F-014

多項分布変化点検出法による Twitter 上のユーザ動向分析

User trends analysis on Twitter by change detection method based on multinomial distribution

藤野まり菜 † 加藤翔子 † 斉藤和巳 † 風間一洋 §
Marina Fujino Shoko Kato Kazumi Saito Kazuhiro Kazama

1. はじめに

近年 Twitter のような SNS をはじめとするソーシャルメディアが様々な分野において利用されており，その研究は盛んに行われている [1, 2, 3]．その中でも私たちは，Twitter の代表的な機能を利用したユーザ動向の分析に関心がある．

本研究では，Twitter 上におけるユーザ動向として，ツイートに高頻度で出現するリプライ先ユーザ，リツイート元ユーザ，及びハッシュタグに着目し，これらを多項分布型変化点検出法 [4] で分析することにより，その利用動向にどのような違いがあるのかを比較する．妥当な変化点数の評価には一つ抜き交差検定を用いる．また，各期間における特徴的なユーザやハッシュタグの選定には Z-スコアを採用する．本研究では，東日本大震災前後のツイートデータを利用することにより，震災時における Twitter の利用動向の変化を分析，評価する．その結果，リプライには明確な変化点が存在する一方で，リツイートとハッシュタグは時間とともに内容が変化していくため，明確な変化点が存在しないことを示す．

本稿の構成は以下の通りである．第 2 章で，本研究に用いた分析手法として変化点検出法，一つ抜き交差検証，Z-スコア，提案分析法について説明する．第 3 章では評価に用いたデータの概要を説明し，分析結果について論じる．最後に第 4 章で本研究のまとめと今後について述べる．

2. 分析手法

2.1. 変化点検出法

本研究では，リプライ先ユーザ，リツイート元ユーザ，ハッシュタグのそれぞれの集合を，一般にワード集合 $W = \{w_1, \cdots, w_m, \cdots, w_M\}$ と呼ぶ．これらワードの実際の出現データを $D = \{(x_1, t_1), \cdots, (x_n, t_n), \cdots, (x_N, t_N)\}$ とする．ここで，x_n は時刻 t_n で出現したワードを表し，$x_n \in W$ である．

一方，ワード w_m の出現データ集合を

$$D_m = \{(x_n, t_n) \mid x_n = w_m\} \tag{1}$$

とし，ワード w_m の出現確率を $P_m = \frac{|D_m|}{N}$ とする．

また，時刻 T_{k-1} から T_k の間でのワード w_m の出現データ集合を

$$D_{k,m}(T_{k-1}, T_k) = \{(x_n, t_n) \in D_m \mid T_{k-1} \leq t_n < T_k\} \tag{2}$$

とし，その出現確率を $P_{k,m} = \frac{|D_{k,m}|}{|D_m|}$ とする．

ここで，変化点の個数が K のとき，$T_0 = t_1$，$T_{K+1} = t_N + 1$ とし，変化点時刻を並べて構成したベクトルを $\mathbf{T}_k = (T_1, \cdots, T_K)$ と定義する．

このとき，変化点ベクトル $\mathbf{T}_K$ に対する対数尤度は次式で定義される．

$$L(\mathbf{T}_k) = \sum_{k=1}^{K} \sum_{m=1}^{M} |D_{k,m}| \log P_{k,m} \tag{3}$$

よって，変化点ベクトル $\mathbf{T}_K$ は式 (3) が最大になるように求める．ここで，変化点ベクトルの探索には，局所改善付貪欲法を用いる [4]．

2.2. 一つ抜き交差検証 (LOOCV)

本研究では，適切な変化点の個数 K を求めるため，次式の一つ抜き交差検証評価を用いる．

$$Loo(\mathbf{T}_K) = \sum_{k=1}^{K} \sum_{m=1}^{M} |D_{k,m}| \log \frac{|D_{k,m}| - 1}{|D_k| - 1} \tag{4}$$

すなわち，$Loo(\mathbf{T}_e)$ を最大とする K を適切な変化点の個数とする．

2.3.Z-スコア

変化点区間 k で有意に多く出現する単語 w_m を抽出するため，次式の Z-スコアを用いる．

$$Z_{k,m} = \frac{|D_{k,m}| - |D_k| P_m}{\sqrt{|D_k| P_m (1 - P_m)}} \tag{5}$$

すなわち，$Z_{k,m}$ が有意に大きければ，時刻 T_{k-1} から T_k の特徴をワード w_m として抽出する．

2.4. 提案分析法

提案分析法の処理手順は以下となる．

1 $K = 1$ から K の最大値 K_{max} まで次の処理を繰り返す．

1.1 式 (3) を最大にする $\mathbf{T}_K$ を求める．

1.2 式 (4) で評価値 $Loo(\mathbf{T}_K)$ を求める．

2 最適変化点数を $\hat{K} = \arg max_{1 \leq K \leq K_{max}} \{Loo(\mathbf{T}_K)\}$ で求める．

