ARG WI2 No.14, 2019
WI2-2019-10

ソーシャルグラフにおけるユーザの中心性と居住地推定の難しさとの関係

廣中 詩織　　吉田 光男　　梅村 恭司

豊橋技術科学大学

s143369@edu.tut.ac.jp yoshida@cs.tut.ac.jp umemura@tut.jp

概要　ユーザの居住地などの属性は，ニュース推薦や広告提供など様々なアプリケーションで必要とされている．しかし，現実のユーザに関わる属性は不明なことが多く，ユーザ間の関係などから推定する必要がある．これまでの研究において，多くのユーザと関わりのある有名人など，居住地を推定しにくいユーザが存在しており，推定しにくいユーザは一部の有名人のみであることがわかっている．我々は，ソーシャルグラフから計算したユーザの中心性をもとに，どのようなユーザが友人と同じ居住地を持つのかを分析した．その結果，PageRank と HITS で計算した中心性の値はユーザが友人と同じ居住地を持つ傾向と関連していること，高い HITS スコアを持つユーザはあまり友人と居住地を共有していないことがわかった．この結果は，居住地を推定しにくいユーザとして，多くの有名人をフォローしている Hub ユーザと有名人である Authority ユーザとの 2 種類がいることを示している．

キーワード　居住地推定，ソーシャルグラフ，中心性

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，トレンド検出 [1] やニュース推薦 [2, 3] など様々なアプリケーションで活用されている．このようなアプリケーションでは，居住地や年齢，職業などのユーザ属性が用いられており，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 [4, 5]．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，多くのユーザとメンションをしている一部の有名なユーザなど，居住地推定のうまくいかないユーザが存在することがわかっている [6]．そのような有名人に着目した分析では，すべての有名人のうち，メンション相手が広範囲に分散している global celebrity だけが居住地を正しく推定されにくいと報告されている [7]．我々は，有名人以外にも友人と同じ居住地を持ちにくく，居住地を推定しにくいユーザがおり，そのようなユーザは何らかのネットワーク的特徴を持つと考える．本論文では，ネットワーク的特徴を測るために，ノードの様々な重要度を比較・評価するための指標である中心性を用いる．次数中心性や PageRank [8], HITS [9] などのネットワーク構造をもとに計算される中心性は，周囲と似た属性値を持っている割合と関係があると考える．

本論文では，様々な中心性の値と，そのユーザが友人と同じ居住地を持つかどうかのあいだに，関連があるのかを分析する．その結果，多くの友人と同じ居住地を持つユーザは，PageRank と，HITS の Authority と Hub スコアの分布に違いがあり，PageRank と HITS の Authority と Hub の値が高いユーザは居住地を推定することが難しいことがわかった．加えて，友人の多数と同じ居住地を持たないユーザとして，Authority と Hub との 2 種類のユーザが存在することがわかった．

2　データ

本節では，分析に用いるユーザの居住地データと，ソーシャルグラフデータの作成方法を説明する．データの作成には先行研究 [10] と同様の方法を用いる．

2.1　居住地

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．ユーザに居住地を付与するために，2014 年 1 月 1 日から 12 月 31 日のあいだの日本を包含する矩形[1]内での位置情報付きツイートを Streaming API[2] により 140,055,452 件集めた．そして，総務省統計局の境界データを用いて，それらのツイートに付与されている緯度経度を含んでいる日本の市区町村（エリア）を照合した．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．境界データでは日本の市区町

[1] 北緯 20 から 50，東経 110 から 160 の範囲．

[2] `https://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter.html` (viewed 2019-05-13)

村は 1901 種類出現しているが，最終的な居住地データには 1873 種類が出現していた．

2.2 ソーシャルグラフ

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の集合とフォローされているユーザの集合を 2015 年 7 月に収集した．これらのデータを用いて，ユーザ A がユーザ B をフォローしているときに，ユーザ A からユーザ B の方向へ向きを持つエッジを作成することで，有向グラフを構築する．

最終的に，471,761 ノードと 8,295,355 エッジを含むソーシャルグラフができた．居住地を付与していないユーザはソーシャルグラフから除外したため，ソーシャルグラフのノードはすべて居住地が付与されているユーザとなっている．各ノードは平均 17.58 の入・出エッジを持ち，平均相互フォロー数は 13.2 であった．ソーシャルグラフには，出エッジを 1 つも持たないユーザが 52,416 ユーザ存在した．

3 分析方法

本節では分析方法の説明をする．まず，用意したソーシャルグラフを用いて各ユーザの中心性の値を計算する．次に，友人と同じ居住地をもとに，ユーザを 3 つのグループに分類する．そして，グループごとに，ユーザ全体と比べたときの中心性の値の偏り度合いを計算する．

