特集論文 「Web インテリジェンスとインタラクション 2017」

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

Analysis of Social Network Generation Methods for Home Location Estimation in Japan

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
s143369@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/

岡部 正幸 Masayuki Okabe
県立広島大学 経営情報学部 経営情報学科
Department of Management Information Systems, Prefectural University of Hiroshima
okabe@pu-hiroshima.ac.jp

梅村 恭司 Kyoji Umemura
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
umemura@tut.jp, http://www.ss.cs.tut.ac.jp/umemura/

keywords: home location estimation, social network, Twitter

Summary

The home locations of Twitter users can be estimated using a social network, which is generated by various relationships between users. There are many network-based location estimation methods with user relationships. However, the estimation accuracy of various methods and relationships is unclear. In this study, we estimate the users' home locations using four network-based location estimation methods on four types of social networks in Japan. We have obtained two results. (1) In the location estimation methods, the method that selects the most frequent location among the friends of the user shows the highest precision and recall. (2) In the four types of social networks, the relationship of follower has the highest precision and recall.

1. は じ め に

ソーシャルメディアのデータは，一般的に，投稿と投稿したユーザとが関連付けられている．この特徴をもとに，ソーシャルメディアの分析および研究開発には居住地などのユーザの属性が利用される [奥村 12]．しかし，自身のプロフィールに居住地を入力しているユーザは少ない [Hecht 11, 山口 13]．そのため，ユーザの居住地を推定する試みが多数なされている．

Twitter*1などのソーシャルメディアには，フォローしている，またはフォローされているなどのユーザ間の関係があり，それらの関係から作成したソーシャルネットワーク（友人関係グラフ）を利用し，ユーザの居住地を推定する研究がある [McGee 13, Rout 13]．ソーシャルネットワークを作成する際に利用するユーザ間の関係を変えると，異なる形のソーシャルネットワークができる．ユーザ間の関係によって地理的に近くにいる友人の割合が変化する [McGee 11] と報告されているが，居住地推定の性能がどのように変化するのかは明らかになっていない．

本研究では，ユーザ間の関係を変えて作成した複数のソーシャルネットワークを用いて，それらが居住地推定に与える影響を調査する．この調査により，フォローされているというユーザ間の関係が居住地推定に最も有効であることを示す．また，代表的な居住地推定手法の推定傾向は，ソーシャルネットワークの形状に影響を受けないことも示す．

2. 関 連 研 究

ソーシャルメディアにおける居住地推定に関する研究は，主に Twitter のデータを用いて検証されている．Twitter の分析および研究開発には居住地などのユーザの属性が利用される [奥村 12] が，自身のプロフィールに居住地を入力しているユーザは少ない [Hecht 11, 山口 13]．そのため，ユーザの居住地を推定する試みが多数なされている．居住地推定手法は，推定に利用する情報の違いから，ユーザの友人関係を利用するネットワークベースの

*1 https://twitter.com/ (viewed 2016-11-04)

手法，投稿内容を利用するコンテンツベースの手法，さらにそれら両方を組み合わせて利用するハイブリッドの手法に分けられる．

Twitter のフォロー関係をもとにしたネットワークベースの手法として，友人の居住地の中で最も出現数の多いものを居住地と推定する手法が提案されている [Davis Jr. 11]．また，Sadilek らは居住地推定とリンク予測を同時に解く手法を提案している [Sadilek 12]．McGee らは友人関係を分析し，決定木によりユーザの信頼度を決め，尤度を用いるモデル [Backstrom 10] を拡張している [McGee 13]．Rout らは，居住地推定をユーザの住んでいる都市の分類問題とみなし，SVM を用いてユーザの居住地を推定している [Rout 13]．Jurgens は，リプライから作成したソーシャルネットワークを利用し，友人の情報のみを利用する推定手法を繰り返し適用することで多くのユーザの居住地が推定できることを示している [Jurgens 13].

コンテンツベースの手法には，Cheng らのツイート本文に含まれる地理的な単語を利用して居住地を推定する手法がある [Cheng 10]．Kinsella らはツイート本文から作成した言語モデルをもとに推定している [Kinsella 11]．ハイブリッドの手法には，Li らのユーザとツイート本文に含まれる地名をノードとするネットワークを用いた手法がある [Li 12b]．さらに複数の居住地を推定する方法も提案している [Li 12a]．Chen らはつながりの強さを考慮するよう Li らの手法を拡張している [Chen 16].

居住地推定のための多くの手法が提案されているが，実験条件が異なるため，論文の情報だけでは結果を比較することができない．そのため，新たな手法の提案はせず，これまでに提案されてきた手法の比較および分析をする研究もある．Jurgens ら [Jurgens 15] はメンションをもとに作成したソーシャルネットワークを利用し，ネットワークベースの手法の統一的な評価をしている．

これまでに提案されてきたネットワークベースの手法ではフォロー関係が使われる傾向にあることから，本研究ではメンション関係ではなくフォロー関係に着目した調査をする．つまり，フォロー関係をもとに作成した 4 種類のソーシャルネットワークを用いて，それらが居住地推定に与える影響を調査する．この調査により，フォローされているというユーザ間の関係が居住地推定に最も有効であることを示す．また，代表的な居住地推定手法の推定傾向は，ソーシャルネットワークの形状に影響を受けないことも示す．Twitter ユーザすべてのソーシャルネットワークを調べることは困難であるため，本研究では位置情報付きツイートを投稿したユーザのソーシャルネットワークで調査する．

3. データセットの作成および特徴

本調査では，Twitter ユーザの居住地データと，フォロー関係をもとにしたソーシャルネットワークとを利用して，居住地推定の性能を調べる．これらのデータ作成方法の詳細について 3·1 節以降で述べる．

3·1 位置情報付きツイートをもとにした居住地

調査に利用するユーザの居住地は位置情報付きツイートをもとに決定する．ユーザは主に居住地周辺で活動していると考えられるため，ユーザが位置情報付きツイートを投稿している主な場所をそのユーザの居住地とする．本研究では，ネットワークベースの手法を提案している主要な先行研究 [Davis Jr. 11] と同様に，居住地を市区町村レベルのエリアとする．このエリアは，森國ら [森國 15] と同様の方法で総務省統計局の境界データから作成する．位置情報付きツイートの地理座標情報（coordinates）からその座標が含まれるエリア（日本国内の市区町村）を求め，ユーザごとに最もツイート数の多いエリアをそのユーザの居住地とする．

