博士論文要旨

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

服部　充典（Mitsunori Hattori）　　指導：西村　昭治

1．はじめに

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・竜巻などの突発的な事象や、それらの事象・地震などによる被害状況（河川の氾濫、停電など）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。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、特定状況可視化の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」、「場所」の観点が不足しており、その結果、「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、そ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、特定地域状況可視化の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。従来手法では、それらについて、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。

2．方法

準備作業として、ツイートデータを取得・蓄積するサーバ環境の構築、各種ツールの整備を実施した。サーバはMac mini late 2012 (Intel Core i5、メモリ4GB、OS X 10.9.4)、文字コードはUTF-8、各種ツール・フィルタの構築にはPython: 2.7.6 とシェルスクリプト、OS標準装備コマンド（grep、awkなど）を使用した。位置情報付きツイートデータの取得にはStreaming API、ツイートデータの保存にはMongoDB: 2.4.10、逆ジオコーダや評価システムのDBにはSQLite: 3.7.13 を使用した。形態素解析には、MeCab: 0.996、RMeCab: 1.0 、mecab-ipadic-neologd 2016-05-02 研究バージョン（辞書）を使用した。

3．開発したノイズリダクション手法

本研究では、特定状況を可視化するためのノイズリダクション手法（以降、特定状況用と表記）、特定地域の状況を可視化するためのノイズリダクション手法（以降、特定地

域状況用と表記）の２つの手法を開発した。

特定状況用では、抽出精度・情報の信ぴょう性の課題に対応するために、従来手法の自然言語処理（単ワード・複数ワード指定による抽出・除去）に、新たに、位置情報と時間軸をベースに考慮した処理、および、機械的処理を加え、次の２つを構築した。１つは、自然言語処理に、ノイズフィルタの基となるデータの機械的収集、ノイズフィルタの機械的生成処理、時間軸（「今朝」、「明日」などのワードの特定）、特定状況に関するストップワードを加えた手法で、本研究では、NLP法と定義した。もう１つは、これまでの自然言語処理とは別に、位置情報、時間軸、複数人の投稿（複数人が同じ話題で投稿）の条件にてフィルタする手法で、本研究では、近距離法と定義した。そして、この２つを組み合わせたのが、特定状況用の手法である。

特定地域状況用では、再現率を重視する上で、事前のコストをかけずに、実行時に機械的に生成されるフィルタを適用するアプローチを採用した。そして、この実現において、名詞の形態素n-gram頻度を適用した。これらに該当するツイートは、ノイズの可能性が高くフィルタの機械的生成の実現が容易であり、さらにn-gramのｎ数で正解ツイートの抽出（再現率）を制御できる可能性が考えられるからである。本手法は、２段階処理となる。第１段階では、日単位などある期間に特定地域から投稿されたツイートの抽出を行う。第２段階では、抽出した可視化対象ツイートから名詞の形態素n-gram解析を行い、次に解析後のn-gramデータから名詞の組合せ頻度が高いものだけフィルタとして生成、つまり、ある期間に名詞の組合せ頻度が高いものをノイズと見なした処理である。

４．評価実験

特定状況用では、偶然のばらつきによる影響を取り除くために、ランダムにフィルタ生成用と検証用に分け、同一データによる手法ごとの実験を計10回実施した。また、客観的評価を行うために、評価システムの構築と、東京都(49,422件）の観測所がある地域の降雨データを抽出した。精度評価は適合率を使用し、フィルタ無し、従来手法、近距離法、本手法（組み合わせ）による多重（４群）比較検定を実施した。その際に、対応あり、ノンパラメトリックを条件とし、その結果、ライアン法による補正を伴うウィルコクソンの符号付順位検定を適用した。

特定地域状況用では、地震や集中豪雨により被災した熊本県（広域:県レベル）、茨城県の常総市と猿島郡境町（狭域:市・町レベル）から投稿されたツイートデータを用いて２つの実験を実施した。各実験では、被災直後から１日単位で計５日分実施した。さらに、実験ごとに3-gram、4-gramの検証を行った。従来手法との比較には、本手法と同程度の精度を得るために、最もツイート量の多い熊本県３日目のデータ（3,088件）を用いてフィルタを生成し、熊本県５日目のデータ（1,522件）に適用した。評価には、適合率、再現率、F尺度、適合率向上を使用した。

５．結果と考察

特定状況用では、次の３点が得られた。（１）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）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、降雨タイプ別検証でも、従来の自然言語処理だけの手法と比較して、より高い適合率が得られ、その抽出数もある程度見込めること。（３）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。

特定地域状況用では、次の３点が得られた。（１）抽出単位が県レベルの地震では、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５において、再現率がより高い結果となること。（２）抽出単位が市レベルでは、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 0.91という高い結果となること。（３）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。

６．結論

特定状況の可視化では、本手法（特定状況用）は、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するという仮説に対して、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、その有効性が明らかになった。特定地域状況の可視化では、本手法（特定地域状況用）は、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可能であ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。

参考文献

総務省消防庁 平成22版消防白書 局地的大雨や集中豪雨に備えて
<http://www.fdma.go.jp/html/hakusho/h22/h22/html/1-5c-2_kakomi08.html> (2018-09-30)