5B-04

位置情報付 SNS データを用いた情報発信拠点の分析

藤本　祥二[†]　　石川　温[†]　　水野　貴之[‡]
金沢学院大学[†]　　国立情報学研究所[‡]

1. はじめに

社会を理解するためには、社会の現状を把握することが不可欠となる。その代表的な方法として、現在、国勢調査などの公的統計調査が、世界各国で実施されている。日本では 5 年に一度国勢調査が行われているが、調査の時間間隔がそれより長い国は少なくない。この頻度を上げることは、社会のリアルタイムな状況を把握するために重要である。本研究の目的は、Twitter や Facebook のようなソーシャルネットワークサービス（SNS）のビッグデータに含まれる全地球測位システム（GPS）データなどの位置情報を用いて、この問題に対する一つのアプローチを示すことである。

我々は、先行研究［1］により、Tweet データの位置情報を用いて、各ユーザが頻繁に Tweet した地域（Tweet 拠点エリア）の分布を観察し、国勢調査の人口分布と比較し、両者の間に強い相関があることを明らかにした。しかし、図 1 のように両者の散布図における分散は大きく、国勢調査による人口よりも Tweet 拠点としている人口の方が多い地域（エリア）が多数存在したのは大きな問題であった。本研究では先行研究の問題点である、国勢調査による人口よりも Tweet 拠点としている人口の方が多い地域が多数存在するという現象を解決する手法を提案し、その分析を行い、どのような結果が得られたかを説明する。そして、本研究における新しい発見とその結果からの将来の展望を示す。

2. Tweet 拠点エリアと自宅エリアの同一性

図 1、2 における分散が大きいということは、ユーザが最も頻繁に Tweet を行っている Tweet 拠点エリアが自宅エリアと異なっている例が多数存在することを示している。例えば、駅、仕事場や学校、あるいは商業施設等での Tweet が自宅の Tweet より多いユーザが居ることは十分にありえる。そこで我々は、「〇〇now」「〇〇なう」「〇〇ナウ」のように、Tweet にユーザが自宅に居ることを主張する言葉が含まれている Tweet に注目し、これらは自宅エリアから発信されているものだと扱うことができる。本研究のポイントは、この自宅エリアの正解が分かっているユーザに対して、Tweet 拠点エリアと自宅エリアにどのような差があるかを観測する事である。

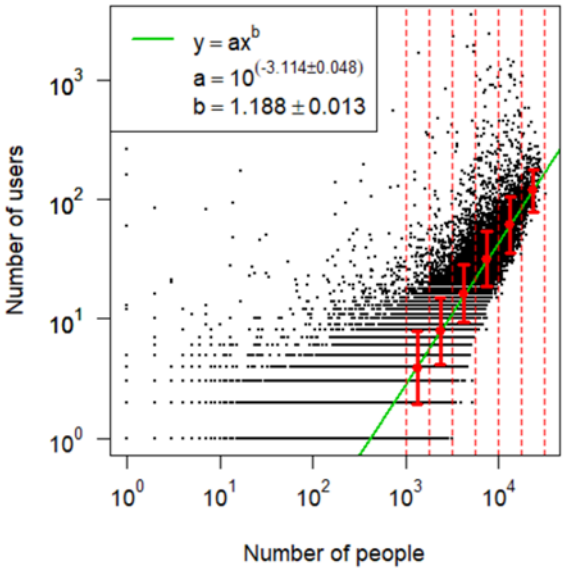

図 1　各エリアの人口と、そのエリアを Tweet 拠点とするユーザ数の散布図

我々は、各ユーザの最頻 Tweet エリアが自宅エリアと一致することを正解とみなすこととして、その正解率が各ユーザの Tweet 数にどのように依存しているかを観測した。我々は、自宅エリアが判明しているユーザの夜間 Tweet 数を対数的に等間隔になる階級に分け、階級別の正解率を調査したところ、正解率は 50%から 80%であり、夜間ツイート数の多い階級であるほど、正解率が高くなる傾向があることが判明した。

また、最頻 Tweet エリアだけではなく、第 2 最頻 Tweet エリアや第 3 最頻 Tweet エリアを順次特定し、自宅エリアと一致するユーザの調査を行った。その結果をまとめたものが表 1 である。

