2H5-5in2

日本における居住地推定のための ソーシャルネットワーク作成方法の調査

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

廣中 詩織 *1 Shiori Hironaka　吉田 光男 *1 Mitsuo Yoshida　岡部 正幸 *1 Masayuki Okabe　梅村 恭司 *1 Kyoji Umemura

*1豊橋技術科学大学
Toyohashi University of Technology

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. 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 the undirected relationship of following either or both improves the recall relative to mutual relationship, which is reported in previous studies.

1. はじめに

Twitter にはフォローしているや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を利用し、Twitter ユーザの居住地を推定する研究がある [Rout 13, McGee 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と先行研究 [McGee 13] で示されているが、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、ソーシャルネットワークの作成方法または居住地推定手法を変えて、居住地推定をおこなうシステムを作成する。このシステムにより、日本の Twitter データから作成したソーシャルネットワークに複数の居住地推定手法を適用した結果を評価する。さらに、利用する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利用し、居住地推定に最も有効なユーザ間の関係を特定する。先行研究 [McGee 13] で利用されている相互フォロー関係から作成したソーシャルネットワークは、居住地推定において、必ずしも良い性能を発揮しないことを示す。また、繰り返し推定手法を適用するときの性能を明らかにする。

2. データセットの作成

本調査では、ユーザの居住地データと、そのユーザ間の関係から作成したソーシャルネットワークとを利用して実験をおこなう。これらのデータ作成方法の詳細について次節以降で述べる。

2.1 居住地の決定方法

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主に投稿している場所をそのユーザの居住地とする。先行研究では、ユーザの居住地を地理座標にするもの [McGee 13] とエリアにするもの [Davis Jr. 11] とがある。本研究では居住地推定を分類問題とみなし、居住地を市区町村レベルのエリアとする。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を求め、ユーザごとに最もツイート数が多いエリアをそのユーザの居住地とする。

Twitter Streaming API を使用し、2014 年に日本を包含する矩形 *1 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。居住地の信頼度を上げるため、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できた。

2.2 ソーシャルネットワーク作成方法

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とユーザをフォローしているユーザの集合との 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 ユーザを実験に使用する。

Twitter でのフォロー関係をもとにする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。本研究でのソーシャルネットワークは、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。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する 4 種類のソーシャルネットワークを作成する。

2.3 ソーシャルネットワークの特徴

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違いを調べるため、基本的なグラフの統計量を調べる。調べる統計量は、作成した有向ソーシャルネットワーク $G(V, E)$ のノード数 $|V|$、エッジ数 $|E|$、平均次数 K、孤立ノード数 $|I|$ である。なお、孤立ノードは入次数、出次数ともに 0 のノードである。

連絡先: 廣中 詩織，豊橋技術科学大学，愛知県豊橋市天伯町雲雀ヶ丘 1-1，hironaka15@ss.cs.tut.ac.jp

*1　北緯 20 から 50、東経 110 から 160 の範囲。

表 1: ソーシャルネットワークの統計量

(a) ネットワーク全体

関係	$\|V\|$	$\|E\|$	K	$\|I\|$
followee	62721054	417334528	6.65382	44200
follower	62721054	417334528	6.65382	44200
mutual	22917760	251625205	10.97949	83300
linked	62721054	583043851	9.29582	44200

(b) 部分グラフ

関係	$\|V\|$	$\|E\|$	K	$\|I\|$
followee	472350	8163069	17.28182	44200
follower	472350	8163069	17.28182	44200
mutual	472350	6226387	13.18172	83300
linked	472350	10099751	21.38192	44200

作成した 4 種類のソーシャルネットワークの統計量を調べた結果が表 1(a) である。作成したソーシャルネットワークには、居住地の付いているノードと、そのノードに関係しているために得られた居住地の付いていないノードとがある。ユーザ間の関係を取得したノードは居住地の付いているノードのみであり、居住地の付いていないノードと比較するとエッジの数に違いがある。そこで、居住地の付いているノードのみを取り出して作成した部分グラフの統計量を調べた結果が表 1(b) である。エッジ数は、取得したユーザ間の関係の数である。followee と follower はユーザ間の関係を反対にとらえて作成したソーシャルネットワークであるため、エッジ数が同じになる。エッジ数が最も多いのは linked のソーシャルネットワークである。居住地の付いているノードのみを取り出して部分グラフを作成すると平均次数が高くなる。ソーシャルネットワークを作成するときに、誰もフォローしておらず、誰からもフォローされていないユーザを取り除いていないため、孤立ノードが存在する。