3 $\hat{K}$ において，式 (5) で特徴ワードを求める．

なお，本分析では，$K_{max} = 20$ に設定した．

†静岡県立大学経営情報学部
§和歌山大学

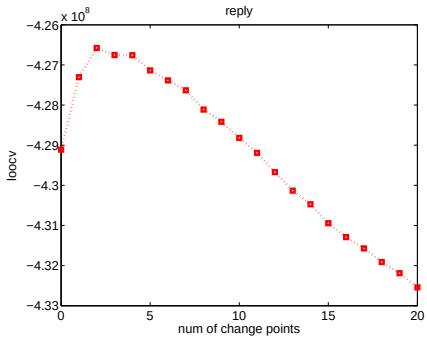

図 1: リプライ

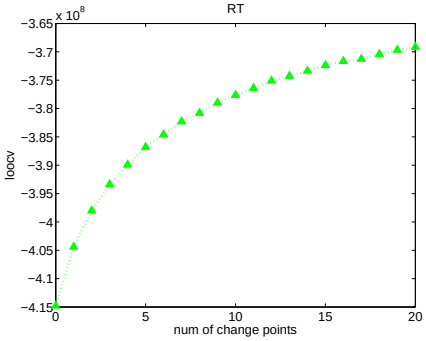

図 2: リツイート

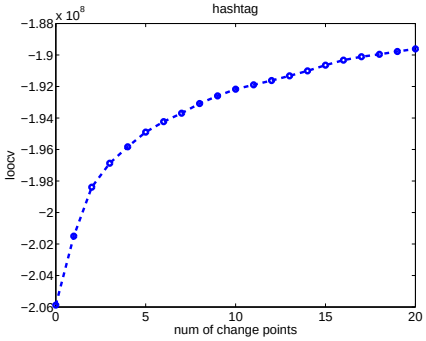

図 3: ハッシュタグ

3. 実験

3.1. データ概要

本研究では，2011 年 3 月 5 日 00:00:00 から同月 24 日 23:59:59 までの間に日本語で投稿されたツイートの中から，文頭が”@user”から始まるツイートをリプライ，”RT @user”で始まるツイートをリツイートとして収集する．また，同様にハッシュタグが使用されているツイートも収集し，ワードデータ (x_n, t_n) として利用する．このように，各ワードには時刻を付与し，時間経過による Twitter の利用動向を分析する．

3.2. 分析結果

式 (4) で求める一つ抜き交差検証での評価結果を図 1，図 2，図 3 にそれぞれ示す．図 1 では，変化点が 2 のときに値が最大となっている．このことから，リプライにおいては本分析法で適切な変化点の個数が存在することがわかる．それに対して図 2，図 3 では右上に増加し続けている．これにより，リツイートとハッシュタグにおいては，適切な変化点の個数が $K \leq 20$ では存在せず，時間とともに著しく変化していることが示唆される．

表 1: リプライトップ 3

	区分 1	区分 2	区分 3
1	shuumai	itsumonoTL	3HRYoshiyuki
2	mariko_dayo	NHK_PR	OfficialTEPCO
3	karashichan	TMR15	AC_popopopon

表 1 に，リプライネットワークにおいて，2 つの変化点により区切られた 3 区分それぞれで，式 (5) で求める Z-スコア $Z_{k,m}$ の高い上位 3 ユーザを示す．区分 1 は震災前の 2011 年 3 月 5 日 00:00:01 から同月 11 日 14:51:00，区分 2 は震災直後の 2011 年 3 月 11 日 14:51:00 から同月 16 日 19:37:39，区分 3 は震災からしばらく後の 2011 年 3 月 16 日 19:37:39 から同月 24 日 23:59:59 の期間に投稿されたリプライである．

震災前は，よく知られた bot である shuumai や karashichan，また，フォロワー数が圧倒的に多い mariko_dayo などが上位に入っている．それに対し震災直後は，itsumonoTL や NHK_PR，TMR15 などの震災関連のリプライが多く，震災からしばらく後には，CM で有名になった AC_popopopon や，震災によって原子力発電所事故が発生した OfficialTEPCO などが上位に入った．

これら結果については，変化点の時刻も各区間での特徴ワードも自然に解釈できる．よって，ワード出現時系列データの分析において，提案法の有効性が示唆されたと考える．

4. おわりに

本研究では，多項分布型変化点検出法により，リプライ，リツイート，及びハッシュタグの利用動向を分析した．その結果，リツイートとハッシュタグについては時間とともに常に内容が変化するが，リプライについては明確な変化の時期が存在することを確認できた．今後は，本分析で得られた知見を用いて，さらに大規模な Twitter データや，平常時における Twitter のユーザ動向の変化の評価および震災時との比較を行い分析を進めていきたい．

謝辞

本研究は，科研費 (No.26330345) の助成を受けた．

参考文献

[1] H.Kwak, C.Lee, H.Park, and S.Moon, “ What is Twitter, a social network or a news media?“ , In Pro-ceedings of the 19th international conference on World wide web, pp.591-600. ACM,2010.

[2] 小出明弘, 斉藤和巳, 大久保誠也, 鳥海不二夫, 風間一洋, ” Twitter の@-message で構成される成長ネットワークの分析” , 情報処理学会 第 74 回全国大会, 2012.

[3] A.Java, X.Song, T.Finin, and B.Tseng, “ Why we twitter: understanding microblog-ging usage and communities“ , In Proceedings of the 9th WebKDD and 1st SNA-KDD2007 workshop on Web mining and social network analysis, pp.56-65. ACM,2007.

[4] 山岸 祐己, 斉藤 和巳, 武藤 伸明, ” 評点時系列データの区間分割法” , 日本データベース学会論文誌, Vol.12 , No.3 , pp.1–6, Feb.2014.