The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Twitterデータに見られる特徴と人間のリツイート行動

Common Characteristics Found in Twitter Data and Human Behavior on Retweet

塩田 茂雄　　中島 圭佑
Shigeo Shioda　　Keisuke Nakajima

千葉大学 大学院融合理工学府 都市環境システムコース
Urban Environment & Systems, Guraduate School of Science and Engineering, Chiba University

In order to see the human behavior on Twitter, we analyzed a large amount of Twitter data collected by the keyword search function of Twitter API. We found some common features in Twitter data; e.g., for any search keywords, the number of retweets follows a power-law distribution. We describe some of these characteristics based on the assumption that Twitter users tend to retweet tweets that have been retweeted many times before.

1. まえがき

実社会で重大な出来事が発生すると，Twitterに様々な書き込みやそのコピー（リツイート）が大量に投稿され，やがて沈静化する様子が見られるが，これは実社会における人々の情報活動がTwitter上に表出した現象であり，それら現象のから人々がどのような行動特性を有するかを分析することができる．

現実のソーシャルメディア上の現象は様々であるが，我々はTwitter APIのキーワード検索機能により収集した大量のツイートを分析した結果，以下の共通の特徴を見出した．

1. 日常的なキーワードで検索するとオリジナルツイートが半数以上を占めるが，重大な出来事に関する非日常的なキーワードで検索するとリツイートが大半を占める．
2. （1つのツイートに着目したときの）単位時間あたりのリツイート数の変化は，急峻なピークを迎えたのち昼夜変動を繰り返しながら減衰するという定型パタンに従う．
3. 検索を行うキーワードに関わらずリツイート数は（べき分布のような）裾の長い分布に従い，最大リツイート数と平均リツイート数の差が非常に大きい．
4. リツイート数と（ツイートを行ったユーザの）フォロワー数との相関は小さい．

リツイートは情報のコピーの拡散であるから，1番目は通常時にはつぶやく（日常の感想を記す）ための道具として使われているTwitterが，ひとたび重大な出来事が発生すると情報拡散用ツールとして機能することを意味する．2番目は，Twitter上の情報拡散の仕方に一定の法則が存在することを示唆している．3番目はごく一部のツイートにリツイートが集中すること，さらには（キーワードに関わらず分布形状が共通であることから）その集中の仕方にやはり一定の法則があることを意味するものと考えられる．4番目はいわゆるインフルエンサーがリツイート数の多寡に影響しないことを示唆している．

本稿では，上述の特徴について実データを用いて詳細に説明を行うとともに，特に3番目の特徴に焦点を当て，3番目の特徴がどのような人間の行動特性（リツイート行動）から生じているかについて考察する．

以下，2章ではキーワード検索で収集したTwitterデータの特徴を詳細に説明する．次いで，3章では，人間のリツイート行動に単純なルールを仮定した確率モデルにより3番目の特徴が再現できること，その結果，人々はツイート内容よりもそれまでにリツイートされている回数を基準としてリツイートするか否かを判断している可能性が高いことを示す．最後に，4章において結論を述べる．

表 1: キーワード毎の総ツイート数，リツイート（RT）数

キーワード	総ツイート数	リツイート数	RT 割合
ハロウィン	6,592,492	4,362,237	66.2%
ゴーン	838,076	616,565	73.6%
miss universe	451,784	388,709	86.0%
紅白	3,526,719	2,351,358	66.7%
嵐	3,971,476	2,338417	58.9%
NGT	329,598	223,896	67.9%
なう	772,841	144,957	18.8%
拡散希望	538,722	395,688	73.4%
インフル	374,564	74,064	19.8%
美しい	573,436	278,379	48.5%

2. キーワード検索で収集されるTwitterデータの特徴

2.1 収集方法

Twitter APIのキーワード検索機能により，指定したキーワードを含む収集日からさかのぼって10日間分のツイート（リツイートを含む）を収集する [*1] とともに，収集した各ツイートの投稿日，投稿ユーザのフォロワー数，リツイート数，お気に入り数などのメタデータをあわせて取得した．表1に，本稿で分析するデータを収集する際に使用したキーワードを示す．

