E-028

Twitter を利用したユーザの状況推定のための学習データの自動収集
Automatic collection of training data for estimating users' situation using Twitter

竹下　知宏† 新妻　弘崇† 太田　学†
Chihiro Takeshita Hirotaka Niitsuma Manabu Ohta

1. はじめに

近年，多くの人々が SNS で情報アクセスや情報発信を行うようになった．そうした SNS の一つである Twitter にはユーザの状況を伝えるツイートが数多くある．それらのツイートからユーザが何をしているかを自動で推定できれば，その状況に合わせた行動を推薦することができる．著者らはこれまで手作業で Twitter ユーザを一般的なユーザかどうか判定し，それらのユーザのツイートからユーザの状況推定を行ってきた．しかし，手作業によるユーザの判定はツイートが膨大なため現実的でない．そこで本稿では Twitter ユーザの状況推定のための学習データを自動で収集するために，まずツイートに基づいてそのユーザが一般的であるかどうかを自動で判定する手法を提案する．

2. 研究背景

著者らは Twitter ユーザが「何を」しているかに着目し，ツイートの分析によってユーザの状況を自動で推定する手法を提案した[1]．この手法ではまずツイートを取得し，そのツイートからユーザが移動しているか，移動していないかの状況を推定した．移動していると推定された場合は，さらに「鉄道」，「飛行機」，「車」，「歩き」，「その他」の 5 種類の移動手段で分類した．移動していないと推定された場合は，総務省統計局が定めた行動の種類の 3 区分に基づいて，睡眠，食事など生理的に必要な活動である「1 次活動」，仕事，買い物といった生活する上で義務的な性格が強い活動である「2 次活動」，これら以外の自由時間における活動である「3 次活動」の 3 種類のいずれかに分類した．この分類では，ツイート群から Paragraph Vector の実装の一つである doc2vec を用いてそれぞれのツイートの分散表現を獲得し，得られたベクトルに状況を付与して分類器で未知のツイートの状況を分類した．

しかし，Twitter には，速報等を伝えるニュースアカウントやおすすめの飲食店を紹介するグルメアカウント等，ユーザ自身の活動を表さないツイートも多く存在する．先に述べたユーザの状況推定では，事前にユーザ自身の活動を示すツイートを多く投稿している一般的なユーザであるかどうかを手作業で判定していた．本稿ではこの判定を自動化することを目的とする．

3. 提案手法

ここでは，Twitter から取得したツイートの文章からそのユーザが一般的であるかどうかを判定する手法について説明する．

3.1 ユーザ判定の概要

まず，ツイート本文をベクトル化し，ツイートの特徴ベクトルとする．そして分類器によりツイートを「活動ツイート」，「非活動ツイート」のいずれかに分類する．なお

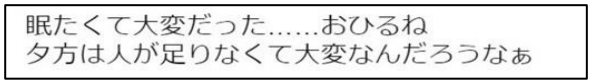

図 1　活動ツイートの一例

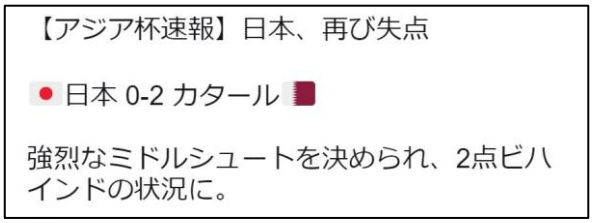

図 2　非活動ツイートの一例

本稿では**活動**を「食事や買い物といった身体的行動あるいはある物事に対する見解の主張や心境の述懐」と定義する．よって「活動ツイート」とは図 1 のようなユーザの活動を含むツイートであり，「非活動ツイート」とは図 2 のようなユーザの活動を含まないツイートである．そして，「活動ツイート」が「非活動ツイート」より多いユーザを「一般ユーザ」，そうでないユーザを「非一般ユーザ」と定義する．

3.2 ノイズツイートの除外

本稿では，「活動ツイート」であるかどうか判定する前にリプライや一部のリツイートをノイズとして除外する．リプライは，ある特定のユーザのツイートに対する返答であり，ツイートの内容が独立していないため本稿ではノイズとする．またリツイートに関しては，他人のツイートを加筆せずそのまま投稿している場合はそのユーザのツイートでないためノイズとするが，ユーザのコメントが付け加えられている場合には，除外しない．

3.3 ツイートの特徴ベクトル

3.2 のノイズを除いて残ったツイートの特徴ベクトルを求める．はじめにツイート文からハイパーリンクの URL や括弧，句読点を取り除く．次にツイートの文章を形態素解析エンジンである MeCab によって形態素に分解する．その後 doc2vec[2]を用いてツイート文をベクトルに変換する．

3.4 ツイートの分類器

ベクトル化したツイートを「活動ツイート」，「非活動ツイート」のいずれかに分類する．分類には Support Vector Machine (SVM)[3]を用いる．

