マイクロブログユーザからの現地被災者抽出の技術的支援

水野淳太[†]　岡崎直観[†‡]　乾健太郎[†]
東北大学情報科学研究科[†]　科学技術推進機構さきがけ[‡]
{junta-m, okazaki, inui}@ecei.tohoku.ac.jp

1 はじめに

東日本大震災における情報伝達手段として，ツイッターをはじめとするマイクロブログは大きな役割を果たした．被災地で発生した問題や要望などは，今後の災害への対策に有用であると考えられている [1]．そこで本研究では，震災当時のツイッターから，被災したユーザを抽出し，そのツイートの収集に取り組む．情報伝達手段や安全上の制約があったにも関わらず，震災直後に被災地から情報を発信していたユーザは少なくない．震災による被災は，火災や津波など様々であるが，本研究では津波による被災者を抽出の対象とする．すなわち，震災当時に東北 3 県の沿岸部に滞在していたユーザを抽出することが本研究の目的である．以下では，抽出対象となるユーザを「津波被災者」と呼ぶ．

ツイッターには，緯度経度情報をツイートに付与する機能が存在するが，この機能の利用者は少数であるため，本研究では取り扱わない．ツイート本文をもとにして発信場所推定手法についても研究が進められているが [2, 3]，いずれの推定精度も限定的であり，東北 3 県の沿岸部という狭い範囲の推定において，有効であるとは考えにくい．そこで本研究では，ツイート本文に含まれる住所情報，画像データを手がかりとすることで，津波被災者を効率よく見つけられることを示す．

2 津波被災者の抽出

本研究で抽出の対象とするツイートデータは，東日本大震災ビッグデータワークショップ[1]において Twitter Japan から提供された，2011 年 3 月 11 日の午前 9 時から 3 月 18 日の午前 9 時までの全ツイート（179,286,297 件）である．

津波で特に大きな被害を受けたのは岩手・宮城・福島の 3 県である．人口の比率を考えると，日本全体のツイッターユーザに対して，この 3 県のユーザが占める割合は小さい．さらに，この 3 県では停電やネットワーク障害が長期間にわたって発生しており，津波被災者からの情報発信が滞っていた可能性もある．このような理由から，ツイートデータの中から単語の頻度や共起頻度を測定し，統計的に顕著な部分に着目したとしても，津波被災者のツイートを発見するのは難しいと想像される．

津波被災者のツイートを発見することの難しさを示す一例として，ツイートデータ全体に対して，「津波」を本文に含むツイートを検索し[2]，検索された 1,545,910 ツイートの中でリツイート数の多いツイート 100 件をまとめたものを表 1 に示す．このツイート群の中では，注意喚起のツイート，情報提供のツイートが 7 割以上を占めており，津波に関する注意や情報を積極的に拡散していることが分かる．しかしながら，津波の被害を自分の体験として報告しているツイートは，この 100 件の中には見つからなかった．この結果から，津波被災者のツイートは，よくリツイートされるとは限らないことが分かる．そこで，ツイート本文に含まれる住所情報，画像データを利用した抽出を試みる．

表 1: RT 数 top100 のツイートの分類結果

ツイートタイプ	ツイート数
注意喚起	39
情報提供	32
賞賛	11
意見	7
救援要請	5
非難	4
ジョーク	2
合計	100

2.1 住所情報に基づく抽出

ツイッターユーザが津波被災者であるかは，そのユーザが被災地域に滞在していたかによって判断することができる．ユーザのプロフィール情報を閲覧すると，そのユーザがどの地域に住んでいるかを判断することができるが，本ワークショップにはプロフィール情報は含まれていない．そこで，ユーザのツイート内容を基にプロファイリングを行い，ユーザの滞在地を推定することが考えられる．

東北 3 県の沿岸部についてよく言及しているユーザは，その地域に居住あるいは滞在している可能性が高いという仮説に基づき，以下の手順によって津波被災者の抽出を行った．

1. 宮城県の主要な沿岸部（南三陸町など）を，町名の粒度で人手で 15 箇所を選択する．
2. 各ユーザのツイート集合に対して，15 箇所の地名の本文中での出現頻度を計る．
3. 15 箇所の地名のうち，20 回以上言及していた地名があるユーザは，その地域に滞在していたと判断する．20 回以上言及していた地名が複数ある場合は，より多く言及していた地域に滞在していたと判断する．
4. 抽出されたユーザのツイート本文を読み，沿岸部に滞在していたかを人手で判断する．

3 までで，723 人のユーザを抽出することができた．それらのユーザに対して，4 で人手で判断したところ，15 人が滞在していたと判断できた．本手法は，4 でかかるコストが問題となる．723 人から 15 人を抽出するのに約 12 時間かかっており，多大な労力を要する．そこで，次節ではツイートに含まれる画像データに着目した抽出手法について述べる．

[1] http://sites.google.com/site/prj311/

[2] 全文検索エンジンには Apache Solr (http://lucene.apache.org/solr/) を用い，全ツイートの本文を，文字 bi-gram で索引付けした．