2.2 リツイートの占める割合

リツイートはオリジナルツイートのコピーをフォロワーに流す行為であり，典型的な情報拡散である．従って，全体のツイートの中にリツイートが占める割合で，Twitterが情報拡散ツールとして機能する程度を測ることができる．

連絡先: 塩田茂雄，千葉大学，〒263-8522 千葉市稲毛区弥生町1-33，Tel/Fax: 043-290-3237，shioda@faculty.chiba-u.jp

*1 Twitter APIでは，(検索時点を起点とする）過去10日間のデータを収集することができる．

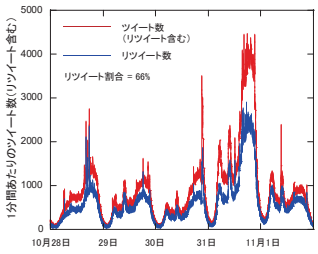

図 1: 1 分当たりのツイート数の時間変化（ハロウィン）

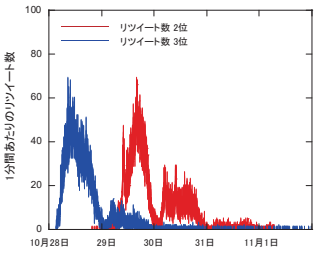

図 2: リツイート数上位のリツイート数変化（ハロウィン）

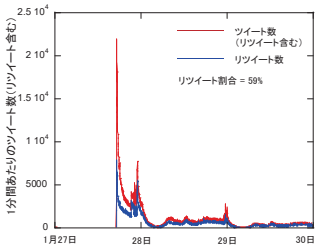

図 3: 1 分当たりのツイート数の時間変化（嵐）

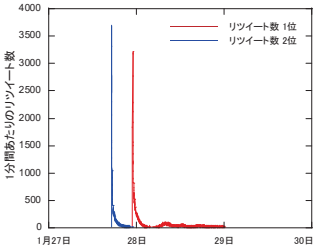

図 4: リツイート数上位のリツイート数変化（嵐）

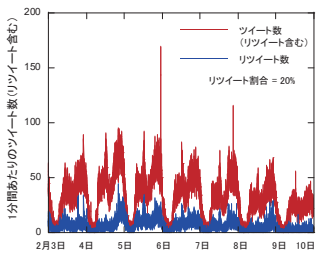

図 5: 1 分当たりのツイート数の時間変化（インフル）

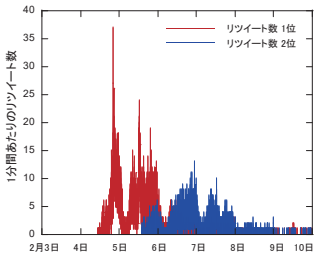

図 6: リツイート数上位のリツイート数変化（インフル）

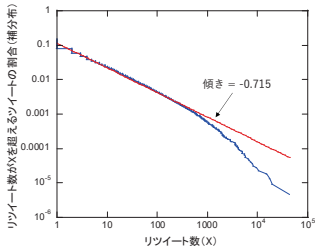

図 7: リツイート数の補分布（ゴーン）

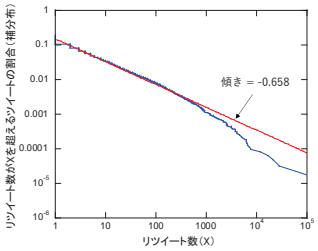

図 8: リツイート数の補分布（miss universe）

表 1 に，キーワード毎に収集した総ツイート数（リツイートを含む），リツイート数，およびリツイートが全体に占める割合を記した．表 1 において,「ハロウィン」から「NGT」までは非日常的な出来事にかかわるキーワードであり，リツイートが半分以上を占める．一方,（日常的なキーワードに相当すると考えられる）「なう」,「インフル」,「美しい」で検索を行うと，リツイートは半分以下となる．

以上のように，普段は日常の感想等を書き込む（つぶやく）ためのツールとして機能している Twitter は，ひとたび非日常的な出来事が発生すると情報拡散ツールとして使われる実態が確認された．

