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.