居住地推定法に基づいた
ソーシャルグラフに関わるプロパティの分析

(Analysis of Social Graph Properties for Home Location Estimation)

2021 年 1 月

博士（工学）

廣中 詩織

豊橋技術科学大学

別紙４－２（課程博士（和文））

2021 年　1 月　8 日

情報・知能工学専攻	学籍番号	第 143369 号
氏名	廣中　詩織	

指導教員	梅村　恭司 北崎　充晃

論文内容の要旨　(博士)

博士学位論文名	居住地推定法に基づいたソーシャルグラフに関わるプロパティの分析

(要旨　1,200 字程度)

人間社会では、似た属性を持つ人とのつながりを持ちやすいことが知られている。ソーシャルメディアが広く使われるようになり、人々の行動がソーシャルメディアを通じて観測できるようになってきた。多くの人々に利用されているソーシャルメディアは、現実の社会を観測し分析するために利用される。様々な分析をする際にはユーザ属性が利用されるが、一部のユーザ属性は欠損していることが多いため、他の情報から推定する必要がある。ソーシャルグラフはソーシャルメディア上のユーザ間の関係をもとに構築することができるが、このとき似た属性を持つユーザ同士のつながりをもとにしたソーシャルグラフを用いると、居住地などのユーザ属性を推定することができる。

ソーシャルグラフは現実の社会を反映していると考えられるが、ソーシャルグラフの性質は明らかではないため、実際のデータをもとにソーシャルグラフの性質を調べるアプローチが必要である。ソーシャルグラフを用いた居住地推定では、ソーシャルグラフの持つ性質により推定性能が変化する。本論文では、ユーザ属性とソーシャルグラフとの関係に着目し、居住地推定を通じてソーシャルグラフのプロパティを分析する。

ソーシャルメディア上でのユーザ間の関係には多くの種類があることから、居住地推定に適したユーザ間の関係の特定に取り組んだ。その結果、居住地推定に用いるソーシャルグラフを構築するために利用するユーザ間の関係について、向きを考慮することで推定性能が向上することを明らかにした。

ソーシャルグラフ上において、ユーザに紐付いている居住地ラベルがグラフ上でどのように分布しているかは、ラベルをどのように伝搬させていくと推定がうまくいくかに関係している。そこで、ソーシャルグラフを構成するノードが持つ居住地ラベルがどのような分布をしているのかを分析した。その結果、88%のユーザは同じラベルを持つユーザが1ホップ以内に存在することを明らかにした。

ソーシャルグラフはユーザが他者と交流する過程で構築されていくものであるため、ソーシャルグラフの形状に関わるプロパティは、ユーザの特徴と関係していると考えられる。そこで、居住地推定が困難なユーザの持つノードのプロパティである、プロフィール属性を分析した。分析に使用した属性は、ユーザ名や自己紹介文などプロフィールのテキストの文字数、グラフの次数に関連するフォロー数、フォロワー数、フォロー／フォロワー比、アクティビティを示すいいね数、総ツイート数、1日あたりのツイート数、公開リストに入れられている数、アカウント作成日からの日数である。また、他ユーザとつながっている度合いを測る中心性も、ノードのプロパティであるため、中心性についても分析した。分析には、ソーシャルグラフの次数、PageRank、HITSのAuthorityとHubを用いた。

本論文では、日本のTwitterユーザによるソーシャルグラフの性質についての発見をまとめた。本論文は日本周辺で投稿された1年分の位置情報付きツイート140,055,452件をもとに大規模な分析をおこなったものである。

別紙４－１（課程博士（英文））

Date of Submission（month day, year）： January 8, 2021

Department of Computer Science and Engineering	Student ID Number	D143369
Applicant's name	Shiori Hironaka	

Supervisors	Kyoji Umemura Michiteru Kitazaki

Abstract（Doctor）

Title of Thesis	Analysis of Social Graph Properties for Home Location Estimation

Approx. 800 words

People tend to interact with others who have similar attributes. Social media, which is widely used worldwide, can be used to analyze real-world social behaviors. Users' attributes are used in the analysis; however, because certain user attributes are not open to the public, it is necessary to estimate them using other sources of information. A social graph is constructed based on the relationships among users on social media. As we use the social graph based on the relationships among users with similar attributes, we can estimate user attributes such as home location.

While a social graph is considered to reflect the real world, the properties of the social graph are not clearly known. These properties need to be analyzed using data that represent the real world. The performance of social graph-based home location estimation varies based on social graph properties. In this thesis, we analyzed social graph properties using home location estimation, which is based on user attributes and social graphs.

There are several types of relationships between users on social media, however the estimation performance of each relationship is unclear. Therefore, we conducted a study to identify the relationship between users, which is helpful for home location estimation. Based on the results, we observed that the estimation performance can be improved by considering the direction of the relationships.

The distribution of the location labels associated with the users on the social graph is related to the success ratio of the estimation, and we analyzed the distribution of home locations on the social graph. From the results, it was observed that 88% of the users had the same home location within one hop (friends and friends of friends).

A social graph is constructed while interacting with others, and its properties are related to user characteristics. We analyzed users whose home locations were difficult to estimate. We focused on the user profile attributes, which is a subset of the social graph properties, and we analyzed the relationship between the degree of difficulty of estimation and user profile attributes. We employed the following profile attributes: length of the profile text, such as name or description; attributes related to the degree of the graph, such as the number of followings and followers or follow ratio; and activity measures, such as the number of likes,

average number of tweets per day, number of lists, or number of days since the account was created. We also conducted an analysis using centrality, which measures the connectivity of other users. We employed the following centralities: the in-/out-degree centrality, PageRank, and Authority and Hub scores of the HITS algorithm.

In this thesis, we summarize our findings on the properties of the Twitter social graph. This was a large-scale analysis based on 140,055,452 geo-tagged tweets posted throughout Japan in 2014.

目次

図目次

表目次

第 1 章

序論：本論文の枠組み

社会的な生物である人間は、相互に影響を与え合いながら暮らしている。人間と人間とのあいだの関係には、家族や友人、恋人、同僚など様々なものがある。関係がある人間同士はその人の持つある面（属性）を共有しているといえる。同僚は同じ会社に勤めていたり、家族は一緒に住んでいることが多かったり、学友は同じ学校に通っていたりする。住んでいる場所や通っている大学などをその人の属性ととらえると、似た属性を持つ人同士のつながりというものは人間社会に多く存在する。

人間同士の関係を分析するために、人間をノード、人間と人間とのあいだの関係をエッジとして表現したソーシャルグラフが用いられる。ソーシャルグラフとしてデータを表現することで、人が持つ属性と人間同士の関係とを調べることができる。多くの人々が日常的にソーシャルメディアを利用するようになってきたことから、ソーシャルメディア上での人間関係がオンラインソーシャルグラフとして観測され、分析に用いられるようになってきた。ソーシャルメディア上での人間関係はユーザの持つ現実世界での人間関係をベースにしていることも多く、オンラインソーシャルグラフと現実世界の人間関係には関連がある。本論文では、基本的にソーシャルグラフと呼ぶ際、オンラインソーシャルグラフのことを指す。

インターネット上、特にソーシャルメディア上において、人々はユーザとして認識される。ソーシャルメディア上では、ユーザ同士が様々な交流をしており、様々な関係が築かれている。ソーシャルメディアのデータは社会の一部を反映しているため、様々な研究開発に用いられている。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。このような研究開発においては、ユーザの属性を組み合わせることでより詳細な分析がおこなえる。わたしたちが普段利用している Web サービスにおいても、様々な場所でユーザごとのパーソナライズがおこなわれており、パーソナライズにはユーザ属性が利用されている。ユーザ属性は様々な場面で必要とされるものであるが、一部のユーザの属性は欠損していることが多いため、欠損したユーザ属性をほかの情報から推定する取り組みがなされ

ている。

似た属性を持つオンライン上でのユーザ同士のつながりを用いると、ユーザ属性を推定できる。例えば、現実の友人関係をベースとしたソーシャルメディアである場合、住んでいる場所の近いユーザ同士や、近い年齢のユーザ同士がつながっていることになり、ユーザのおおまかな位置や年齢の推定に利用できる。また、ニュースを購読するためソーシャルメディアを利用している場合でも、住んでいる地域に関するニュースがよく読まれると考えられ、ユーザの位置の推定に利用できる可能性がある。このように、オンライン上のユーザ間の関係から属性を推定できる可能性があるが、ユーザ間の関係により推定できる属性が異なる。

ソーシャルメディア上には様々なユーザ間の関係が存在するため、ソーシャルグラフを用いてユーザ属性を推定することを考えると、それぞれのユーザ間の関係が属性推定においてどのような働きをするのか調べる方法が必要である。ソーシャルグラフがどのようにできるのかを定義したモデルを作り、ソーシャルグラフの生成をシミュレーションするアプローチも考えられるが、実際のデータをもとに性質を調べるアプローチも必要である。本研究では、ユーザ属性とソーシャルグラフとの関係に着目して、ソーシャルグラフのプロパティを分析する。

ソーシャルメディア上の様々な関係をもとにソーシャルグラフを作ることができるが、ユーザ間の関係の種類によってどのようなソーシャルグラフとなるかが変わる。関係の種類によりソーシャルグラフの形状が異なるため、その関係が属性推定に適したユーザ間の関係であるのかを調べる必要がある。第 3 章では、ソーシャルグラフの構築に用いられるユーザ間の関係が推定性能に与える影響について分析した。

ソーシャルグラフ上にユーザの属性（ノードのラベル）がどのように位置しているかによって、属性を推定する方法が変わってくる。ソーシャルグラフ上の属性の分布によって、直接つながりがある部分グラフを使うだけで予測できるのか、ソーシャルグラフ全体を使うことで性能を向上させられるのかを調べる必要がある。第 4 章では、同じ値を持つラベル間の距離について分析した。

ユーザが構築した他ユーザとの関係と、他ユーザからそのユーザがどのような関係を持たれているかによって、ソーシャルグラフの形状が決まると考えられる。すなわち、推定性能に影響するソーシャルグラフの形状は、関係を作ったユーザのプロパティと関連していると考えられる。第 5 章では、推定対象のユーザがどのような関係を築いているかを、ソーシャルメディアのプロフィールをもとに分析した。また、第 6 章では、グラフネットワークの一般的な解析手法を用いて、推定対象のユーザがソーシャルグラフ上でどのような特徴を持ったユーザかを分析した。

第 2 章

ソーシャルグラフとプロパティ

本章では、属性推定に必要な、ソーシャルメディアと関連する諸概念について記述する。ソーシャルメディアとは、ユーザが情報を投稿・閲覧し、交流をするような双方向のメディアである。平成 30 年版情報通信白書では、「ブログ、ソーシャルネットワーキングサービス（SNS）、動画共有サイトなど、利用者が情報を発信し、形成していくメディア」をソーシャルメディアの定義としている [総 18]。ソーシャルメディアでは、テレビやラジオのようなマスメディアとは異なり、各ユーザが情報の発信者となれるため、人々は各自のニュースや考えを投稿している。このことから、ソーシャルメディアのデータを分析することで、世論や流行など社会を分析できる。

社会を分析する上で、社会を構成する人のあいだの関係を調べることは重要である。ソーシャルメディア上でのユーザ間の関係をデータとして取得することで、人間同士の関係を表したソーシャルグラフが得られる。ここで述べるソーシャルメディアとは、インターネット上で人々が活動する場であるため、得られたソーシャルグラフは現実の人間同士の関係を表現したものではなく、あくまでソーシャルメディア上での関係を表現したものである。しかしながら、ソーシャルメディアを利用しているユーザは現実に存在する人間であり、現実の知り合いとソーシャルメディア上で交流することも多い。そのため、オンラインソーシャルグラフには現実世界での関係が反映されている。

以下では、まず実在するソーシャルメディアについて 2.1 節で述べる。次に、本論文での分析に用いたソーシャルメディアである Twitter と、ソーシャルグラフ構築に用いられる Twitter 上でのユーザ間の関係について 2.2 節で説明する。さらに、ユーザ属性のうちの居住地について 2.3 節で述べる。ソーシャルグラフとユーザ属性との関係について 2.4 節で述べる。そして、ユーザ間の関係と属性の類似性の強さについて 2.5 節で述べる。ソーシャルグラフの中心性について 2.6 節で述べる。

2.1　実在するソーシャルメディア

ソーシャルメディアには多くの種類がある。例えば、Twitter や Facebook などのソーシャルネットワーキングサービスや、掲示板、ブログ、メッセージングサービス、コンテンツの投稿と閲覧を主にしたサービスなどがあり、多くのソーシャルメディアはこれらのカテゴリに当てはまる。多くのソーシャルメディアは共通した以下の機能を持つ：コンテンツの投稿、コンテンツの閲覧、ユーザの購読、コンテンツへの評価、コンテンツの共有（シェア）。投稿されるコンテンツの種類や利用目的により、どのようなユーザと関係を持つかなどに違いがあるが、共通している点もある。

本研究では、種々のソーシャルメディアの中から Twitter に注目し、Twitter のソーシャルグラフを対象に分析をしている。平成 30 年版情報通信白書によると、日本では 40% の人が Twitter を利用していると答えた [総 18]。Twitter は API を提供しているためデータが収集しやすく、かつ多くの人が利用しており、分析の対象とするのに適している。また、ソーシャルメディアの多くは似た機能を持っているため、本論文の分析方法は他のソーシャルメディアを対象とした場合にも参考になると考える。

2.2　対象とするソーシャルメディア：Twitter

Twitter は人々がいま起きていることを比較的短い文章で投稿するメディアである。ツイートと呼ばれる短文を投稿することが特徴で、ツイートは日本語で 140 文字以内という制限がある[*1]。投稿が短文であるため、他のソーシャルメディアと比べ、比較的リアルタイムに情報が投稿されるという特徴がある [吉田 16]。投稿されたツイートは各ユーザのタイムラインと呼ばれる場所に時系列順に表示される。各ユーザは、自身が過去に投稿したツイートが新しいものから順に表示されるユーザタイムラインと、フォロー（購読）しているユーザの投稿が新しい順に表示されるホームタイムラインを持つ。

各ユーザはプロフィール情報を入力できる。ユーザの情報の詳細は 2.2.1 節で説明する。また、各ユーザが投稿するツイートの情報については 2.2.2 節で説明する。さらに、フォローなどのユーザ間の関係について 2.2.3 節で述べる。

[*1] 2017 年 9 月から 11 月に、日本語、中国語、韓国語以外の言語では、ツイートの文字数制限が 280 文字に緩和されている。https://blog.twitter.com/official/en_us/topics/product/2017/Giving-you-more-characters-to-express-yourself.html (viewed 2020-12-02)

2.2.1　ユーザのプロフィール情報

Twitter を利用してツイートを投稿するユーザはアカウントを作る必要がある。アカウント登録の際には、他人と重複していないユーザ名（アカウント名；例：@twitter）を決める必要がある。また、このときユーザは公開アカウントにするか非公開アカウントにするかを選べる。非公開アカウントでは、そのアカウントがフォローを許可したユーザのみにプロフィールとツイートの閲覧を許可する。非公開アカウントのデータは許可なしに取得できないため、本研究では公開アカウントのデータのみを対象としている。

プロフィールには名前（表示名）、場所、自己紹介などを入力できる。名前は 50 文字まで入力することができ、タイムラインにツイートが表示される際、共に表示される。場所には位置情報などをテキストで入力するが、位置情報に関係のないテキストを入力することもできる。自己紹介には 160 文字以内のテキストを入力でき、場所とともにユーザのプロフィール画面に表示される。また、アイコン画像も設定できる。

図 2.1 に示すように、ユーザのプロフィール画面にはユーザ名、名前（表示名）、アイコン画像、自己紹介、場所、Twitter 登録日、フォロー数、フォロワー数などが表示されている。API によりツイートのデータを取得した際にも、各ツイートには投稿者のプロフィール情報が紐付けられている。

2.2.2　ツイートの情報

Twitter Web では、投稿するためのテキストを入力する欄に「いまどうしてる？」と表示されており、多くのユーザがリアルタイムにいま起こったことを投稿している。ツイートには、そのツイートが投稿された日時と、そのツイートを投稿したユーザの情報、ツイートの内容が紐付けられている。ツイートによっては、内容が自動的に解析され、ツイートに含まれている URL やハッシュタグ、言及（メンション）されたユーザの情報がメタデータとして付与されている。また、画像や動画、位置情報をツイートに含めることができる。