なお,「拡散希望」は日常的なキーワードの 1 つと考えられるが,（拡散を希望するという内容のツイートであるためか）リツイート率は例外的に高かった．

2.3 1 分間あたりツイート数の時間変化

図 1 は「ハロウィン」というキーワードで収集したツイートの 1 分間あたりのツイート数の時間変化を示したものである．図で赤線はツイートとリツイートを合わせた数，青線はリツイートのみの数を表す．図から，ハロウィン（10 月 31 日）より前の 10 月 28 頃から多数のツイートが投稿されており，昼夜変動を繰り返しながら，ハロウィン当日にツイート数のピークが生じている様子が明瞭に見える．10 月 28 日は日曜日であり，渋谷で暴徒と化した集団が軽トラックを横転させるという事件が起きた日である．ハロウィン終了後も多数のツイートの投稿が確認される．図 2 はリツイート数が第 2 位と第 3 位（6 万 6 千件，5 万 8 千件）のツイートのリツイート数の変化である．リツイート数最上位（10 万 7 千件）のリツイートの時間変化データは，途中までしか取れていなかったため，図には掲載しなかった．なお，リツイート数が最大のツイートは交番で警察官がお菓子をあげている所に遭遇した，という内容のツイートであった．

「嵐」というキーワードで取得した Twitter データに関する結果を図 3 と図 4 に示す．データはアイドルグループ「嵐」が 2020 年での活動休止を発表した時期に取得したものである．「嵐」の活動休止は「嵐」の公式サイトで流れ，その直後に 1 分間に 20,000 件を超える大量のツイートが書き込まれている．その少し直後に別のピークが見られるが，これは「嵐」の会見が TV で流れているときのものと思われる．しかし，Twitter への書き込みは 27 日の二つのピーク以降，急速に減少している．なおリツイート数最上位（16 万 4 千件）は「嵐」の会見に関連して青木源太アナウンサーが書き込んだツイート，第 2 位（10 万件）はテレ朝 news の「嵐」の活動休止を報告するツイートであった．

図 5 と図 6 は「インフル」というキーワードで 2019 年 2 月 3 日から取得した結果である．主として，インフルエンザに関するツイートが収集されている．インフルエンザは必ずしも日常的なキーワードではないが，今冬は全国的にインフルエンザが広く流行し，インフルエンザにかかることが珍しくない（日常的な）出来事になっているためか，1 分間あたりのツイート数には日変動はあまり観察されず，昼夜変動を繰り返すだけの定常的な時間変化を示している．実際，リツイートの占める割合も少なく，インフルエンザに関するごく日常的な書き込みが多数を占めている．リツイート数 1 位のツイートもごく日常的な出来事に関するものであった．

2.4 リツイート数の補分布

次に，キーワード毎に，オリジナルツイートがそれぞれ何回リツイートされているかを確認し，その分布の特徴を分析した．一部のキーワードについての結果を図 7 から図 10 に示す．図の横軸はリツイート数（X），縦軸はリツイート数が X を超えるツイートの割合（リツイート数の補分布）を示す．いずれも両対数グラフである．どのキーワードについても，リツイート数の補分布は両対数グラフで直線状にプロットされ，べき分布を連想させる，典型的な裾の長い分布に従う．

表 2 には，それぞれのキーワードに対する補分布の傾きとべき分布で近似したときのべき指数を示した．なお，べき指数は補分布の傾きの符号を反転し（正の値とし），1 を加えたものに等しい．非日常的なキーワードに対するべき指数と日常的なキーワードに対するべき指数は若干異なり，前者の方がやや

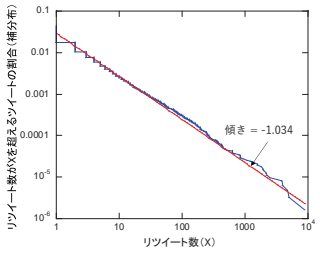

図 9: リツイート数の補分布（なう）

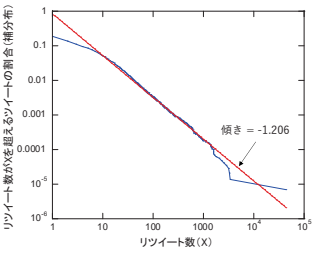

図 10: リツイート数の補分布（拡散希望）

表 2: キーワード毎の補分布の傾きとべき指数

キーワード	補分布の傾き	べき指数
ハロウィン	-0.765	1.765
ゴーン	-0.715	1.715
miss universe	-0.658	1.658
紅白	-0.680	1.680
嵐	-0.715	1.715
NGT	-0.698	1.698
なう	-1.034	2.034
拡散希望	-1.206	2.206
インフル	-0.931	1.931
美しい	-0.817	1.817