3.1 中心性指標

中心性として，入・出次数中心性, PageRank [8], HITS [9] の Authority と Hub を用いる．入次数中心性は，フォロワー数が多いユーザほど大きな値を持ち，出次数中心性はフォロー数が多いユーザほど大きな値を持つ．PageRank は入次数中心性に近く，多くのフォロワーを持つユーザほど高い値を持つ傾向にある．しかし，多くのフォロワーを持つユーザだけではなく，多くのフォロワーを持つユーザにフォローされているときにより大きな値を持つ．HITS は，多くの有名人 (Authority) をフォローしている Hub となるユーザと多くのユーザにフォローされている Authority となるユーザがいると仮定し，Authority と Hub のスコアを計算する．多くの有名人をフォローしているユーザだと Hub が，多くのユーザにフォローされているユーザだと Authority の値が高くなりやすい．PageRank と HITS の値は 2.2 節で構築したソーシャルグラフ上で計算した．ライブラリ NetworkX[3] の `networkx.pagerank_scipy`，`networkx.hits_scipy` と，デフォルトパラメータを用いて計算した．

3.2 ユーザグループ

Davis Jr. らの提案した居住地推定手法 [11] がうまくいくかどうかを，友人の居住地との類似度の指標とする．この推定手法は，友人の持つ居住地の中で最も出現頻度の高いものをユーザの居住地と推定する．この手法による推定結果を用いて，次の 3 種類にユーザを分類する：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，手がかりがなく居住地を推定することができなかったユーザ．これらユーザはすなわち，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，手がかりがなく類似度を測ることができないユーザ，となる．

本論文では相互フォローであるユーザを友人とみなす．すなわち，2.2 節で構築したソーシャルグラフのエッジのうち，相互フォローの部分のみを取り出したグラフを用いる．そして，leave-one-out 交差検証による推定結果によりユーザをグループに分割する．

3.3 偏り度合いの計算方法

ユーザ集合 U に対するある中心性の値の分布を次のように計算する．ユーザ総数を $|U| = N$，ユーザの中心性の値がある区間 $i : [x_i, x_{i+1})$ に含まれるユーザの数を n_i とする．そのとき，ユーザ全体のうち中心性の値が区間 i の中に存在しているユーザの割合は n_i/N である．縦軸をその割合，横軸を区間としてプロットした，$f(i) = n_i/N$ をスコア分布とする．$\sum_i n_i/N = 1$ となるように区間を決めた．

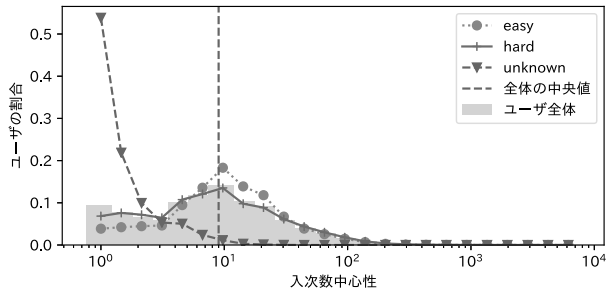
我々は，各ユーザグループに対してスコア分布を計算する．そして，ユーザ全体に対しての各グループの偏り度合いを調べるために，差を次の方法で計算する．ユーザ全体のうち区間 i に存在するユーザの割合を x，比較するユーザユーザグループのうち区間 i に存在する割合を y とするとき，全体の分布との偏り度合いを $\log_{10}(y/x)$ と計算する．全体の分布でその区間にいるユーザの割合が多いとき負の値になり，少ないとき正の値になる．比較した分布間で差が大きいほど値の絶対値が大きくなる．

4 結果と考察

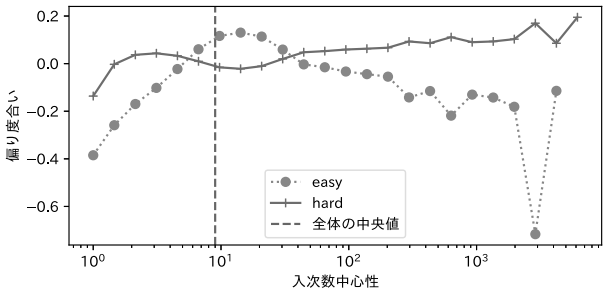
居住地を付与したユーザすべてを 3 つのグループに分類した．その結果，友人の多数と同じ居住地を持つのは 121,275 ユーザ（easy），友人の多数と異なる居住地を持つのは 267,809 ユーザ（hard），友人に居住地を持つユーザが存在しなかったのは 82,677 ユーザ（unknown）であった．そして，ソーシャルグラフに含まれる全ユーザと各グループとのスコア分布と，各グループの偏り度合いを計算した[4]．入・出次数中心性, PageRank, HITS の Authority と Hub の結果を図 1 と図 2 に示す．

