ソーシャルグラフによる居住地推定のための ユーザプロフィール分析

Analysis of User Profile for Home Location Estimation from Online Social Graph

廣中 詩織
Shiori Hironaka

豊橋技術科学大学 情報・知能工学系
Department of Computer Science and Engineering, Toyohashi University of Technology
sl43369@edu.tut.ac.jp, <http://www.ss.cs.tut.ac.jp/~hironaka15/>

吉田 光男
Mitsuo Yoshida

(同上)
yoshida@cs.tut.ac.jp, <http://www.cs.tut.ac.jp/~yoshida/>

梅村 恭司
Kyoji Umemura

(同上)
umemura@tut.jp, <http://www.ss.cs.tut.ac.jp/umemura/>

keywords: social graph, home location estimation, user profile, Twitter.

Summary

Users' attributes, such as home location, are necessary for various applications, such as news recommendations and event detections. However, most real user attributes (e.g., home location) are not open to the public. Therefore, their attributes are estimated by relationships between users. A social graph constructed from relationships between users can help estimate home locations, but it is difficult to collect many relationships, such as followers' relationships. We focus on users whose home locations are difficult to estimate, so that we can select users whose locations can be accurately estimated before collecting relationships. In this paper, we use their profiles which can be collected before collecting relationships. Then, we analyze difficult users with their profiles. As a result, we found that users whose home locations incorrectly estimated had a longer duration since the date their account was created, longer name, and longer description. In addition, the results indicated that the users whose home locations were incorrectly estimated differed from those whose home locations could not be estimated.

1. はじめに

ソーシャルメディアは友人との交流や様々な情報を得るために、多くの人々に使われているツールである。ソーシャルメディアに投稿されたデータは人々の嗜好や社会の様子を表していると考えられており、様々なアプリケーションで活用されている。例えば、マーケティングのためのトレンド検出 [Benhardus 13] やユーザ体験向上のためのニュース推薦 [Jonnalagedda 13, Phelan 09], 疾患の流行観測 [Signorini 11], 現実世界で起きたイベントの検出 [Sakaki 10] が挙げられる。このようなアプリケーションには、居住地や年齢、職業などのユーザ属性が用いられる。居住地などの現実世界に関わるユーザ属性は重要であるものの、そのような属性は提供されていないことが多く、ユーザ属性の推定に関する研究開発が盛んに行われている。本研究ではユーザの居住地を推定する問題に取り組む。

ユーザの居住地を推定するために、ユーザ間の関係を表したソーシャルグラフを手がかりとする方法がある [Ju-

rgens 15, Zheng 18]. この方法は、ソーシャルグラフ上でつながっているユーザ同士の地理的な距離が近いという仮定を用いる。しかし、フォロー関係などのユーザ間の関係は収集が難しいという問題がある [Vesdapunt 16]. 本研究では、居住地を正しく推定しにくいユーザを特定する問題を考える。これによって、そのユーザらをソーシャルグラフの収集対象から除外することで、効率よくデータを収集し、かつ高精度に居住地を推定することを意図する。

我々は、店の公式アカウントやアーティストのアカウントなど、居住地を推定しにくいユーザが存在することに着目する。これまでの研究では、多くのユーザと会話（リプライ）しているユーザは有名人であり、現実の友人ではないユーザとのつながりが多いため、推定の前にソーシャルグラフから取り除く処理が行われていた [Rahimi 15]. 我々は、何らかの理由で他にも居住地を推定しにくいユーザが存在すると考え、推定が難しいユーザをソーシャルグラフ収集前に判別することで、効率的な収集ができると考えた。そこで、ソーシャルグラフ収集前に居

住地の推定可能性を予測するため、プロフィールに注目する。名前や自己紹介文など、プロフィールはユーザの行動により公開されているものであるため、プロフィールにはユーザの特徴が現れる。そして、この特徴は推定のしやすさに関連があると考えられる。加えて、ユーザのプロフィールはフォロー関係やユーザ間の会話に比べて収集しやすく、ソーシャルグラフとは別にユーザごとに取得できるため、ソーシャルグラフを収集する前にユーザを選択する目的に適していると考ええる。

本論文では、ユーザのプロフィールをもとに居住地推定の対象を絞り込むことで、居住地を推定しにくいユーザの特徴を調べる。実験の結果、ユーザのアカウント作成日が古いユーザほど居住地を正しく推定しにくいこと、名前や自己紹介文が長いユーザほど居住地を正しく推定しにくいことなどを明らかにした。

2. 関 連 研 究

居住地推定手法は、推定に使う情報の違いによって、主にコンテンツベースとネットワークベースとに分けられる [Zheng 18]。コンテンツベースの手法ではツイートの内容を用いてユーザの居住地を推定する一方、ネットワークベースの手法ではユーザ同士の関係を用いてユーザの居住地を推定する。本研究の対象ではないため、コンテンツベースの手法には言及しない。

本研究で注目するネットワークベースの手法では、地理的近接性を示すユーザ同士の関係をもとに構築したソーシャルグラフが推定に用いられる。直接の関係を持つユーザらの既知な居住地を用いる手法には、ある地理的距離の相手との関係が存在する確率をモデリングする手法 [Backstrom 10] や、関わりがあるユーザの居住地の中から最もよく出現するものを選ぶ手法 [Davis Jr. 11] がある。McGee らは、同じフォロワーを持つユーザや相互フォローのユーザなど関係の強いユーザが存在することを考慮し、関係の強さを Backstrom らのモデルに組み込んだ [McGee 13]。他にも、位置情報を投稿しているユーザはわずかであることから、ネットワーク上にあるラベル（居住地）の付与されていないユーザを活用し、ラベルを伝搬させ居住地を推定する手法も提案されている [Compton 14, Jurgens 13]。このように、ネットワークベースの手法では、ユーザ同士の関係は地理的近接性を持つという仮定を主に用いている。