小さい値を取る．

2.5 リツイート数と発信者のフォロワー数との相関

最後に各ツイートのリツイート数と発信者のフォロワー数との相関係数を確認した結果を表 3 に示す．やや意外なことに，どのキーワードについても，リツイート数と発信者のフォロワー数との相関係数は小さく，(フォロワー数という尺度での) インフルエンサーが Twitter 上の情報拡散に必ずしも大きな影響力を持つものではないという結果となった．参考までに，各ツイートのリツイート数とお気に入り数の相関係数も示した．リツイート数とお気に入り数の間には明らかな相関がある．

3. リツイート数分布のべき則性の出現モデル

3.1 優先的選択ルールに基づくリツイートモデル

リツイート数がべき分布に従う傾向は，Barabási-Albert モデル [Barábasi 99] で用いられた優先的選択ルールによって説明できると考えられる．以下，優先的選択ルールの要素を取り入れた単純なリツイートモデルを説明する．時刻 t までに書き込まれたツイートの総数を $N_0(t)$，時刻 t までの総リツイート回数を $N_1(t)$，n 番目に書き込まれたツイートの時刻 t でのリツイート数を $r_n(t)$，n 番目に書き込まれたツイートの内容を点数化したものを a_n とする．

1. 時刻 0 以降，頻度 λ_0 でツイートが書き込まれる．
2. 時刻 0 以降，頻度 λ_1 で，その時刻までに書き込まれた全ツイートの中から 1 つツイートが選ばれて，リツイートされる．
3. 2 において，n 番目に書き込まれたツイートは確率 $(a_n + r_n(t))/\sum_{i=1}^{N_0(t)}(a_i + r_i(t))$ で選択される．a_n は n 番目のツイートの価値（魅力）を数値化したものに相当する．

簡単のために，$a_1 = a_2 = \cdots = a$ とする．リツイートされる機会は単位時間あたり λ_1 回存在し，1 回の機会あたり，n 番目に書き込まれたツイートがリツイートされる確率は $(a + r_n(t))/\sum_{i=1}^{N_0(t)}(a + r_i(t))$ であることから，次が成り立つ．

$$\frac{dr_n(t)}{dt} = \lambda_1 \frac{a + r_n(t)}{\sum_{i=1}^{N_0(t)}(a + r_i(t))}$$

ここで，時刻 t までに書き込まれたツイートの総数はおよそ $\lambda_0 t$ に等しいこと，また時刻 t までのリツイート総数はおよそ $\lambda_1 t$ に等しいことより

$$\sum_{i=1}^{N_0(t)}(a + r_i(t)) = aN_0(t) + N_1(t) \approx (a\lambda_0 + \lambda_1)t.$$

したがって

$$\frac{dr_n(t)}{dt} \approx \lambda_1 \frac{a + r_n(t)}{(a\lambda_0 + \lambda_1)t}. \quad (1)$$

n 番目に書き込まれたツイートの書き込み時刻を t_n とする．$r_n(t_n) = 0$ の初期条件のもとで (1) を解くと，次が得られる．

$$r_n(t) = a\left(\left(\frac{t}{t_n}\right)^{1/\gamma} - 1\right), \quad \gamma \stackrel{\text{def}}{=} \frac{a\lambda_0 + \lambda_1}{\lambda_1}.$$