ツイートに付与される位置情報には、エリア情報（place；豊橋市や愛知県など）と詳細な座標情報（coordinates；緯度経度）との 2 種類がある。座標情報が付けられている場合、Twitter のサーバによってその座標に対応するエリア情報が追加されるため、詳細な座標情報が付与されているツイートの場合は両方の位置情報がツイートに付与されている。

Twitter では、他のユーザが投稿したツイートを再投稿（リツイート）することができる。リツイートは、自身のフォロワーに対してツイートを共有する行為である。これによ

図 2.1: Twitter Web のプロフィール画面の例（2020 年 12 月 2 日閲覧）

り、投稿者のフォロワーにしか閲覧されていなかったツイートが、リツイートをしたユーザのフォロワーに拡散し、多くのユーザの目に触れることになる。リツイートした場合には、リツイートをしたユーザのプロフィール情報と、リツイートの元ツイートの情報、元ツイートの投稿者の情報が共に得られる。

Twitter は情報の取得・投稿のための API を公開しているため、自動でツイートすることができる。自動投稿が主体のアカウントを Bot アカウントと呼ぶ。ユーザ属性推定などソーシャルメディアデータを利用したアプリケーションを作る際に、Bot の存在を無視することはできない。

Twitter には、今日の天気やサイトの新着ニュースを投稿する Bot など、様々な Bot アカウントがある。また、普段は手動でツイートを投稿しているユーザであっても、Twitter クライアントの機能やその他のツールの機能を利用して、自動投稿をすることがある。例

えば、Foursquare（Swarm）[*2]という位置情報の共有を目的としたソーシャルメディアでは、いまどこにいるかを友人と共有でき、Foursquare 上で共有した場所を Twitter にも同時に投稿できる。このときある程度決まった文面が投稿される。他にも、Web サイトにシェアボタンが設置されており、ボタンをクリックすると決まった文面（タイトルやコマーシャル文など）が入力済みの投稿画面が出るというものもある。これを利用すると、同じような文面が入ったツイートを複数のユーザが投稿することになる。自動生成された投稿が分析の障害となるときは、このような特徴をもとに、データに対する前処理の段階で Bot アカウントや Bot 投稿を除外する。

2.2.3　ユーザ間の関係

ユーザとユーザとのあいだの関係を表現するものが**ソーシャルグラフ**である。ソーシャルメディア上では、各ユーザは他ユーザやツイートに対して行動を起こすことができ、これらの行動をもとにユーザ間の関係を得ることができる。

ユーザに対する行動として、フォロー、リスト、メンション（@ツイート）がある。フォロー機能は、他ユーザのツイートを購読して自分自身のホームタイムラインを作るために使われる。リスト機能はホームタイムライン以外に仮想的なタイムラインを作るために使われる。ホームタイムラインは 1 ユーザにつき 1 つしか持つことができないが、リストは複数作成できる。また、公開リストと非公開リストがあり、他ユーザが作成した公開リストを購読できる。各ユーザは、自分が追加されている公開リストの一覧をプロフィール画面から確認できる。公開リストはあるトピックに関連するユーザを集めるなどして作成されることが多いため、公開リストに付けられた名前をもとに、リストに追加されているユーザに対するタグ付けとみなされることがある [Yamaguchi 15]。メンションは、ツイートを投稿する際に他ユーザのユーザ名をテキスト中に含めることである。メンションが投稿された際、言及されたユーザは通知を受け取る。ユーザをフォローする、ユーザをリストに入れる、ユーザをツイートで言及（メンション）するという行為は、相手に許可を求めずにおこなうことができる。

ツイートに対する行動として、お気に入り[*3]とリツイートがある。ユーザのツイートをお気に入りに入れる行為と、ユーザのツイートをリツイートする行為も、ツイートをしたユーザの許可は必要ない。ツイートにはそれを投稿したユーザの情報が紐付いているため、ユーザからツイートへの関係はユーザからユーザへの関係とみなせる。加えて、ツ

[*2] `https://www.swarmapp.com/` (viewed 2020-12-02)

[*3] 2015 年 11 月にお気に入りからいいねへ名称が変更された。同時に UI のボタンの形も星形からハート型へ変更された。`https://blog.twitter.com/official/ja_jp/a/ja/2015/1104heart.html`, `https://twitter.com/TwitterJP/status/661659581832015878` (viewed 2020-12-02)

イートからツイートへの行動として、リプライ（返信）がある。メンションはツイートを対象とせずユーザを言及する行動であるが、リプライは特定のツイートに対して返信のツイートを投稿するものである。

ユーザがフォローをするということは、対象のユーザの投稿を購読するということである。ユーザがどのような理由でフォローをするかを分析した研究がいくつかある [Kwak 10, Barbieri 14, Tanaka 14, Yamaguchi 15]。大きく分けると、ツイートの内容やトピックに興味を持ちフォローをするか、知り合いであるからフォローをするかの 2 つに分けられる。このように、オンラインソーシャルグラフには友人関係が含まれていることがわかっているが、それだけとは限らない。

2.3　推定対象の属性としての居住地

ユーザの年齢や性別、嗜好、居住地などの属性はソーシャルメディアデータを利用するアプリケーションにおいてよく用いられているが、特に情報の整理に利用される属性がある。情報を整理する際によく使われる軸として、時間、空間、名前の 3 軸がある。例えば、年表は情報を時間の軸で整理することであり、地図上へのマッピングは情報を空間の軸で整理することである。また、本の末尾などに用意されている索引は名前の軸による整理である。これらの軸は、組み合わせることでさらに情報を特定しやすくなる。時間と空間が得られれば場所を特定できることから、本論文では時間と空間の軸として利用できる属性に着目している。

ソーシャルメディアデータにおける時間と空間に該当する属性について考える。Twitter には投稿の即時性が高いという特徴があると報告されている [吉田 16]。このことから、ツイートの投稿時間をみることで、そのツイートに記載されている事柄が起きた時間を推測できる。空間の情報については、ツイートに付与されている位置情報などが利用できると考えられる。しかしながら、位置情報はまれにしか付与されていないため、欠損している空間についての情報を推定する必要がある。我々はユーザに着目しているため、ユーザの位置に注目する。

ユーザの位置は、ユーザの特徴を表す重要な属性である。ユーザが Twitter を利用するときに主に滞在している場所がわかれば、ユーザはその周辺でツイートをすると考えると、ツイートの位置を補完することに使える。また、ユーザの位置が利用されているものとして、検索エンジンや推薦結果のパーソナライズがある。様々な場所で検索をしたとき、検索結果がただ表示されるのではなく、ユーザの属性を利用して結果を表示することで、ユーザがいる場所周辺での結果を提示できる。商品やユーザの推薦においても同様である。他にも、ソーシャルメディアデータをマーケティングに利用することを考えると、ユーザがどこに住んでいるかは重要な要素である。ある特定の商品に興味のあるユーザに

対してメッセージを送るためにユーザを集めたが、サービスの提供範囲があるエリアに限られていたとき、ユーザがサービスの提供範囲内にいるかどうかの情報を活用することで、マーケティングを有利に進めることができる。また、ソーシャルメディア上でのあるトピックへの言及を集めたとき、それを分析するためにユーザの位置がわかると、そのトピックがどのエリアに住むユーザのあいだで流行しているのかがわかる。以上の例ではユーザの位置として、ユーザが主に滞在している場所や、住んでいる場所などを考えた。他に勤務先などもユーザの位置と考えられる。

Twitter のデータから得られる具体的なユーザの位置に関するデータとしては、ユーザのプロフィールに入力されている「場所」のテキストと、ユーザが投稿した位置情報付きツイートに付与されている地理座標とがある。自身のプロフィールに場所を入力しているユーザは少ないこと [Hecht 11, 山口 13] や、全ツイート中に位置情報付きツイートが占める割合は低いこと [Sloan 15] が知られており、これらの位置情報を多く利用するためには推定が必要である。我々はユーザが Twitter を利用するとき主に滞在している場所（エリア）に注目しているため、本論文では位置情報付きツイートをもとに求めた、ユーザが主に滞在している場所を**居住地**としている。

2.4　ソーシャルグラフと属性の伝搬

ソーシャルグラフのノードおよびエッジに付属する情報を**プロパティ**という。例えば、ソーシャルメディアユーザは年齢や嗜好、居住地など様々な属性を持っている。これらユーザ属性は、ソーシャルグラフのノードに付属するものであるため、ソーシャルグラフのプロパティであるといえる。ソーシャルメディアデータを用いるアプリケーションではユーザの属性が利用されるが、これらの属性は公開されているとは限らないため、欠損している属性を他の情報から推定する必要があると述べた。本論文では、属性の推定のために、似た属性を持つユーザ間の関係を利用することを考える。

人は似た好みを持つ他者との関係を持ちやすいという特徴があるが、このような性質はソーシャルメディア上での関係においても存在する [McPherson 01]。この「似た好みを持つユーザ同士はそれ以外のユーザと比べてつながりを持ちやすい」という特徴は、ユーザ間の関係からユーザ属性の推定を可能にする。例えば、ユーザ間の関係と地理的距離との関係を調査した研究があり [Kulshrestha 12]、ソーシャルメディア上の関係を用いてユーザの位置を推定する試みが多数なされている [Jurgens 15, Zheng 18, Luo 20]。居住地以外の属性においても、職場や通っている大学、学部などのユーザ属性を、ソーシャルグラフを用いて推定する研究がされている [Mislove 10]。このようなソーシャルグラフを用いる属性推定手法には、言語に依存せず適用できるという利点がある。

属性推定にソーシャルグラフを用いる場合、どのようなソーシャルグラフであるかが推

定性能に影響する。そのため、ソーシャルグラフのそれぞれのプロパティが属性推定においてどのような働きをするのかを調べる必要がある。ソーシャルグラフのプロパティには、ノードに付属するものに加え、ソーシャルグラフの形状に関わるものがある。具体的には、ソーシャルグラフを構築するもととなったユーザ間の関係や、ソーシャルグラフ上でのユーザが持つ属性の分布などもソーシャルグラフのプロパティと考えられる。

2.5　伝搬の強さとユーザのプロパティ

ソーシャルグラフを用いる属性推定手法は、似た属性を持つユーザ同士の関係を利用する。このような関係を利用してユーザ属性を推定することは、ノードの属性がエッジによって伝搬しているという解釈ができる。しかし、すべてのユーザ間の関係が同じ強さであるとは考えられないため、伝搬の強さは関係やユーザなど、ソーシャルグラフのプロパティによって変わると考えられる。

ユーザの持つつながりはユーザを中心として生まれるものであるため、ユーザのタイプによって伝搬の強さが変わると考えることができる。ソーシャルメディア上でのユーザのタイプを考える上で、直接得られる情報としてユーザのプロフィールがある。プロフィールから得られる情報やそこに記入されているテキストは、ユーザのタイプを知る手がかりになると考えられる。

2.6　グラフの中心性

ソーシャルグラフに限らず、ネットワーク構造は様々な研究に用いられてきた。例えば、共著ネットワークによる研究者の分析や、引用ネットワークによる論文の分析、Web グラフによる Web ページの評価などがある。これらのネットワークにおいて、重要なノードを発見するというタスクは一般的であり、汎用的なネットワーク分析手法が開発されてきた。そこで、本論文でも、ノード（ユーザ）の重要度を測る指標である中心性をオンラインソーシャルグラフに導入する。ここでのソーシャルグラフにおける中心性とは、多くのノードとつながっている度合いを測るものである。中心性はソーシャルグラフのプロパティの一種であると考えられる。ノードの重要性がわかれば、それを推定に利用できると考えられる。

ソーシャルグラフのノードに対して計算される中心性について以下の小節で述べる。説明では、以下に定義する変数を用いる。ソーシャルグラフは有向単純グラフとする。ソーシャルグラフの頂点をノード、辺をエッジと呼ぶ。ソーシャルグラフの隣接行列表現を A とする。隣接行列 A の要素は、ノード j からノード i の向きにエッジがあるとき $A_{ij} = 1$、エッジがないとき $A_{ij} = 0$ の値を持つ。あるノード i に入るエッジの総数を入

次数 k_i^{in}、あるノード i から出るエッジの総数を出次数 k_i^{out} とする。

2.6.1　次数中心性

次数中心性とは、あるノードが持つエッジの数をそのまま重要度の指標とするものである。有向グラフではエッジの向きがあるため、入次数中心性と出次数中心性とがある。次数とはどれだけ多くのノードとの関係を持っているかを表す値であるため、多くのノードとつながりを持っているほど重要とみなす指標である。

入次数中心性 x_i^{in} と出次数中心性 x_i^{out} は次の式で定義される。

$$x_i^{\text{in}} = \sum_j A_{ij} = k_i^{\text{in}} \tag{2.1}$$

$$x_i^{\text{out}} = \sum_j A_{ji} = k_i^{\text{out}} \tag{2.2}$$

Twitter のフォロー関係をもとにしたソーシャルグラフ[*4]においては、入次数中心性はそのユーザが持つフォロワー数、出次数中心性はそのユーザのフォロー数に該当する。多くのフォロワーを持つユーザを重要なユーザとみなすのが入次数中心性であり、多くのユーザをフォローしているユーザを重要なユーザとみなすのが出次数中心性である。高い入次数中心性を持つユーザは多くのユーザに知られている有名なユーザ、高い出次数中心性を持つユーザは多くの情報源を購読しているユーザであると考えることができる。

2.6.2　PageRank

PageRank [Page 98] は、単純に多くのノードと関係を持っているノードではなく、重要なノードからのエッジを持っているほど重要なノードであると考える中心性である。あるノードのスコアは、そのノードを指すエッジを持つノードのスコアの和とする。ただし、各ノードから伝搬するスコアは、そのノードが持つ出エッジの数で割ったものとすることで、大きな値を持つノードが多くの出エッジを持っていたときにスコアが伝搬しすぎないようにする。

PageRank は次の式で表される。

$$x_i = \alpha \sum_j A_{ij} \frac{x_j}{k_j^{\text{out}}} + \beta \tag{2.3}$$

ここで、α と β は定数である。β の項は、スコアが 0 であるユーザ（入次数が 0 のユーザ）も初期値としていくらかの重要度を持っていることを表している。

[*4] ユーザ u がユーザ v をフォローしているとき u から v の方向へエッジを作り構築したソーシャルグラフ。

PageRank は Google の Web ページ用ランキングアルゴリズムとして使うために提案された [Page 98]。その後、Web ページ間のリンクをネットワークとして表現した Web ネットワークを対象として使われるだけでなく、様々なネットワークを対象とした応用に利用されている [Gleich 15]。

Twitter のフォローをもとにしたソーシャルグラフでは、PageRank は入次数中心性と同じように、有名なユーザに高い値を与える。すなわち、多くのユーザにフォローされているほど高い値を持ちやすい。しかし、それだけでなく、高いスコアを持つユーザがフォローしているユーザも高いスコアを持つ。これは、フォローされることによって、フォローをおこなったユーザの権威がフォロー先のユーザへ受け渡されると理解することができる。例えば、多くのユーザにフォローされている有名人がフォローしているユーザが数名しかいない場合、それらのフォローされているユーザは、その有名人と親しい有名人やマネージャーなどとても関わりが強いユーザであると考えられ、重要度が高い（ソーシャルメディア上で影響力がある）ユーザであると考える。

2.6.3　HITS

HITS（Hyperlink-Induced Topic Search）[Kleinberg 99] は、各ノードに対して 2 種類の重要度を考える指標である。PageRank のように高い中心性を持つノードに指されていると高い値となる Authority スコアとともに、高い中心性を持つノードを指しているノードが高い値を持つ Hub スコアを考える。すなわち、高い Hub スコアを持つノードに指されていると高い値となる Authority スコアと、高い Authority スコアを持つノードを指していると高い値となる Hub スコアを定義する。