[3] `https://networkx.github.io/` (viewed 2019-04-15)

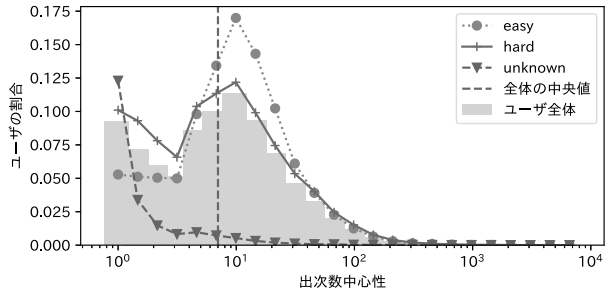
[4] unknown は類似度が測れなかったユーザであるため，除外した．



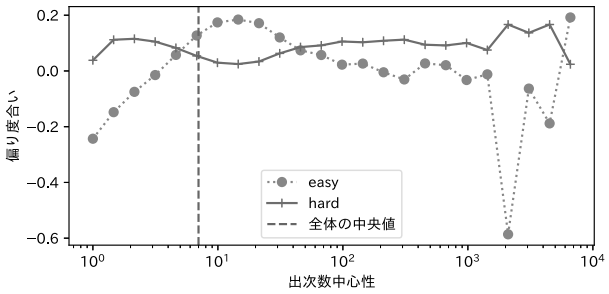
(a) スコア分布（入次数中心性）



(b) 偏り度合いの分布（入次数中心性）



(c) スコア分布（出次数中心性）



(d) 偏り度合いの分布（出次数中心性）

図 1 入・出次数中心性の結果

図 1 にある次数中心性の結果を見ると，easy と hard の分布の中心位置にはあまり差が見られない．各グループに含まれるユーザの全体に対する割合で見ると，値が 20 あたりで easy の分布する割合が増えていることがわかる．中心性の値が大きいほど easy のユーザが減っているため，フォローしているユーザ数やフォロワー数が多いユーザを有名人だと考えると，有名人は居住地を正しく推定することが難しいといえる．この結果は，Davis Jr. ら [11] の相互フォロー数が 20 以上 200 以下のユーザのみを利用したときに最も適合率が高くなるという報告とも矛盾しない．しかし，相互フォロー数の数え方が本論文と異なる可能性があるため，数値を単純に比べることはできない．

PageRank と HITS の結果を図 2 に示す．easy のユーザは hard のユーザより高い PageRank の値を持つ傾向にあることがわかる．HITS は Authority と Hub との両方の結果で，easy と hard とのスコア分布のピークの位置に差があるように見える．これらは中央値を中心として，hard の分布が中心性の値の大きい右側に，easy の分布が値の小さい左側に位置している．easy と hard のユーザについてみると，HITS で得られた結果のほうが PageRank と比べて山の重なりが小さいため，Twitter ユーザが友人と居住地を共有する傾向があるかは HITS の値をみるほうがよくわかると考えられる．この結果は，多くの有名人をフォローする Hub となるユーザと，有名である Authority のユーザとの 2 種類のタイプが存在することを示唆していると考えられる．HITS の両方のスコアの結果で，偏り度合いの分布を見ると，スコアが高いほど easy のユーザが少ない．この結果から，Hub と Authority の両方のスコアが高いユーザだけでなく，どちらか片方のスコアが高いユーザは友人と居住地を共有している割合が少ないことがわかった．

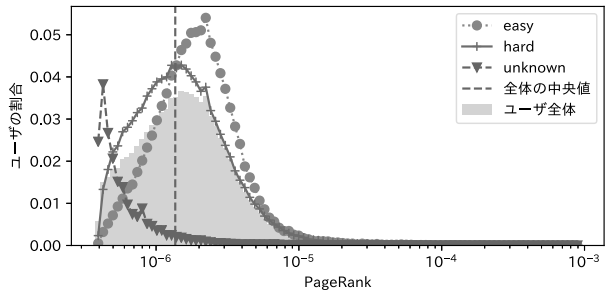
5 おわりに

本論文では，入・出次数中心性, PageRank, HITS を用いて，友人と同じ居住地を持つかどうかとそれら中心性との関係を調べた．その結果，PageRank と HITS の Authority または Hub の値が大きいユーザは，多数の友人と同じ居住地を持っていないことがわかった．つまり，居住地の推定が難しいことが明らかになった．加えて，Twitter には多くの有名人をフォローする Hub となるユーザと，有名である Authority のユーザとの 2 種類のタイプが存在することがわかった．

