一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
NS2018-290 (2019-03)

Twitterデータの特徴分析と人間の行動モデル

南川　雅人†　　中島　圭佑†　　塩田　茂雄†

† 千葉大学大学院融合理工学府地球環境科学専攻, 〒 263-8522 千葉市稲毛区弥生町 1-33
E-mail: †{adda2062,afwa5455}@chiba-u.jp, ††shioda@faculty.chiba-u.jp

あらまし　マーケティング活動や災害などの特異な出来事が発生した際，Twitter などの SNS は貴重な情報源になりうる．そこで，Twitter の特徴や情報拡散過程を明らかにすることが期待されている．本稿では，Twitter API のキーワード検索機能により収集した Twitter データを分析し，(1) 非日常的な出来事に関するキーワードで検索を行うと，リツイートが大半を占める，(2) 単位時間あたりのリツイート数の変化は，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，リツイート数は裾の長い分布に従う，(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，といった幾つかの共通の特徴が見いだされることを述べる．さらに，3 番目の特徴に焦点を当て詳細な分析を行うとともに，3 番目の特徴を再現する確率モデルについて考察する．

キーワード　Twitter，リツイート，情報拡散，べき分布，確率モデル

Characteristics of Twitter Data and Modelling of Human Behavior

Masato MINAMIKAWA†, Keisuke NAKAJIMA†, and Shigeo SHIODA†

† Guraduate School of Science and Engineering, Chiba University, 1-33 Yayoi, Inage, Chiba, 263-8522 Japan
E-mail: †{adda2062,afwa5455}@chiba-u.jp, ††shioda@faculty.chiba-u.jp

Abstract　SNS such as Twitter can be a valuable information source when marketing or when a special event such as disaster occurs. Therefore, it is expected to clarify the characteristics of Twitter and information diffusion process. We analyze the tweet data collected using the keyword search function of Twitter API, so that 1. What kind of topics are spreading information such that retweets account for the majority 2. The number of retweets will follow the fixed patterns 3. To follow the long distribution of the hem that the number of retweets concentrates on some of the tweets, mostly not retweeted. We found four characteristics that the correlation between the number of followers and the number of retweets is small. We also explain that the features of 2 and 3 can be reproduced by a probabilistic model with simple assumptions about human diffusion behavior.

Key words　Twitter，retweet，information diffusion，power law, probabilistic model

1. まえがき

対面型コミュニケーションが中心であった時代の想像をはるかに超える規模の情報が，Twitter のようなソーシャルメディアを介して急速に拡散し，我々の生活に大きな影響を及ぼす時代である．災害や重大な事件などが発生すると，ソーシャルメディアに様々な書き込みやそのコピー（リツイートなど）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ソーシャルメディア上に表出した現象とみなすことができ，それら現象の差異から現実の各事象に関する情報の拡散の速さや拡散範囲，つまりは各事象の社会への影響度を測ることができる．

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，以下の共通の特徴が見いだされることを発見した [1].

（1）現実の事象と直接の関わりのない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，現実の事象に関するキーワードで検察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

表 1: キーワード毎の総ツイート数，リツイート数

キーワード	総ツイート数	リツイート数	リツイート割合
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%
都立& 町田& 高校	27,396	26,034	95.0%
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%

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，重大事象発生時には情報拡散のためのツールとして機能していることを意味する．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，文献 [1] で分析に用いたデータに加えて，2018 年 11 月から 2019 年 2 月頃にかけて新たに取得した Twitter データについても解析を行い，上述の特徴が確認されるかを分析する．また，文献 [1] では 2 番目の特徴（リツイート数の定型パタンによる変動）がソーシャルネットワーク上の情報拡散モデル [2–5] で再現できること等を確認したが，本稿では特に 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モデルから生じているかについて考察する．

以下，**2.** 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.** 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.** において結論を述べる．

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 に本稿で分析するツイート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．

我々は，日常的なキーワードの場合はリツイートが占める割合は半分以下であるが，非日常的な出来事に関わるキーワードの場合，リツイートが半分以上を占めることを文献 [1] で報告した．今回調査したキーワードについても同様の傾向が見られるかを示す．表 1 には，キーワード毎に収集した総ツイート数