Authority スコア x_i と Hub スコア y_i は次の式で定義される。

$$x_i = \alpha \sum_j A_{ij} y_j \tag{2.4}$$

$$y_i = \beta \sum_j A_{ji} x_j \tag{2.5}$$

ここで、α と β は定数である。

Twitter のフォローをもとにしたソーシャルグラフでは、多くのフォロワーを持つユーザは高い Authority スコアを持ち、多くのユーザをフォローしているユーザは高い Hub スコアを持つ傾向にある。しかし、それだけではなく、フォロワーが少ないユーザでも高い Hub スコアを持つユーザにフォローされていると高い Authority スコアを持ち、フォローが少なくても高い Authority スコアを持つユーザをフォローしていると高い Hub スコアを持つ。PageRank と異なり、各ノードから伝搬するスコアは Authority の場合は Hub、Hub の場合は反対の Authority となるため、有名人（Authority スコアが高い）が

少数の知人をフォローしていたとしても、そのユーザを重要なノードであるとはみなさない。代わりに、多くの有名人をフォローしているようなユーザ（Authority スコアが低く、Hub スコアが高い）がフォローしているユーザは、有名なユーザであるとみなして高い Authority スコアを与える。例えば、Authority の値によって Twitter を利用してプロモーションをおこなっているユーザが特定でき、Hub の値によって情報収集を目的としているユーザが特定できると考えられる。

第 3 章

居住地推定法の分析

3.1　本章の背景

ソーシャルグラフを用いてユーザの居住地を推定する試みが多数なされている。ソーシャルグラフ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グラフができる。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。

本章では、ユーザ間の関係を変えて作成した複数のソーシャルグラフを用いて、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、ソーシャルグラフの形状に影響を受けないことも示す。

3.2　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 [Dominic 13]。Jurgens は、リプライから作成したソーシャルグラフを利用し、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グラフを利用し、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、本章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 種類のソーシャルグラフを用いて、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、ソーシャルグラフ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グラフを調べることは困難であるため、本章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グラフで調査する。

3.3　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グラフとを利用して、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 3.3.1 節以降で述べる。

3.3.1　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[*1]を使用し、2014 年に日本を包含する矩形[*2]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。

3.3.2　フォロー関係をもとにしたソーシャルグラフ

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グラフを作成する。ユーザがフォローしているユーザの集合[*3]とユーザをフォローしているユーザの集合[*4]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 種類のソーシャルグラフを作成した。本研究でのソーシャルグラフは、図 3.1 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グラフにおいて、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）にあるノードである。図 3.1 では、ノード B はノード A の隣接ノードとなる。

3.3.3　ソーシャルグラフの特徴

本節では、作成したソーシャルグラフの統計量を調べ、ユーザ間の関係を変えて作成したソーシャルグラフ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グラフの特徴を明らかにするため、グラフ

[*1] `https://dev.twitter.com/streaming/reference/post/statuses/filter` (viewed 2016-11-04)
[*2] 北緯 20 度から 50 度、東経 110 度から 160 度の範囲。
[*3] `https://dev.twitter.com/rest/reference/get/friends/ids` (viewed 2016-11-04)
[*4] `https://dev.twitter.com/rest/reference/get/followers/ids` (viewed 2016-11-04)

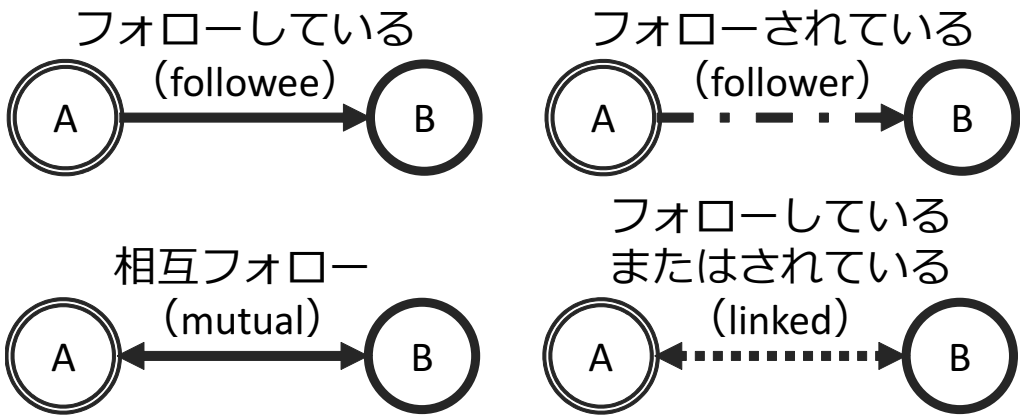


図 3.1: フォロー関係をもとにした 4 種類のユーザ間の関係

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グラフ $G(V, E)$ 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 $|I_{\text{out}}|$ 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、3.3.2 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、$|V| \geq |V'| + |I_{\text{out}}|$ 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。

3.3.2 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 種類のソーシャルグラフを作成した。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グラフには、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グラフと、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グラフとを区別して統計量を調べる。表 3.1 に調べた統計量を示す。なお、3.4 節で後述するように、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グラフ（表 3.1b）を使用することとなる。

ソーシャルグラフを作成する際、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グラフでは、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 3.1b 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グラフでは、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 3.1 の統計量から、もとにした関係によるソーシャルグラフ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 $|I_{\text{out}}|$ が大きくなる。このため、mutual をもとにしたソーシャルグラフは、推定できないユーザ数 $|I_{\text{out}}|$ がほかのソーシャルグラフ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グラフの統計量（表 3.1a）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グラフの統計量（表 3.1b）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% であることがわかる。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グラフには友人関係が比較的多いと考えられる。

3.4　調査する居住地推定手法

ソーシャルグラフを利用する居住地推定手法は、ソーシャルグラフ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グラフは、3.3.2 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

表 3.1: ソーシャルグラフの統計量

(a) 収集したネットワーク全体

関係	$\|V'\|$	$\|E\|$	K_{out}	S_{out}	M_{out}	$\|I_{\text{out}}\|$
followee	62 676 854	417 334 528	385.9831	2059.3849	205	3301
follower	62 676 854	417 334 528	514.8269	4761.8717	194	2780
mutual	22 834 460	251 625 205	272.9455	1837.2806	129	8710
linked	62 676 854	583 043 851	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}	$\|I_{\text{out}}\|$
followee	428 150	8 163 069	17.2818	53.7495	7	54 618
follower	428 150	8 163 069	17.2818	61.1349	6	65 838
mutual	389 050	6 226 387	13.1817	45.1985	5	83 300
linked	428 150	10 099 751	21.3819	68.5991	9	44 200

ベルとして表現する。3.4.1 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、ソーシャルグラフ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 3.4.1 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、$\text{dist}(l_1, l_2)$ はラベル l_1 とラベル l_2 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[*5] [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。

[*5] 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。

3.4.1 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 (3.1) 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[*6]。このモデル式を利用し、式 (3.2) でノード u の居住地を推定する[*7]。推定に必要な計算量は $O({K_{\text{out}}}^2)$ である。

$$p(d) = a(d+b)^c \tag{3.1}$$

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

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

$$\text{ProbabilityModel}(u) = \underset{l \in \{l_n | n \in N_u \cap L\}}{\arg\max} \gamma(l, u) \tag{3.2}$$

3.4.2 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グラフ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 (3.3) であり、計算量は $O(K_{\text{out}})$ である。ここで、$\arg\max^*$ は同値の集合を返すものと定義する。

$$S_u = \underset{l \in \{l_n | n \in N_u \cap L\}}{\arg\max^*} |\{x | x \in N_u \cap L, l = l_x\}|$$

$$\text{MajorityVote}(u) = \underset{l \in S_u}{\arg\max} |\{n | n \in L, l = l_n\}| \tag{3.3}$$

[*6] 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。

[*7] 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。

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。

3.4.3　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[*8] [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 (3.4) であり、計算量は $O({K_{\text{out}}}^2)$ である。

$$\text{GeometricMedian}(u) = \underset{l \in \{l_n | n \in N_u \cap L\}}{\arg\min} \sum_{x \in N_u \cap L, n \neq x} \text{dist}(l, l_x) \tag{3.4}$$

3.4.4　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[*9]。この手法を表現したものが式 (3.5) であり、計算量は $O(1)$ である。ここで、$\text{choice}(S)$ 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。

$$\text{RandomNeighbor}(u) = l_{\text{choice}(N_u \cap L)} \tag{3.5}$$

3.5　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グラフ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

[*8] Fermat–Weber Problem や L1 Median とも呼ばれる。

[*9] Jurgens らのベースラインとは、繰り返しの有無が異なる。

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[*10]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。

$$\mathrm{Precision}(T, X) = \frac{|\{u | u \in T \cap X, l_u = e_u\}|}{|T \cap X|} \tag{3.6}$$

$$\mathrm{Recall}(T, X) = \frac{|\{u | u \in T \cap X, l_u = e_u\}|}{|T|} \tag{3.7}$$

$$\mathrm{F1}(T, X) = \frac{2 \cdot \mathrm{Precision}(T, X) \cdot \mathrm{Recall}(T, X)}{\mathrm{Precision}(T, X) + \mathrm{Recall}(T, X)} \tag{3.8}$$

$$\mathrm{Coverage}(T, X) = \frac{|T \cap X|}{|T|} \tag{3.9}$$

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。

3.5.1　居住地推定の評価

4 種類の居住地推定手法と 4 種類のソーシャルグラフとをそれぞれ変えて、居住地推定をおこなった結果を表 3.2 と表 3.3 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、3.5.3 節で後述する Mean ED と Median ED は小さいほど良い。表 3.3 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 3.2 と表 3.3 から、日本のソーシャルグラフでは Majority Vote が最も精度良く居住地を推定できることがわかる。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、ソーシャルグラフ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グラフ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 3.2 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グラフに対して 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

[*10] 該当するノードの数は表 3.1b に示した $|I_{\mathrm{out}}|$ である。

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 3.3 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[*11]。なお、Probability Model と Majority Vote との計算量はそれぞれ $O(K_{\mathrm{out}}^2)$ と $O(K_{\mathrm{out}})$ 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。

3.5.2　ユーザ間の距離の分布と居住地推定性能の関係

本章で作成した 4 種類のソーシャルグラフでの、ユーザ間の地理的な距離の分布を図 3.2 に示す。図 3.2a 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グラフにおいても、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 3.2b 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グラフは、他の関係をもとにしたソーシャルグラフと比べ、得られるユーザ間の関係数が少ないことがわかる。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 3.2c 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

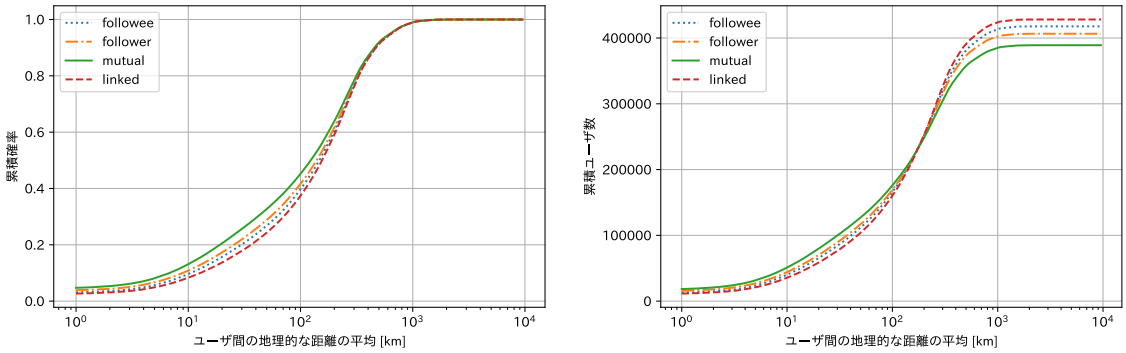
[*11] t 検定での p 値は 0.769 であった。

表 3.2: 居住地推定性能（leave-one-out 交差検証）

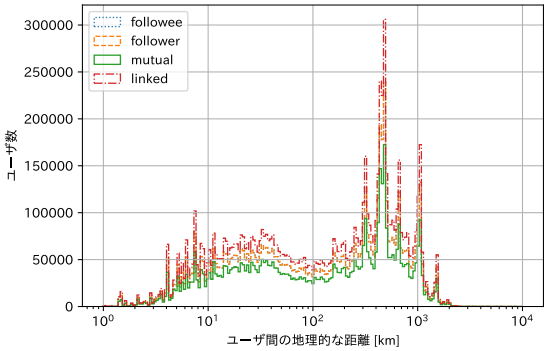
手法	関係	適合率	再現率	F 値	カバー率	Mean ED [km]	Median ED [km]
Probability Model	followee	0.29598	0.26176	0.27782	0.88437	153.212	19.610
	follower	0.30695	0.26416	0.28395	0.86062	146.955	18.534
	mutual	0.30116	0.24805	0.27204	0.82365	146.606	18.930
	linked	0.30451	0.27602	0.28957	0.90643	151.292	18.780
Majority 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 3.3: 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



(a) 正規化累積分布　　(b) 累積分布



(c) 密度分布

図 3.2: ユーザ間の地理的な距離の分布

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 3.2a および図 3.2b 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 3.2 および表 3.3 からわかるとおり、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グラフ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グラフ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グラフにおいて、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グラフ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グラフ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。

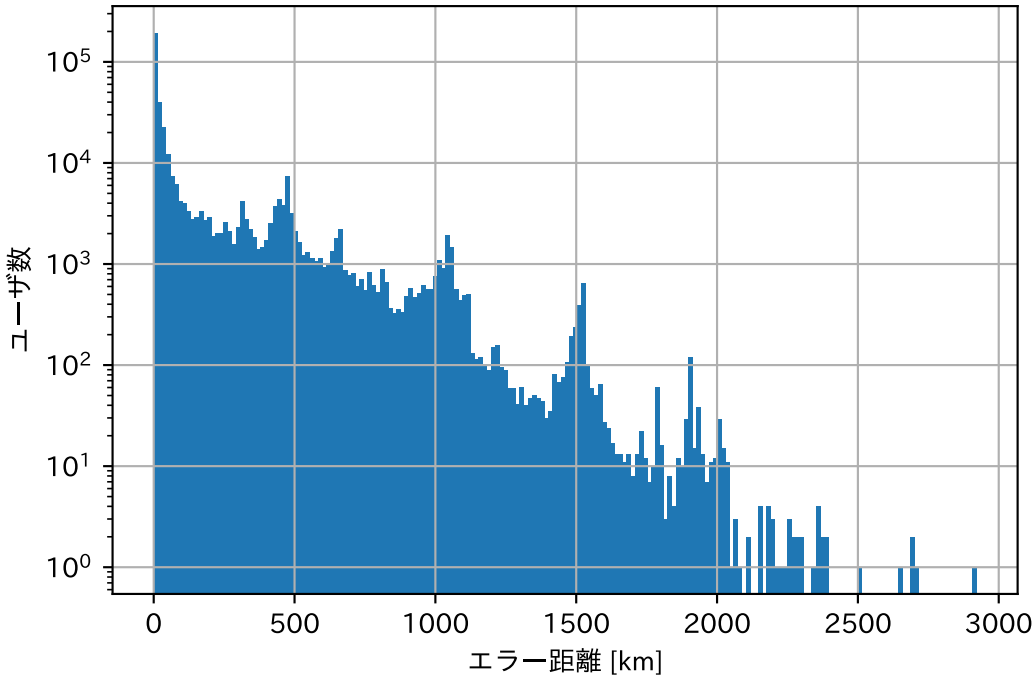


図 3.3: エラー距離の分布（Majority Vote + follower）

3.5.3　エラー距離での評価

3.5.1 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おこなう。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 [Dominic 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 3.3 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 $\mathrm{dist}(l_u, e_u)$ 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 3.2 と表 3.3 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を t 検定で確認した。なお、表 3.2 と表 3.3 の読み取り方および検定方法の詳細は 3.5.1 節を参照されたい。