Twitter Streaming API*2を使用し，2014 年に日本を包含する矩形 *3の中で投稿された位置情報付きツイート（250,564,317 件）を集めた．森國ら [森國 15] と同様に bot による投稿を除外したうえで，2014 年に 5 回以上位置情報付きツイートを投稿しているユーザという条件を設定し，614,440 ユーザへ居住地を付与した．

3·2 フォロー関係をもとにしたソーシャルネットワーク

本研究では，ユーザ間のフォロー関係を利用してソーシャルネットワークを作成する．ユーザがフォローしているユーザの集合 *4とユーザをフォローしているユーザの集合 *5との 2 種類の情報を取得し，これらを合わせてユーザ間のフォロー関係として利用する．居住地を付与できた 614,440 ユーザの周りのフォロー関係を 2015 年 7 月に取得した．必要な情報をすべて取得することができた 472,350 ユーザを調査に使用する．

Twitter でのフォロー関係をもとにしたユーザ間の関係として，フォローしている関係（*followee*），フォローされている関係（*follower*），相互にフォローしている関係（*mutual*），フォローしているまたはされている関係（*linked*）の 4 種類が考えられる．居住地推定に最も有効な関係を特定するため，それぞれの関係をもとにした 4 種類のソーシャルネットワークを作成した．本研究でのソーシャルネットワークは，図 1 に示すように，ユーザをノード，ユーザ間の関係を有向エッジとして作成する単純有向グラフである．作成したソーシャルネットワークにおいて，あるノードの隣接ノードとは，あるノードからその関係（followee や follower など）にあるノード

*2 https://dev.twitter.com/streaming/reference/post/statuses/filter (viewed 2016-11-04)
*3 北緯 20 から 50，東経 110 から 160 の範囲．
*4 https://dev.twitter.com/rest/reference/get/friends/ids (viewed 2016-11-04)
*5 https://dev.twitter.com/rest/reference/get/followers/ids (viewed 2016-11-04)

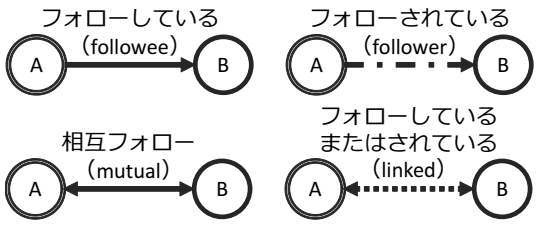


図 1 フォロー関係をもとにした 4 種類のユーザ間の関係

である．図 1 では，ノード B はノード A の隣接ノードとなる．

3·3 ソーシャルネットワークの特徴

本節では，作成したソーシャルネットワークの統計量を調べ，ユーザ間の関係を変えて作成したソーシャルネットワークの特徴を明らかにする．さらに，居住地の付与されているユーザとされていないユーザとの違いについて調査する．

ユーザ間の関係を変えて作成したソーシャルネットワークの特徴を明らかにするため，グラフの基本的な統計量を調べる．ネットワークの大きさをみるために，作成した有向ソーシャルネットワーク $G(V,E)$ の次数が 1 以上のノード数 $|V'|$，エッジ数 $|E|$ を調べる．さらに，推定には隣接ノード（友人）を利用するため，居住地を付与したノードの出次数（隣接ノード数）の平均 K_{out} と標準偏差 S_{out}，中央値 M_{out} を調べる．加えて，隣接ノードのみを利用する手法では推定できないユーザの数となる，居住地を付与したノードのうち出次数が 0 のノード数 $|I_{out}|$ を調べる．なお，次数が 1 以上のノード集合 V' のほかに，次数が 0 以上のノード集合を V として仮定するが，3·2 節で述べたように居住地を付与したノードの隣接ノードしか取得していない都合上，観測できないノードが存在する．そのため，$|V| \geq |V'| + |I_{out}|$ の関係が成立するものの，$|V|$ の正確な値は算出不能であるため，本稿では V の議論はしない．

3·2 節で述べたように，居住地を付与したユーザの周りのフォロー関係を取得し，4 種類のソーシャルネットワークを作成した．居住地を付与したユーザとフォロー関係にあるユーザには，居住地の付与されているユーザとされていないユーザとがある．つまり，収集したすべてのデータから作成したソーシャルネットワークには，居住地の付与されているノードとされていないノードとが含まれている．しかし，フォロー関係を取得する起点としたノードは居住地が付与されたノードのみであり，居住地が付与されていないノード同士の関係は取得できていない．以上のような制約があることから，取得した関係すべてを利用して作成したソーシャルネットワークと，取得した関係のうち居住地を付与したノード同士の関係のみから作成したソーシャルネットワークとを区別して統計量を調べる．表 1 に調べた統計量を示す．なお，4 章で後述するように，本研究では隣接ノードのみを利用する手法で居住地推定性能を評価するため，実験では，居住地を付与したノード同士の関係のみから作成したソーシャルネットワーク（表 1(b)）を使用することとなる．

ソーシャルネットワークを作成する際，フォローしている関係とフォローされている関係とをそれぞれ取得し，それらを合わせたデータを利用している．また，followee をもとにしたネットワークは follower をもとにしたネットワークの有向エッジを逆向きにしたものと同じである．これらにより，followee をもとにしたネットワークと，follower をもとにしたネットワークとでは，エッジ数 $|E|$ が等しくなる．さらに，ノードすべてのフォロー関係を取得できたソーシャルネットワークでは，あるユーザがフォローしているとき，フォローされているユーザが必ず存在する．そのため，表 1(b) に示すとおり，居住地を付与したノードに絞ったソーシャルネットワークでは，followee をもとにしたネットワークと follower をもとにしたネットワークとで平均出次数 K_{out} が一致する．ただし，フォローされやすいユーザやされにくいユーザが存在するため，出次数の標準偏差 S_{out} は異なる．

表 1 の統計量から，もとにした関係によるソーシャルネットワークの差異について述べる．エッジ数 $|E|$ からネットワークの規模をみると，linked をもとにしたネットワークが最も大きく，mutual をもとにしたネットワークが最も小さい．エッジ数 $|E|$ が小さいネットワークでは，推定に利用できる隣接ノードが少なく，推定できないユーザ数である $|I_{out}|$ が大きくなる．このため，mutual をもとにしたソーシャルネットワークは，推定できないユーザ数 $|I_{out}|$ がほかのソーシャルネットワークよりも大きくなっている．