4. 評価実験

提案手法によるツイート分類及びユーザ分類を行い，その精度を評価する．

† 岡山大学　Okayama University

4.1 実験データ

実験では，TwitterAPI を用いて無作為に選んだ Twitter ユーザの 2019 年 6 月 7 日時点での最新 400 件のツイートを取得し，各ツイートに「活動ツイート」，「非活動ツイート」のラベルを付与した．その中から「活動ツイート」が多い「一般ユーザ」30 ユーザと，「非活動ツイート」が多い「非一般ユーザ」30 ユーザの計 60 ユーザを選出した．また，取得したツイートからノイズツイートを除いた結果，「一般ユーザ」のツイートは 12,000 件中 6,902 件が残り，その内「活動ツイート」は 6,843 件，「非活動ツイート」は 59 件となった．また，「非一般ユーザ」のツイートは 12,000 件中 8,801 件が残り，その内「活動ツイート」は 298 件，「非活動ツイート」は 8,503 件となった．この計 15,703 件のツイートを実験では分類する．

4.2 評価方法

まず実験データのツイートを，SVM によって「活動ツイート」，「非活動ツイート」のいずれかに分類する．評価には 5 分割交差検証を用いる．5 回の検証ともにテストデータは「一般ユーザ」，「非一般ユーザ」がそれぞれ 6 ユーザずつの計 12 ユーザのツイート，学習データは「一般ユーザ」，「非一般ユーザ」がそれぞれ 24 ユーザずつの計 48 ユーザのツイートで構成されている．実験データ作成時にツイートに付与したラベルを正解とし，ツイート分類の精度は式(1)で算出する．

$$\text{ツイート分類の精度} = \frac{\text{正しく分類されたツイートの数}}{\text{テストデータのツイートの数}} \quad (1)$$

次にツイート分類の結果を基にユーザ分類を行う．「活動ツイート」に分類されたツイートの方が多いユーザを「一般ユーザ」，「非活動ツイート」に分類されたツイートの方が多いユーザを「非一般ユーザ」とする．ユーザ分類の精度は，実験データ作成時に決めたユーザが「一般ユーザ」であるか「非一般ユーザ」であるかの判断を正解とし，式(2)で算出する．

$$\text{ユーザ分類の精度} = \frac{\text{正しく分類されたユーザの数}}{\text{テストデータの 12 ユーザ}} \quad (2)$$

4.3 実験結果

ツイート分類とユーザ分類を行ったところ，表 1 のような結果となった．表 1 はツイート分類，ユーザ分類の 5 回の検証における精度の平均を表している．

表 1 に示すようにツイート分類の精度は 0.854 だった．表 2 に 5 分割交差検証のある回のツイート分類の結果を示す．表 2 ではツイートの多くは正しく分類されているが，「活動ツイート」，「非活動ツイート」ともに約 1 割は正しく分類されていない．

一方，表 1 に示すようにツイート分類の結果に基づいてユーザを分類した結果の精度は 0.967 だった．ユーザ分類では，5 回の検証のうち 3 回ですべてのユーザが正しく分類されたのに対し，2 回の検証では，「一般ユーザ」と誤って分類された「非一般ユーザ」が 1 ユーザいた．しかし，ユーザ分類の精度は比較的高いといえる．

表 1　ツイート分類とユーザ分類の精度

ツイート分類	ユーザ分類
0.854	0.967

表 2　ツイート分類の結果

		分類結果		計
		活動ツイート	非活動ツイート	
正解	活動ツイート	1145	188	1333
	非活動ツイート	131	1633	1764
計		1276	1821	3097

4.4 考察

実験データとして集めたツイートのうち，「一般ユーザ」のツイート 5,098 件，「非一般ユーザ」のツイート 3,199 件がノイズツイートだった．このことから，「一般ユーザ」は「非一般ユーザ」よりもリプライ，リツイートを投稿する頻度が高い傾向にあることがわかる．

本研究では Twitter ユーザの状況推定のための学習データを収集するために，「一般ユーザ」であるかどうかを判定したが，「一般ユーザ」の中にも活動を含まないツイートが存在する．このようなツイートではユーザが何をしているかが述べられていないため状況が推定できない．ゆえに状況推定の学習データには「一般ユーザ」の「活動ツイート」を利用する．その分類のために，本稿の提案は利用できる．

「非一般ユーザ」のツイートに関しては状況推定の学習データとして利用しないが，それらの中には「一般ユーザ」にとって有益な情報が含まれていることがある．ユーザの状況推定をした後に，それらの情報を用いてユーザに行動の推薦をすることができれば，「非一般ユーザ」のツイートも有効に活用できる．これを実現するためには，「非一般ユーザ」がどのような種類の情報を投稿し，それがユーザの行動推薦にどのように利用できるかさらに検討する必要がある．

5. おわりに

Twitter からユーザの状況推定のための学習データを自動収集するためにユーザが一般的かどうかを自動で判定する手法を提案した．提案手法では，doc2vec 及び SVM を用いてツイートをユーザの活動を含んでいるかどうかで分類し，活動を含むツイートが含んでいないツイートよりも多いとき一般的なユーザであると判定した．その結果，ユーザ判定の精度は 0.967 となった．今後はこの手法を利用し，学習データを増やすことでユーザの状況推定の精度向上を目指す．

参考文献

[1] 竹下知宏，新妻弘崇，太田学，“Twitter を利用したユーザの移動等の状況推定”，第 11 回データ工学と情報マネジメントに関するフォーラム，H8-1 (2019).

[2] Le，Q. and Mikolov，T.，“Distributed Representations of Sentences and Documents”，Proceedings of the 31st International Conference on Machine Learning，pp. 1188-1196 (2014).

[3] Cortes，C. and Vapnik，V.，“Support-Vector Networks”，Machine Learning，vol. 20，no. 3，pp. 273-297 (1995).