ネットワークベースの手法などで使われるソーシャルグラフの構築には、Twitter のリプライ関係 [Jurgens 13] やフォロー関係 [Davis Jr. 11]、Facebook の友人関係 [Backstrom 10] などのユーザ同士の関係が使われている。Twitter のリプライ関係はフォロー関係に比べて多くのユーザにまたがる関係が取得できるものの、すべてのリプライを取得することは困難なため、データはサンプリングされたものとなる。一方、フォロー関係を用いる場合は、ユー

ザ同士のつながりが完全に取得できる。しかし、ユーザ数を増やしてネットワークの規模を大きくすることは API 制限のため困難である [Vesdapunt 16]。このような状況下では、フォロー関係などの関連するデータの収集を、居住地の推定しやすいユーザから順に収集する必要が現実的に生じる。

ソーシャルグラフを用いる居住地推定では関係の地理的近接性を仮定しているが、すべてのエッジが同じような地理的近接性を有しているとは考えられていない。Twitter でのフォローには友人関係と購読関係とがある [Kwak 10] といわれており、ソーシャルグラフのエッジの種類を分類する研究 [Li 14] が行われている。Barbieri らは説明付きリンク予測に取り組んでおり、リンクは social か topical かであるとしている [Barbieri 14]。廣中らはフォローをもとにした居住地推定のための有向ソーシャルグラフ構築には、フォローされている関係が最も適していると述べている [廣中 17]。居住地推定に有効な関係を選ぶことができれば居住地推定の性能を向上させることができると考えられるが、ソーシャルグラフの規模は大きく、エッジごとにフィルタリングをすることは現実的でない。

ソーシャルグラフに対するユーザごとのフィルタリングは、ソーシャルグラフを用いる居住地推定において一般的に用いられている。Rahimi らや Miura らなどは、リプライ関係をもとに構築したソーシャルグラフを用いる際に、多くのユーザとリプライ関係にあるユーザを *celebrity* として事前に取り除いている [Miura 17, Rahimi 18]。さらに、Ebrahimi らは、*celebrity* の中には *global celebrity* と *local celebrity* がいるとして、位置情報付きツイートを投稿した場所をクラスタリングすることで、*celebrity* とされていたユーザを *global celebrity* と *local celebrity* とに分類し、*local celebrity* は推定に役立つことを明らかにした [Ebrahimi 18]。Davis Jr. らも、相互フォローをもとに構築したソーシャルグラフを用いて推定する際、相互フォロー数によってユーザを絞り込むことで適合率が向上することを示している [Davis Jr. 11]。多くのユーザとリプライ関係にあるユーザの一部や、相互フォロー数が少ないユーザなどの居住地が推定しにくいことはわかっているが、リプライ関係をもとにしたソーシャルグラフでのリプライ関係にあるユーザ数と、相互フォロー関係をもとにしたソーシャルグラフでの相互フォロー数は、ともにそれぞれのソーシャルグラフにおける次数である。各ユーザの属性として得られるフォロー数やフォロワー数以外の次数は、ソーシャルグラフ全体を収集した後でなければ計算することができないものであるため、ソーシャルグラフ収集のコストを減らすために利用することはできない。ソーシャルグラフの収集対象とするユーザを選択するために、ソーシャルグラフ収集に取りかかる前に手に入るプロフィール情報を用いて居住地を推定しにくいユーザの特徴を調べるところが、本論文の特徴である。

3. データ

本論文では、Twitter ユーザを対象に、ソーシャルグラフを用いた居住地推定が困難なユーザの特徴を分析する。まず、居住地を推定する際に用いるソーシャルグラフと各ノード（ユーザ）のラベル（居住地）のデータを用意する。ソーシャルグラフと居住地データの作成には先行研究[廣中 17]と同様の方法を用いる。さらに、ユーザの特徴を調べるために、プロフィールをもとにした各ユーザの属性データを用意する。

3.1 居住地データ

ユーザは主に居住地周辺で活動していると考え、投稿された位置情報付きツイートをもとにして各ユーザに居住地を付与する。居住地の付与は次の手順で行う。まず、ユーザに居住地を付与するために、Twitter の Streaming API^{*1}を利用して日本を包含する矩形^{*2}内での位置情報付きツイートを収集する。このうち、地理座標情報 (coordinates) が付与されているツイートのみを利用し、付与されていなかったツイートは除外する。また、Bot アカウントによるツイートの影響を減らすために、森國ら[森國 15]と同様の方法で、位置情報付きツイート集合から Bot アカウントによるツイートを取り除く。その後、総務省統計局による平成 22 年度国勢調査の境界データを用いて、ツイートを付与されている地理座標を含んでいる日本の市区町村（エリア）を照合する。このとき、エリアと照合できなかったツイート（日本国外からのツイート）は除外する。そして、ユーザごとに最も多くのツイートを投稿しているエリアを居住地として付与する。

2014 年 1 月 1 日から 12 月 31 日のあいだの位置情報付きツイートを Streaming API により 140,055,452 件集めた。投稿回数が極端に少ないユーザを除外するために、同じエリアで 5 回以上投稿しているユーザに絞り込み、ユーザごとに最も多くのツイートを投稿しているエリアを居住地として付与した。その結果、610,891 ユーザに居住地を付与できた。

3.2 ソーシャルグラフデータ

ソーシャルグラフを構築するために、居住地を付与したユーザらのフォロー関係を用いる。居住地を付与したユーザらそれぞれがフォローしているユーザ集合とフォローされているユーザ集合とを 2015 年 7 月に収集した。これらのデータを用いて、次の手順でソーシャルグラフを構築する。まず、居住地を付与したユーザ集合から、フォローしているユーザ集合、フォローされているユーザ集合、または 3.3 節で述べるプロフィール情報のうち、1 つ