取得した関係すべてを利用して作成したソーシャルネットワークの統計量（表 1(a)）と居住地を付与したノード同士の関係のみから作成したソーシャルネットワークの統計量（表 1(b)）とを比べると，ユーザが持つ友人の数の平均である K_{out} に差がある．居住地が付与されている友人は，すべての友人のうち，最小では 3.36%，最大でも 4.83%であることが分かる．このことから，位置情報付きツイートをもとに居住地を付与できるユーザは，Twitter における全ユーザの 5%未満であることが示唆される．

linked をもとにしたネットワークと mutual をもとにしたネットワークとの $|E|$ の比は相互フォロー率を表す．収集したネットワーク全体では，フォローのうち約 43%が相互フォローである．一方，居住地が付与されたノードのみのネットワークでは，フォローのうち約 62%が相互フォローである．相互フォロー率の大小と，フォローが購読関係（subscription）と友人関係（friendship）とのどちらを表すかどうかには関連があるため [Yamaguchi 15]，実験で用いるソーシャルネットワークには友人関係が比較的多いと考えられる．

表 1 ソーシャルネットワークの統計量

(a) 収集したネットワーク全体

関係	$\|V'\|$	$\|E\|$	K_{out}	S_{out}	M_{out}	$\|I_{out}\|$
followee	62676854	417334528	385.9831	2059.3849	205	3301
follower	62676854	417334528	514.8269	4761.8717	194	2780
mutual	22834460	251625205	272.9455	1837.2806	129	8710
linked	62676854	583043851	627.8645	4941.1793	271	1085

(b) 居住地が付与されたノードのみのネットワーク（実験に使用するネットワーク）

関係	$\|V'\|$	$\|E\|$	K_{out}	S_{out}	M_{out}	$\|I_{out}\|$
followee	428150	8163069	17.2818	53.7495	7	54618
follower	428150	8163069	17.2818	61.1349	6	65838
mutual	389050	6226387	13.1817	45.1985	5	83300
linked	428150	10099751	21.3819	68.5991	9	44200

4. 調査する居住地推定手法

ソーシャルネットワークを利用する居住地推定手法は，ソーシャルネットワークとその一部のユーザに付与された居住地とをもとに，その他のユーザの居住地を推定する手法である．ここでのソーシャルネットワークは，3·2 節で述べたように，ユーザをノード，ユーザ間の関係をエッジとする単純有向グラフである．また，あるユーザの居住地はノードへ付けられたラベルとして表現する．4·1 節以降で説明する居住地推定手法は，推定対象ノード u，推定対象ノード u の隣接ノード集合 N_u とそれらのラベルのみを利用して推定を行うため，ノード u のラベルの推定はラベルを返す推定関数 $f(u)$ で表せる．本研究では，ソーシャルネットワークのもととなるユーザ間の関係が居住地推定にどのような影響を与えるのかを調査するために，隣接ノードをそのまま利用する手法のうち，Jurgens らによる性能評価 [Jurgens 15] で良好な結果を示していた 3 手法およびベースラインの計 4 手法を実装する．これらの手法の詳細は 4·1 節以降で説明する．

手法の説明では次の変数を用いる．L は学習データ集合，N_u はノード u の隣接ノード集合，A は推定対象ラベル集合（エリア集合），l_u はノード u の正解ラベル，$dist(a,b)$ はラベル a とラベル b との間の距離，K_{out} は隣接ノード数の平均値である．学習データ集合はノードの集合であり，ラベル間の距離はラベルに対応付けられる居住地（エリア）の重心間の地理的な距離をヒュベニの式 *6[Hubeny 54] で計算したものである．ノード間の距離は，ノードに付けられたラベル間の距離とする．

4·1 Probability Model

Probability Model は，ノード間がある地理的距離のときにエッジが存在する確率のモデルを作り，推定対象のノードのラベル（居住地）である確率が最も高いラベルを推定する手法である [Backstrom 10]．この手法は Facebook のデータセットに対して提案された手法であるものの，Twitter のデータセットを対象とする研究でも使われている [McGee 13]．あるノード間の距離が d のときに，そのノード間にエッジが存在する確率 $p(d)$ を表すモデルが式 (1) である．a，b，c は実数のパラメータであり，実験の際には，文献 [Backstrom 10] に書かれている値 $a = 0.0019$，$b = 0.196$，$c = -1.05$ を使う *7．このモデル式を利用し，式 (2) でノード u の居住地を推定する *8．推定に必要な計算量は $O({K_{out}}^2)$ である．

$$p(d) = a(d+b)^c \tag{1}$$

$$\gamma_l(l) = \prod_{n \in L} [1 - p(dist(l, l_n))]$$

$$\gamma(l,u) = \prod_{n \in N_u \cap L} \frac{p(dist(l,l_n))}{1 - p(dist(l,l_n))} \gamma_l(l)$$

$$ProbabilityModel(u) = \underset{l \in \{l_n | n \in N_u \cap L\}}{\arg\max} \gamma(l,u) \tag{2}$$

4·2 Majority Vote

Majority Vote 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する手法である [Davis Jr. 11]．この手法のもととなる仮定は，同じ居住地（ラベル）に住んでいる友人（隣接ノード）が最も多いというものである．文献 [Davis Jr. 11] には，隣接ノードが持つラベルの中で出現頻度が最大のラベルが複数存在する場合の処理が明記されていないため，本研究ではソーシャルネットワーク全体での出現頻度が高いラベルを優先的に選択する．この手法を表現したものが式 (3) であり，計算量は $O(K_{out})$ である．ここで，arg max*

*6 処理速度向上のため，実際の距離計算にはヒュベニの式の第 1 項のみを用いた簡略式を使用した．地球を楕円体とするための定数には WGS84 の値を用いた．

*7 実験で利用するデータをもとにパラメータを探索したが，より良い推定性能を示すパラメータが見つからなかった．

*8 オリジナル [Backstrom 10] の式に誤りがあると考えられるため，$\gamma(l,u)$ の式に $\gamma_l(l)$ を補っている．

は同値の集合を返すものと定義する．

$$S_u = \underset{l \in \{l_n | n \in N_u \cap L\}}{\arg\max}^{*} |\{x | x \in N_u \cap L, l = l_x\}|$$
$$MajorityVote(u) = \underset{l \in S_u}{\arg\max} |\{n | n \in L, l = l_n\}| \quad (3)$$