3. 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。ソーシャルネットワークは、ユーザをノード、ユーザ間の関係をエッジとするグラフである。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。3.1 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おこなうため、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を利用する居住地推定手法のうち、3.1 節以降で説明する 4 手法を実装する。

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（学習データから得られる）、$dist(a,b)$ はラベル a とラベル b との間の距離である。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で計算したものである。

3.1 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 [Backstrom 10]。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る。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。

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

$$\gamma(l,u) = \prod_{v \in N_u \cap L} p(dist(l,l_v)) \prod_{v \notin N_u \cap L} [1 - p(dist(l,l_v))]$$

$$ProbabilityModel(u) = \arg\max_{l \in A} \gamma(l,u) \tag{2}$$

尤度 $\gamma(l,u)$ の計算量が大きいため、文献 [Backstrom 10] に書かれているとおり、計算の最適化をする。ノード u の正解ラベルは隣接ノードの持つラベルの中に存在すると報告されているため、ノード u の尤度を最大にするラベルの探索範囲は、隣接ノードのラベルに存在するラベルのみとする。さらに、$\gamma(l,u)$ の代わりに、式変形した $\gamma'(l,u)$ を用いる。最終的に推定に使用するのは式 (3) である。

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

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

$$ProbabilityModel'(u) = \arg\max_{l \in \{l_n | n \in N_u \cap L\}} \gamma'(l,u) \tag{3}$$

推定にはパラメータ a、b、c が必要である。実験の際のパラメータには、文献 [Backstrom 10] に書かれている値、$a = 0.0019$、$b = 0.196$、$c = -1.05$ を使う。

3.2 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。隣接ノードの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選択する。この手法を表現したものが式 (4) である。

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

$$MajorityVote(u) = \arg\max_{l \in S_u} |\{n | n \in L, l = l_n\}| \tag{4}$$

この手法には、ノードの隣接ノード数の取りうる範囲、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。文献 [Davis Jr. 11] で示されているパラメータであるユーザの持つ友人の数は、推定対象ノードの隣接ノード数と考える。今回の実験ではパラメータを設定しないため、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 である。

3.3 Geometric Median

Geometric Median は、推定対象のノードの隣接ノードの中から地理的な中央値に近いノードのラベルを選択し、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 (5) である。

$$GeometricMedian(u)$$

$$= \arg\min_{l \in \{l_n | n \in N_u \cap L\}} \sum_{x \in N_u \cap L, n \neq x} dist(l,l_x) \tag{5}$$

3.4 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 Random Neighbor を用いている。本研究の 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。この手法を表現したものが式 (6) である。ここで、$choice(S)$ は集合 S からランダムに要素をひとつ選択する関数である。

$$RandomNeighbor(u) = l_{choice(N_u \cap L)} \tag{6}$$

4. 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 交差検証により推定環境が最も良いときの性能を検証し、10 分割交差検証により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解したユーザの割合、再現率はテストデータのうち正解したユーザの割合、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を用いる。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。式 (7) に 4 つの評価指標を示す。ここで、X は推定したノードの集合、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。

$$\begin{aligned}
Precision(T, X) &= \frac{|\{u | u \in X \cap T, l_u = e_u\}|}{|X \cap T|} \\
Recall(T, X) &= \frac{|\{u | u \in X \cap T, l_u = e_u\}|}{|T|} \\
F1(T, X) &= \frac{2 * Precision(T, X) * Recall(T, X)}{Precision(T, X) + Recall(T, X)} \\
Coverage(T, X) &= \frac{|X \cap T|}{|T|}
\end{aligned} \tag{7}$$