以上が取得できなかったユーザを除外する。そして、残ったユーザ集合をソーシャルグラフのノードとする。次に、ソーシャルグラフのノードとなるユーザ同士が相互フォローであるときにエッジを作る。できあがったソーシャルグラフは無向グラフとなる。

最終的に、471,761 ノードと 3,112,137 エッジを含むソーシャルグラフができた。ソーシャルグラフのすべてのノードには居住地が付与されている。総務省統計局の境界データでは日本の市区町村は 1,901 種類あり、作成したデータには 1,873 種類が出現していた。ソーシャルグラフには、エッジを 1 つも持たないユーザ（孤立ノード）が 82,677 ユーザ存在する。本研究ではソーシャルグラフの情報のみを使う居住地推定手法を用いるため、これらユーザの居住地は推定することができない。推定できないユーザも居住地推定の対象とするが、分析の際にこれらユーザを考慮する。

3.3 ユーザの属性データ

ユーザの特徴を調べるため、ユーザの属性データを用意する。ユーザの特徴として、ユーザのプロフィール^{*3}から得られる値を用いる。ソーシャルグラフ構築のためにフォロー関係を収集したときと同時期の 2015 年 7 月に、居住地を付与したユーザらのプロフィール情報を取得した。プロフィールをもとにした属性は、ユーザ名 (screen.name) の文字数、名前 (name) の文字数、ロケーションの文字数、自己紹介文の文字数、フォロー数、フォロワー数、いいね数、公開リストに入れられている数、総ツイート数である。日本語の文字も英数字もそれぞれ 1 文字とカウントした。

その他の属性として、ユーザのプロフィールの値から計算した、アカウント作成日からの日数、1 日あたりのツイート数、フォロー／フォロワー比がある。アカウント作成日から 2015 年 7 月 1 日までの日数をアカウント作成日からの日数とする。1 日あたりのツイート数は、総ツイート数をアカウント作成日からの日数で割ったものとする。フォロー／フォロワー比は $\text{フォロワー数} / (\text{フォロー数} + 1)$ と定義する。以上、プロフィールから有効な可能性のあるものを広く調査対象とした。

4. 実験設定

居住地を推定しにくいユーザの特徴を明らかにするため、ユーザの属性値をもとに居住地推定の対象とするユーザを選択し、そのときの居住地推定の性能を分析する。居住地を付与した 471,761 ユーザを対象に、付与されているラベルを 1 つずつ隠して推定を繰り返す leave-one-out 交差検証により性能を調査する。

^{*1} <https://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter.html> (viewed 2019-05-13)

^{*2} 北緯 20 から 50, 東経 110 から 160 の範囲。

^{*3} <https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/user-object.html> (viewed 2019-05-13)

4.1 居住地推定手法

居住地推定は、無向ソーシャルグラフ $G(V, E)$ とラベルの付与されたユーザの集合 M を使って、居住地のわからない（ラベルの付いていない）ユーザ u の居住地 $\hat{l}_u = \text{infer}(V, E, M)$ を推定するタスクとする。 V はユーザ集合であり、 E はエッジ集合である。

本論文では、ネットワークベースの居住地推定手法のうち隣接ノードの情報のみを使う手法を実験に用いる。すなわち、 $\text{infer}(V, E, M)$ の代わりに V と E から計算できるあるユーザ u の隣接ノード集合 $N_u = \{v | (u, v) \in E\}$ を利用し、 $\text{infer}(N_u, M)$ とする。ソーシャルグラフの最もローカルな情報だけを用いて居住地推定をする手法は、フォロー関係をもとにしたソーシャルグラフなど比較的低密度が高い場合には十分有効に機能することがわかっていて [廣中 17]。

本論文では、[Jurgens 15] で良好な性能を示していた [Davis Jr. 11] を推定手法として用いる。この手法は、推定対象ノードの隣接ノードが持つラベルの中で最もよく現れるラベルを選択し推定値とする手法である。この手法を表現したものが式 (1) である。ここで、 $\arg \max^*$ は同値の集合を返すものと定義する。本論文では、最もよく現れるラベルが複数ある場合は、学習データ中での出現頻度が高いものを選択する。

$$S_u = \arg \max_{l \in \{l_n | n \in N_u\}} |\{v | v \in N_u, l = l_v\}|$$

$$\text{Infer}(N_u, M) = \arg \max_{l \in S_u} |\{n | n \in M, l = l_n\}| \quad (1)$$

Davis Jr. ら [Jurgens 15] は、居住地を推定するユーザを選択するためのパラメータとして、最小友人数、最大友人数、最小投票数を用いている。最小友人数と最大友人数は、相互フォロー数（Davis Jr. らが用いたソーシャルグラフの次数）の最小値と最大値であり、最小投票数は $|N_u \cap M| = |N_u|$ である。本論文では、プロフィールから得られるユーザの属性データを用いて推定が難しいユーザの特徴を調べるため、最小友人数と最大友人数は考慮しない。最小投票数は 1 として、推定できるすべてのユーザのラベルを推定する。

ソーシャルグラフを用いる多くの居住地推定手法では、関係のあるユーザは近くに住んでいるという仮定を用いている。そのため本論文の分析結果は、同様の仮定をしている、ソーシャルグラフを用いる他の推定手法にも参考となる結果であると考えられる。

4.2 評価方法

評価には、推定したユーザのうちラベルを正しく割り当てることができた割合である適合率（Precision）と、すべてのユーザのうちラベルを正しく割り当てることができた割合である再現率（Recall）、すべてのユーザのうち居住地を推定したユーザの割合であるカバー率（Coverage）