この手法には，推定対象ノードの隣接ノード数の範囲，多数決の際の最低投票数という 2 つのパラメータが存在する．今回の実験では他の手法と条件をそろえるため，推定対象ノードの隣接ノード数の範囲は 0 から無限大，最低投票数は 0 とする．

4·3 Geometric Median

2 次元の点集合の中から，主な点を選択する手法の一つとして Geometric Median*9 [Eftelioglu 15, Vardi 00] があり，標本点集合の中で他の点との距離の和が最小になる点と定義されている．本研究で用いる手法 *Geometric Median* は，推定対象のノードの隣接ノードのラベルの中から，その他のラベルとの距離の和が最小になるラベルを選択し，推定対象ノードのラベルと推定する手法である [Jurgens 13]．この手法を表現したものが式 (4) であり，計算量は $O({K_{out}}^2)$ である．

$$GeometricMedian(u) = \underset{l \in \{l_n | n \in N_u \cap L\}}{\arg\min} \sum_{x \in N_u \cap L, n \neq x} dist(l, l_x) \quad (4)$$

4·4 Random Neighbor

Jurgens ら [Jurgens 15] は，手法の性能を比較する際のベースラインとしてランダムに選択する手法を用いている．*Random Neighbor* は，ラベルの付いた隣接ノードをランダムに選択し，そのノードのラベルを推定ラベルとする手法である *10．この手法を表現したものが式 (5) であり，計算量は $O(1)$ である．ここで，$choice(S)$ は集合 S からランダムに要素を 1 つ選択する関数である．

$$RandomNeighbor(u) = l_{choice(N_u \cap L)} \quad (5)$$

5. 実　　験

leave-one-out 交差検証と 10 分割交差検証により，居住地推定手法とソーシャルネットワーク作成方法とをそれぞれ変えたときの推定性能を比較する．leave-one-out 交差検証では推定環境が最も良いときの性能を検証し，10 分割交差検証では学習データによって性能が大幅に変化しないことを検証する．

推定性能は適合率（Precision），再現率（Recall），F 値（F1）の 3 つの指標で評価する．適合率は推定されたユーザのうち正しいエリアを推定できたユーザの割合，再現率はテストデータのうち正しいエリアを推定できたユーザの割合，F 値は適合率と再現率の調和平均である．加えて，分析のために，推定可能なユーザの割合を表すカバー率（Coverage）を用いる．本実験でのテストデータに含まれるユーザには，出次数が 0，つまり隣接ノード数が 0 のノード *11が存在するため，カバー率の最大値は 100%にならない．これらの評価指標を次の式で計算する．

$$Precision(T, X) = \frac{|\{u | u \in T \cap X, l_u = e_u\}|}{|T \cap X|}$$
$$Recall(T, X) = \frac{|\{u | u \in T \cap X, l_u = e_u\}|}{|T|}$$
$$F1(T, X) = \frac{2 * Precision(T, X) * Recall(T, X)}{Precision(T, X) + Recall(T, X)}$$
$$Coverage(T, X) = \frac{|T \cap X|}{|T|}$$

ここで，X は推定されたノードの集合，T はテストデータ集合，l_u はノード u の正解居住地，e_u はノード u の推定された居住地である．10 分割交差検証では，それぞれのテストデータでの評価指標の平均値を評価値とする．

5·1 居住地推定の評価

4 種類の居住地推定手法と 4 種類のソーシャルネットワークとをそれぞれ変えて，居住地推定を行った結果を表 2，表 3 に示す．これらの表における下線（一重下線および二重下線）は，その指標の中で最も良い結果であることを示す．適合率，再現率，F 値は大きいほど良く，5·3 節で後述する Mean ED と Median ED は小さいほど良い．表 3 における下線のうち二重下線は，t 検定により，二重下線の結果とその他すべての結果との間に危険率 1%で有意に差があることを示す．

表 2，表 3 から，日本のソーシャルネットワークでは Majority Vote が最も精度良く居住地を推定できることが分かる．最も性能が良かった Majority Vote を用いて居住地を推定するとき，ソーシャルネットワーク作成のためのユーザ間の関係として follower と mutual を利用すると適合率が高くなり，follower と linked を利用すると再現率が高くなる．F 値が最も高くなるソーシャルネットワーク作成のためのユーザ間の関係は follower（フォローされている関係）である．

表 2 に示す leave-one-out 交差検証では，follower をもとに作成したソーシャルネットワークに対して Majority Vote を用いて居住地推定をした（以降，手法とネットワークの組み合わせを Majority Vote + follower のように表記する）結果と，その他すべての推定結果は，危険率 1%で有意に差がある（推定結果が異なる）ことを McNemar 検定で確認した．表 3 に示す 10 分割交差検証では，適合率および F 値の平均に関して，Majority Vote + follower の

*9 Fermat–Weber Problem や L1 Median とも呼ばれる．
*10 Jurgens らのベースラインとは，繰り返しの有無が異なる．
*11 該当するノードの数は表 1(b) に示した $|I_{out}|$ である．

表 2 居住地推定性能（leave-one-out 交差検証）

手法	関係	適合率	再現率	F 値	カバー率	Mean ED [km]	Median ED [km]
Probability Model	followee	0.29598	0.26176	0.27782	0.88437	153.212	19.610
	follower	0.30695	0.26416	0.28395	0.86062	146.955	18.534
	mutual	0.30116	0.24805	0.27204	0.82365	146.606	18.930
	linked	0.30451	0.27602	0.28957	0.90643	151.292	18.780
Majority Vote	followee	0.29807	0.26361	0.27978	0.88437	165.936	19.355
	follower	0.31615	0.27208	0.29246	0.86062	156.137	17.218
	mutual	0.31214	0.25709	0.28195	0.82365	155.090	17.466
	linked	0.30581	0.27719	0.29080	0.90643	163.698	18.596
Geometric Median	followee	0.25560	0.22605	0.23992	0.88437	149.481	20.852
	follower	0.27061	0.23289	0.25034	0.86062	140.649	19.352
	mutual	0.27387	0.22557	0.24738	0.82365	139.659	18.835
	linked	0.25661	0.23260	0.24402	0.90643	147.993	20.793
Random Neighbor	followee	0.17782	0.15726	0.16691	0.88437	216.106	41.827
	follower	0.18849	0.16222	0.17437	0.86062	207.806	39.843
	mutual	0.19582	0.16129	0.17688	0.82365	199.462	36.224
	linked	0.17444	0.15812	0.16588	0.90643	220.241	44.919