3.5.4　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。

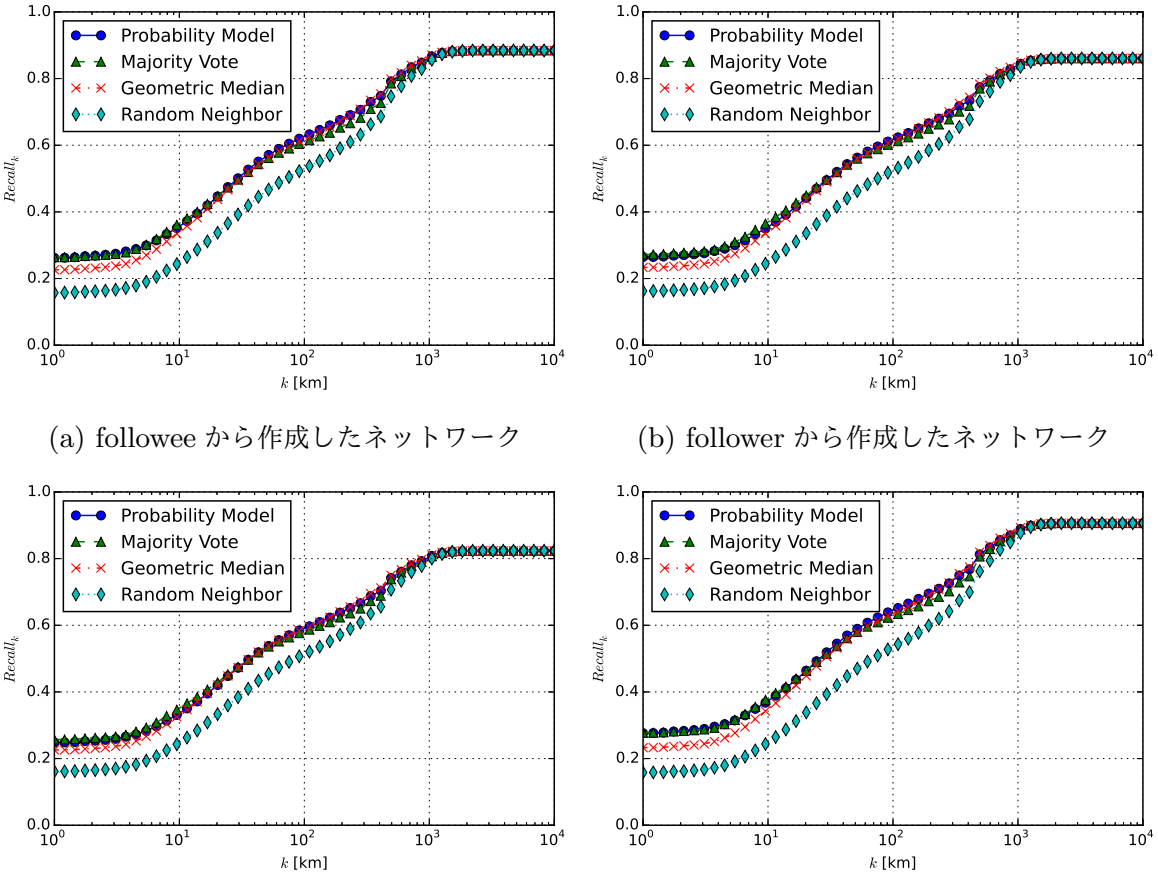
$$\mathrm{Recall}_k(T, X, k) = \frac{\{u | u \in T \cap X, \mathrm{dist}(l_u, e_u) < k\}}{|T|} \tag{3.10}$$

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わかっているため、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 3.4 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 3.4 に示す。表 3.4 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わかる。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



(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 3.4: 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

との間には、有意差を確認できなかった[*12]。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。

3.6　考察と限界

本章では、フォロー関係に着目し、フォロー関係をもとにした 4 種類のソーシャルグラフを用いて、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おこなった。同様の統一的な評価は Jurgens ら [Jurgens 15] もおこなっているが、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作成したソーシャルグラフのみを用いている。対して本研究では、ソーシャルグラフ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、ソーシャルグラフを作成した。さ

[*12] McNemar 検定での p 値は 0.607 であった。

表 3.4: 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

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 種類のソーシャルグラフを利用し、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グラフを利用し、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。本章では、フォローの関係から生成することができる 4 種類のソーシャルグラフ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、3.5.4 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 3.2 と表 3.4 とを比較すればわかるとおり、有効な手法および関係の組み合わせが 3.5.1 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。

3.7　本章のまとめ

本章では、フォロー関係をもとに作成した 4 種類のソーシャルグラフを用いて、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グラフ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 がソーシャルグラフ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。

第 4 章

ソーシャルグラフ上での距離と居住地

4.1　本章の背景

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した、多くの居住地推定手法が提案されている。ソーシャルグラフを用いて居住地を推定するアプローチとして、直接の友人であるユーザの居住地を利用して居住地を推定する方法が提案されている [Backstrom 10, Davis Jr. 11, McGee 13]。ここで、直接の友人の居住地が不明である場合を考える。この場合、直接の友人の持つ情報を伝搬する方法では推定ができない。そのため、直接の友人以外の情報を利用するために、推定を複数回繰り返す方法が提案されている [Jurgens 13, Kong 14]。ソーシャルグラフは友人関係とみなされる関係から構築されるため、Twitter のフォロー関係をもとに構築されることが多い。フォロー関係をもとに構築したソーシャルグラフにおける、繰り返しを用いる居住地推定手法の性能はまだ報告されていない。

本章では、フォロー関係をもとに構築したソーシャルグラフを用いて、繰り返しを採用しているネットワークベースの居住地推定手法の性能を調査する。繰り返し回数と繰り返し適用する居住地推定手法を変えて推定性能を調査した結果、フォロー関係から構築したソーシャルグラフにおける繰り返し回数は、ソーシャルグラフの特徴を分析した結果から、2 回で十分であることを示す。

4.2　分析に用いる居住地推定手法

ネットワークベースの居住地推定手法は、ソーシャルグラフ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。本章では、ソーシャルグラフは、ユーザをノード、ユーザ間の関係をエッジとした単純有向グラフとする。ユーザの居住地はノードに付与された 1 つのラベルとする。

Spatial Label Propagation (SLP) [Jurgens 13] はグラフネットワークを用いるラベル

伝搬法 [Zhu 03] をネットワークベースの居住地推定に応用した手法である。SLP は隣接ノード集合とそれらユーザのラベルをもとに居住地を推定する推定関数を繰り返し適用することでラベルを伝搬させ、多くのノードの居住地を推定できる手法である。SLP を用いると、隣接していないノードの持つラベルも推定に利用できるため、隣接ノードにラベルがないユーザの居住地も推定できる。

SLP のパラメータには推定関数と繰り返し回数がある。我々は隣接ノードの情報を用いる居住地推定手法を推定関数として用いる。本章では、推定関数として 4.2.1 節から 4.2.4 節で説明する 4 手法を用いる。ユーザ u の推定関数は $\mathrm{Select}(u; N_u)$ と表す。推定関数は、推定対象ノード u の隣接ノード集合 N_u と、それらのラベルを利用して推定をおこなう。

推定関数の説明では次の変数を用いる。L は学習データに含まれるノードの集合、N_u はノード u の隣接ノード集合、A はラベル（居住地）の集合、l_u はノード u の正解ラベル、$\mathrm{dist}(l_1, l_2)$ はラベル l_1 とラベル l_2 との間の距離である。ラベル間の距離はラベルに対応づけられる居住地（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。

4.2.1　Probability Model

Probability Model [Backstrom 10] は、ノード間がある地理的距離のときに友人である確率のモデルを作成し、隣接ノードが持つラベルの中で最も尤度が高いラベルを選択する手法である。ノード間の地理的距離が d であるとき、そのノード間にエッジが存在する確率 $p(d)$ は式 (4.1) と表される。ここで、a, b, c は定数である。元論文 [Backstrom 10] で使われていた $a = 0.0019$, $b = 0.196$, $c = -1.05$ を本章の実験でも用いる。

ユーザ u のラベルを予測する推定関数は式 (4.4) である[*1]。

$$p(d) = a(d + b)^c \tag{4.1}$$

$$\gamma_l(l) = \prod_{v \in L} [1 - p(\mathrm{dist}(l, l_v))] \tag{4.2}$$

$$\gamma(l, u) = \prod_{v \in N_u \cap L} \frac{p(\mathrm{dist}(l, l_v))}{1 - p(\mathrm{dist}(l, l_v))} \gamma_l(l) \tag{4.3}$$

$$\mathrm{ProbabilityModel}(u) = \underset{l \in \{l_n | n \in N_u \cap L\}}{\arg\max} \gamma(l, u) \tag{4.4}$$

[*1] 元論文の式が間違っていると考えられるため、$\gamma(l, u)$ の式に $\gamma_l(l)$ を補っている。

4.2.2 Majority Vote

Majority Vote [Davis Jr. 11] は隣接ノードの持つラベルの中で出現頻度が最も高いものを選択する手法である。この手法は推定対象のユーザと同じ居住地に住んでいるユーザが多数派であることを仮定している。元論文では無向グラフが用いられていたが、本章の実験で用いるソーシャルグラフは有向グラフであるため、ユーザが持つ友人を有向グラフの隣接ノードとみなして拡張した。さらに、最頻値が複数あったときの処理も不明であったため、最頻値が複数あった場合、本論文では学習データ全体で出現頻度が高い順にラベルを選択することとする。

ユーザ u のラベルを推定する推定関数は式 (4.6) で表される。

$$S_u = \underset{l \in \{l_n | n \in N_u \cap L\}}{\arg\max^*} |\{v | v \in N_u \cap L, l = l_v\}| \tag{4.5}$$

$$\mathrm{MajorityVote}(u) = \underset{l \in S_u}{\arg\max} |\{n | n \in L, l = l_n\}| \tag{4.6}$$

ここで、$\arg\max^*$ は関数が最大となる引数の集合を返すものと定義する。

Majority Vote は推定する隣接ノード数の範囲と、最低投票数という 2 つのパラメータが存在する。他の推定関数では推定対象とする隣接ノード数に条件を設けていないことから、この推定関数でも同様とする（隣接ノード数の範囲を $(0, \inf)$ とする）。また、最低投票数は 0 とする。

4.2.3 Geometric Median

Geometric Median [Jurgens 13] は隣接ノードの持つラベルとの距離の総和が最小になるラベルを選択する手法である。ユーザ u のラベルを推定する推定関数は式 (4.7) である。

$$\mathrm{GeometricMedian}(u) = \underset{l \in \{l_n | n \in N_u \cap L\}}{\arg\min} \sum_{x \in N_u \cap L, n \neq x} \mathrm{dist}(l, l_x) \tag{4.7}$$

4.2.4 Random Neighbor

Random Neighbor は隣接ノードの持つ居住地の中からランダムに選択して推定値とする、ベースラインの推定関数である。推定関数は式 (4.8) で表される。

$$\mathrm{RandomNeighbor}(u) = l_{\mathrm{choice}(N_u \cap L)} \tag{4.8}$$

ここで、$\mathrm{choice}(S)$ は集合 S からランダムに要素を 1 つ返す関数である。

4.3　データ

Twitter ユーザの居住地データと、フォロー関係をもとに構築したソーシャルグラフを実験に用いる。これらのデータの作成方法について次節以降で述べる。最終的には、52,508 ユーザを含む居住地データと、8,003,858 ノードと 40,453,444 エッジを含むソーシャルグラフデータが用意できた。

4.3.1　位置情報付きツイートをもとにした居住地

実験で用いる居住地は、市区町村レベルのエリアを表すラベルであるとする。また、各ユーザには 1 つの居住地を付与する。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。

各ユーザの居住地は、各ユーザが投稿した位置情報付きツイートもとに次の方法で付与する。まず、各位置情報付きツイートに付与されている位置情報（coordinates）を含む市区町村レベルのエリアを照合する。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。次に、ユーザごとに投稿回数が最も多いエリアを求めて、居住地として付与する。

Twitter Streaming API[*2]を用いて、2014 年の 1 年間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。収集したツイートから、Bot アカウントからの投稿とみられるツイートを除外した。さらに、居住地を付与する際に、最低でも 1 つのエリアで 5 回以上位置情報付きツイートを投稿しており、位置情報付きツイートの投稿回数が 365 回以上であるという条件を設けて、ユーザを絞り込んだ。その結果、71,166 ユーザに居住地を付与できた。

4.3.2　フォロー関係をもとにしたソーシャルグラフ

フォロー関係をもとに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のフォローしているユーザの集合[*4]と、フォローされているユーザの集合[*5]とを収集する。そして、それらのデータをもとに、相互フォロー関係にあるユーザ間にエッジを作ることで、ソーシャルグラフを構築する。構築されたソーシャルグラフは無向単純グラフと

[*2] https://dev.twitter.com/streaming/reference/post/statuses/filter
[*3] 北緯 20 度から 50 度、東経 110 度から 160 度の範囲。
[*4] https://dev.twitter.com/rest/reference/get/friends/ids
[*5] https://dev.twitter.com/rest/reference/get/followers/ids

なる。

居住地を付与できた 71,166 ユーザ周辺のフォロー関係を 2015 年 7 月に取得した。完全にデータを取得できたのは 52,508 ユーザであった。取得したデータをもとに構築したソーシャルグラフには 8,003,858 ユーザと 40,453,444 エッジが含まれている。このうち、居住地が付与されているのは 52,508 ユーザである。

4.4　実験と考察

推定関数を比較するため、実験をおこなう。本節では、まず実験方法を述べた後、結果とその分析について述べ、その後結果についての議論をおこなう。

4.4.1　実験設定

4 つの推定関数を比較するため、ラベル付きの 52,508 ユーザを用いて 10 分割交差検証により評価する。

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。適合率は居住地を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、再現率は居住地を付与されている全ユーザ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、推定可能だったユーザの割合を表すカバー率（Coverage）を用いる。ソーシャルグラフには孤立ノード（エッジを持たないノード）が含まれるため、カバー率の最大値は 100% にならない。さらに、推定された居住地と正解居住地とのエラー距離を用いて、平均エラー距離（MeanErrorDistance）と中央値エラー距離（MedianErrorDistance）を算出する。

6 個の評価指標は次の式で計算される。

$$
\begin{aligned}
\mathrm{Precision}(T, X) &= \frac{|\{u | u \in T \cap X, l_u = e_u\}|}{|T \cap X|} \\
\mathrm{Recall}(T, X) &= \frac{|\{u | u \in T \cap X, l_u = e_u\}|}{|T|} \\
\mathrm{F1}(T, X) &= \frac{2 \cdot \mathrm{Precision}(T, X) \cdot \mathrm{Recall}(T, X)}{\mathrm{Precision}(T, X) + \mathrm{Recall}(T, X)} \\
\mathrm{Coverage}(T, X) &= \frac{|T \cap X|}{|T|} \\
\mathrm{MeanErrorDistance}(T, X) &= \mathrm{mean}(D_{T \cap X}) \\
\mathrm{MedianErrorDistance}(T, X) &= \mathrm{median}(D_{T \cap X})
\end{aligned}
\tag{4.9}
$$

ここで、X は居住地が推定されたユーザの集合、T はテストデータに含まれるユーザの集

合、l_u はユーザ u の正解居住地、e_u はユーザ u の推定された居住地、D_U はユーザ集合 U に含まれる全ユーザのエラー距離（$\mathrm{dist}(l_u, e_u), u \in U$）のリスト、$\mathrm{mean}(A)$ はリスト A の要素の平均値を返す関数、$\mathrm{median}(A)$ はリスト A の要素の中央値を返す関数である。10 分割交差検証の際の評価値は、各テストデータで計算した評価指標の平均値とする。