を用いる。適合率とカバー率の積が再現率になる。それぞれの算出式を次に示す。

$$\text{Precision} = \frac{\text{正しく推定できたユーザ数}}{\text{居住地を推定できたユーザ数}}$$

$$\text{Recall} = \frac{\text{正しく推定できたユーザ数}}{\text{居住地を付与したユーザ数}}$$

$$\text{Coverage} = \frac{\text{居住地を推定できたユーザ数}}{\text{居住地を付与したユーザ数}}$$

ここで、居住地を推定できたユーザとは、居住地の判明している友人が 1 人以上いるユーザの数であり、正しく推定できたユーザ数とは、居住地を推定できたユーザのうち、正解ラベルと同じ居住地を推定できたユーザの数である。居住地を付与したユーザ数は 471,761 ユーザである。

4.3 推定対象ユーザの選択方法

ユーザのある属性の値により、そのユーザを推定対象とするかを選択する。推定対象のユーザを選ぶ条件として、ある属性値がしきい値より小さいユーザを除外する LowCut と、ある属性値がしきい値より大きいユーザを除去する HighCut との 2 種類のフィルタ条件を用意する。しきい値を θ としたときのユーザ集合 V に対するフィルタは次の式で表される。ここで、 $a(u)$ はあるユーザ u の属性値を返す関数である。 $a(\cdot)$ はフォロー数を返す関数や、フォロワー数を返す関数などになる。

$$\text{LowCut}(V; a) = \{u | u \in V, a(u) > \theta\}$$

$$\text{HighCut}(V; a) = \{u | u \in V, a(u) < \theta\}$$

しきい値には、各属性値の取りうる値のおおよそ最小値から最大値までの範囲を 200 個に分割した値を用いる。フォロー数、フォロワー数、いいね数、総ツイート数、リストに入れられている数、1 日あたりのツイート数、フォロー／フォロワー比は、対数に変換したあと等間隔に分割する。

5. 結果・考察

ユーザのプロフィールから得た属性を用いてユーザを選択したときの推定性能を調べることで、居住地推定をしにくいユーザの特徴を分析する。まず、プロフィールから得た属性によってそのようなユーザが判別できるかを調べ、その後どのようなユーザの居住地が推定しにくいのかを明らかにする。

5.1 居住地推定に寄与するユーザ属性

各ユーザの属性値に対して様々なしきい値を設定したフィルタを適用して推定対象のユーザを絞り込み、居住地を正しく推定できるユーザの割合を調べることで、居住地を推定しにくいユーザをプロフィールから得た属性

表 1 適合率が最大になったときのしきい値と性能

属性	フィルタ	θ	Precision	Recall	Coverage	n
フォロワー数	LowCut	190	<u>0.3416</u>	0.1738	0.5089	240094
	HighCut	612	<u>0.3230</u>	0.2265	0.7012	330810
フォロワー数	LowCut	210	<u>0.3537</u>	0.1621	0.4582	216170
	HighCut	830	<u>0.3189</u>	0.2345	0.7355	346960
総ツイート数	LowCut	1	0.3116	0.2568	0.8242	388807
	HighCut	5478	<u>0.3350</u>	0.1745	0.5210	245805
1日あたりのツイート数	LowCut	1.804	<u>0.3209</u>	0.1783	0.5556	262130
	HighCut	13.509	<u>0.3238</u>	0.2203	0.6802	320899
公開リストに入れている数	LowCut	35	0.2458	0.0130	0.0529	24936
	HighCut	1	<u>0.3888</u>	0.1232	0.3169	149507
いいね数	LowCut	450	<u>0.3444</u>	0.1268	0.3682	173694
	HighCut	7752	<u>0.3157</u>	0.2440	0.7728	364574
フォロワー／フォロワー比	LowCut	1.02	<u>0.3692</u>	0.1197	0.3241	152913
	HighCut	2.35	<u>0.3140</u>	0.2481	0.7901	372726
ユーザ名の文字数	LowCut	8	<u>0.3144</u>	0.1746	0.5552	261906
	HighCut	14	0.3115	0.2301	0.7386	348457
名前の文字数	LowCut	0	0.3117	0.2571	0.8247	389077
	HighCut	6	<u>0.3461</u>	0.1557	0.4499	212264
ロケーションの文字数	LowCut	15	0.2856	0.0165	0.0579	27336
	HighCut	1	<u>0.3673</u>	0.1593	0.4337	204600
自己紹介文の文字数	LowCut	0	0.3116	0.2279	0.7314	345035
	HighCut	28	<u>0.3592</u>	0.1245	0.3466	163519
アカウント作成日からの日数	LowCut	176	0.3117	0.2571	0.8247	389084
	HighCut	498	<u>0.4011</u>	0.0522	0.1301	61398
フィルタなし			0.3117	0.2571	0.8247	389084

で判別できるかを調べる．それぞれの属性を用いたフィルタを適用して居住地推定を行った結果のうち、カバー率が 0.05 以上のとき *4 に適合率が最大になったしきい値と、そのときの適合率、再現率、カバー率を表 1 に示す．推定対象のユーザの選択には HighCut と LowCut との 2 種類の条件を用いた．表中の n は推定対象のユーザ数であり、下線はフィルタなしの場合より推定精度が改善したことを意味する．

推定精度が改善したかどうかは、次の手順で判定した．推定対象の n ユーザのうち居住地を正しく推定できたユーザ数を x とするとき、適合率を $p = x/n$ と表す． n ユーザを無作為に取り出すことを繰り返すとき、適合率の分布は正規分布 $N(p, \sqrt{p(1-p)/n})$ で近似される． n ユーザをフィルタによって選択し x ユーザの居住地を正しく推定できたときの信頼区間を求め、フィルタなしの場合 ($n = 389084$) の信頼区間と比べて、区間が重ならず適合率が上回ったとき、推定精度が改善したと判定する．

