、図 3 に示す。(アカウントの特定を避けるため、一部を“[省略]”という記号で置換する。)

湊生まれ[省略]/高専出身/[省略]ソフトテニス、バスケット/miwa [省略]
[省略]Web デザインをしています。[省略]Web アプリ [省略]作っています。
茨城出身[省略]都内在住[省略]です。[省略]東京[省略]茨城[省略]

図 2 プロフィール文の例

[省略]腹減りすぎて寝れてない
[省略]節約する。[省略]広島行く。
[省略]リビングに兄が転がってる
[省略]コロッケ食べたい
[省略]バスに乗らないから、たまに乗ると良い[省略]

図 3 ツイート本文の例

表 5 属性の種類ごとの推定成功被験者数

	ユーザ グラフ	α	大学	専門分野	研究室	学年	生活地域	居住地域	出身校	部活動	サークル	趣味	職業
提案手法	フォロー グラフ	0.05	10	7	0	1	6	7	5	3	6	12	1
		0.20	11	8	1	2	4	5	4	3	6	11	3
		0.50	10	8	1	2	3	4	5	3	4	10	3
		0.80	10	8	1	1	4	4	5	3	5	11	3
		1.00	10	8	1	1	4	4	6	3	4	11	2
		平均	10.2	7.8	0.8	1.4	4.2	4.8	5.0	3.0	5.0	11.0	2.4
	メンション グラフ	0.05	11	8	0	1	7	6	7	2	6	12	3
		0.20	10	7	0	1	4	4	7	2	5	11	4
		0.50	10	7	0	1	3	4	7	2	5	11	3
		0.80	10	7	0	1	4	4	6	1	5	10	4
		1.00	10	7	0	2	5	4	6	1	5	9	2
		平均	10.2	7.2	0.0	1.2	4.6	4.4	6.6	1.6	5.2	10.6	3.2
	提案手法全体の平均		10.2	7.5	0.4	1.3	4.4	4.6	5.8	2.3	5.1	10.8	2.8
	ベースライン手法		6	6	0	1	1	1	5	2	3	11	3

(ベースライン手法を 2.0 以上上回っている提案手法全体の平均推定成功被験者数を太字で表記)

プロフィール文は、ユーザ自身が様々な属性に関する情報を公開する目的で記述されるため、図 2 に示した例のように、多様な属性に関連する単語が含まれている上に、属性情報と関連しない話題が記述されることが少ない。それに対して、ツイートは主に日常の出来事や自らの考えを公開する目的で記述される。そのため、図 3 の例のように、属性情報と関連のない話題や単語がより多く含まれており、ツイート内で属性情報に関連する単語が利用されていても、これを正確に抽出することは困難である。このような理由から、提案手法では複数のユーザのプロフィール文を単語抽出の対象とすることで、本人のツイートを利用するベースライン手法と比較して、より多様な属性に関連する単語を正確に抽出できていると考えられる。

また、今回の実験を通して推定することができた属性の種類を手作業で分類すると、職業、所属大学、専門分野、所属研究室、学年、出身校、活動地域、居住地域、部活動、所属サークル、趣味の 11 種類であった。提案手法及びベースライン手法によって推定することができた属性の種類を手作業で分類し、それぞれの種類の属性の推定に成功した被験者数を表 5 に示す。表 5 より、ベースライン手法と比較して、提案手法では、所属大学、所属サークル、生活地域、及び居住地域の 4 種類の属性において、平均で 2.0 人以上多くの被験者に対する推定に成功した。このことから、提案手法では複数ユーザのプロフィール文を利用することによって、属性情報以外の多様なトピックを含んでいるツイート本文では埋もれてしまいやすい種類の属性に対して、より正確な抽出を可能にしていると考えられる。

パラメータ α と推定結果の関係

Personalized PageRank のパラメータ α は、小さな値であるほど、ユーザグラフにおいて推定対象ユーザからの距離が遠いユーザにランダムウォークが到達し、より広範囲のユーザのプロフィール文が単語抽出の対象