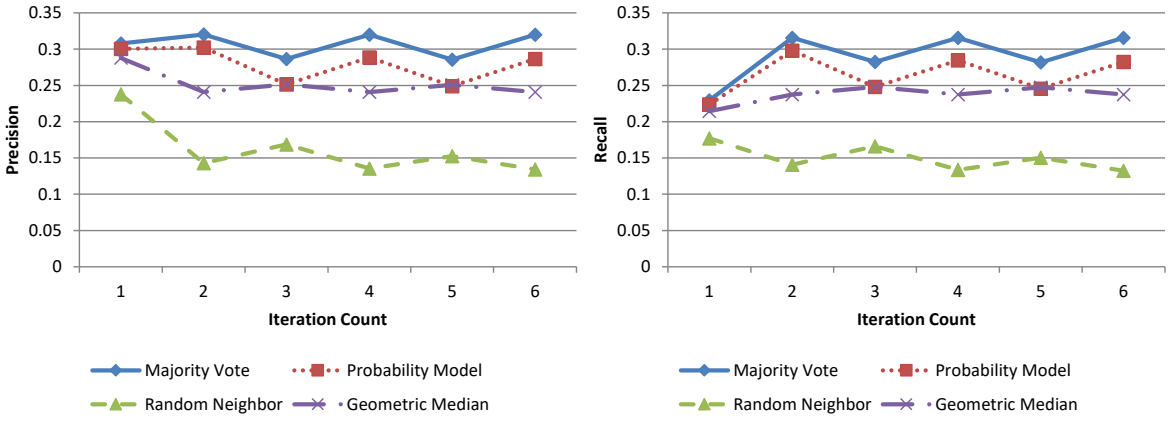
4.4.2　結果と分析

4 つの推定関数をフォロー関係をもとにしたソーシャルグラフを用いて比較する。繰り返し回数を 1 回から 6 回にしたときの、推定関数ごとの推定性能を図 4.1 に示す。Majority Vote はすべての繰り返しにおいて適合率と再現率が最も高かった。Probability Model は 2 番目に高い適合率と再現率を示した。適合率と再現率が最大値となったのは、繰り返し回数が 2 回のときだった。Geometric Median は、適合率と再現率の値の繰り返し回数による違いが最も小さかった。繰り返し回数が 1 回のときのカバー率は 0.745 で、繰り返し回数が 2 回以降のときのカバー率は 0.986 であった。本章での推定関数の比較結果は、相互メンション関係をもとに構築したソーシャルグラフを用いている先行研究 [Jurgens 13] の結果とは一致しない結果であった。

平均エラー距離と中央値エラー距離による評価結果を図 4.2 に示す。エラー距離による評価では、値が小さいほど良い結果となる。平均エラー距離による評価では、Geometric Median がすべての繰り返しで最も小さい値を示した。このことから、Geometric Median は大きく間違えない（特に大きなエラー距離を取らない）性質があると考えられる。Majority Vote と Probability Model は、繰り返し回数が 2 回のとき、中央値エラー距離が最も小さくなった。この結果は適合率と再現率による評価と同じ傾向である。

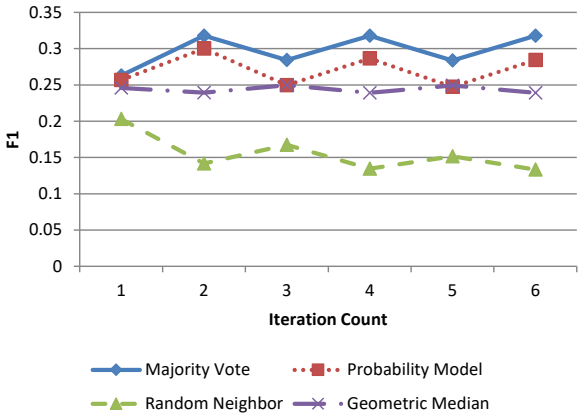
SLP は隣接ノードの情報のみを用いる手法による推定結果を学習データに追加して再推定する手法であるといえる。性能が向上するためには、ある程度推定した居住地が信頼できる必要がある。各推定関数は、推定対象とする隣接ノード数のパラメータを設定することができ、これにより適合率を向上させられると考えられる。しかしながら、再現率と適合率はトレードオフの関係となるため、適合率が向上すると再現率は低下すると考えられる。推定関数単体の適合率が高いほど推定した居住地の信頼性が高くなり、低下した再現率は SLP の繰り返しによって補えると考えられるため、適合率を重視した推定関数を用いることで全体的に良い性能が得られる可能性がある。各推定関数の適合率に関する調査は今後の課題である。

SLP は繰り返すことで居住地ラベルを伝搬させていく。繰り返し回数はどの距離まで居住地を伝搬させるかを決めるパラメータである。繰り返し回数について分析するため、同じ居住地を持つユーザへのソーシャルグラフ上での距離を調べる。ユーザごとに同じ居住地を持つユーザへの距離を計算し、距離ごとに全ユーザに占める割合を計算した結果



(a) 適合率

(b) 再現率



(c) F 値

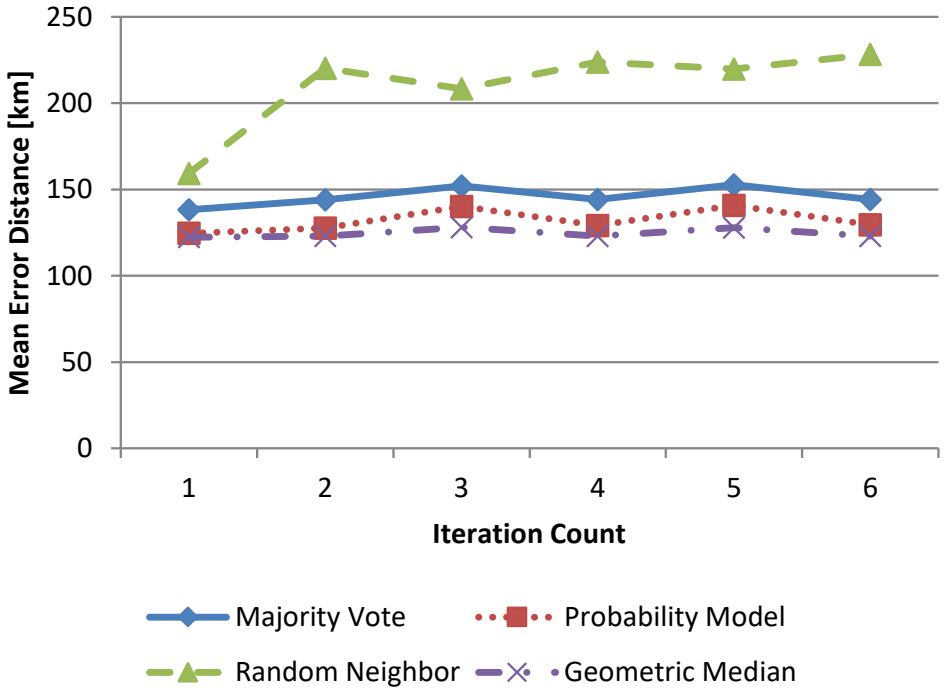
図 4.1: 繰り返し回数と 4 種類の推定関数を変えての評価結果（適合率、再現率、F 値）

を図 4.3 に示す。ここで、‘-’ は同じ居住地を持つユーザが見つからなかったユーザである。この結果から、約 88% のユーザは同じ居住地を持つユーザが 1 ホップ以内（友人と友人の友人）に存在していることがわかった。再現率向上のためには、繰り返し回数が 2 回までのあいだに、より多くのユーザに正しい居住地を推定できるようにすることが重要であると考えられる。

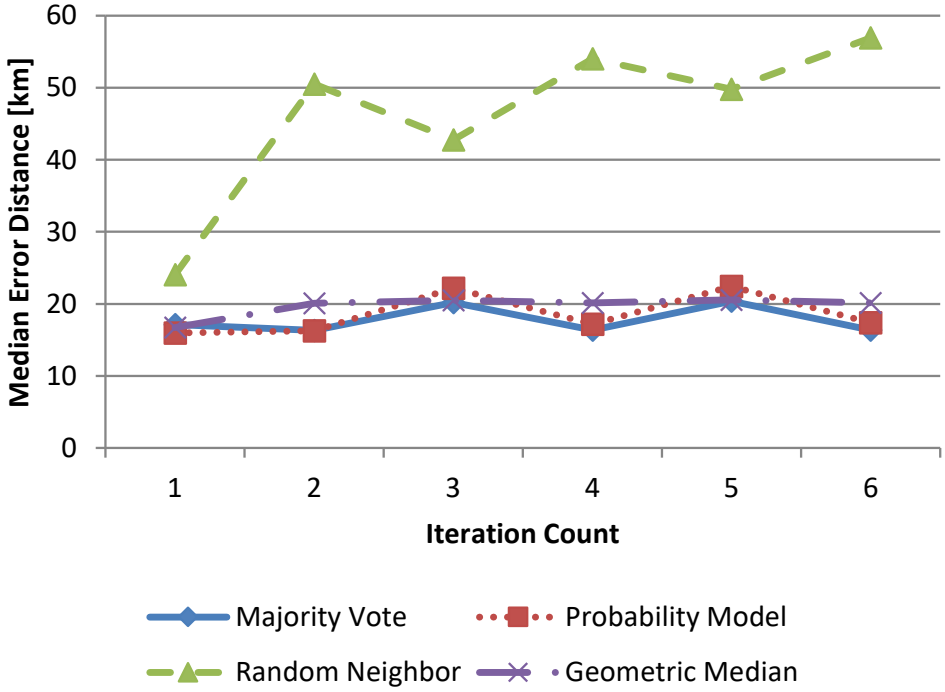
4.4.3　議論と考察

図 4.1 で示したように、我々の実験では Majority Vote が最高性能であった。Jurgens ら [Jurgens 13] は Geometric Median が、Probability Model を除く 3 つの推定関数の中で最も高い性能を示したと報告している[*6]。このような結果が得られた理由が 2 つ考え

[*6] Probability Model は比較に使われていなかった。



(a) 平均エラー距離



(b) 中央値エラー距離

図 4.2: 繰り返し回数と 4 種類の推定関数を変えての評価結果（平均エラー距離、中央値エラー距離）

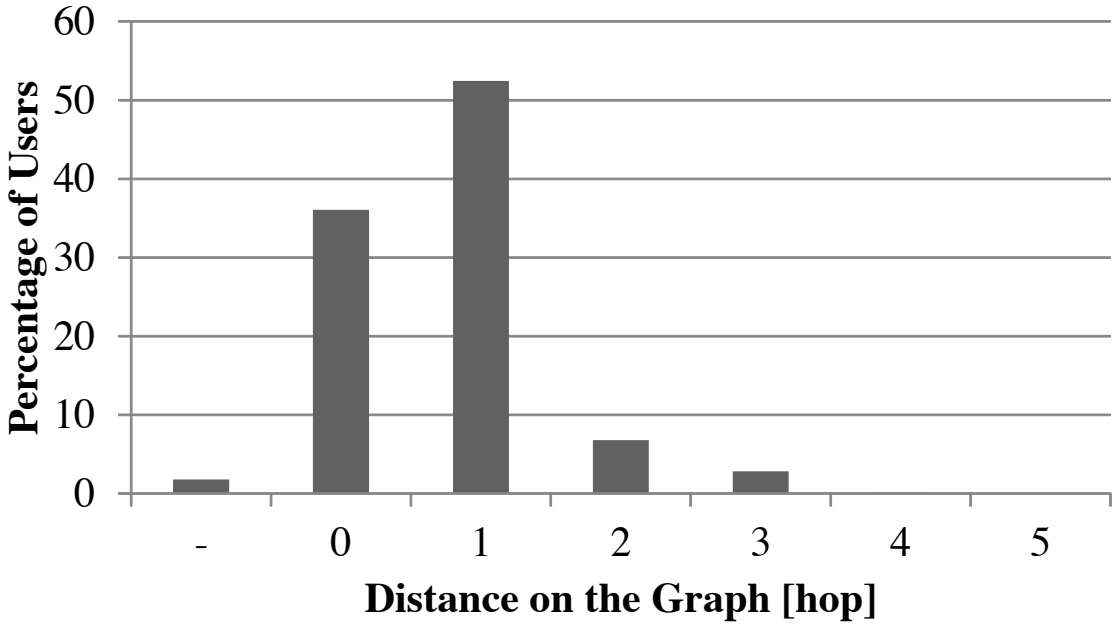


図 4.3: 同じラベルを持つユーザまでの最短距離の分布

られる。1 つ目は居住地のサイズとタイプが異なること、2 つ目はソーシャルグラフのもととなる関係が異なることである。

1 つ目に、居住地のサイズとタイプが異なることが理由として考えられる。Majority Vote は居住地をただのラベルと見なして利用する。推定に利用する投票数がエリアの大きさに依存しやすいため、他の推定関数より居住地の大きさと形に影響されると考えられる。意味のある推定とするためには、ある程度の投票数が必要となる。本研究では、居住地を市区町村レベルのエリアとしているため、Majority Vote の性能が良くなった可能性がある。

2 つ目に、ソーシャルグラフを構築するもととなるユーザ間の関係が異なることが考えられる。本章では、先行研究でも多く使われている [Davis Jr. 11, McGee 13, Dominic 13]、フォローをもとにした関係をソーシャルグラフ構築に利用している。Jurgens らは我々と異なり、相互メンション関係を用いている [Jurgens 13, Jurgens 15]。McGee らは、地理的に近くに存在する友人の割合はユーザ間の関係によって異なると報告している [McGee 11]。そのため、ユーザ間の関係が異なることにより、適した推定関数が変わったのだと考えられる。

4.5　本章のまとめ

本章では、SLP の推定関数をフォローをもとにしたソーシャルグラフを用いて比較した。その結果、隣接ノードの持つラベルの中で出現頻度が最も高いラベルを選ぶ手法である Majority Vote を用いるときに適合率と再現率が最も高くなることがわかった。さらに、SLP の繰り返し回数が 2 回のときに最高性能となった。88% のユーザは 1 ホップ（友人と、友人の友人）以内に同じ居住地を持つユーザが存在するため、繰り返し回数は 2 回で十分であることがわかった。

第 5 章

ソーシャルグラフを用いた居住地推定の性能とユーザプロフィール

5.1　本章の背景

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 [Jurgens 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が、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]。そこで、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おこなわれていた [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の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る。

本章では、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。

5.2　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] がおこなわれている。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られると考えられるが、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、本研究の特徴である。

5.3　データ

本章では、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。

5.3.1　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、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おこなう。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、先行研究 [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、ツイートに付与されている地理座標を含む日本の市区町村（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。

5.3.2　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のあいだ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 5.3.3 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単純無向グラフとなる。

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には日本の市区町村は 1,901 種類含まれており、作成したデータにはそのうち 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

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

[*2] 北緯 20 度から 50 度、東経 110 度から 160 度の範囲。

も居住地推定の対象とするが、分析の際にこれらのユーザを考慮する。

5.3.3　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、場所の文字数、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 フォロワー数/(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。

5.4　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。推定の対象とするとは、つまりそのユーザの持つ関係をソーシャルグラフ構築のために収集するということである。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。

5.4.1　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 とを使って、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 $\hat{l_u} = \mathrm{infer}(V, E, M)$ 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。

本章では、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

[*3] `https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/user-object.html` (viewed 2019-05-13)

手法を実験に用いる。すなわち、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、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る [廣中 17]。

本章では、[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 (5.1) である。ここで、$\arg\max^*$ は同値の集合を返すものと定義する。本章では、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。

$$
\begin{aligned}
S_u &= \underset{l \in \{l_n | n \in N_u\}}{\arg\max^*} |\{v|v \in N_u, l = l_v\}| \\
\mathrm{Infer}(N_u, M) &= \underset{l \in S_u}{\arg\max} |\{n|n \in M, l = l_n\}|
\end{aligned}
\tag{5.1}
$$

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。本章では、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。そのため本章の分析結果は、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。

5.4.2　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。

5.4.3　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。

5.5　結果と考察

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。

表 5.1: 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
フォロー数	LowCut	190	0.3416	0.1738	0.5089	240094
	HighCut	612	0.3230	0.2265	0.7012	330810
フォロワー数	LowCut	210	0.3537	0.1621	0.4582	216170
	HighCut	830	0.3189	0.2345	0.7355	346960
総ツイート数	LowCut	1	0.3116	0.2568	0.8242	388807
	HighCut	5478	0.3350	0.1745	0.5210	245805
1 日あたりのツイート	LowCut	1.804	0.3209	0.1783	0.5556	262130
数	HighCut	13.509	0.3238	0.2203	0.6802	320899
公開リストに入れら	LowCut	35	0.2458	0.0130	0.0529	24936
れている数	HighCut	1	0.3888	0.1232	0.3169	149507
いいね数	LowCut	450	0.3444	0.1268	0.3682	173694
	HighCut	7752	0.3157	0.2440	0.7728	364574
フォロー／フォロワー	LowCut	1.02	0.3692	0.1197	0.3241	152913
比	HighCut	2.35	0.3140	0.2481	0.7901	372726
ユーザ名の文字数	LowCut	8	0.3144	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	0.3461	0.1557	0.4499	212264
場所の文字数	LowCut	15	0.2856	0.0165	0.0579	27336
	HighCut	1	0.3673	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	0.3592	0.1245	0.3466	163519
アカウント作成日から	LowCut	176	0.3117	0.2571	0.8247	389084
の日数	HighCut	498	0.4011	0.0522	0.1301	61398
フィルタなし			0.3117	0.2571	0.8247	389084