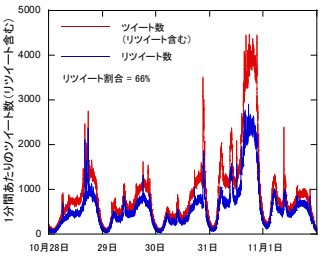

図 1: 1 分当たりのツイート数の時間変化（ハロウィン）

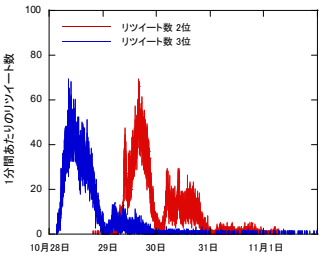

図 2: リツイート数上位のツイート数変化（ハロウィン）

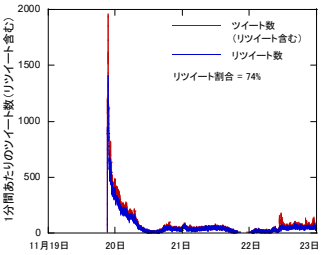

図 3: 1 分当たりのツイート数の時間変化（ゴーン）

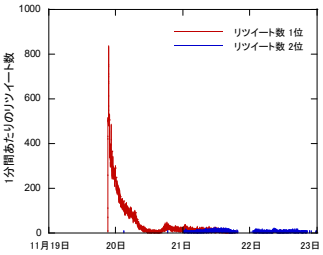

図 4: リツイート数上位のツイート数時間変化（ゴーン）

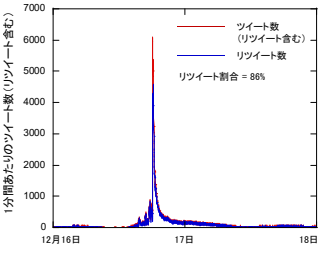

図 5: 1 分当たりのツイート数の時間変化（miss universe）

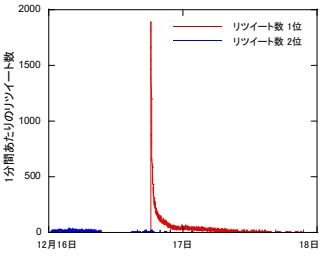

図 6: リツイート数上位のツイート数変化（miss universe）

（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（詳細は 2.3 項を参照のこと），やはりリツイートが半分以上を占めている．また，例えば，キーワードが「ゴーン」の場合，ツイートにおける頻出単語は，「日産」4 万回，「逮捕」3 万回，「容疑」1．5 万回，「ルノー」1．1 万回等であり，ゴーン氏逮捕に関する情報を Twitter 上で発信・拡散しようとしていることが伺える．

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をあらためて確認された．

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

1 分間あたりのツイート数の時間変化を図示したものをキーワード毎に示す．

2.3.1 イベント，事件

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，青線はリツイート数のみの数を表す．図から，ハロウィン（10 月 31 日）