表 3 居住地推定性能（10 分割交差検証）

手法	関係	適合率	再現率	F 値	カバー率	Mean ED [km]	Median ED [km]
Probability Model	followee	0.29261	0.25607	0.27312	0.87514	154.597	20.006
	follower	0.30331	0.25800	0.27883	0.85061	148.172	18.882
	mutual	0.29772	0.24192	0.26694	0.81259	147.615	19.233
	linked	0.30079	0.27016	0.28465	0.89817	152.714	19.169
Majority Vote	followee	0.29304	0.25645	0.27353	0.87514	168.018	20.152
	follower	0.31109	0.26461	0.28597	0.85061	158.054	17.839
	mutual	0.30716	0.24959	0.27540	0.81259	156.738	18.000
	linked	0.30059	0.26998	0.28446	0.89817	165.762	19.233
Geometric Median	followee	0.25373	0.22205	0.23684	0.87514	150.763	21.115
	follower	0.26800	0.22796	0.24637	0.85061	142.143	19.734
	mutual	0.27130	0.22045	0.24325	0.81259	141.199	19.181
	linked	0.25480	0.22885	0.24113	0.89817	149.326	21.135
Random Neighbor	followee	0.17766	0.15547	0.16583	0.87514	216.398	42.259
	follower	0.18840	0.16025	0.17319	0.85061	207.721	39.788
	mutual	0.19692	0.16002	0.17656	0.81259	199.577	36.026
	linked	0.17483	0.15703	0.16546	0.89817	219.951	44.484

結果と，その他すべての結果との間に危険率 1%で有意に差があることを t 検定で確認した．また，再現率に関して，Probability Model + linked および Majority Vote + linked の結果と，その他すべての結果との間に危険率 1%で有意に差があることを t 検定で確認した．Probability Model + linked と Majority Vote + linked との間には有意差を確認できなかった [*12]．なお，Probability Model と Majority Vote との計算量はそれぞれ $O(K_{out}^2)$ と $O(K_{out})$ であり，計算量に差がある．双方の推定性能に有意差がないため，計算量の小さい Majority Vote の方が有効に機能すると考えられる．以上の検定では，着目している群とそれ以外の群との 2 群間検定を繰り返し，そのすべての組み合わせにおいて危険率 1%で有意差があるかどうかを確認した．

5·2 ユーザ間の距離の分布と居住地推定性能の関係

本研究で作成した 4 種類のソーシャルネットワークでの，ユーザ間の地理的な距離の分布を図 2 に示す．図 2(a) は，友人（隣接ノード）との地理的な距離の平均が k [km] 以下であるユーザの割合のグラフである．日本のユーザから取得したデータをもとに作成したソーシャルネットワークにおいても，McGee らの調査 [McGee 11] と同様に，相互にフォローしている関係（mutual）のとき，近くに友人のいる割合が最も高くなる．しかし，居住地推定性能で比較すると，F 値が最も高くなっている関係は follower である．図 2(b) は，友人（隣接ノード）との地理的な距離の平均が k [km] 以下であるユーザ数のグラフである．mutual をもとにしたソーシャルネットワークは，他の関係をもとにしたソーシャルネットワークと比べ，得られるユーザ間の関係数が少ないことが分かる．そのため，カバー率が低くなり，再現率も低くなっていると考えられる．

図 2(c) はユーザの友人（隣接ノード）との地理的な距離の分布である．この分布には，1 [km] から 100 [km] の部分の近くにある山と，200 [km] 以降の部分の遠くにある山とがある． 近くの山は友人であるユーザが，遠くの山には有名人や企業の公式アカウントなど購読しているユーザが含まれるといわれている [McGee 13]．Twitter においてユーザをフォローする目的は，友人と購読との 2 種類に大きく分けられる [Kwak 10]．あるユーザがフォローしているユーザ集合をみたとき，その集合には友人と購読目的のアカウントが混ざっている．また，有名人などの一部のユーザが多くのフォロワーを持つ傾向がある [Kwak 10]．これらのことから，有名人は購読目的で多くのユーザにフォローされてフォロワーが多くなる一方，大多数の一般ユーザは購読目的でフォローされないため，一般ユーザのフォロワーには友人が多くなると考えられる．このことは，図 2(a) および図 2(b) において，k が小さいときに follower が followee を上回っていることからも裏付けられる．以上より，友人が多く含まれる follower の関係が居住地推定に適し，適合率が高くなると考えられる．

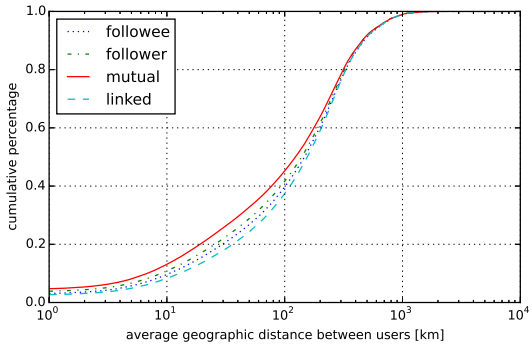
表 2 および表 3 から分かるとおり，フォローされている関係（follower）をもとに作成したソーシャルネットワークと相互にフォローしている関係（mutual）をもとに作成したソーシャルネットワークとで適合率は同等である．mutual をもとに作成したソーシャルネットワークにおいて，あるユーザの隣接ノード集合は，follower をもとに作成したソーシャルネットワークでのそのユーザの隣接ノード集合の部分集合である．つまり，mutual をもとに作成したソーシャルネットワークで適合率が高くなるのは，follower が居住地推定に有効な関係であり，mutual にも follower と同様に，隣接ノード集合に友人であるユーザが多く含まれているからであると考えられる．

