4ZD-05

沖縄県内のツイートデータを用いた移動パターンの分析手法の研究

宮城　圭介†　　中里　収‡

名桜大学大学院国際文化研究科†　　名桜大学国際学群国際学類‡

1. はじめに

Twitter における日本国内の月間アクティブユーザー数は3500万人（2015年6月）[1]である．2010 年のデータによれば，日本は世界のツイートのうち約15%を占める第2位の国である．そのうち位置情報付きツイートデータ（以下，ツイートデータ）は，日本国内の全ツイートのうち，約7%であることが明らかとなっている[2]．近年はSNSやスマートフォンの普及をきっかけに，場所を問わず情報の発信が気軽に出来るようになった．加えて，Twitterをはじめとする SNS は，位置情報を付加できる．この特性を活かし，リアルワールドのイベントや状況を抽出する研究が行われている[3]．ツイートデータは，投稿日時，ツイート位置，ユーザ名，ツイート本文から構成されており，同一のユーザのツイートを追跡することで，1 日の移動パターンを特定することが可能である．

2. 先行研究

SNS 内でのリアルタイムな位置情報を取得することで，リアルワールドと結びつけた観察が可能となる．那須野・松尾は，Twitter における選挙候補者の情報拡散に着目して研究を行なった．当選者予測を行い，候補者がTwitterを利用する際の，当選しやすい状態について明らかにしている[4]．さらに Twitter における情報拡散性に着目した荒牧らは，ツイートをもとに言語処理技術を用いた機械学習で，これまでの検索履歴よりも精度の高いインフルエンザの予測に成功している[5]．以上のように，Twitter を利用し，バーチャルワールドの情報から，リアルワールドの状況を抽出する試みは複数の角度からなされている．

3. 研究目的

本研究は，パターンモデルの構築を目的とする．毎日のツイートデータ構造をもとに，ツイート数，ユーザの移動距離，ツイート速度から移動パターンを明らかにする．ユーザの移動によって現れるデータ構造から，要因をパターンモデルに組み込む．データ構造をもとに作成したモデルを活用することによって，リアルワールドの人の移動や混雑状況などを予め推測することが可能となる．

4. 研究方法

沖縄県内で収集されたツイートデータをもとにし，本研究のベースデータとして用いた．本研究で用いたツイートデータは収集を行うプログラムの不具合などのために若干の漏れが生じたものの，2016年6月1日から2016年12月8日までの約半年間で収集したツイートデータは約11 万件である．ツイートデータには，収集された番号（通し番号），投稿日時，緯度・経度（位置情報），ユーザ ID（ユーザ名），ツイート本文（写真リンクのURL添付等含む）が収集されている．つまり1ツイートあたり6つの要素が収集されている．本研究では，ツイート本文を用いた研究は行わないため，上記のデータから投稿日時，緯度・経度，ユーザ ID を利用し，組み合わせることでパターン抽出に用いた．本研究を進める過程で，天気や地震情報などの自動ツイートが収集されていることが判明した．本研究ではそれらの移動を伴わないツイートはフィルタリングを行い，パターン抽出においては使用しなかった．

収集されるツイートデータの位置情報である緯度・経度は角度で表記されている分析に際して，基準点からの距離として，m 単位系に統一した．次に，同一ユーザの2ツイートから，移動距離を求める．求めた移動距離を2点間の移動時間で割り，ツイート速度を求めた．

5. 研究結果

ツイートデータの分析から，データ構造のツイート件数の推移に曜日によるパターンが現れていることが確認できた．ツイート件数の曜日平均算出方法は，全ツイートデータの同曜日のみを取り出して算出した．図1以降に提示する各要因の平均値は，項目ごとの曜日別平均値を用いて算出を行った．収集されたツイートデータのツイート件数推移の様子を以下に示す（図 1）．