4.1 隣接ノードのみを利用する手法での性能評価

先行研究でよく用いられる mutual のソーシャルネットワークを利用し、居住地推定手法の推定性能を比較した結果が表 2 である。日本のソーシャルネットワークでは Majority Vote が最も精度よく居住地を推定できることが分かる。先行研究 [Jurgens 15] でも Majority Vote だけ異なる性質の結果が出ていることから、Majority Vote の性能には多数決の際に利用するエリアの分割方法が関係している可能性がある。

最も性能が良かった Majority Vote を用いて、利用するユーザ間の関係を変えて作成した 4 種類のソーシャルネットワークでの推定性能を評価した結果が表 3 である。follower を利用すると適合率が高くなり、linked を利用すると再現率が高くなる。カバー率をみると、推定できるユーザの数が最も多い関係は linked である。推定できるユーザの数が多くなる linked を利用すると再現率が上がり、F 値も高くなることが明らかになった。

4.2 ラベルを伝搬させた場合の性能評価

Spatial Label Propagation (SLP) [Jurgens 13] は隣接ノードのみを利用する手法を複数回適用することでラベルを伝搬させ、多くのノードを推定できるようにする手法である。SLP を用いると、隣接していないノードのもつラベルも利用して居住地を推定することができ、居住地推定に適するユーザ間の関係が変化すると考えられる。そのため、SLP でも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を利用するときの推定性能を比較する。

表 2: 居住地推定手法の性能比較

(a) leave-one-out 交差検証

推定手法	適合率	再現率	F 値
ProbabilityModel	0.12146	0.10004	0.10972
MajorityVote	0.31213	<u>0.25709</u>	0.28195
GeometricMedian	0.27386	0.22557	0.24738
RandomNeighbor	0.19675	0.16205	0.17772

(b) 10 分割交差検証

推定手法	適合率	再現率	F 値
Probability Model	0.12660	0.10288	0.11351
Majority Vote	<u>0.30717</u>	<u>0.24960</u>	<u>0.27541</u>
Geometric Median	0.27130	0.22045	0.24325
Random Neighbor	0.19692	0.16002	0.17656

表 3: 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での居住地推定性能の比較

(a) leave-one-out 交差検証

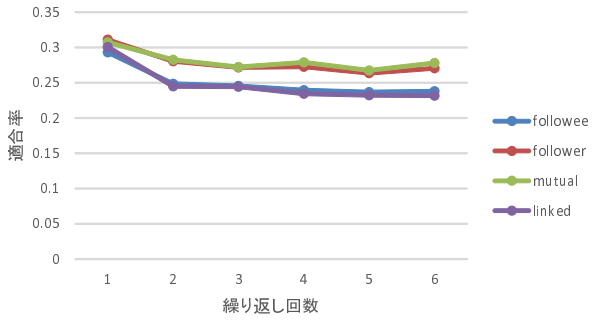
関係	適合率	再現率	F 値	カバー率
followee	0.29807	0.26361	0.27978	0.88437
follower	0.31614	0.27208	0.29246	0.86062
mutual	0.31214	0.25709	0.28195	0.82365
linked	0.30581	<u>0.27719</u>	0.29080	0.90643

(b) 10 分割交差検証

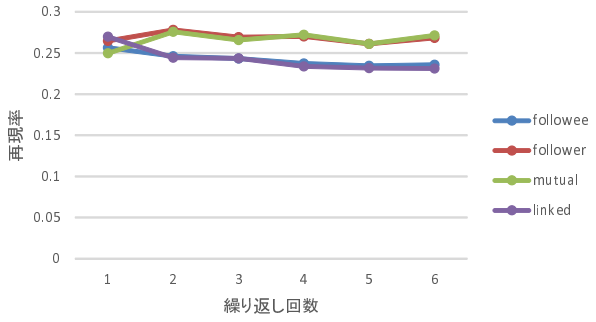
関係	適合率	再現率	F 値	カバー率
followee	0.29303	0.25644	0.27352	0.87514
follower	<u>0.31109</u>	0.26461	<u>0.28597</u>	0.85061
mutual	0.30717	0.24960	0.27541	0.81259
linked	0.30058	<u>0.26997</u>	0.28446	<u>0.89817</u>