5·3 エラー距離での評価

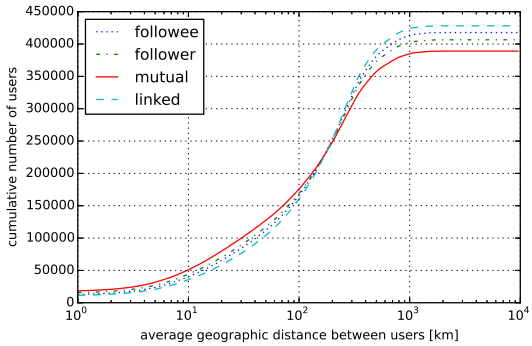
5·1 節では，厳密に正しい居住地を推定できるか否かを評価した．しかし，正しい居住地を推定できなくとも，正しい居住地の近くに推定できていれば有用だと考えられる．本節では，エラー距離での評価を行う．エラー距離は，正解居住地と推定居住地との距離とする．ユーザ間の地理的な距離の分布はべき乗分布である [Rout 13] ため，友人の居住地の中から居住地を選択する手法での推定結果において，エラー距離の平均は一部の大きく間違った（エラー距離の大きい）結果に引きずられると考えられる．図 3 は Majority Vote + follower による leave-one-out 交差検証での推定結果におけるエラー距離の分布であり，エラー距離の偏りを確認できる．これらにより，テストデータにおけるエラー距離の平均とする平均エラー距離（Mean ED）のほかに，エラー距離の中央値とする中央値エラー距離（Median ED）も評価に用いる．ユーザ集合 U に含まれるユーザ u のエラー距離 $dist(l_u, e_u)$ のリストを D_U とするとき，平均エラー距離は $D_{T \cap X}$ の平均，中央値エラー距離は $D_{T \cap X}$ の中央値と計算する．

表 2，表 3 に示すように，中央値エラー距離において，follower および mutual ならびに Probability Model および Majority Vote は厳密に正しい居住地を推定する場合と同様に，良い性能を達成する傾向がある．しかし，平均エラー距離の評価では Geometric Median が Probability Model や Majority Vote を上回っている．先に述べたように，エラー距離の平均は一部の大きく間違った結果に引きずられる．このことから，Geometric Median は Probability Model や Majority Vote よりも大きく間違えない可能性が示唆される．10 分割交差検証では，中央値エラー距離に関して，Majority Vote + follower の結果はその他すべての結果と比べて，危険率 1%で有意に差があることを t 検定で確認した．また，同様に平均エラー距離に関して，Geometric Median + mutual の結果はその他すべての結果と比べて，危険率 1%で有意に差があること

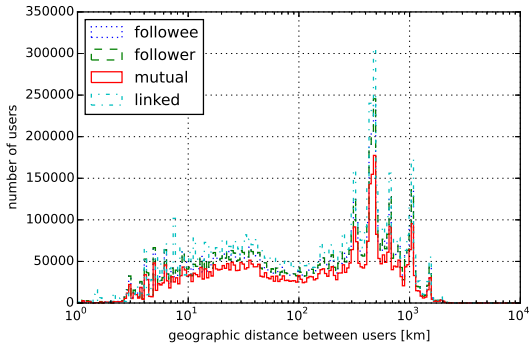
*12 t 検定での p 値は 0.769 であった．



(a) 正規化累積分布



(b) 累積分布



(c) 密度分布

図 2 ユーザ間の地理的な距離の分布

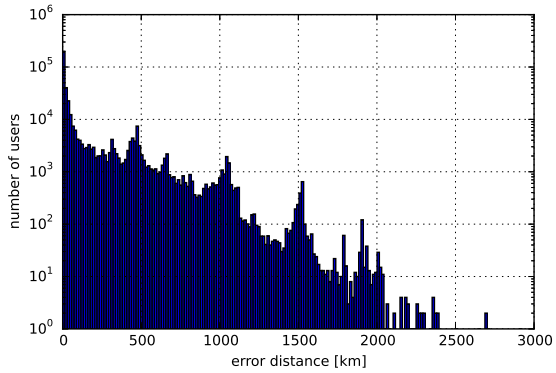


図 3 エラー距離の分布（Majority Vote + follower）

を t 検定で確認した．なお，表 2，表 3 の読み取り方および検定方法の詳細は 5·1 節を参照されたい．

5·4 正解とする距離や正解粒度を変えての評価

実際には，許容されるエラー距離はアプリケーションによって変化すると考えられるため，正解とする距離を変えて再現率を評価する．k [km] 以内を正解とするときの再現率（$Recall_k$）の式を次に示す．

$$Recall_k(T, X, k) = \frac{\{u | u \in T \cap X, dist(l_u, e_u) < k\}}{|T|}$$

実験に利用するユーザには居住地として日本の市区町村がラベル付けされており，最大エラー距離は日本の全長より小さいことが分かっているため，k を 1 [km] から 10^4 [km] まで変化させて評価する．

leave-one-out 交差検証での，正解とする距離 k を変えたときの評価結果を図 4 に示す．k が 20 [km] より近くのときは Majority Vote の推定性能が高い．また，100 [km] から 400 [km] の付近では Geometric Median が Majority Vote を上回る．これは前節で述べたように，Geometric Median が大きく間違えないことを裏付けている．

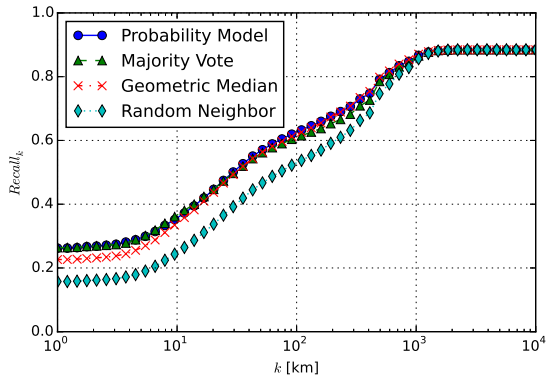
アプリケーションによっては，市区町村レベルより大きな都道府県レベルでの居住地情報を活用したい場合がある．そこで，正解粒度を変更し，都道府県レベルでの適合率，再現率を評価する．市区町村レベルで推定したエリアが，正解である居住地と同じ都道府県である場合に正解であるとみなし，各指標を計算する．

leave-one-out 交差検証での，正解エリアの粒度を都道府県レベルとしたときの評価結果を表 4 に示す．表 4 における下線は，その指標の中で最も良い結果であることを示す．正解を都道府県レベルとみなせば，5 割程度のユーザの居住地を当てることができると分かる．Probability Model + linked の推定結果とその他すべての推定結果との間で，McNemar 検定により，危険率 1%で有意に差がある（推定結果が異なる）ことを確認した．Geometric Median + mutual の推定結果と，Majority Vote + mutual の推定結果との間には，有意差を確認できなかった [*13]．以上の検定では，着目している群とそれ以外の群との 2 群間検定を繰り返し，そのすべての組み合わせにおいて危険率 1%で有意差があるかどうかを確認した．