表 1　最頻 Tweet エリアと正解ユーザ率

	第 1 最頻	第 2 最頻	第 3 最頻
正解率	70. 0%	82. 1%	84. 8%

表 1 より、第 1 最頻 Tweet エリアだけを考えると正解率が 70.0%であるが、それに第 2 最頻エリアを加えると正解率が 82.1%と大きくジャンプする事が分かる。第 3 最頻エリア以降を加えると正解率は確かに上昇するが、このような大きなジャンプは見られない。従って、ここでは最頻 Tweet メッシュは正解ではないが、第 2 最頻 Tweet メッシュが正解であるユーザを詳しく調べることによって、自宅エリア判定の正解率を上げる手法を検討することが可能である。このようなユーザの最頻 Twee エリアと第 2 最頻エリア、そして自宅エリアの位置関係を調べたところ、2 つの問題点が明らかになった。

まず 1 つ目は最頻 Twee エリアと第 2 最頻エリアが隣接していることにより、最頻 Tweet エリアではなく第 2 最頻エリアが正解となるユーザの存在が確認された。このようなユーザは、自宅エリア近くで Tweet を行っていたが、エリアの境界の切り方の問題により、最頻 Tweet エリアと自宅エリアがずれてしまったと考えられる。このようなケースに当てはまる問題が修正されれば、正解率を 70%から 73.4%に上げることができることが明らかになった。

もう 1 つの原因として、主要な駅や公共施設、あるいは商業施設などを含むエリアが最頻 Tweet エリアと判定されているため、自宅エリアと一致しないケースも確認された。図 2 は、あるユーザの夜間ツイート地点（小さな点）を地図上にプロットしたものである。実線はエリアの境界を示しており、図 2 の中央のエリアは、このユーザの最頻 Tweet エリアである。図中の円は、JR 町田駅と小田急町田駅を中心とする半径 15 秒角の円である。このユーザは、夜間に自宅で Tweet するよりも駅近くの繁華街等で多くの Tweet を発していると考えられる。このようなケースに当てはまるユーザの存在が確認されたことにより、自宅以外で夜間に Tweet を発する可能性のあるエリアを分析対象から除外する事でこの問題が修正されれば、成果率を上げることができることが分かった。

ここで検討すべきは、分析対象から除外するエリアをどのように決定するかである。そのために、主要な駅や公共施設周辺の Tweet を取り除くことで正解率がどのように上がるのかを、除外するエリアを円としてその半径を変えて、このケースに当てはまる例を全て同時に対象として分析を行い、半径 600m～800m ほどで、正解率が 78%程度に上昇することが確認された。

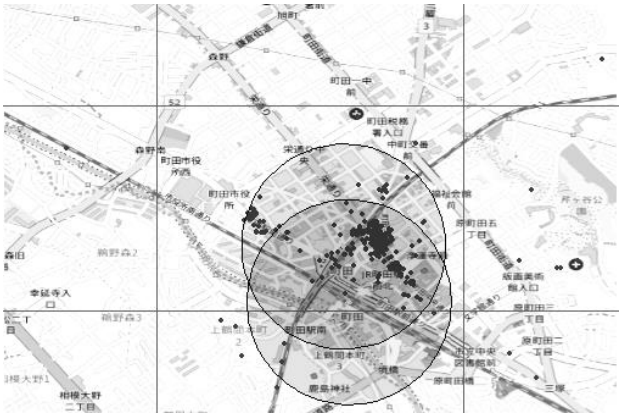

図 3 ある Twitter ユーザの夜間 Tweet 地点

3. まとめと今後の課題

本稿では、SNS 上のビッグデータを用いて、高い頻度で、そして社会への負荷が少ない形で社会統計調査を実行する手法を開発するために、Twitter の位置情報データと Tweet 内容を組み合わせ、ユーザの自宅エリアを特定する精度を向上させる手法について検討した。

本研究により、従来のように単純に Tweet の位置情報のみより自宅エリアを特定する分析手法を大きく改善することが可能であることが明らかにされた。今後、Twitter などのような SNS データを分析する際には、本稿で提案した手法を取り入れることで、分析の精度が大きく上がると期待される。これは、現在、深刻な問題となっている、分断社会の現状把握、特に都市の移民コミュニティのネットワーク解析では力を発揮すると考えられる。これについては現在、研究を進めているところであり、近い将来、報告する予定である。

謝辞

本研究は JSPS 科研費 17K01277、19K22852、16H02872、国立情報学研究所、大林財団、大川情報通信基金の助成を受けています。

参考文献

[1] A. Ishikawa, S. Fujimoto, and T. Mizuno, “Comparison between Spatial Distributions of Tweet Base and Population in Japan,” 2017 IEEE International Conference on Big Data (2017) 3052 - 3057

Analysis for Information Transmission Base Area Using SNS Data with Location Information

†FUJIMOTO Shouji, ISHIKAWA Atsushi, Kanazawa Gakuin University

‡MIZUNO Takayuki, National Institute of Informatics