上式より，n 番目に書き込まれたオリジナルツイートのリツイート数が x を超える，つまり $a\left(\left(\frac{t}{t_n}\right)^{1/\alpha} - 1\right) > x$ であることは，

$$t_n < t\left(\frac{a}{a + x}\right)^{\gamma}$$

であること，つまりそのツイートが時刻 $t\left(\frac{a}{a+x}\right)^{\gamma}$ 以前に書き込まれたことを意味する．オリジナルツイートは一定の頻度で書き込まれているので，時刻 t までに書き込まれたツイートのうち，時刻 $t\left(\frac{a}{a+x}\right)^{\gamma}$ 以前に書き込まれたツイートの割合は $\left(\frac{a}{a+x}\right)^{\gamma}$ に等しい．したがって，リツイート数が x を超えるオリジナルツイートの割合 $P(r > x)$ は

$$P(r > x) = \left(\frac{a}{a + x}\right)^{\gamma} \approx \left(\frac{x}{a}\right)^{-\gamma}$$

つまり，リツイート数の分布はべき則に従い，そのべき指数は $\gamma + 1$ に等しい．このモデルでは $\gamma \geq 1$ であり，従ってべき指数の値は 2 以上である．

表 2 で示したように，日常的なキーワードの場合，リツイート数の分布のべき指数は 2 付近もしくはそれ以上の値を取るため，このモデルで説明できる．また，平均リツイート数は λ_1/λ_0 で与えられる点に注意すると

$$a = (\gamma - 1)\frac{\lambda_1}{\lambda_0} = (\gamma - 1) \times \text{平均リツイート数}$$

表 3: リツイート数と（発信者の）フォロワー数，お気に入り数との相関係数

キーワード	対フォロワー数	対お気に入り数
ハロウィン	0.05	0.86
ゴーン	0.05	0.82
miss universe	0.03	0.95
紅白	0.19	0.94
嵐	0.12	0.91
NGT	0.01	0.96
なう	0.23	0.88
拡散希望	0.05	0.98
インフル	0.03	0.95
美しい	0.17	0.83

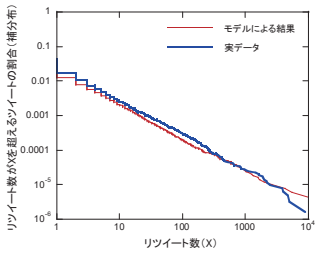

図 11: 提案モデルによるリツイート数の補分布と実データとの比較（なう）

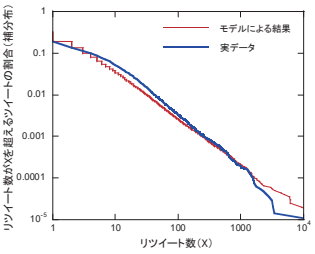

図 12: 提案モデルによるリツイート数の補分布と実データとの比較（拡散希望）

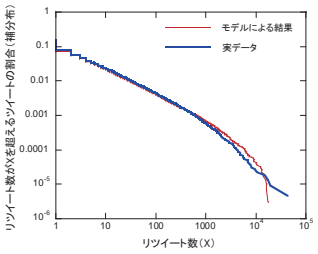

図 13: 提案モデルによるリツイート数の補分布と実データとの比較（ゴーン）

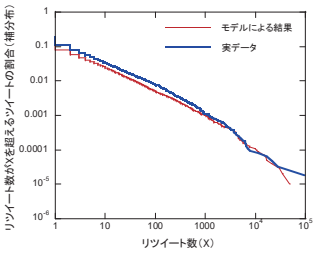

図 14: 提案モデルによるリツイート数の補分布と実データとの比較（miss universe）

が成り立つので，べき指数の値と平均リツイート数から a の値も定まる．キーワードが「拡散希望」のときの平均リツイート数は 2.99，γ の値は 1.206 であるので，$a = 0.616$ である．また，キーワードが「なう」のときの平均リツイート数は 0.27，γ の値は 1.034 であるので，$a = 0.009$ となる．いずれも a の値は非常に小さい．

図 11 は $\lambda_1/\lambda_0 = 0.27$，$a = 0.03$ の設定で，同様に提案モデルを用いてシミュレーションを行い，リツイート数の補分布を図示し，「なう」というキーワードで取得した Twitter データのリツイート数の補分布と比較したものである．補分布はおよそ一致しており，再現性が得られていることが確認できる．一方，図 12 は $\lambda_1/\lambda_0 = 2.99$，$a = 0.616$ の設定で，提案モデルを用いてシミュレーションにより各ツイートのリツイート数を生成し，「拡散希望」というキーワードで取得した Twitter データのリツイート数の補分布と比較したものである．やはり，補分布はおよそ一致する．両ケースとも，1 以上の値を取るリツイート数に比べて a の値は非常に小さく，ほぼリツイート数がリツイートされるか否かを決める要因となっている．すなわち，このモデルは，人々がツイートの内容ではなく，主としてリツイート回数に基づいてリツイートするか否かを決めていることを意味している．

