早稲田大学審査学位論文
博士（人間科学）
概要書

位置情報付きツイートを利用した
気象・災害状況の可視化におけるノイズリダクション

Noise reduction for visualization of meteorological phenomena
and disaster damages using geo-tagged tweets

２０１９年７月

早稲田大学大学院 人間科学研究科

服部 充典
HATTORI, Mitsunori

研究指導担当教員： 西村 昭治 教授

近年、台風の大型化や集中豪雨、ゲリラ豪雨、そして、地震などの事象により、日本各地において頻繁に被害が生じている。総務省消防庁（2010）によると、集中豪雨などによる人的被害を軽減させるために、迅速な状況の把握と、早めの避難行動の重要性が指摘されている。そのような中、ゲリラ豪雨、突風･竜巻、雹などの突発的な事象や、それらの事象および地震などによる被害状況（河川の氾濫、交通機関の乱れ、停電、建物の倒壊など）の可視化については、速報性･網羅性の観点から十分とは言えない状況にあった。

しかし、昨今の SNS の利用、スマートフォンの普及により、その場に居合わせた人々が、その場におけるさまざまな「状況」を発信しやすい環境が整備されてきた。その中で、Twitter には、スマートフォンに付いている GPS 機能を利用し、位置情報を付与したツイートが可能であり、これにより、ツイートの投稿場所に関連した分析が可能となった。そのため、位置情報付きツイートによる「状況」可視化のアプローチは、速報性･網羅性の課題を解決する 1 つのソリューションとなり得ることが考えられる。

一般的に、位置情報を利用して「状況」の可視化を行うには、ツイートからその「状況」を示すワードを指定して抽出し、そこに付加されている位置情報を用いた可視化が考えられる。しかし、単純なワード指定による抽出方法では、精度に問題が生じる場合がある。このような抽出方法では、その「状況」が生じていた場所･時間から離れた投稿、たとえば、振り返りや感想、予報などのツイートも抽出してしまうからである。

そのため、特定状況の可視化におけるノイズリダクションでは、従来、事前にノイズとなるキーワードを定義してノイズを除去する方法や、bot などの非個人ユーザーのアカウントを特定･除去するアプローチが取られてきた。しかし、その「状況」が、今現在もその場所で続いているのか考慮するための「時間」と「場所」との両側面からのアプローチは考慮されておらず、「明日は雨かな？」、「今朝の雨は酷かった」、「埼玉では豪雨らしいね」などのツイート除去には対応できず、これによる抽出精度の課題が存在していた。また、従来手法では、ツイート内のテキストを処理する自然言語処理のみでのアプローチであり、そこには、情報の信ぴょう性の課題が存在していた。さらに、従来手法では、ノイズフィルタの機械的生成は行われておらず、そのため、少なからず時間と労力（コスト）がかかり、それにより、主観的なアプローチが入り込む可能性も考えられ、その結果、抽出精度への影響が懸念された。

一方、特定地域の状況可視化におけるノイズリダクションでは、多様な状況を捉えるために、ノイズ除去を行いながらも、より多くの正解データを抽出する、つまり、再現率を重視したアプローチが考えられる。そして、再現率を重視する上では、フィルタ生成において、事前に、より多くのノイズツイートを集め、そこから、ノイズツイートだけに該当する条件設定が重要となる。しかし、従来手法では、より多くのノイズツイートの識別･収集とノイ

ズツイートだけに該当する条件設定においては、事前の人手によるフィルタ生成が行われており、少なからず時間と労力（コスト）の課題が存在していた。

そこで、本研究では、特定状況の可視化、および、特定地域状況を可視化するための２つのノイズリダクション手法について、抽出精度、情報の信ぴょう性、人手によるフィルタ生成のための時間と労力（コスト）の課題を解決するために、従来手法と異なるアプローチに取り組んだ。特定状況の可視化においては、降雨状況を用いて、抽出精度、情報の信ぴょう性の課題解決を目的とした。また、特定地域の状況可視化においては、地震や集中豪雨により被災した地域を対象に、人手によるフィルタ生成のための時間と労力（コスト）の課題解決を目的とした。実験の結果、特定状況可視化のためのノイズリダクション手法においては、次の３点が得られた。１点目は、都市部などのツイート数が多く見込まれる地域では、ノイズツイートが多く混入する可能性があること。２点目は、従来の自然言語処理だけの手法と比較して、適合率での有意差が認められ、また、抽出数において一定量のツイートが得られること。さらに、降雨タイプ別検証においても、従来の自然言語処理だけの手法と比較して、より高い抽出精度（適合率）が得られ、その抽出数もある程度見込めること。３点目は、本手法は、近距離法を含んでおり、また、評価システムを用いた客観的検証による高い抽出精度からも、従来の手法と比べて、より情報の信ぴょう性の課題に対応可能であること。これにより、特定状況可視化のためのノイズリダクションでは、本手法を用いると、従来手法と比べて、情報の信ぴょう性の解決と抽出精度が向上できるといった仮説に対して、都市部などのツイート数が多く見込まれる地域においては、その有効性が明らかになった。

また、特定地域の状況可視化のためのノイズリダクション手法においては、実験の結果、次の３点が得られた。１点目は、抽出単位が県レベルの地震においては、抽出範囲も広いためノイズ混入の割合も高く、一定程度のフィルタ適用後の適合率向上と F 尺度を維持しながら、形態素 4-gram において、再現率が高い結果(頻度閾値５:0.95～1.00、頻度閾値３:0.91～0.96）となること。さらに、頻度閾値に着目すると、頻度数５において、再現率がより高い結果となること。２点目は、抽出単位が市レベルにおいては、ツイート数が少ない可能性があり、ノイズ混入の割合も、県レベルでの抽出と比べて高くはないが、処理するツイート数が 110 件程度以上であれば、4-gram、頻度閾値３もしくは５の条件で再現率が 0.91 という高い結果となること。３点目は、処理するツイート数が極端に少ない状況では、フィルタ効果が弱い結果となること。これにより、特定地域状況可視化のためのノイズリダクションでは、本手法を用いると、従来の人手によるフィルタ生成の手法と比べてフィルタ生成のための時間と労力（コスト）の課題を解決し、また、高い再現率を維持した抽出ができるという仮説に対して、処理するツイート数を、少なくとも 110 件程度以上とした場合に、形態素名詞 4-gram、頻度閾値３もしくは５の条件において、その有効性が示唆された。