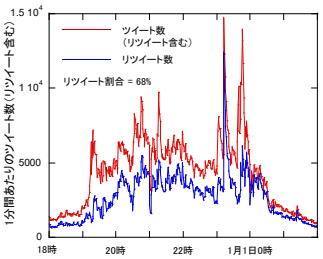

図 7: 1 分当たりのツイート数の時間変化（紅白）

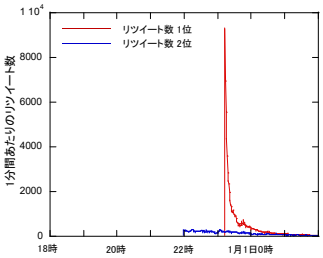

図 8: リツイート数上位のツイート数時間変化（紅白）

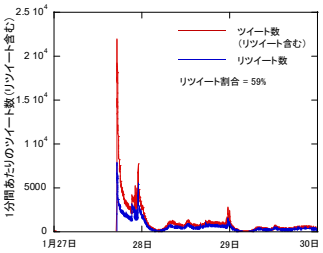

図 9: 1 分当たりのツイート数の時間変化（嵐）

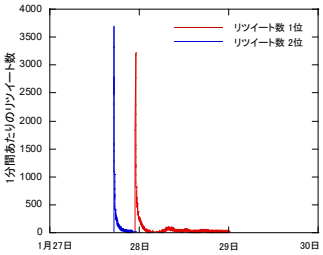

図 10: リツイート数上位のツイート数時間変化（嵐）

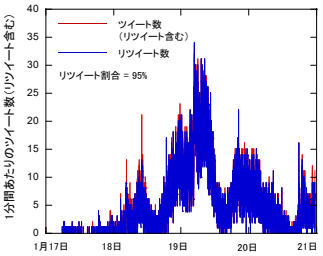

図 11: 1 分当たりのツイート数の時間変化（都立, 町田, 高校）

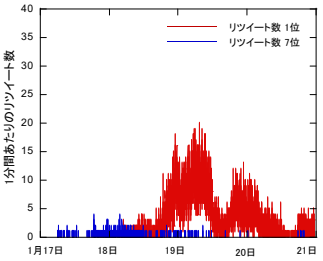

図 12: リツイート数上位のツイート数変化（都立, 町田, 高校）

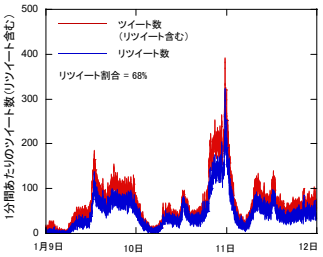

図 13: 1 分当たりのツイート数の時間変化（NGT）

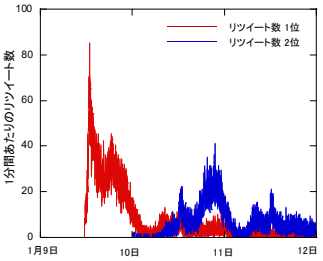

図 14: リツイート数上位のツイート数変化（NGT）

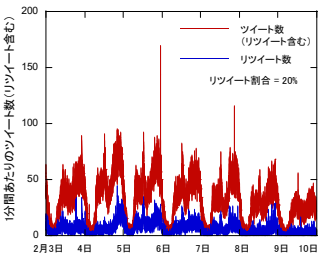

図 15: 1 分当たりのツイート数の時間変化（インフル）

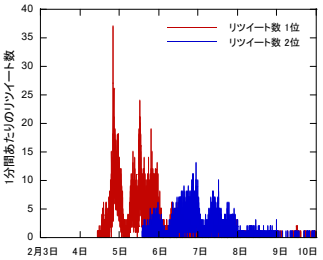

図 16: リツイート数上位のツイート数変化（インフル）

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．

図 3 は「ゴーン」というキーワードで，日産自動車の元社長カルロス・ゴーン氏が逮捕された時期に収集したツイートに関する結果である．ゴーン氏が任意同行されたという報道があった 11 月 19 日に急峻なピークが見られ，その後すぐに書き込みは減少している．図 4 はやはり上位 2 件（4 万 4 千件，2 万件）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はゴーン氏の逮捕を報道するものではなく，自民党税制調査会最高顧問の野田毅氏が「消費税率は 20%を上限と考えたい」と日本記者クラブで発言したことを受けて，ゴーン氏の逮捕よりもこちらの報道の方が重要だとマスコミを揶揄する内容のツイートであった．

「Miss Universe」というキーワード取得した Twitter データに関する結果を図 5 と図 6 に示す．データはミスユニバース 2018 世界大会の優勝報道があった時期に取得したものであり，ツイートは大半が日本語以外の言語で書かれている．やはり優勝者が発表された直後にピークが見える．リツイート数最上位（12 万 1 千件）は，ミスユニバースの主催者からの優勝者を発表する公式ツイートであり，2 番目（2 万 8 千件）はシエラレオネ共和国の代表者に関するツイートであった．米代表の差別的発言が波紋を呼んだと報道されていたが，それに関するツイートはリツイート数の上位ではなかった．

2. 3. 2 芸　　能

図 7 は「紅白」というキーワードでちょうど紅白歌合戦が行われていた時間帯に収集したツイートに関する結果である．12 月 31 日の 23 時頃に二つのピークが見られる．1 つは丁度米津玄師が歌い終わった頃のピークであり，2 番目はサザンオールスターズが歌っていた頃（もしくはフィナーレの頃）の時間帯である．図 8 はやはり上位 2 件（14 万 7 千件，10 万 6 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は米津玄師が自分が歌い終わった直後に投稿したツイートであり，そのツイートに反応してファンがほぼ同時刻にリツイートを行ったことによると考えられるピークが明瞭に見える．

「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 9 と図 10 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．

2. 3. 3 炎　　上

「都立，町田，高校」という 3 つのキーワードで取得した Twitter データに関する結果を図 11 と図 12 に示す．データは「都立町田総合高校」の教師が生徒を殴った場面を別の生徒が撮影し，Twitter で拡散させた事件の報道があったときに取得したものである．図 12 で示したリツイート数 7 位（リツイー