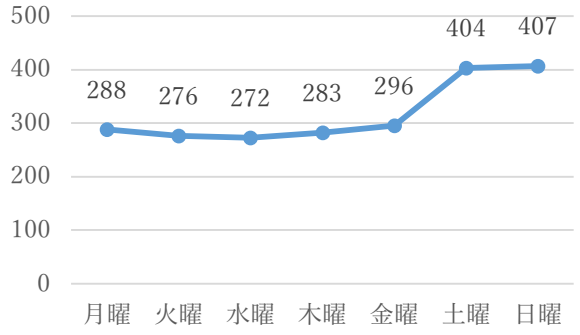

図1　ツイート件数　曜日別平均推移

次に，曜日の要因があることを前提とし，曜日平均を差し引いた場合の他要因について特定を試みた．その要因がユーザの移動距離とツイート速度である．ツイート件数と同様に各曜日の平均を移動距離，ツイート速度から取り除き，1σ範囲内・外の日付を抽出した．この結果から，他の同曜日より突出した日を特定した．本研究において突出している値と述べるのは，標準偏差の 1σより上回ったもしくは下回ったデータと定義した．凡例として，Hは1σを上回った値，Mは1σの範囲内におさまった値，Lは1σを下回った値を意味している．表1は，ツイート件数，移動距離，ツイート速度の関連結果を以下に示す（表1，2，3）．

表1　ツイート件数Lの移動距離とツイート速度日数

速度＼距離	H	M	L
H	6日(A)	1日(B)	1日(C)
M	0日(D)	2日(E)	3日(F)
L	2日(G)	3日(H)	2日(I)

表2　ツイート件数Mの移動距離とツイート速度日数

速度＼距離	H	M	L
H	2日(J)	3日(K)	0日(L)
M	5日(M)	4日(N)	2日(O)
L	2日(P)	4日(Q)	0日(R)

表3　ツイート件数Hの移動距離とツイート速度日数

速度＼距離	H	M	L
H	4日(S)	6日(T)	1日(U)
M	1日(V)	5日(W)	4日(X)
L	1日(Y)	3日(Z)	0日(AA)

6.　考察

本研究で得られた結果から，各項目に対して便宜上アルファベットを括弧内に割り当てた．最も多く現れたパターンは，AとTの6日であった．反対に全く現れなかったパターンは，D，L，R，AA であった．A はツイート数が少ない時移動距離の平均，ツイート速度の平均が高いパターンが最も現れたことを意味する．加えてTはツイート数が多い時，移動距離は平凡だがツイート速度は速いことを意味している．A，T ともに，共通していたことは，ある1ヶ月に値が集中していることであった．A の場合，6/3，7，10，11，24，7/19となっており，ほとんどが6月に集中していることが確認された．Tの場合も9/4，5，12，13，22，10/30 という内訳となっていた．M の場合も同様に，6/16，17，18，19，7/1 とほとんどが同月に集中していた．これらの考察から，移動距離とツイート速度には関連性が見出せる．さらには，1 度パターンが現れるとそのパターンが複数日同月で繰り返される可能性が推察された．今後は機械による気象データなどの自動ツイートを活用することで，より正確なパターンモデル設計を想定している．全ツイートを完全に用いることが本研究の課題である．今後は，パターンを判別する要素を追加していくことで，さらにモデルの充実をはかりたいと考えている．

7.　引用・参考文献

[1]THE HUFFINGTON POST　2016　「Twitterが国内ユーザー数を初公表『増加率は世界一』」http://www.huffingtonpost.jp/2016/02/18/twitter-japan_n_9260630.html　[2016/07/21閲覧]
[2]semiocast. 2010　Only 30% of messages on twitter are from the u.s. https://semiocast.com/downloads/Semiocast_Only_30_percent_of_messages_on_Twitter_are_from_the_U.S._20100331.pdf　[2017/01/13閲覧]
[3]榊剛史・柳原正・那和一成・松尾豊　2015「Twitter を用いた道路交通情報の抽出」　『電子情報通信学会論文誌』D Vol.　J98-D　No.6 pp.1019-1032
[4]那須野薫・松尾豊　2014　「Twitter における候補者の情報拡散に着目した国政選挙当選者予測」　『人工知能学会全国大会論文集』Vol.28　pp.1-4
[5]荒牧英治・増川佐知子・森田瑞樹　2011「Twitter Catches the Flu: 事実性判定を用いたインフルエンザ流行予測」　『情報処理学会研究報告』2011-NL-201 No.1