表 1 のフィルタを適用したすべての結果をフィルタを適用しなかった場合の結果と比較すると、すべての属性

で HighCut と LowCut とのどちらかまたは両方で適合率が向上した．適合率が最も高くなったのは、アカウント作成日からの日数が 498 日より長いユーザを除外したときであった．すなわち、アカウント作成日からの日数が 498 日より長いユーザは居住地を推定しにくいということがわかった．また、アカウントを作成したばかりのユーザを除外しても適合率は向上せず、LowCut はフィルタなしと同じ結果となった．フォロワー数、フォロワー数、いいね数、1日あたりのツイート数、フォロワー／フォロワー比では、値が大きいユーザを除外する場合でも小さいユーザを除外する場合でも適合率が向上した．このことから、値が大きすぎるユーザも小さすぎるユーザも居住地を正しく推定しにくいことが示唆される．総ツイート数、公開リストに入れている数、名前の文字数、ロケーションに入力している文字数、自己紹介文の文字数が多いユーザ、ユーザ名の文字数が少ないユーザを除外した場合も適合率が向上した．

次に、しきい値を変えたときの適合率とカバー率の変化をみる．表 1 で最も適合率が高かったアカウント作成日からの日数を用いて、しきい値より高いユーザを推定対象から除外したときの結果を図 1(a) に示す．図におけ

*4 5%のユーザは外れ値となる値を持っていると考えて除外した．

(a) アカウント作成日からの日数が長いユーザを除外するとき

(b) 公開リストに入れられている数が高いユーザを除外するとき

(c) フォロー／フォロワー比が低いユーザを除外するとき

図 1 しきい値による適合率の変化

る x 軸に水平な破線は、全ユーザを推定対象とする、フィルタなしの場合の適合率である。しきい値を 2000 あたりから小さくしていくほど適合率が高くなり、498 日以上のユーザを推定対象から除外するとき適合率が最大になる。Twitter を利用し始めてからの日数が多いユーザほどソーシャルメディア上で知り合ったユーザとの関係や古い関係が増えるため、そのような地理的距離を反映しない関係が推定精度に影響を与えていると解釈できる。

表 1 で適合率が 2 番目に高かった、リストに入れられている数を用いてしきい値より高いユーザを推定対象か

ら除外した HighCut のときの、しきい値を変えた結果を図 1(b) に示す。しきい値を上げていくほど適合率が下がっていくため、リストに入れられている数が多いほど居住地を正しく推定しにくいことがわかる。Twitter で公開リストを利用するユーザは、タイムラインとは別のタイムラインがほしいと考え、ツイートを見たいユーザをリストに追加する。リストに入れられるということは友人以外のユーザからも有益な情報を得られるアカウントであると認識されているため、友人以外のアカウントとつながっている可能性が高くなり、居住地を正しく推定しにくくなったと解釈できる。

表 1 で 3 番目に適合率が高かったフォロー／フォロワー比が低いユーザを除外するときのしきい値による結果の変化を図 1(c) に示す。まず、しきい値を 0.01 から 1.02 まで大きくしていくほど適合率が上がる。その後はしきい値が約 10 になるまで適合率が低下していき、その後再び適合率が上昇する。傾きに着目したときに適合率が上昇するポイントが複数あるため、フォロー比により居住地を正しく推定できるかを判別するためには、HighCut と LowCut 単体のフィルタでは不十分であると考えられる。

5.2 居住地推定が困難なユーザ

前節では、居住地を正しく推定できなかったユーザを居住地を推定しにくいユーザとみなして、適合率を用いて居住地を推定しにくいユーザを判別できる属性について分析した。居住地を正しく推定できなかったユーザには、居住地を誤って推定してしまったユーザと、そもそも推定のための関係が得られず居住地を推定できなかったユーザとが含まれている。本節では、推定しにくいユーザをさらに 2 種類に分けて分析する。

分析では、まず、居住地推定の結果をもとにユーザ集合を分割し、各ユーザ集合に対して属性値の分布を計算する。そして、ユーザ全体に対しての各ユーザ集合の偏り度合いを計算する。ユーザを次の 3 種類に分けてそれらのユーザ集合ごとに属性値の分布を調べる：正しく居住地を推定できたユーザ (easy)、誤って居住地を推定したユーザ (hard)、そもそも居住地が推定できなかったユーザ (unknown)。これらはすなわち、多数の友人と同じ居住地を持つユーザ、多数の友人と同じ居住地を持たないユーザ、居住地の判明している友人が存在せず推定の手がかりがないユーザとなり、推定の難しさを表すクラスとなる。これら 3 種類のユーザ集合に含まれるユーザ数をそれぞれ A , B , C とするとき、適合率と再現率は次のように計算される値と等しい。

$$\text{Precision} = \frac{A}{(A+B)}$$

$$\text{Recall} = \frac{A}{(A+B+C)}$$

ユーザ集合 $U \subseteq V$ に対するある属性値の分布は次のように計算する。ユーザ総数を $|U|$, そのユーザの属性値

がある区間 $i: [x_i, x_{i+1})$ に含まれているユーザの数を n_i とする。そのとき、ユーザ全体のうちある属性値が区間 i の中に存在しているユーザの割合は $n_i/|U|$ である。縦軸をユーザの属性値がその区間に存在する割合、横軸を区間としてプロットした、 $f(i; U) = n_i/|U|$ を属性分布とする。 $\sum_i n_i/|U|$ が 1 となるように区間を決めた。しきい値を変化させたときと同じように、フォロワー数、フォロワー数、いいね数、総ツイート数、リストに入れられている数、1 日あたりのツイート数、フォロワー／フォロワー比の属性では、区間は等間隔でなく対数スケールとした。

ユーザ全体 (V) のうち区間 i に存在する割合を $f(i; V)$ 、比較するユーザ集合 U のうち区間 i に存在する割合を $f(i; U)$ とするとき、全体の分布に対する偏り度合いを $\log_{10} \frac{f(i; U)}{f(i; V)}$ と計算する。ユーザ全体の分布に対して、その区間にいるユーザの割合が多いとき正の値になり、少ないとき負の値になる。比較した分布間で差が大きいほど値の絶対値が大きくなる。