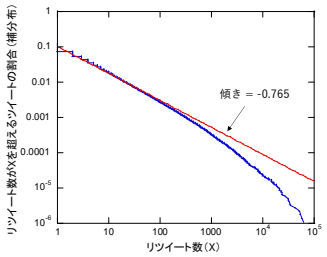

図 17: リツイート数の補分布（ハロウィン）

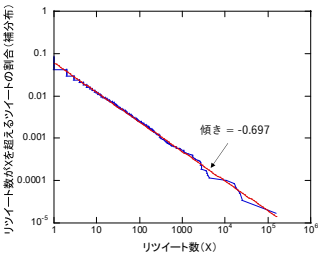

図 18: リツイート数の補分布（ゴーン）

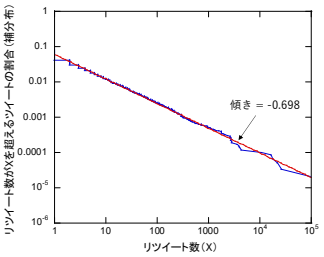

図 23: リツイート数の補分布（NGT）

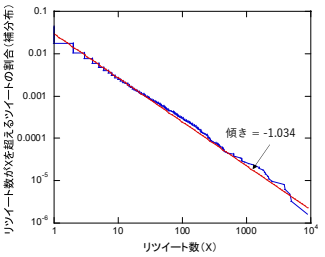

図 24: リツイート数の補分布（なう）

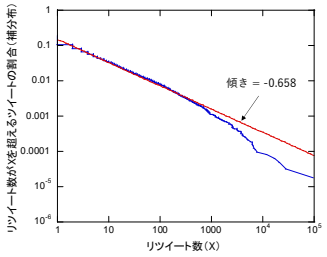

図 19: リツイート数の補分布（miss universe）

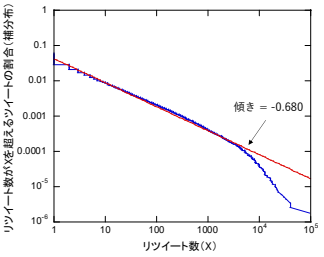

図 20: リツイート数の補分布（紅白）

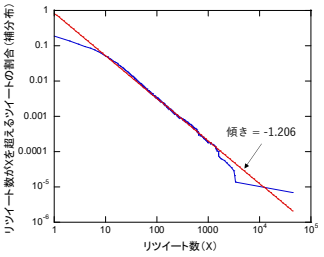

図 25: リツイート数の補分布（拡散希望）

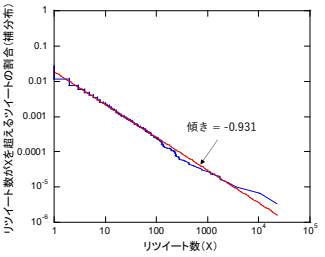

図 26: リツイート数の補分布（インフル）

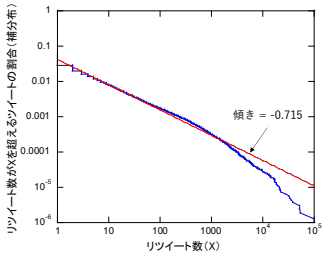

図 21: リツイート数の補分布（嵐）

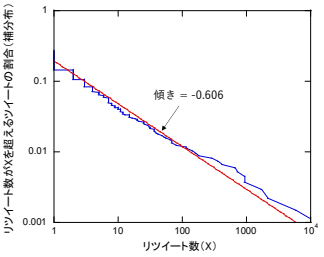

図 22: リツイート数の補分布（都立, 町田, 高校）

ト数 721 件）のツイートは最初に Twitter に流れた動画を含むツイートであり，1 位（リツイート数 2 万件超）は生徒が教師を挑発シーンから撮った動画を含むツイートである．最初のツイートはさほど拡散せず，生徒の行為を非難するきっかけとなったツイッターの方が多数拡散し，生徒を非難するいわゆる「炎上」状態となっている．

今年の 1 月に，NGT の山口真帆さんが，公演後の帰宅時に暴行事件に巻き込まれそうになったことを Twitter に書き込んだことがきっかけに騒動となる事件が起きた．図 13 は「NGT」というキーワードでこの時期の Twitter データを取得した結果である．山口真帆さん自身の Twitter は既に削除されており，リツイート数 1 位（4 万 6 千件）は「週刊ジャーナリズム」からのこの事件を伝えるツイートであった．Twitter 上では，NGT の運営側を非難するある種の「炎上」が生じている．

2.3.4 日　　常

図 15 と図 16 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．

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	-0.697	1.697
miss universe	-0.658	1.658
紅白	-0.680	1.680
嵐	-0.715	1.715
都立& 町田& 高校	-0.606	1.606
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