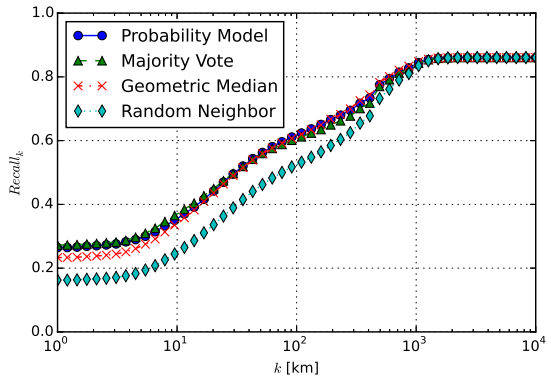
6. 考 察 と 限 界

本研究では，フォロー関係に着目し，フォロー関係をもとにした 4 種類のソーシャルネットワークを用いて，ネットワークベースの居住地推定手法の統一的な評価を行った．同様の統一的な評価は Jurgens ら [Jurgens 15] も行っているが，彼らはツイート内のメンション（リプライ）に着目し，相互にメンションしている関係をもとに

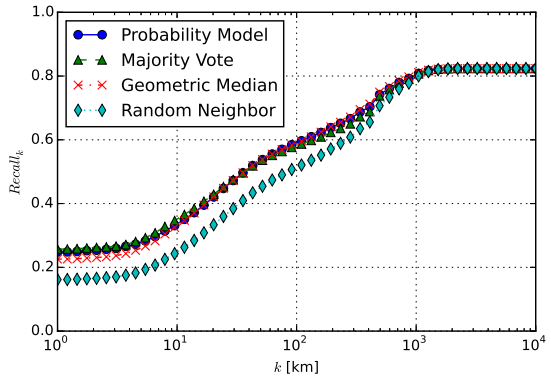
*13 McNemar 検定での p 値は 0.607 であった．



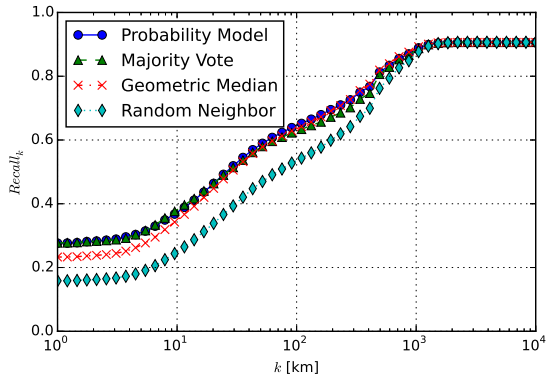
(a) followee から作成したネットワーク



(b) follower から作成したネットワーク



(c) mutual から作成したネットワーク



(d) linked から作成したネットワーク

図 4 4 種類の手法での推定性能を k を変えた $Recall_k$ で評価した結果

作成したソーシャルネットワークのみを用いている．対して本研究では，ソーシャルネットワーク上でのユーザ関係を捉える，より一般的な方法であるフォロー関係に着目し，ソーシャルネットワークを作成した．さらに，これまでの研究では相互にフォローしている関係がよく利用されているが [Davis Jr. 11, McGee 13]，本研究ではフォロー関係から生成することができる 4 種類のユーザ間の関係をもとにした 4 種類のソーシャルネットワークを利用し，ユーザ間の関係が居住地推定にどのような影響を与えるのかを調査した．ソーシャルメディアにおけるユーザ間の関係は，フォロー関係やメンション関係以外にも，いいね関係（お気に入りに入れたか否か）やリツイート関係（投稿を他のユーザに拡散したか否か）も存在する．これらを組み合わせたり，横断したりしての評価は今後の課題である．

McGee ら [McGee 13] は，フォロー関係とメンション関係のソーシャルネットワークを利用し，隣接ノード（友人）との地理的な距離の変化について調査しているが，居住地推定に与える影響は明らかにされていなかった．本研究では，フォローの関係から生成することができる 4 種類のソーシャルネットワークを利用し居住地推定に与える影響を調査し，近傍となる確率の高まる相互にフォローしている関係（mutual）が居住地推定に必ずしも有効ではないことを明らかにした．データから観察できるユーザとの地理的な距離は大都市間の距離が影響するという報告 [Takhteyev 12] や，居住地の人口密度が友人との地理的な距離に影響するという報告 [松本 05] があるなど，友人との地理的な距離には，様々な外的要因がある．このような外的要因が居住地推定に与える影響の調査は今後の課題である．

本研究では，日本国内で投稿された位置情報付きツイートをもとにユーザを抽出し，フォロー関係を取得している．そのため，大半のユーザは日本に居住する日本人であると考えられ，今回の調査結果が国をまたぐデータに適用可能であるかどうかは明らかではない．より大規模な実験は今後の課題である．また，日本国内での地域間における比較や，国間における比較も重要だと考えている．このような比較により，実社会での人間関係をインターネット（ソーシャルメディア）上でも構築しうる文化的背景が明らかにできる可能性もある．

本研究では，主に市区町村レベルでの居住地推定性能を評価しているが，5·4 節では正解粒度を都道府県レベルに変更して評価した．その結果，表 2 と表 4 とを比較すれば分かるとおり，有効な手法および関係の組み合わせが 5·1 節で述べた組み合わせと異なる結果を得た．しかし，評価では正解粒度のみを変更しており，推定粒度，つまり居住地推定に使用するエリアの粒度を変更しておらず，エリアの粒度を変えた場合の評価は今後の課題である．使用するエリアの粒度が市区町村レベルなのか都道府県レベルなのか，全世界的には州レベルなのか国レ