5.1 節でのフィルタなしの場合の推定結果を用いて、ユーザを 3 種類のグループに分けた。その結果、正しく居住地を推定できたのは 121,275 ユーザ、誤って居住地を推定したのは 267,809 ユーザ、そもそも居住地を推定できなかったのは 82,677 ユーザであった。

ユーザ属性ごとに、ソーシャルグラフに含まれるユーザ全体の集合、ソーシャルグラフに含まれるユーザのうち居住地を正しく推定できたユーザの集合、誤って推定したユーザの集合、居住地を推定することができなかったユーザの集合の属性分布を計算した。そして、計算した属性分布を用いて、居住地を正しく推定できたユーザの集合、誤って推定したユーザの集合、居住地を推定することができなかったユーザの集合の、全体の分布に対する偏り度合いの分布を計算した。ここでは紙面の都合上、表 1 で適合率が向上した属性のうち、わかりやすい傾向の結果が得られた、アカウント作成日からの日数、名前の文字数、自己紹介文の文字数の結果を図 2 に示す。図 2(a)、図 2(c)、図 2(e) はこれら分布を重ねてプロットしたものである。ソーシャルグラフに含まれるユーザ全体の集合の分布は棒グラフでプロットしてある。偏り度合いの分布をプロットしたものが図 2(b)、図 2(d)、図 2(f) である。

図 2(a) と図 2(b) がアカウント作成日からの日数の結果である。まず図 2(a) をみると、アカウント作成日からの日数が 500 から 900 のユーザが多いとわかる。次に図 2(b) をみる。ユーザ全体に対する誤って居住地を推定するユーザの割合は、その他 2 つのユーザ集合に影響されず、アカウント作成日からの日数が多くなるほど増えていく。ユーザ全体に対する正しく居住地を推定したユーザの割合は、アカウント作成日からの日数が多くなるほど減っている。居住地推定が正解するユーザの割合は、アカウント作成日からの日数が 500 から中央値に近

づく 900 あたりで多くなっている。さらに、アカウント作成日からの日数が小さいユーザには居住地を推定できないユーザが多いことがわかる。日数が増えるほどその割合は減っていくが、アカウント作成日からの日数が約 1400 日のところでまた増える。アカウント作成日からの日数の平均値は 1091 日、中央値は 939 日であった。

図 2(c) と図 2(e) はユーザの名前と自己紹介文の長さの分布を示す。ユーザ集合全体で見ると、名前の文字数が 4 文字であるユーザが最も多い。自己紹介文の文字数は 10 文字あたりにピークがあり、文字数が増えるほどユーザの割合が減っていき、最大の 160 文字でまたピークとなる。図 2(d) をみると、ユーザの名前の長さが中央値である 5 のとき、ユーザ全体のうち正しく居住地を推定できるユーザの割合が最も多くなっていること、名前が短すぎても長すぎても居住地を推定できないユーザの割合が増えることがわかる。居住地を誤って推定するユーザの割合は、他の 2 つのユーザ集合に影響されず、名前の文字数が長くなるほど多くなる。図 2(f) に示す自己紹介文の結果では、文字数が長くなるほど正しく居住地を推定できるユーザと有効な居住地を推定できないユーザの割合が減り、誤って居住地を推定するユーザの割合が増える。特に、名前の文字数または自己紹介文の文字数が 1 文字増加するごとに、居住地を誤って推定するユーザのユーザ全体に対する割合が一定の割合で増加することがわかった。名前の文字数と自己紹介文の文字数の偏り度合いの結果を見ても、アカウント作成日からの日数と同様に、ユーザが多く分布している部分で正しく居住地を推定できるユーザの割合が増えていた。この傾向は、ユーザ名の文字数とロケーションの文字数以外の属性でも確認できた。

5.3 考察と限界

5.1 節では、フォロワー数やフォロワー数、名前の文字数、自己紹介文の文字数などで居住地の推定しやすさが判別できることがわかった。5.2 節では、アカウント作成日からの日数や名前の文字数、自己紹介文の文字数によって、居住地を正しく推定できるユーザの割合や、誤って推定するユーザの割合が変化することがわかった。これらの属性が性能に影響した理由として、推定手法によるものと、ユーザのソーシャルメディア利用方法によるものとがあると考えられる。

ソーシャルグラフの構造に直接関係のあるプロフィール情報、具体的にはフォロワー数とフォロワー数からは、以下の知見が得られた。5.1 節では、フォロワー関係をもとにしたソーシャルグラフを用いて居住地推定をするときには、フォロワー数やフォロワー数が少なすぎるまたは多すぎるユーザは居住地を正しく推定しにくい、特にフォロワー数とフォロワー数が少ないユーザは居住地を正しく推定しにくいという結果を得た。フォロワー／フォロワー比を用いたときは、フォロワー数に比べてフォロワー数が

(a) アカウント作成日からの日数 (分布)

(b) アカウント作成日からの日数 (偏り度合い)

(c) 名前の文字数 (分布)

(d) 名前の文字数 (偏り度合い)

(e) 自己紹介文の文字数 (分布)

(f) 自己紹介文の文字数 (偏り度合い)

図 2 ユーザプロフィールの値の分布

相対的に多いユーザである値が小さいユーザを除外することで、適合率が大きく向上した。[Davis Jr. 11] では相互フォローをもとに構築したソーシャルグラフを推定に用いたとき、相互フォロー数によってユーザを制限することで適合率が向上することが報告されていた。また、[Rahimi 15] らは相互プライをもとにソーシャルグラフを構築したとき、相互プライ数でユーザを制限する前処理を行っていた。すべての関係を収集できた完全なソーシャルグラフの次数であるフォロー数とフォロワー数によってユーザを制限することで適合率を向上させら