となる。

表 2 から、フォローグラフ、メンショングラフのどちらのユーザグラフを利用した場合も、 $\alpha=0.05$, 0.20 の α の値が小さい場合に、推定成功属性数が大きくなっており、多様な属性が推定できていることがわかる。 α の値が小さい場合には、より広範囲のユーザのプロフィール文から単語を抽出することができるため、より多様な属性に関連する単語の抽出が行えていると考えられる。

一方で表 3 からは、フォローグラフを利用したとき、 α の値が小さい場合には MAP@30 の値が小さくなっていることがわかる。原因としては、 α の値が小さい場合には、ユーザグラフにおいて推定対象ユーザからの距離がより大きいユーザのプロフィール文からも単語抽出を行うため、実際には推定対象ユーザと属性を共有していないユーザのプロフィール文から単語を抽出してしまっていることが考えられる。なお、メンショングラフを利用した場合の MAP@30 と α の値との関係は、12 名の被験者に対する実験の結果からは読み取ることが困難であるため、より大規模な被験者に対して実験を行った上で考察を行うことが今後の課題となる。

フォローグラフとメンショングラフの比較

表 2, 表 3 から、 $\alpha=0.05$ の時を除き、推定成功属性数及び MAP@30 のどちらにおいても、フォローグラフを利用した場合、メンショングラフを利用した場合と比較して高い評価を得ている。表 1 に示した通り、フォローグラフは、メンショングラフより多くのエッジを張っている。そのため、ランダムウォークが、より少ないホップ数で多くのユーザに訪問することが可能となり、より多様な属性に関連する単語を抽出することが可能となっていると考えられる。

また、フォローグラフを利用したときに、メンショングラフを利用したときと比較して MAP@30 の値が高くなっていることから、フォローグラフでは、より

正確に属性を共有しているユーザ同士をエッジで接続していることが考えられる。そのため、属性を共有しているユーザ同士では、相互メンションに比べ、相互フォローの関係もつ可能性が高いことが予想される。

また、各被験者に対して算出した、フォローグラフを利用した際に正しく抽出することができた単語の集合とメンショングラフを利用した際に正しく抽出することができた単語の集合との Jaccard 係数の平均は 0.422 であった。同様に、フォローグラフを利用した際に推定に成功した属性の種類(表 5 に示した手作業による分類)の集合とメンショングラフを利用した際の推定属性の種類(集合)との Jaccard 係数の平均は 0.834 であった。抽出単語の Jaccard 係数の値から、抽出に成功した単語全体の半数程度が、フォローグラフ、メンショングラフのどちらかを利用した場合にのみ抽出されていることがわかる。それに対して、推定に成功した属性の種類においては、全体の 80%以上がどちらのユーザグラフを利用した場合にも正しく推定されている。これらのことから、フォローグラフを利用する場合とメンショングラフを利用する場合では、抽出される単語は異なるが、推定属性の種類には比較的大きな差異がないと考えられる。

5. おわりに

本稿では、推定対象ユーザ以外のユーザのプロフィール文から推定対象ユーザの属性情報に関連する単語を複数抽出することで、既存の SNS を対象としたユーザ属性推定手法の「推定属性の種類が限定的である」という問題点を解決し、多様な属性を推定する手法を提案した。

12 名の学生を被験者として属性推定実験を行った結果、提案手法では、推定対象ユーザ本人のツイートを利用するベースライン手法と比較して、より多様な属性をより正確に推定することができた。さらに、実験結果から、提案手法の中で利用している **Personalized PageRank** のパラメータ α の値を変化させることで、推定属性の多様性と推定の正確性のトレードオフをコントロールすることができると考えられる。

また、今回の実験では、推定属性の多様性及び単語関連度スコアに基づく出力単語順位の正確性について評価を行った。しかし、同じ属性に関連する単語であっても、それぞれの単語がどれほど詳細な属性情報を示す単語であるかは、評価できていない。例えば、推定対象ユーザが早稲田大学の基幹理工学部(工学部)に所属する学生であれば、「早稲田大学」、「基幹理工」は、ともに通っている大学という同一の属性に関連する単語であるが、「基幹理工」は「早稲田大学」と比較して、当該ユーザのより詳細な所属情報を示している。詳細な属