5.5.1　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

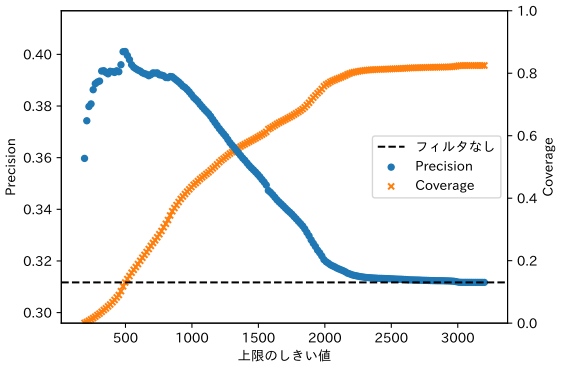
推定しにくいユーザをプロフィールから得た属性で判別できるかを調べる。それぞれの属性を用いたフィルタを適用して居住地推定をおこなった結果のうち、カバー率が 0.05 以上のとき[*4]に適合率が最大になったしきい値と、そのときの適合率、再現率、カバー率を表 5.1 に示す。推定対象のユーザの選択には HighCut と LowCut との 2 種類の条件を用いた。表中の n は推定対象のユーザ数であり、下線はフィルタなしの場合より推定精度が改善したことを意味する。

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、適合率の分布は平均 p、分散 $p(1-p)/n$ の正規分布 $N(p, p(1-p)/n)$ 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、区間が重ならずに適合率が上回ったとき、推定精度が改善したと判定する。

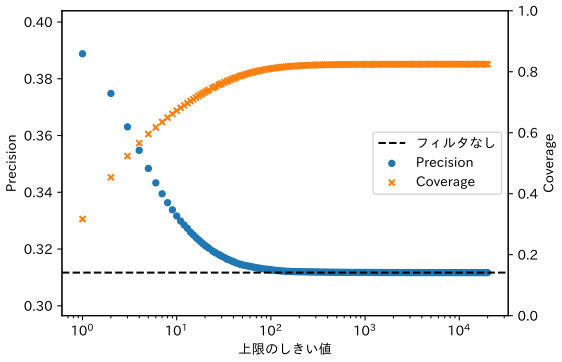
表 5.1 のフィルタを適用したすべての結果をフィルタを適用しなかった場合の結果と比較すると、すべての属性で HighCut と LowCut とのどちらかまたは両方で適合率が向上した。適合率が最も高くなったのは、アカウント作成日からの日数が 498 日より長いユーザを除外したときであった。すなわち、アカウント作成日からの日数が 498 日より長いユーザは居住地を推定しにくいということがわかった。また、アカウントを作成したばかりのユーザを除外しても適合率は向上せず、LowCut はフィルタなしと同じ結果となった。フォロー数、フォロワー数、いいね数、1 日あたりのツイート数、フォロー／フォロワー比では、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、公開リストに入れられている数、名前の文字数、場所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 5.1 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、しきい値より大きな値を持つユーザを推定対象から除外したときの結果を図 5.1a に示す。図における 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長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。

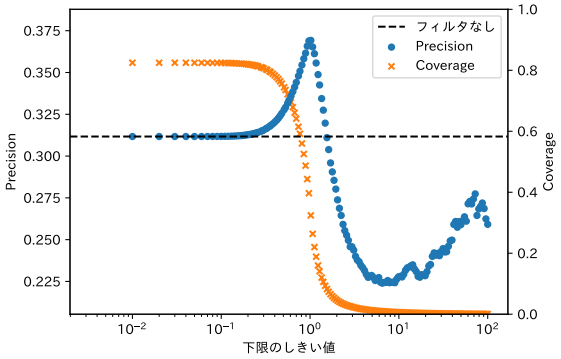
[*4] 5% のユーザは外れ値となる値を持っていると考えて除外した。



(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大きいユーザを除外するとき



(c) フォロー／フォロワー比が小さいユーザを除外するとき

図 5.1: しきい値による適合率の変化

表 5.1 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大きい値を持つ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 5.1b 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 5.1 で 3 番目に適合率が高かったフォロー／フォロワー比が小さ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 5.1c 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。

5.5.2　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、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大

きいとき正の値になり、小さ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。

5.5.1 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 5.1 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 5.2 に示す。図 5.2a、図 5.2c、図 5.2e 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 5.2b、図 5.2d、図 5.2f である。

図 5.2a と図 5.2b がアカウント作成日からの日数の結果である。まず図 5.2a をみると、アカウント作成日からの日数が 500 から 900 までのユーザが多いとわかる。次に図 5.2b 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長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長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 あたりで大きくなっている。さらに、アカウント作成日からの日数が短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 5.2c と図 5.2e へ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 5.2d 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、名前の文字数が多くなるほど大きくなる。図 5.2f に示す自己紹介文の結果では、文字数が多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

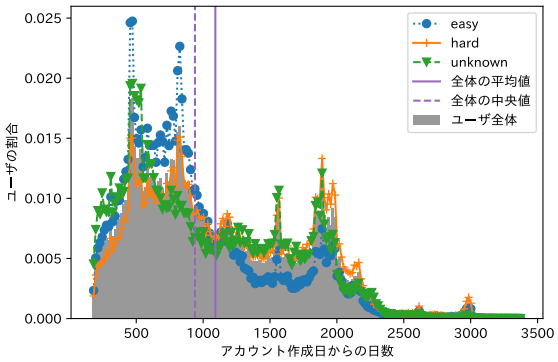
り、誤って居住地を推定するユーザの割合が増える。特に、名前の文字数または自己紹介文の文字数が 1 文字増加するごとに、居住地を誤って推定するユーザのユーザ全体に対する割合が一定の割合で増加することがわかった。名前の文字数と自己紹介文の文字数の偏り度合いの結果を見ても、アカウント作成日からの日数と同様に、ユーザが多く分布している部分で正しく居住地を推定できるユーザの割合が増えていた。この傾向は、ユーザ名の文字数と場所の文字数以外の属性でも確認できた。

5.5.3　考察と限界

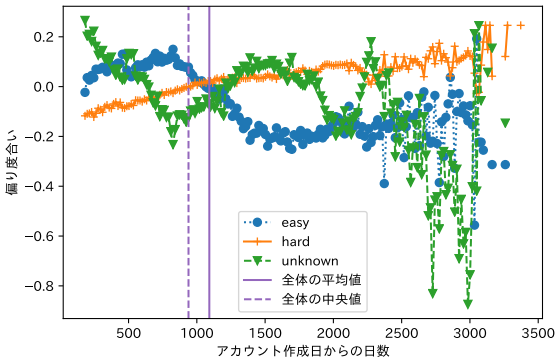
5.5.1 節では、フォロー数やフォロワー数、名前の文字数、自己紹介文の文字数などで居住地の推定しやすさが判別できることがわかった。5.5.2 節では、アカウント作成日からの日数や名前の文字数、自己紹介文の文字数によって、居住地を正しく推定できるユーザの割合や、誤って推定するユーザの割合が変化することがわかった。これらの属性が性能に影響した理由として、推定手法によるものと、ユーザのソーシャルメディア利用方法によるものとがあると考える。

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、具体的にはフォロー数とフォロワー数からは、以下の知見が得られた。5.5.1 節では、フォロ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ー数が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] らは相互リプライをもとにソーシャルグラフを構築したとき、相互リプライ数でユーザを制限する前処理をおこなっていた。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章の結果でも示された。

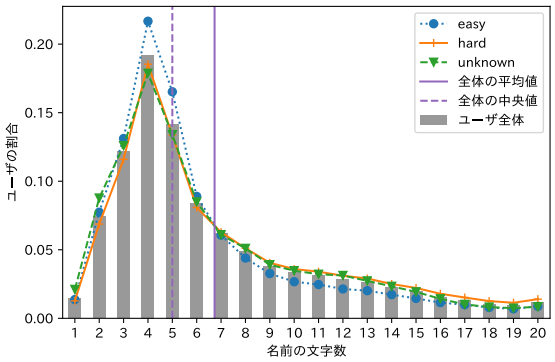
フォロー・フォロワー関係は時間の経過によって変化する [Hironaka 18]。5.5.1 節では、アカウント作成日からの日数が長くなるほど推定が難しくなるという結果が得られた。このことから、Twitter を長く利用するほど地理的近接性を示さないフォロー関係が蓄積していき、居住地が推定しにくくなっていることが示唆される。また、ユーザがアカウントを作成したときにはフォロー数とフォロワー数は共に 0 であるため、アカウント作成日からの日数が短いときはフォロー数とフォロワー数が少なく居住地を推定しにくいと考えられた。しかし、アカウント作成日からの日数が少ないユーザを除外しても推定精度



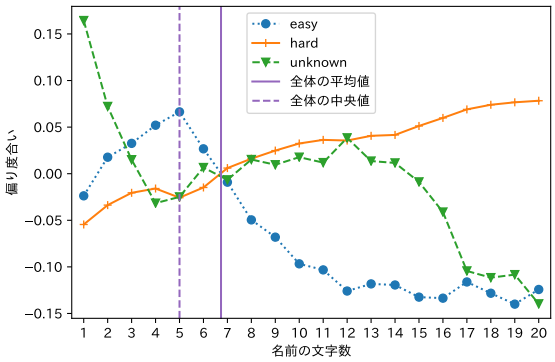
(a) アカウント作成日からの日数（分布）



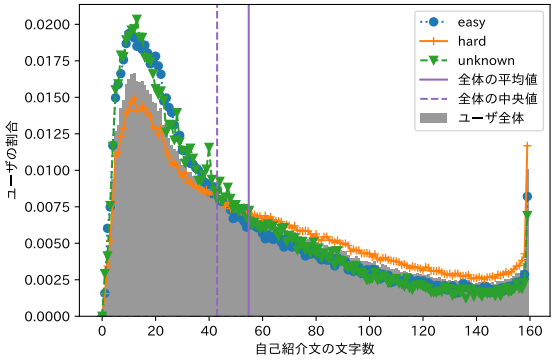
(b) アカウント作成日からの日数（偏り度合い）



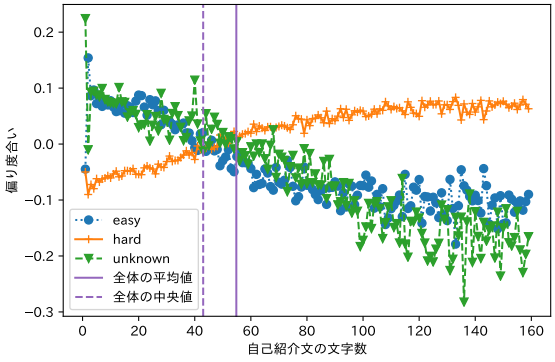
(c) 名前の文字数（分布）



(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 5.2: ユーザプロフィールの値の分布

が上がらなかった。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。5.5.2 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、場所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。

本章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。本章では日本の 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、本研究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、本章で用いた 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 5.3 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。5.5.2 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。

[*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。

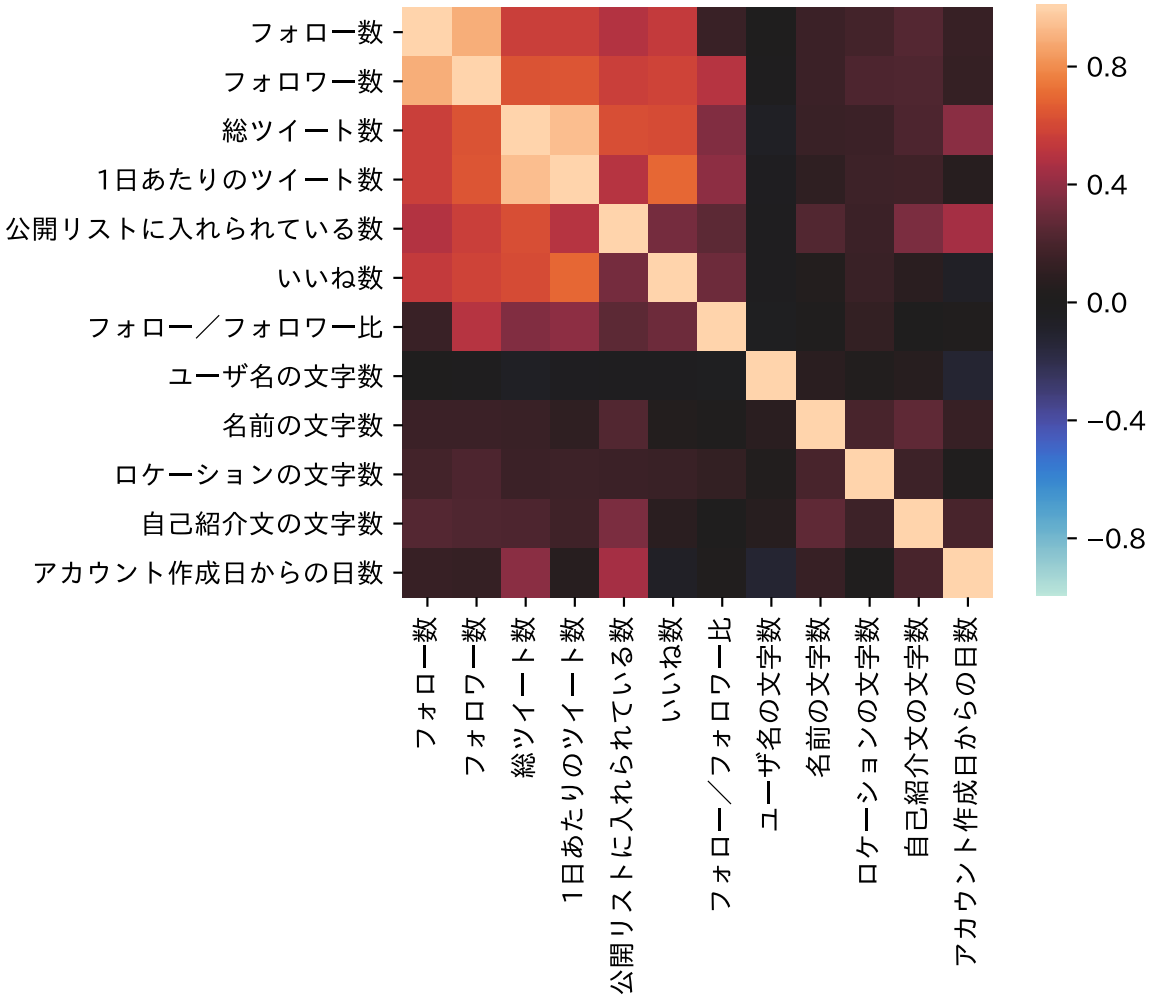


図 5.3: ユーザ属性間のスピアマンの順位相関係数

5.6　本章のまとめ

本章では、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大きいこと、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大きいこと、自己紹介文が長いほど居住地を誤って推定するユーザの割合が大きいことなども明らかになった。本章では、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できるため、本章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できる。

第 6 章

ソーシャルグラフの中心性と居住地推定性能

6.1 本章の背景

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が利用されている [Jurgens 15]。もととなる仮定は、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ものである。しかしながら、Rahimi らによって、多くのユーザとメンションしているユーザなど、居住地の推定がうまくいかないユーザが存在することが明らかになっている [Rahimi 15]。このような推定が難しいユーザのことは Celebrity と呼ばれている。さらに、Ebrahimi らは、すべての Celebrity の推定が難しいのではなく、メンション相手が地理的に広く散らばっている Global celebrity の推定が難しいことを、Celebrity のクラスタリングによって明らかにした [Ebrahimi 18]。我々は、ソーシャルグラフ上で近くに位置しているが、地理的には近くに位置していないユーザとの関係を多く持つユーザには、何らかのネットワーク的特徴があると考える。本章では、ネットワーク的特徴をとらえるために中心性指標を用いて、居住地推定が難しいユーザの特徴を分析する。次数中心性や PageRank [Page 98], HITS [Kleinberg 99] などの中心性指標は、ユーザのランキングなどに用いられており [Weng 10, 山口 11, Chien 14]、周囲のユーザと似た属性値を持っているかどうかと関連していると考えられる。