れることが本論文でも示された。

フォロー・フォロワー関係は時間の経過によって変化する [Hironaka 18]。5.1 節では、アカウント作成日からの日数が長くなるほど推定が難しくなるという結果が得られた。このことから、Twitter を長く利用するほど地理的近接性を示さないフォロー関係が蓄積していき、居住地が推定しにくくなっていることが示唆される。また、ユーザがアカウントを作成したときにはフォロー数とフォロワー数は共に 0 であるため、アカウント作成日からの日数が短いときはフォロー数とフォロワー数が少なく居住

地を推定しにくいと考えられた。しかし、アカウント作成日からの日数が少ないユーザを除外しても推定精度が上がりなかった。したがって、居住地を正しく推定できるユーザは、アカウントを作成してすぐに友人と相互フォローになっていると考えられる。

ソーシャルグラフの構造に直接関係しないプロフィール情報からは、以下の知見が得られた。5.2節では、名前の文字数や自己紹介文の文字数が居住地を正しく推定できるかを判別するのに有用な手がかりとなりそうであるという結果を得た。先行研究にはユーザ名や名前から性別が予測できる [Burger 11] という研究があり、一定のユーザはユーザ名や名前に本名を入力していると考えられる。さらに、一部のユーザは、名前フィールドに名前以外の情報を入力していることもわかっている [Shima 17]。Shima らはユーザが名前や自己紹介文、ロケーションに入力している情報を変更する行動について分析し、他ユーザへのお知らせなど名前以外の情報を名前フィールドへ入力するユーザが存在すると報告している。我々は、自己紹介文が長いユーザも同様に、自分がどのようなユーザであるかが他のユーザへ伝わるように説明文を書いており、Twitter 上で知り合うことを考慮していると考えられる。そのため、自己紹介文や名前の長さがソーシャルメディアをどれだけ多くの現実の友人と一緒に使っているかを表す手がかりとなり、居住地推定の難しさに影響したと解釈できる。

本論文で得られた結果を使えば、居住地を推定しやすいユーザを事前にプロフィール情報のみから選ぶことができ、それらユーザのソーシャルグラフを収集し、居住地を推定することができる。本論文は日本の Twitter ユーザのデータを用いて日本国内で活動しているユーザを対象に分析している。そのため、本論文で得られたユーザの特徴は、他の国のユーザには適用できない可能性がある。先行研究に位置情報を投稿するユーザの特徴を分析した研究があり [Sloan 15]、日本のユーザは他の国のユーザに比べて位置情報サービスを有効にしている割合が低く、また位置情報付きツイートの割合も低いことがわかっている。これらから、日本のユーザは他の国のユーザとはプライバシー意識が異なり、ソーシャルメディアの使い方が異なることが示唆される。なお、得られた分析結果は居住地推定手法に依存するものの、本論文で用いた Davis Jr. らの推定手法 [Davis Jr. 11] は隣接ノードの持つラベルの中から最頻値を選ぶというシンプルなものであるため、その他のソーシャルグラフを使う推定手法を用いる場合にも通用する結果であると考えられる。

ユーザ属性の中には相関のあるものが含まれていると考えられる。そこで、各ユーザ属性間のスピアマンの順位相関係数^{*5}を計算する。ユーザ属性の値よりユーザを並べ替え、その並びから相関係数を計算した結果を図3

図3 ユーザ属性間のスピアマンの順位相関係数

に示す。最も高い相関係数は、フォロー数とフォロワー数とのあいだの約 0.89 であった。5.2節で詳述した属性である、名前の文字数、自己紹介文の文字数、アカウント作成日からの日数とのあいだの相関係数はそれぞれ 0.5 未満と小さいものであり、結果の解釈へは影響しないと考えられる。

6. ま と め

本論文では、ソーシャルグラフ収集の前に入手できる情報として、ユーザのプロフィールから得られる属性情報でユーザを選択したときに、居住地推定の適合率が上がることを明らかにした。さらに、居住地を正しく推定できたユーザ、居住地を誤って推定したユーザ、有効な居住地を推定できなかったユーザの持つユーザ属性の分布を調査した。その結果、アカウント作成日からの日数により居住地を推定しやすいユーザを選択できることがわかった。加えて、アカウント作成日からの日数が長いほど居住地を誤って推定するユーザの割合が多いこと、ユーザ名の文字数が多いほど居住地を誤って推定するユーザの割合が多いこと、自己紹介文の長いほど居住地を誤って推定するユーザの割合が多いことなども明らかになった。本論文では、ユーザのプロフィールから計算できる属性を分析に用いた。プロフィールはソーシャルグラフを構築する前に取得することができるため、本論文で得られた結果を用いることで、居住地を正しく推定できる可能性が高いユーザだけに絞ってデータを収集することができる。

◇ 参 考 文 献 ◇

- [Backstrom 10] Backstrom, L., Sun, E., and Marlow, C.: Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity, in *Proceedings of the 19th International Conference on*