性情報の推定が行えることで、推薦システムなどにおいてより詳細にそれぞれのユーザの嗜好を捉えられることが期待される。そのため、推定属性の詳細度の評価方法を考案し、多様かつ詳細な属性推定を行うことが今後の課題である。

謝辞

本研究は JSPS 科研費 25280113 の助成を受けたものである。

参考文献

- [1] Hecht, B., Hong, L., Suh, B., and Chi, E. H.: “Tweets from Justin Bieber’s heart: The dynamic of the location field of user profiles.” In Proc. of the CHI 2011, pp. 237-246, 2011.
- [2] Gong W., Lim E. P., and Zhu, F.: “Characterizing silent users in social media communities.” In Proc. of the 9th ICWSM, pp. 140-149, 2015.
- [3] Huang, W., Weber, I., and Vieweg, S.: “Inferring nationalities of twitter users and studying inter-national linking.” In Proc. of the 25th ACM Conference on Hypertext and Social Media, pp. 237-242, 2014.
- [4] Jurgens, D., Finethy, T., McCorriston, J., Xu, Y. T., and Ruths, D.: “Geolocation prediction in twitter using social networks: A critical analysis and review of current practice.” In Proc. of 9th ICWSM, pp. 188-197, 2015.
- [5] Jurgens, D.: “That’s what friends are for: Inferring location in online social media platforms based on social relationships.” In Proc. of the 7th ICWSM, pp. 273-282, 2013.
- [6] Li, R., Wang, C., and Chang, K. C. C.: “User profiling in an ego network: Co-profiling attributes and relationships.” In Proc. of the WWW 2014, pp. 819-829, 2014.
- [7] Li, R., Wang, S., Deng, H., Wang, R., and Chang, K. C. C.: “Towards social user profiling: Unified and discriminative influence model for inferring home locations.” In Proc. of the KDD 2012, pp. 1023-1031, 2012.
- [8] McGee, J., Caverlee, J., and Cheng, Z.: “Location prediction in social media based on tie strength.” In Proc. of the CIKM 2013, pp. 459-467, 2013.
- [9] Mislove, A., Viswanath, B., Gummadi, K. P., and Druschel, P.: “You are who you know: Inferring user profiles in online social networks.” In Proc. of the WSDM 2010, pp. 251-260, 2010.
- [10] Pennacchiotti, M., and Popescu, A. M.: “Democrats, republicans and starbucks aficionados: User classification in twitter.” In Proc. of the KDD 2011, pp. 430-438, 2011.
- [11] Rout, D., Bontcheva, K., PreoŃiuc-Pietro, D., and Cohn, T.: “Where’s @wally? A classification approach to geolocating users based on their social ties.” In Proc. of the 24th ACM Conference on Hypertext and Social Media, pp. 11-20, 2013.
- [12] Page, L., Brin, S., Motwani, R., and Winograd, T.: “The PageRank citation ranking: Bringing order to the Web.” Tech. Rep. 1999-66, Stanford University, 1998.

- [13] Uesato, K., Asai, H., and Yamana, H.: "Predicting various types of user attributes in twitter by using personalized PageRank." In Proc. of the IEEE International Conference on Big Data 2015, 2015.
- [14] Alowibdi, J. S., Buy, U. A., and Yu. P.: "Language independent gender classification on twitter." In Proc. of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 739-743, 2013.
- [15] Burger, J. D., Henderson, J., Kim, G., and Zerrella, G.: "Discriminating gender on twitter." In Proc. of the EMNLP 2011, pp. 1301-1309, 2011.
- [16] Li, C., Sun, A.: "Fine-grained location extraction from tweets with temporal awareness." In Proc. of the SIGIR 2014, pp. 43-52, 2014.
- [17] Zhao, Z., Cheng, Z., Hong, L., and Chi, E. H.: "Improving user topic interest profiles by behavior factorization." In Proc. of the WWW 2015, pp. 1406-1416, 2015.