表 4 都道府県レベルでの居住地推定性能（leave-one-out 交差検証）

手法	関係	適合率	再現率	F 値	カバー率
Probability Model	followee	0.55960	0.49489	0.52526	0.88437
	follower	0.57166	0.49198	0.52883	0.86062
	mutual	0.57088	0.47020	0.51567	0.82365
	linked	0.56525	0.51236	0.53750	0.90643
Majority Vote	followee	0.55147	0.48770	0.51763	0.88437
	follower	0.57301	0.49314	0.53008	0.86062
	mutual	0.57307	0.47200	0.51765	0.82365
	linked	0.55741	0.50525	0.53005	0.90643
Geometric Median	followee	0.55052	0.48687	0.51674	0.88437
	follower	0.56871	0.48944	0.52611	0.86062
	mutual	0.57335	0.47224	0.51790	0.82365
	linked	0.55188	0.50024	0.52480	0.90643
Random Neighbor	followee	0.44533	0.39384	0.41800	0.88437
	follower	0.45985	0.39575	0.42540	0.86062
	mutual	0.47357	0.39005	0.42777	0.82365
	linked	0.43696	0.39607	0.41552	0.90643

ベルなのか，あるいは地理座標の矩形サイズの大小など，それぞれの推定粒度で有効な手法および関係が異なる可能性がある．

7. お わ り に

本研究では，フォロー関係をもとに作成した 4 種類のソーシャルネットワークを用いて，それらが居住地推定に与える影響を調査した．この調査により，フォローされているというユーザ間の関係から作成したソーシャルネットワークが居住地推定に最も有効であることを示した．このことは，従来手法でよく用いられる相互にフォローしている関係を準備せずとも同等以上に居住地を推定できることを意味する．また，居住地推定手法に着目すると，友人の居住地の中から最頻のものを選択する Majority Vote がソーシャルネットワークの形状に影響を受けず，最も精度良く居住地を推定できることを示した．

謝 辞

本研究は JSPS 科研費 JP16K16155 の助成を受けたものです．

◇ 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 World Wide Web*, pp. 61–70 (2010)

[Chen 16] Chen, J., Liu, Y., and Zou, M.: Home Location Profiling for Users in Social Media, *Information & Management*, Vol. 53, No. 1, pp. 135–143 (2016)

[Cheng 10] Cheng, Z., Caverlee, J., and Lee, K.: You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users, in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pp. 759–768 (2010)

[Davis Jr. 11] Davis Jr., C. A., Pappa, G. L., de Oliveira, 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)

[Eftelioglu 15] Eftelioglu, E.: Geometric Median, in *Encyclopedia of GIS*, Springer International Publishing, 10 February 2016 edition (2015)

[Hecht 11] Hecht, B., Hong, L., Suh, B., and Chi, E. H.: Tweets from Justin Bieber's Heart: The Dynamics of the "Location" Field in User Profiles, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 237–246 (2011)

[Hubeny 54] Hubeny, K.: Zur Entwicklung der Gauss'schen Mittelbreitenformeln, *Österreichische Zeitschrift für Vermessungswesen*, Vol. 42, No. 1, pp. 8–17 (1954)

[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)

[Kinsella 11] Kinsella, S., Murdock, V., and O'Hare, N.: "I'm Eating a Sandwich in Glasgow": Modeling Locations with Tweets, in *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, pp. 61–68 (2011)

[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 12a] Li, R., Wang, S., and Chang, K. C.-C.: Multiple Location Profiling for Users and Relationships from Social Network and Content, *Proceedings of the VLDB Endowment*, Vol. 5, No. 11, pp. 1603–1614 (2012)

[Li 12b] 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 *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1023–1031 (2012)

[松本 05] 松本 康：都市度と友人関係, 社会学評論, Vol. 56, No. 1, pp. 147–164 (2005)

[McGee 11] McGee, J., Caverlee, J. A., and Cheng, Z.: A Geographic Study of Tie Strength in Social Media, in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pp. 2333–2336 (2011)

[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)

[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)

[奥村 12] 奥村 学：マイクロブログマイニングの現在, 電子情報通信学会技術研究報告. NLC, 言語理解とコミュニケーション, Vol. 111, No. 427, pp. 19–24 (2012)

[Rout 13] Rout, D., Bontcheva, K., Preoiuc-Pietro, D., and Cohn, T.: Where's @wally?: A Classification Approach to Geolocating Users Based on their Social Ties, in *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, pp. 11–20 (2013)

[Sadilek 12] Sadilek, A., Kautz, H., and Bigham, J. P.: Finding Your Friends and Following Them to Where You Are, in *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, pp. 723–732 (2012)

[Takhteyev 12] Takhteyev, Y., Gruzd, A., and Wellman, B.: Geography of Twitter Networks, *Social Networks*, Vol. 34, No. 1, pp. 73–81 (2012)

[Vardi 00] Vardi, Y. and Zhang, C.-H.: The multivariate L1-median and associated data depth, *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 97, No. 4, pp. 1423–1426 (2000)

[山口 13] 山口 祐人, 伊川 洋平, 天笠 俊之, 北川 博之：ソーシャルメディアにおけるローカルイベントを用いたユーザ位置推定手法, 情報処理学会論文誌：データベース, Vol. 6, No. 5, pp. 23–37 (2013)

[Yamaguchi 15] Yamaguchi, Y., Yoshida, M., Faloutsos, C., and Kitagawa, H.: Patterns in Interactive Tagging Networks, in *Proceedings of the 9th International AAAI Conference on Web and Social Media*, pp. 513–522 (2015)

〔担当委員：奥 健太〕

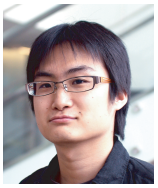
2016 年 5 月 10 日　受理

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．同年，同大学院工学研究科情報・知能工学専攻博士前期課程進学．

著　者　紹　介



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．

著　者　紹　介

岡部　正幸（正会員）

2001 年東京工業大学大学院総合理工学研究科知能システム科学専攻博士課程修了．博士（工学）．同年，科学技術振興機構（CREST）研究員．2003 年豊橋技術科学大学情報メディア基盤センター助手，2007 年同助教．2016 年県立広島大学経営情報学部経営情報学科講師．知的情報検索，インタラクティブデータマイニングに関する研究に従事．

著　者　紹　介

梅村　恭司

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，ACM の各会員．