隣接ノードのみを利用する手法の中で最も性能が良かった Majority Vote を推定関数とする SLP で、4 種類のユーザ間の関係から作成したソーシャルネットワークの比較をする。図 1 に 10 分割交差検証での繰り返し回数ごとの評価結果を示す。図 1(a) より、繰り返すことで適合率が低くなることが分かる。図 1(b) をみると、繰り返すことで、follower と mutual の再現率が高くなることが分かる。図 1(c) は繰り返したときに F 値が高い関係は follower と mutual であることを示している。カバー率は繰り返し回数が増えると高くなるが、繰り返し回数が 2 回の時点でほぼ上限になる。関係 follower から作成したソーシャルネットワークを用いて、繰り返し推定しないときに最も F 値が高くなっている。

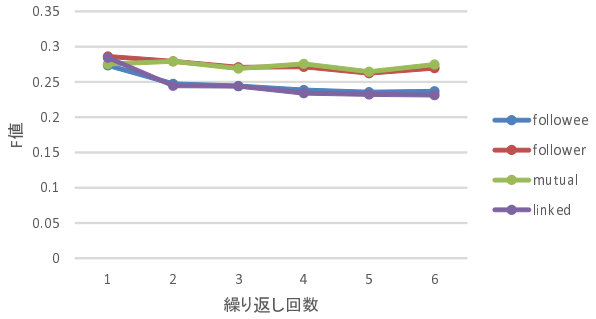
SLP で性能が良くなることの前提となっている仮説は、推定した居住地も学習データに含まれる居住地並みに信頼できるというものである。繰り返したときに性能が良くならないため、推定した居住地が学習データの居住地ほど信頼できないと考えられる。また、繰り返すことで再現率が上がることが示されている。適合率が高いほど信頼性が高くなるため、繰り返す際に性能が良くなると考えられる。一般的には適合率と再現率



(a) 適合率



(b) 再現率



(c) F 値

図 1: SLP の繰り返し回数による推定性能の推移

はトレードオフであるが、適合率が上がるようにパラメータを設定することで、性能を上げることができると考えられる。

4.3 推定に成功するノードの特徴

Majority Vote は、隣接ノード数のしきい値を決めることで適合率を上げることができる。Majority Vote 以外の隣接ノードのみを利用する手法でも同様のしきい値を決めることができる。Majority Vote を用いた leave-one-out 交差検証での推定結果から、一部のノードを取り出して評価したとき、推定性能が変化することを調べる。

隣接ノード数が 200 から 800 のノードのみを取り出して評価した結果が表 4 である。隣接ノード数にしきい値を設けると、適合率が上がり、再現率とカバー率が低下することが分かる。そのため、しきい値を設定すると SLP の推定性能向上が期待できる。

表 4: MajorityVote で隣接ノード数にしきい値を設けたときの推定性能

関係	適合率	再現率	F 値	カバー率
followee	0.36158	0.14998	0.21202	0.41481
follower	0.38615	0.14811	0.21410	0.38356
mutual	0.38610	0.11809	0.18086	0.30585
linked	0.36866	0.16818	0.23098	0.45619

5. おわりに

日本のソーシャルネットワークを用いてソーシャルネットワークを利用する推定手法の統一的な評価をおこない、Majority Vote を使うと最も精度よく居住地を推定できることを示した。さらに、ソーシャルネットワーク作成に利用するユーザ間の関係を変えることで、推定性能が変化することを示した。フォローされている関係である follower を利用して作成したソーシャルネットワークでは適合率が高くなり、どちらかまたは両方がフォローしている関係である linked から作成したソーシャルネットワークでは再現率が高くなることを明らかにした。また、応用的な手法である SLP を用いて居住地推定をするときは、推定するノードを制限し適合率を上げることで性能向上を見込めることが分かった。

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)

[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)

[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* (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)

[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)

[Rout 13] Rout, D., Bontcheva, K., Preoţiuc-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)