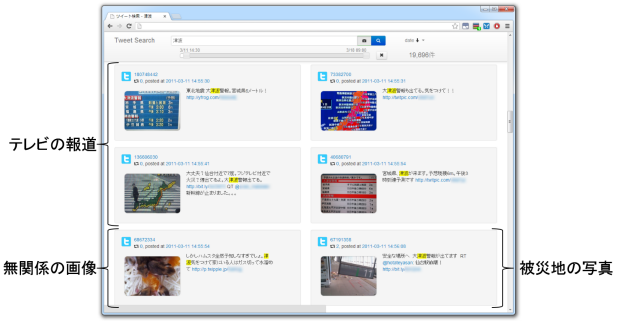

図 1: 「津波」を本文に含む，画像付きツイート

図 2: 横須賀で撮影された写真付きツイート

図 3: 仙台港で撮影された写真付きツイート

2.2 画像データに基づく抽出

津波被災者のツイートを効率よく選び出す方法として，我々はツイート本文中に含まれる画像データへのリンクに着目した．震災当時のツイッター上では，被災状況や安否不明者のリストなどが，画像データとして拡散していた．そこで，津波の状況が添付されているツイートに着目することで，津波被災者の選別が出来るのではないかと考えた．

東日本大震災ビッグデータワークショップのツイートデータの中で,「津波」を本文に含み，かつ画像付きのツイートは 19,696 件であった．その一部を図 1 に示す．なお，ツイートに画像が添付されているかどうかは，本文に含まれる URL が代表的な画像投稿サービス（Twitpic や yfrog など）のものであるかによって判別した．図 1 を見ると，テレビでの報道の一部を撮影して投稿されたツイートが目立つが，津波の被害状況を撮影した写真も少なからず存在する．これらの画像は，以下のように大別できる．

被災地の写真 津波の到達前・到達時・到達後の様子，津波による被害などを撮影したもの

テレビの報道 テレビの報道番組の画面を撮影したもの

無関係の画像 被災地の応援を目的としたイラストや，津波とは無関係の写真など

このうち，テレビの報道は画面の映り込みや回転，L 字型画面，テロップなどを手がかりに，容易に判別可能である．無関係の画像は，津波以外の被害状況の写真やイラストなどが該当する．これらも人間には容易に判別

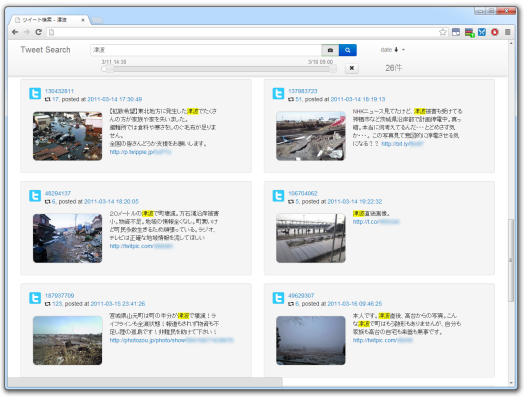

図 4: 人手で抽出した結果

可能である．このように，ツイートに添付されている画像を人間が目視確認することで，津波被災者が津波の状況を撮影した写真かどうか，迅速に判定できる．

ただし，今回の東日本大震災では広範囲の沿岸部に津波が襲来したため，津波被害を撮影した写真かどうかを判別するだけでは，東北 3 県のツイートに限定することはできない．例えば，図 2 のツイートは,「横須賀」で津波を撮影したものである．一方で，図 3 のツイートは,「仙台港」であることが明記されている．そこで，ツイートに添付されている画像に加えて，本文に含まれる地名を手がかりとし，人手で 19,696 件のツイートをチェックした．約 3 時間の作業時間で，全てのツイートに対するチェックを行うことができ，津波被災者が津波の被害状況を撮影したと思われる 28 件のツイート（28 ユーザ）を抽出できた．その一部を図 4 に示した．

本手法は，画像データを投稿したユーザのみに限定した抽出しか行えないが，迅速に判断できるというメリットがある．住所情報に基づく手法で抽出された 15 ユーザと，画像データに基づく手法で抽出された 28 ユーザに重複はなかったことから，その他の情報に着目することによって，新たな津波被災者を抽出できる可能性が示唆される．

3 おわりに

本稿では，東日本大震災当時のツイートデータに対して，本文中の住所情報や画像データを利用することで，津波被災者を抽出するための技術的支援手法について述べた．これらの支援技術により，合計で 43 名の津波被災者を抽出することができた．今後は，抽出されたユーザのツイートを用いて，新たな津波被災者をマイニングしていくことが考えられる．

謝辞

本研究は，文部科学省科研費（23240018，23700159），および JST 戦略的創造研究推進事業さきがけの一環として行われた．

参考文献

[1] 今村文彦, 佐藤翔輔, 柴山明寛. みちのく震録伝：産学官民の力を結集して東日本大震災のアーカイブに挑む. 情報管理, Vol. 55, No. 4, pp. 241–252, 2012.

[2] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *Proc. of CIKM 2010*, pp. 759–768, 2010.

[3] Yohei Ikawa, Miki Enoki, and Michiaki Tatsubori. Location inference using microblog messages. In *Proc. of WWW 2012*, pp. 687–690, 2012.