本章では、日本の Twitter ユーザを対象に、複数の中心性指標とそのユーザが持つ友人の居住地の類似性とが、どのような関係にあるのかを分析する。その結果、多数の友人と居住地を共有しているユーザは、PageRank と HITS の Authority と Hub の分布に違いがあることがわかった。加えて、Authority と Hub となるユーザとの 2 種類が存在するという HITS の仮定は、全てのユーザが同質であると考えるより、Twitter のソーシャルグラフの性質と合うことがわかった。

6.2　データ

本節では、分析に用いるデータセットについて述べる。データセットは居住地データとソーシャルグラフデータからなる。

6.2.1　居住地データ

ユーザは主に居住地周辺でツイートを投稿すると考え、主に位置情報付きツイートを投稿する場所をそのユーザの居住地とする。具体的には、位置情報付きツイートに付与されている地理座標（coordinates）を市区町村レベルのエリアと照合し、最も投稿回数が多いエリアをそのユーザの居住地とする。Twitter Streaming API を用いて 2014 年に投稿された日本を包含する矩形[*1]内の位置情報付きツイートを収集した。そして、総務省統計局の境界データ[*2]を用いて、それぞれの位置情報付きツイートに含まれている地理座標（coordinates）を含む日本の市区町村を照合した。

付与する居住地の正確さを上げるため、同じエリアで 5 回以上投稿しているユーザだけに絞り込み、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、471,761 ユーザに対して 1873 種類の居住地を付与できた。

6.2.2　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のフォローしているユーザ集合とフォロワーの集合を 2015 年 7 月に収集した。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 の方向へ有向エッジを作ることで、ソーシャルグラフを構築し、居住地を付与されていないユーザへのエッジは除外する。

収集したデータをもとに、471,761 ノード（ユーザ）と 8,295,355 エッジを含むソーシャルグラフが構築できた。このソーシャルグラフに含まれるすべてのユーザは居住地を付与されている。各ノードの平均エッジ数は 17.58 であり、居住地を付与されている相手との平均相互フォロー数は 13.2 であった。また、471,761 ユーザのうち 42,316 ユーザはエッジを持たない孤立ノードであった。

[*1] 北緯 20 度から 50 度、東経 110 度から 160 度の範囲。

[*2] `https://www.e-stat.go.jp/` (viewed 2020-12-02)

6.3　分析方法

ユーザの持つ中心性の値と友人と同じ値を持つ傾向との関係を分析する。まず中心性の値をどのように計算するかを説明し、次に傾向の測り方について説明する。そして、同じ値を持つ傾向ごとに中心性の値に偏りがあるかを計算する。

6.3.1　中心性指標

中心性指標として、入次数中心性、出次数中心性、PageRank [Page 99]、HITS アルゴリズム [Kleinberg 99] で計算される Authority と Hub を用いる。中心性指標はユーザ（ノード）ごとに計算される値である。

入次数中心性は、各ユーザのフォロワーのうち居住地を付与されているユーザの数であり、フォロワー数が多いほど大きな値を持ちやすくなる。出次数中心性は入次数中心性の反対で、各ユーザがフォローしているユーザのうち居住地を付与されているユーザの数であり、フォロー数が多いほど大きな値を持ちやすくなる。無向グラフの次数は先行研究 [Rahimi 15, Ebrahimi 18] において Celebrity（有名人）を除外する際に用いられてきた指標であるため、有向グラフを用いる本研究では入次数と出次数を分析に用いる。

PageRank は入次数中心性と似ているが、より大きな値を持つユーザにフォローされているほどそのユーザの値は大きくなる。PageRank では、フォロワーが多いユーザだけでなく、そのようなユーザにフォローされているユーザの中心性の値も高くなることを期待している。先行研究 [Kwak 10] において、影響力のあるユーザを発見するために PageRank が使われている。

Authority と Hub は、HITS アルゴリズムによって同時に計算される値である。Authority は Hub が高いユーザにフォローされていると高くなり、Hub は Authority が高いユーザをフォローしていると高くなるよう定義された指標である。一般的に、Authority は多くのユーザにフォローされていると高くなりやすく、Hub は多くのユーザをフォローしていると高くなりやすい。HITS は PageRank と共にユーザランキングをする際に用いられている [Chien 14]。我々は情報配信元となっており多くのユーザのフォローされているユーザ（Authority が高い）と、情報収集を目的として良い情報源を多く知っているユーザ（Hub が高い）が存在すると仮定して HITS を用いる。これらの中心性指標は、6.2.2 節で構築したソーシャルグラフを用いて、NetworkX[*3]によって計算した。計算の際のパラメータはデフォルト値を用いた。

[*3] `https://networkx.org/` (viewed 2020-12-02)

6.3.2　友人との居住地の類似性

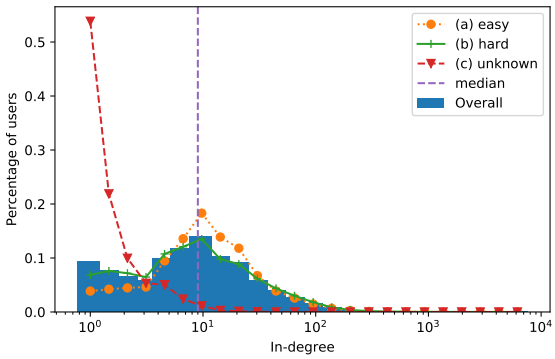
友人と同じ値を持つ傾向を測るために、友人の居住地との類似性を用いる。友人の居住地との類似性は、ソーシャルグラフを用いる居住地推定手法 [Davis Jr. 11] によって居住地が正しく推定できたかどうかで判定する。Davis Jr. らの提案した居住地推定手法は、友人の持つ居住地の中で最も出現頻度が高いものをそのユーザの居住地として推定する。この手法による推定結果を用いて、次の 3 つのグループにユーザを分類する：(a) 居住地を正しく推定できたユーザらの easy グループ、(b) 居住地を誤って推定したユーザらの hard グループ、(c) 手がかりがなく居住地を推定することができなかったユーザらの unknown グループ。これらのグループはすなわち、(a) 多数の友人と同じ居住地を持っているユーザのグループ、(b) 多数の友人と同じ居住地を持っていないユーザのグループ、(c) 手がかり（友人）がなく類似度を測ることができなかったユーザのグループとなる。

本章では、相互フォロー関係にあるユーザを友人とみなすことにする。すなわち、居住地推定には、6.2.2 節で構築したソーシャルグラフのうち、相互にエッジが存在する場合のみを取り出した無向グラフを利用する。居住地推定の正しさはラベルが正確に一致するかどうかで判定し、評価対象としているユーザのラベルのみを隠してその他のユーザすべてを推定に用いる leave-one-out 交差検証によって評価する。推定したラベルが本来のラベルと正確に一致したユーザを easy グループ、誤ったラベルを推定したユーザを hard グループ、手がかりがなくラベルを推定することができなかったユーザを unknown グループに分類する。

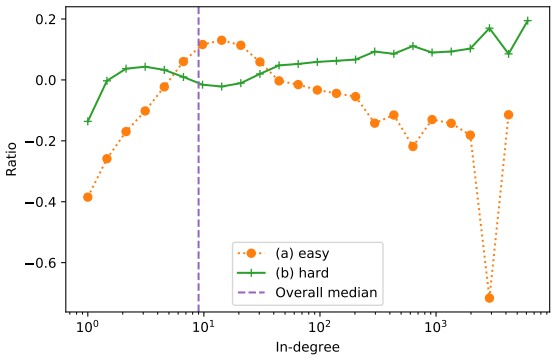
6.3.3　偏り度合い

ユーザ集合 $U \subseteq V$ に対して中心性の分布は次のように計算する。まず、総ユーザ数を $N = |U|$、中心性の値が区間 $i : [x_i, x_{i+1})$ に含まれるユーザの数を n_i とする。そのとき、区間 i に含まれるユーザの割合 $f(i;U)$ は n_i/N である。$f(i;U)$ をスコア分布と呼ぶ。

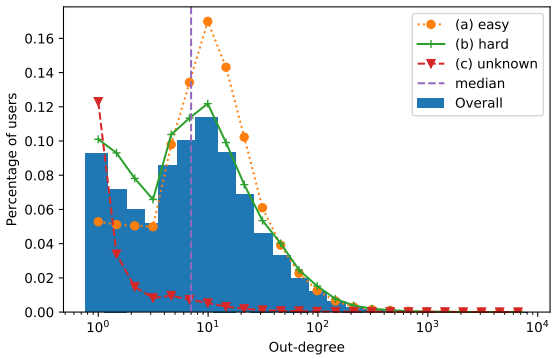
まず、すべてのユーザ集合 V に対してスコア分布を計算する。また、友人との類似度によって分けたユーザグループそれぞれに対してもスコア分布を計算する。そして、友人との類似度によって分けたユーザグループごとの、すべてのユーザ集合に対する偏りを明らかにするため、これらの分布の差を次の方法で計算する。区間 i に対応するすべてのユーザの集合 V のスコア分布の値を $f(i;V)$、あるユーザグループ U の区間 i に対応するスコア分布の値を $f(i;U)$ とする。このとき、偏り度合いの分布を $\log_{10}(f(i;U)/f(i;V))$ と定義する。あるユーザグループの偏り度合いの分布の値は、区間 i において、そのユーザグループのユーザがその区間にいる割合が、すべてのユーザ集合をもとに計算された割



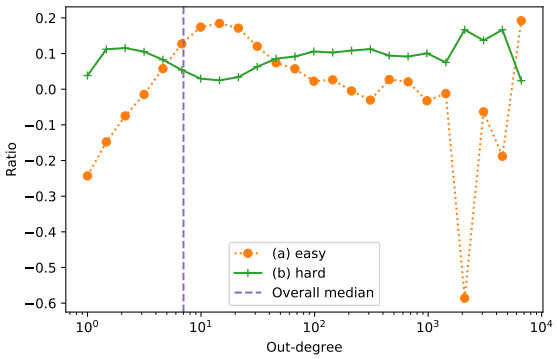
(a) スコア分布（入次数中心性）



(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）



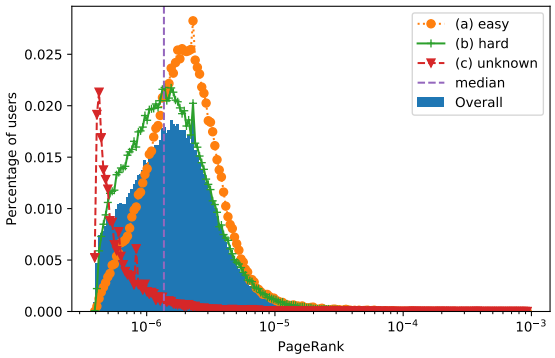
(d) 偏り度合いの分布（出次数中心性）

図 6.1: 入次数中心性と出次数中心性の分布

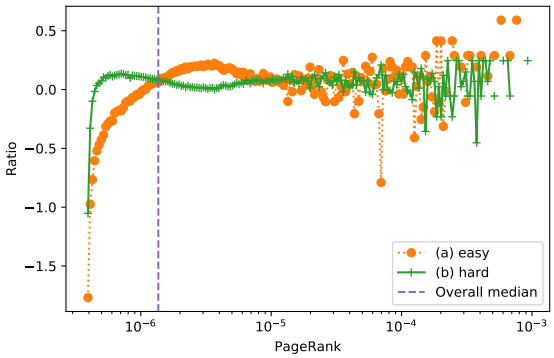
合に対して大きいとき正の値、小さいとき負の値をとる。偏り度合いの分布の値の絶対値は、比較している分布間の差が大きいほど大きくなる。

6.4　結果と考察

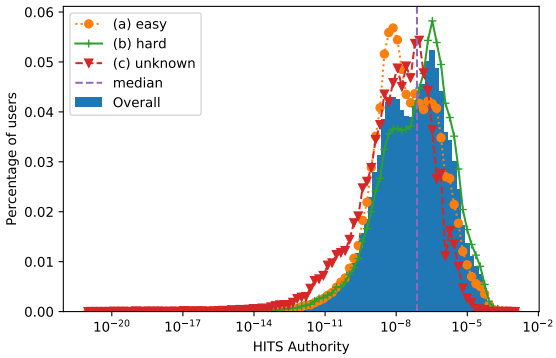
居住地を付与されているすべてのユーザの集合（471,761 ユーザ）を次の 3 つのユーザグループへ分類した。(a) 友人の多数と同じ居住地を持つ 121,275 ユーザは easy グループ、(b) 友人の多数と同じ居住地を持たない 267,809 ユーザは hard グループ、(c) 友人の中に居住地が判明しているユーザが存在しなかった 82,677 ユーザは unknown グループとした。すべてのユーザの集合（Overall）と 3 つのユーザグループとのそれぞれの中心性をもとに、スコア分布を計算した。そして、easy グループと hard グループとの Overall に対する偏り度合いの分布を計算した。unknown グループは友人との類似度が計算できなかったユーザであるため、結果から除外した。入次数中心性、出次数中心性、PageRank、HITS の Authority と Hub の結果を図 6.1 と図 6.2 に示す。



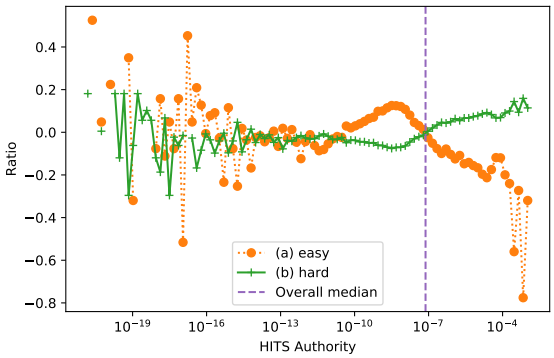
(a) スコア分布（PageRank）



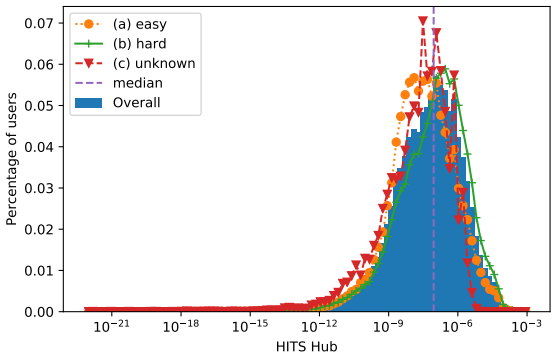
(b) 偏り度合いの分布（PageRank）



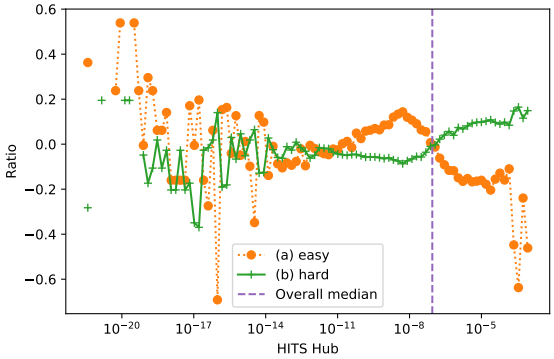
(c) スコア分布（HITS Authority）



(d) 偏り度合いの分布（HITS Authority）



(e) スコア分布（HITS Hub）



(f) 偏り度合いの分布（HITS Hub）

図 6.2: PageRank と HITS Authority, Hub の分布

図 6.1 の入次数中心性と出次数中心性のスコア分布を見ると、easy グループと hard グループとの分布のピーク位置に大きな差はない。偏り度合いの分布をみると、次数中心性の値が 20 付近に easy グループのユーザが多く位置していることがわかる。そして、中心性の値が 20 から大きくなっていくほど、easy グループのユーザの割合が減少し、hard グループのユーザの割合が増えている。もし、単純に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推定が難しい Celebrity だとみなすと、これら Celebrity ユーザたちは推定が難しいユーザであるといえる。Davis Jr. らは、相互フォロー数が 20 から 200 までのユーザを推定に使うときに、適合率が最も高くなったと報告している [Davis Jr. 11]。相互フォロー数の数え方が異なるため単純に比較することはできないが、相互フォロー数を相互フォローをエッジとしたネットワークの次数中心性だと考えると、中心性の値が小さすぎるユーザと大きすぎるユーザとは友人と類似していない居住地を持っている、つまり推定が難しいという結果は一致している。