^{*5} フォロワー数などはずれ値を含む属性があるため、スピアマンの順位相関係数を選択した。

- World Wide Web*, pp. 61–70 (2010)
- [Barbieri 14] Barbieri, N., Bonchi, F., and Manco, G.: Who to Follow and Why: Link Prediction with Explanations, in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1266–1275 (2014)
- [Benhardus 13] Benhardus, J. and Kalita, J.: Streaming Trend Detection in Twitter, *International Journal of Web Based Communities*, Vol. 9, No. 1, pp. 122–139 (2013)
- [Burger 11] Burger, J. D., Henderson, J., Kim, G., and Zarrella, G.: Discriminating Gender on Twitter, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1301–1309 (2011)
- [Compton 14] Compton, R., Jurgens, D., and Allen, D.: Geotagging One Hundred Million Twitter Accounts with Total Variation Minimization, in *Proceedings of the 2014 IEEE International Conference on Big Data*, pp. 393–401 (2014)
- [Davis Jr. 11] Davis Jr., C. A., Pappa, G. L., Oliveira, de D. R. R., and de L. Arcanjo, F.: Inferring the Location of Twitter Messages Based on User Relationships, *Transactions in GIS*, Vol. 15, No. 6, pp. 735–751 (2011)
- [Ebrahimi 18] Ebrahimi, M., ShafieiBavani, E., Wong, R., and Chen, F.: Twitter User Geolocation by Filtering of Highly Mentioned Users, *Journal of the Association for Information Science and Technology*, Vol. 69, No. 7, pp. 879–889 (2018)
- [廣中 17] 廣中 詩織, 吉田 光男, 岡部 正幸, 梅村 恭司: 日本における居住地推定に利用するためのフォロー関係の調査, 人工知能学会論文誌, Vol. 32, No. 1, pp. VII–M.1–11 (2017)
- [Hironaka 18] Hironaka, S., Yoshida, M., and Umemura, K.: Temporal Analysis of Online Social Graph by Home Location, in *Proceedings of the ACM IUI 2018 Workshop on Web Intelligence and Interaction* (2018)
- [Jonnalagedda 13] Jonnalagedda, N. and Gauch, S.: Personalized News Recommendation Using Twitter, in *Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies*, pp. 21–25 (2013)
- [Jurgens 13] Jurgens, D.: That's What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships, in *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, pp. 273–282 (2013)
- [Jurgens 15] 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 *Proceedings of the 9th International AAAI Conference on Web and Social Media*, pp. 188–197 (2015)
- [Kwak 10] Kwak, H., Lee, C., Park, H., and Moon, S.: What is Twitter, a Social Network or a News Media?, in *Proceedings of the 19th International Conference on World Wide Web*, pp. 591–600 (2010)
- [Li 14] Li, R., Wang, C., and Chang, K. C.-C.: User Profiling in an Ego Network: Co-profiling Attributes and Relationships, in *Proceedings of the 23rd International Conference on World Wide Web*, pp. 819–830 (2014)
- [McGee 13] McGee, J., Caverlee, J., and Cheng, Z.: Location Prediction in Social Media Based on Tie Strength, in *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, pp. 459–468 (2013)
- [Miura 17] Miura, Y., Taniguchi, M., Taniguchi, T., and Ohkuma, T.: Unifying Text, Metadata, and User Network Representations with a Neural Network for Geolocation Prediction, in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1260–1272 (2017)
- [森國 15] 森國 泰平, 吉田 光男, 岡部 正幸, 梅村 恭司: ツイート投稿位置推定のための単語フィルタリング手法, 情報処理学会論文誌 データベース, Vol. 8, No. 4, pp. 16–26 (2015)
- [Phelan 09] Phelan, O., McCarthy, K., and Smyth, B.: Using Twitter to Recommend Real-Time Topical News, in *Proceedings of the 3rd ACM Conference on Recommender Systems*, pp. 385–388 (2009)
- [Rahimi 15] Rahimi, A., Cohn, T., and Baldwin, T.: Twitter User Geolocation Using a Unified Text and Network Prediction Model, in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp. 630–636 (2015)
- [Rahimi 18] Rahimi, A., Cohn, T., and Baldwin, T.: Semi-supervised User Geolocation via Graph Convolutional Networks, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2009–2019 (2018)
- [Sakaki 10] Sakaki, T., Okazaki, M., and Matsuo, Y.: Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, in *Proceedings of the 19th International Conference on World Wide Web*, pp. 851–860 (2010)
- [Shima 17] Shima, J., Yoshida, M., and Umemura, K.: When Do Users Change Their Profile Information on Twitter?, in *Proceedings of the 2017 IEEE International Conference on Big Data*, pp. 3119–3122 (2017)
- [Signorini 11] Signorini, A., Segre, A. M., and Polgreen, P. M.: The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic, *PLoS ONE*, Vol. 6, No. 5, p. e19467 (2011)
- [Sloan 15] Sloan, L. and Morgan, J.: Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter, *PLOS ONE*, Vol. 10, No. 11, p. e0142209 (2015)
- [Vesdapunt 16] Vesdapunt, N. and Garcia-Molina, H.: Updating an Existing Social Graph Snapshot via a Limited API, in *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, pp. 1693–1702 (2016)
- [Zheng 18] Zheng, X., Han, J., and Sun, A.: A Survey of Location Prediction on Twitter, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 30, No. 9, pp. 1652–1671 (2018)

〔担当委員：岡 瑞 起〕

2019 年 7 月 3 日 受理

著 者 紹 介

廣中 詩織(学生会員)

2016 年豊橋技術科学大学工学部情報・知能工学課程卒業。2018 年同大学院工学研究科情報・知能工学専攻博士前期課程修了。同年、同博士後期課程進学。

吉田 光男(正会員)

2014 年筑波大学大学院システム情報工学研究科コンピュータサイエンス専攻博士後期課程修了。博士(工学)。同年、豊橋技術科学大学大学院工学研究科(情報・知能工学系)助教。ウェブ工学, 自然言語処理, 計算社会科学に関する研究に従事。言語処理学会, 情報処理学会, 日本データベース学会の各会員。

梅村 恭司(正会員)

1983 年東京大学大学院工学系研究科情報工学専攻修士課程修了。博士(工学)。同年、日本電信電話公社電気通信研究所入所。1995 年豊橋技術科学大学工学部情報工学系助教授, 2003 年同教授。自然言語処理, システムプログラム, 記号処理に関する研究に従事。情報処理学会, 情報電子通信学会, 日本ソフトウェア科学会, 言語処理学会, 計量国語学会, 日本教育工学会, ACM の各会員。