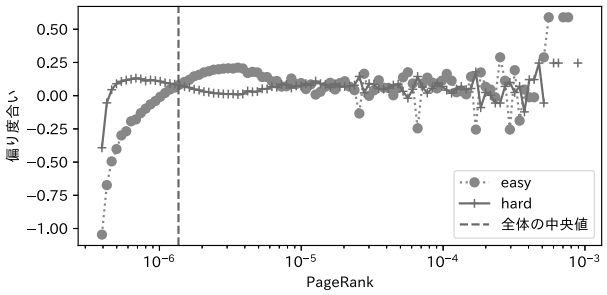
参考文献

[1] Benhardus, J. and Kalita, J.: Streaming Trend Detection in Twitter, *International Journal of Web Based Communities*, Vol. 9, No. 1, pp. 122–139 (2013).

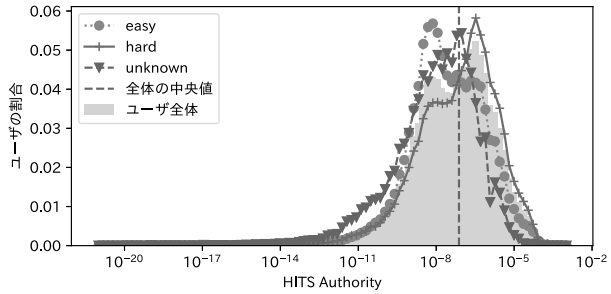
[2] Jonnalagedda, N. and Gauch, S.: Personalized News Recommendation Using Twitter, *Proceedings of the 2013 IEEE/WIC/ACM International Joint Confer-*



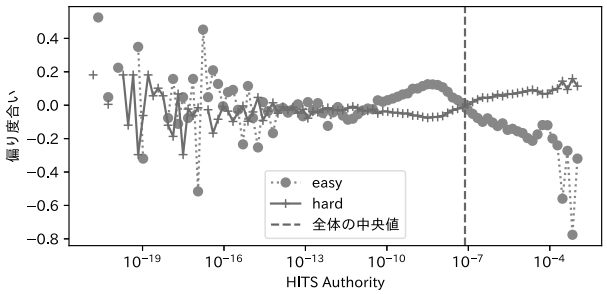
(a) スコア分布（PageRank）



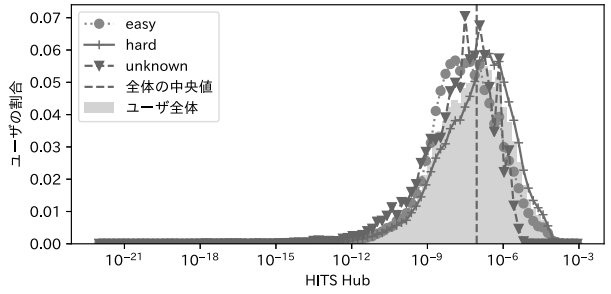
(b) 偏り度合いの分布（PageRank）



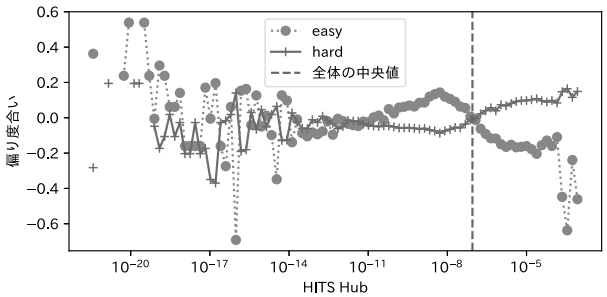
(c) スコア分布（HITS の Authority）



(d) 偏り度合いの分布（HITS の Authority）



(e) スコア分布（HITS の Hub）



(f) 偏り度合いの分布（HITS の Hub）

図 2 PageRank と HITS の結果（PageRank の最小値はグラフに含まれていない）

ences on Web Intelligence and Intelligent Agent Technologies, pp. 21–25 (2013).

[3] Phelan, O., McCarthy, K. and Smyth, B.: Using Twitter to Recommend Real-Time Topical News, *Proceedings of the 3rd ACM Conference on Recommender Systems*, pp. 385–388 (2009).

[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, *Proceedings of the 9th International AAAI Conference on Web and Social Media*, pp. 188–197 (2015).

[5] 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).

[6] Rahimi, A., Cohn, T. and Baldwin, T.: Twitter User Geolocation Using a Unified Text and Network Prediction Model, *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).

[7] 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).

[8] Page, L., Brin, S., Motwani, R. and Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web., Technical report, Stanford InfoLab (1999).

[9] Kleinberg, J. M.: Authoritative sources in a hyperlinked environment, *Journal of the ACM*, Vol. 46, No. 5, pp. 604–632 (1999).

[10] 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. WII–M_1–11 (2017).

[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).