3.2 リツイートモデルの改良

3.1 節のモデルではべき指数は 2 以上の値しか取らないが，表 2 で示したように，非日常的なキーワードの場合，リツイート数の分布のべき指数は 2 未満の値になるため，このモデルでは説明ができない．

べき指数を 2 未満にする一つの方法は，ツイート頻度，リツイート頻度に時間依存性を持たせることである．例えば，

$$N_0(t) = \lambda_0 t^{\eta-1}, \quad N_1(t) = \lambda_1 t^{\eta}, \quad (\eta > 1)$$

とする．このとき 3.1 節のモデルの 1 番目及び 2 番目の項目は以下に変わる．

1*. 時刻 t において，頻度 $(\eta-1)\lambda_0 t^{\eta-2}$ でツイートが書き込まれる．

2*. 時刻 t において，頻度 $\eta\lambda_1 t^{\eta-1}$ で，その時刻までに書き込まれた全ツイートの中から 1 つツイートが選ばれて，リツイートされる．

簡単のために，$a_1 = a_2 = \cdots = a$ とする．このとき，同様の考察から以下が得られる．

$$P(r_n > x) = \frac{\left(\left(\frac{a}{a+x}\right)^{1/\eta}\left(t + \frac{a\lambda_0}{\lambda_1}\right) - \frac{a\lambda_0}{\lambda_1}\right)^{\eta-1}}{t^{\eta-1}}$$

$$\approx \left(\frac{x}{a}\right)^{-\frac{\eta-1}{\eta}}. \quad (x \to \infty)$$

従って η の値を選ぶことによって，べき指数を 1 から 2 の間の自由な値に設定できる．

図 13 は $\eta = 3.3$，$\lambda_1/\lambda_0 = 4$，$a = 0.1$ の設定で，改良リツイートモデルに基づくシミュレーションを行って得られたリツイート数の補分布を図示し，「ゴーン」というキーワードで取得した Twitter データのリツイート数の補分布と比較したものである．補分布はおよそ一致しており，再現性が得られていることが確認できる．一方，図 14 は $\eta = 3$，$\lambda_1/\lambda_0 = 6.6$，$a = 0.1$ の設定で，同様のシミュレーションを行って得られたリツイート数の補分布を図示し，「miss universe」というキーワードで取得した Twitter データのリツイート数の補分布と比較したものである．やはり，補分布はおよそ一致する．

4. むすび

本稿では，Twitter API により収集したツイートデータを分析し，幾つかの共通の特徴が見いだされることを示すとともに， 特にリツイート数分布に関する特徴を，人々のリツイート行動に関するシンプルな確率モデルにより再現できることを示した．なお，1 章で述べた特徴のうち，リツイート数の時間変化に関する（2 番目の）特徴ついては，比較的単純な情報拡散モデル [塩田 18b] により再現できることを確認している [塩田 18a]．本稿で用いた確率モデルはまだ単純であり，今後は Twitter データのより詳細な分析を行ってモデルの精緻化を進めるとともに， Twitter 上の現象の予測法への活用についても検討を進めたい．

参考文献

[Barábasi 99] Barábasi, A.-L. and Albert, R.: Emergence of Scaling in Random Networks, *Science*, Vol. 286, pp. 509–512 (1999)

[塩田 18a] 塩田茂雄, 南川雅人, 中島圭佑：キーワード検索で収集される Twitter データの特徴と Twitter 上の情報拡散過程, 電子情報通信学会情報ネットワーク研究会, IN2018-64, pp. 31–36 (2018)

[塩田 18b] 塩田茂雄，南川雅人，中島圭佑：SNS 投稿件数推移分析のための情報拡散モデルと強相関近似解析, 第二回計算社会科学ワークショップ, pp. 1–10 (2018)