PageRank と HITS の結果を図 6.2 に示す。PageRank のスコア分布をみると、easy グループのユーザは hard グループのユーザよりも高い値を持つ傾向にあることがわかる。HITS の Authority と Hub のスコア分布では、easy グループと hard グループとで分布のピーク位置が異なっている。Authority と Hub との両方の結果において、easy ユーザで中央値を中心として左側（値が小さい）、hard ユーザで右側（値が大きい）にピーク位置がある。スコア分布において、HITS アルゴリズムで計算した Authority と Hub では、easy ユーザと hard ユーザとの分布の山の中心がずれていることから、Twitter ユーザが友人と同じ居住地を持っているかを判別するには、HITS の値を用いるほうが、PageRank より良い解釈であると考えられる。この結果は、Twitter のソーシャルグラフには、多くの有名人をフォローしているハブとなっているユーザ（Hub の値が高い）と、多くの良い読者を抱えている有名人のようなユーザ（Authority の値が高い）が存在していることを示唆している。また、偏り度合いの分布をみると、Authority と Hub の両方の結果において、高い値の部分で easy グループのユーザの割合が減っている。PageRank の値が高い部分ではこのような結果は見られなかった。すなわち、Hub の値が高いユーザであるかまたは Authority の値が高いユーザであるかのどちらかの条件を満たすことで、友人と同じ居住地を持つユーザが減るということを示しており、両方の条件を同時に満たす必要はないことがわかる。

6.5　本章のまとめ

本章では、ユーザが友人と同じ居住地を持つ傾向と、ユーザの中心性スコアの値との関係を分析した。中心性として、入次数中心性、出次数中心性、PageRank、HITS の Authority と Hub を用いた。その結果、多数の友人と同じ居住地を持つ easy グループ

に分類されたユーザは、高い PageRank スコアを持つことがわかった。さらに、推定が難しい hard グループに分類されたユーザをみたときに、HITS の Authority または Hub が高いユーザが多いことがわかった。このことから、推定が難しいユーザには、高い Authority を持つ有名なユーザと、高い Hub を持つ購読目的のユーザとの 2 種類がいると解釈できる。また、PageRank の結果と比較することで、両方の条件ではなく、どちらか片方の条件を満たせば推定が難しいユーザとなると考えられる。

第 7 章

結論

本論文では、居住地推定法を用いてソーシャルグラフのプロパティの分析に取り組んだ。ソーシャルグラフのプロパティとして、ソーシャルグラフのエッジとするユーザ間の関係、ソーシャルグラフ上でのユーザ属性の分布、ユーザのプロフィール、ソーシャルグラフ上でのユーザの中心性に着目した。

ソーシャルグラフは、ユーザをノード、ユーザ間の関係をエッジとして表現したものである。エッジのリストが与えられればソーシャルグラフの形が定まるため、ソーシャルグラフはユーザ間の関係により構成されていると考えられる。あるソーシャルグラフを用いて居住地推定をしたとき、推定に利用した関係が地理的な近さを示す関係であれば、その性能が良くなると考えられる。居住地推定が正解するということはすなわち居住地が同じユーザとつながりがあるということであるため、推定が簡単なユーザとは、つながっているユーザとの地理的な距離が近いことを意味する。地理的な近さには対面（オフライン）での交流の有無が大きく関わっていると考えられる。

第 3 章では、4 種類のユーザ間の関係をもとに作成した 4 種類のソーシャルグラフを用いて、ユーザ間の関係が居住地推定に与える影響を調査した。その結果、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グラフが居住地推定に最も有効であることを示した。加えて、日本のソーシャルグラフを用いる居住地推定においては、友人の居住地の中から最頻のものを選択する推定手法が、ソーシャルグラフの形状に影響を受けず最も精度良く居住地が推定できることを示した。

ソーシャルグラフを用いる居住地推定では、向きのあるエッジで指し示している相手のユーザに対して、お互いの居住地が地理的に近いと仮定している。しかしながら、ソーシャルグラフには地理的な近さを示さない関係も含まれており、このような関係は推定においてノイズとなる。これらをもとに第 3 章で得られた結果を解釈すると、フォローされている関係にある相手と最も近さを示すことがデータから明らかになったと考えられる。

得られた結果で興味深い点は、ユーザ A がユーザ B をフォローしているときに、ユーザ B の推定の手がかりとしてユーザ A を利用することが良いという点である。フォローする相手は選ぶことができるが、フォローされる相手は基本的に選ぶことができない。フォ

ロー関係は人間関係の一種と考えられる。人間関係に向きがあることは一般に知られていることであるが、ソーシャルグラフを用いた居住地推定において関係の向きを適切に考慮した研究がこれまでおこなわれていなかった。本研究では、向きを考慮したソーシャルグラフを用いて居住地推定の性能を調べたことで、Twitter 以外のソーシャルメディア上のソーシャルグラフにおいても、向きのある関係を考慮することの重要性を明らかにした。

第 4 章での実験により、SLP による居住地推定では、繰り返し回数が 2 回のときに適合率と再現率が最も高くなることがわかった。また、隣接ノードの持つラベルの中で最頻のものを選ぶ手法を SLP の推定関数として用いたときに、適合率と再現率が最も高くなることがわかった。さらに、同じ居住地を持つユーザ間のグラフ上での距離について分析し、88% のユーザは 1 ホップ以内（友人と友人の友人の範囲内）に同じ居住地を持つユーザが存在することがわかった。

隣接ノードの情報を用いた推定を繰り返す SLP は、推定した居住地を新たな学習ラベルとして採用するため、直接つながっているユーザ間だけでなく一定の範囲内の近さを考慮する。得られた結果から、フォローをもとにしたソーシャルグラフにおいては、あるユーザと近さを示す関係は 1 ホップ以内に存在すると解釈できる。

第 5 章では、ソーシャルグラフの形状に影響を与えると考えられる、ユーザのプロフィールと居住地推定性能との関係を分析した。その結果、プロフィールから得られる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。特に、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高いこと、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高いことなどが明らかになった。地理的な近さを示す関係を持つかどうかとユーザのプロフィール属性との関連が明らかになったことで、ソーシャルメディア上のソーシャルグラフにおいて近さを示す関係の持ちやすさが、プロフィールから得られるユーザの特徴によって異なることが明らかになった。

得られた結果からは、推定が難しい理由として次のものが考えられた：関係が古い、有名人であるために友人以外のつながりが多い、アクティビティが少なすぎて推定に使える手がかりが少ない、オンラインのつながりが多い。以上の理由を整理すると次のようになる。まず最初に、ソーシャルグラフを用いた推定における問題には、(A) 推定の手がかりとなる関係の有無がある。この理由は、推定の手がかりとなるソーシャルメディア上でのアクティビティがないことにある。次に、(B) 関係が推定のもととなる仮定を満たしているかどうかが問題となる。推定のもととなる仮定を満たしていない理由として、(B-1) 関係が古い、(B-2) オンライン上のつながりが多い、(B-3) 有名人であるの 3 つが考えられた。有名人はオンライン上のつながりを多く持つと考えられるが、有名人でなくとも、オンライン上でのつながりを求めているユーザであれば、オンライン上のつながりを多く持つと考えられる。これらの条件を満たす推定が難しいユーザの持つ関係は、地理的な近さ

を示す可能性が低いと解釈できる。

第 6 章では、ユーザのソーシャルグラフ上での中心性と居住地推定の性能との関係を調べた。高い PageRank や HITS の Authority、Hub のスコアを持つユーザは居住地推定が難しいことがわかった。第 6 章の分析は第 5 章の分析と比べて、よりソーシャルグラフの形状に着目した分析となっている。第 5 章では、ソーシャルグラフの形状を表す属性としてフォロー数、フォロワー数、フォロー／フォロワー比の 3 つを扱った。これに対して、第 6 章ではフォロー数とフォロワー数それぞれに対応する中心性として出次数中心性と入次数中心性があった。さらに、フォロワーのフォロワーなど直接のつながり以外も考慮した場合のフォロワー数をもとにした中心性である PageRank に加えて、フォローを重点的にする Hub ユーザと、フォローされる対象となっている Authority ユーザを仮定した、HITS の値を比較した。その結果、推定が難しいユーザとして、フォロワー数が多い Authority ユーザと、フォロー数が多い Hub ユーザとがいることがわかった。ただし、PageRank の値が大きくなっても推定が難しいユーザは大きく増えないことから、フォロワー数が多いユーザの中でも推定が比較的難しくないユーザが存在することが明らかになった。

つながりを多く持っている有名人のようなユーザは推定が難しいことがわかっており、有名人とはフォロワー数が多いユーザだと考えられる。しかし、フォロー数が多いユーザも推定が難しい。その理由は、フォロー数が多いほどフォロワー数も多くなる傾向にあるからである。しかし、フォロー数が多くフォロワー数が少ないユーザも推定が難しいことが HITS の結果からわかった。第 5 章のフォロー比の結果を見ると、フォロー数と比べてフォロワー数が多すぎるユーザ、フォロワー数と比べてフォロー数が多すぎるユーザの推定が難しかった。ソーシャルグラフの形状から計算したユーザの特徴を見た場合でも、これらの特徴を持つユーザは地理的な近さを示す関係を持ちにくいことがわかったと解釈できる。

本論文では、日本の Twitter ユーザを対象に分析をした。この分析は、Twitter に投稿された丸 1 年分の日本の位置情報付きツイート 140,055,452 件をもとにしたものであり、他に類を見ない大規模なものである。ソーシャルグラフのプロパティとして考えられるものは他にも存在するが、本論文では定量的に評価できるものを選んで、属性推定との関係を調べた。

謝辞

本論文をまとめるにあたり、梅村恭司教授には丁寧にご指導いただきました。また、北岡教英教授、土屋雅稔准教授には、本論文の審査を通じ、多くのご助言をいただきました。感謝いたします。

吉田光男助教には、研究の初歩から丁寧にご指導いただき、成長を見守っていただきました。数多くのアドバイスやご指導は今後の活動の糧となると思います。深く感謝いたします。

県立広島大学経済情報学部の岡部正幸准教授にも、共著者として研究に対するご指導をいただきました。感謝いたします。

本研究では、計算の大部分に豊橋技術科学大学広域連携教育研究用クラスタシステムを利用しました。計算資源の心配をせず、大規模な分析をおこなえたことに感謝いたします。

博士前期課程 2 年のときには、産業技術総合研究所メディアインタラクション研究グループに技術研修生として受け入れていただきました。その際、メンターとして研究のご指導をいただいた佃洸摂さんには、研究に関する多くのご指導・ご助言をいただきました。また、後藤真孝さん、濱崎雅弘さんをはじめ、メディアインタラクション研究グループの皆様にも大変お世話になりました。感謝の意を表します。

所属する梅村研究室の皆様にも多くのご支援をいただきました。最後に、ここまで育てて、見守ってくれた家族に感謝します。ありがとうございました。

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)

[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)

[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)

[Chien 14] Chien, O. K., Hoong, P. K., and Ho, C. C.: A Comparative Study of HITS vs PageRank Algorithms for Twitter Users Analysis, in *Proceedings of the 2014 International Conference on Computational Science and Technology*, pp. 1–6 (2014)

[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)

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

[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)

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

[Gleich 15] Gleich, D. F.: PageRank Beyond the Web, *SIAM Review*, Vol. 57, No. 3, pp. 321–363 (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)

[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. WII–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)

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

[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)

[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)

[Kleinberg 99] Kleinberg, J. M.: Authoritative sources in a hyperlinked environment, *Journal of the ACM*, Vol. 46, No. 5, pp. 604–632 (1999)

[Kong 14] Kong, L., Liu, Z., and Huang, Y.: SPOT: Locating Social Media Users Based on Social Network Context, *Proceedings of the VLDB Endowment*, Vol. 7, No. 13, pp. 1681–1684 (2014)

[Kulshrestha 12] Kulshrestha, J., Kooti, F., Nikravesh, A., and Gummadi, K.: Geographic Dissection of the Twitter Network, in *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, pp. 202–209 (2012)

[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)

[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)

[Luo 20] Luo, X., Qiao, Y., Li, C., Ma, J., and Liu, Y.: An overview of microblog user geolocation methods, *Information Processing and Management*, p. 102375 (2020)

[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)

[McPherson 01] McPherson, M., Smith-Lovin, L., and Cook, J. M.: Birds of a Feather: Homophily in Social Networks, *Annual Review of Sociology*, Vol. 27, No. 1, pp. 415–444 (2001)

[Mislove 10] Mislove, A., Viswanath, B., Gummadi, K. P., and Druschel, P.: You Are Who You Know: Inferring User Profiles in Online Social Networks, in *Proceedings*

of the 3rd ACM International Conference on Web Search and Data Mining, pp. 251–260 (2010)

[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)

[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)

[Page 98] Page, L. and Brin, S.: The anatomy of a large-scale hypertextual Web search engine, *Computer Networks*, Vol. 30, No. 1-7, pp. 107–117 (1998)

[Page 99] Page, L., Brin, S., Motwani, R., and Winograd, T.: The PageRank Citation Ranking: Bringing Order to the Web., Technical report, Stanford InfoLab (1999)

[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)

[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)

[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)

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

[Tanaka 14] Tanaka, A., Takemura, H., and Tajima, K.: Why You Follow: A Classification Scheme for Twitter Follow Links, in *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, pp. 324–326 (2014)

[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)

[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 on Conference on Information and Knowledge Management*, pp. 1693–1702 (2016)

[Weng 10] Weng, J., Lim, E.-P., Jiang, J., and He, Q.: TwitterRank: Finding Topic-sensitive Influential Twitterers, in *Proceedings of the third ACM International Conference on Web Search and Data Mining*, p. 261 (2010)

[山口 11] 山口 祐人, 天笠 俊之, 高橋 翼, 北川 博之：情報伝搬を考慮したグラフ分析による Twitter ユーザランキング手法, 情報処理学会論文誌 データベース, Vol. 4, No. 2, pp. 142–157 (2011)

[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)

[吉田 16] 吉田 光男, 荒瀬 由紀：トレンドキーワードに関するウェブリソースの横断的分析, 情報処理学会論文誌 データベース, Vol. 9, No. 1, pp. 20–30 (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)

[Zhu 03] Zhu, X., Ghahramani, Z., and Lafferty, J.: Semi-Supervised Learning Using

Gaussian Fields and Harmonic Functions, in *Proceedings of the 20th International Conference on Machine Learning*, pp. 912–919 (2003)

[総 18] 総務省：平成 30 年版情報通信白書 (2018)

博士論文に関する論文

本論文における第 3 章の内容は次の査読付き学術雑誌論文として公表済みである。

- 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)
  doi:10.1527/tjsai.WII-M

本論文における第 4 章の内容は次の査読付き国際会議論文として公表済みである。

- Shiori Hironaka, Mitsuo Yoshida, Kyoji Umemura.
  Analysis of Home Location Estimation with Iteration on Twitter Following Relationship.
  The 2016 International Conference on Advanced Informatics: Concepts, Theory and Application (ICAICTA 2016), Penang, Malaysia, August 2016.
  doi:10.1109/ICAICTA.2016.7803100

本論文における第 5 章の内容は次の査読付き学術雑誌論文として公表済みである。

- 廣中 詩織, 吉田 光男, 梅村 恭司.
  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.
  人工知能学会論文誌, Vol. 35, No. 1, pp. E-J71_1–10 (2020)
  doi:10.1527/tjsai.E-J71

本論文における第 6 章の内容は次の査読付き国際会議論文として公表済みである。

- Shiori Hironaka, Mitsuo Yoshida, Kyoji Umemura.
  User's Centrality Analysis for Home Location Estimation.
  The 4th International Workshop on Application of Big Data for Computational Social Science (ABCSS 2019), Thessaloniki, Greece, October 2019.
  doi:10.1145/3358695.3360930