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．我々は，キーワードに依らずリツイート数の補分布がべき分布のような裾の長い分布に従うことを文献 [1] で報告した．今回調査したキーワードについても同様の傾向が見られるかを確認する．結果を図 17 から図 26 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．

文献 [1] で報告したように，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

一般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は小さく，多くの場合 0.2 未満であることを文献 [1] で報告した．本研究で調べた各キーワードについても同様の傾向があるかを確認した．表 3 はその結果である．やはり，ど

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
都立& 町田& 高校	0.12	0.99
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

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，今回のキーワードについても，リツイート数とフォロワー数との間に有意な因果関係は見いだされず,（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

文献 [1] では，2.4 節で説明したリツイート数がべき分布に従う傾向が，Barabási-Albert モデル [6] のような優先的選択ルールを取り入れたモデルによって再現できることを示した．以下，文献 [1] で説明したモデルを再掲する．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

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

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 $\gamma \geqq 1$ 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

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

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 27 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，結果を図示したものである．参考に「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も示す．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 27 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．これについても，補分布はおよそ一致している．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，3.1 節のモデルの

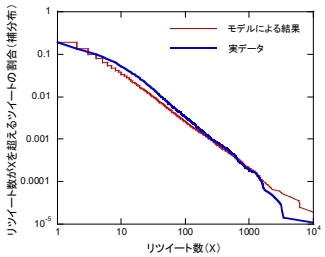

図 27: 提案モデルによるリツイート数の補分布と実データとの比較（拡散希望）

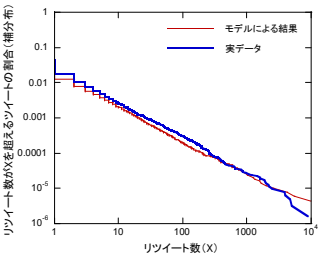

図 28: 提案モデルによるリツイート数の補分布と実データとの比較（なう）

3 番目の項目を以下に変え，さらに a_n の値を負にすることである．

3*. 2 において，n 番目に書き込まれたツイートは確率 $\max\{a_n + r_n(t), b\} / \sum_{i=1}^{N_0(t)} \max\{a_i + r_i(t), b\}$ 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（$b(>0)$ はパラメタ）．

a が負の値を取る場合，リツイート数が $|a|$ を超えて初めて，リツイート回数が有意差を持つことになる（リツイート回数が $|a|$ 未満の場合はリツイート回数は b 回とみなされる．

もう一つの方法は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，3.1 節のモデルの 2 番目の項目を以下に変える．

2*. 時刻 t において，頻度 $\lambda_1 t$ 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．リツイートされる機会は単位時間あたり $\lambda_1 t$ 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．

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

ここで，

$$\sum_{i=1}^{N_0(t)} (a + r_i(t)) = aN_0(t) + N_1(t) \approx a\lambda_0 t + \frac{1}{2}\lambda_1 t^2.$$

したがって

$$\frac{dr_n(t)}{dt} \approx \frac{\lambda_1 (a + r_n(t)) t}{a\lambda_0 t + \frac{1}{2}\lambda_1 t^2}. \qquad (2)$$

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 (2) を解くと，次が得られる．

$$r_n(t) = a \left\{ \left(\frac{t + \delta}{t_n + \delta} \right)^2 - 1 \right\}, \quad \delta \stackrel{\text{def}}{=} \frac{2a\lambda_0}{\lambda_1}.$$

$r_n > x$ であることは

$$t_n < (t + \delta) \left(\frac{a}{a + x} \right)^{1/2} - \delta,$$

と等価であることが導かれるので，

$$P(r_n > x) = \left(1 + \frac{\delta}{t} \right) \left(\frac{a}{a + x} \right)^{1/2} - \frac{\delta}{t}.$$

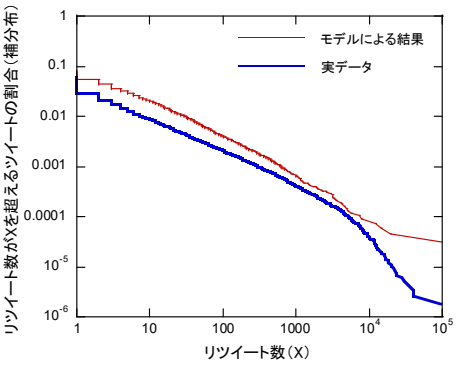

図 29: 提案モデルによるリツイート数の補分布と実データとの比較（拡散希望）

したがって t が十分大きくなれば

$$P(r > x) \approx \left(\frac{a}{a + x} \right)^{1/2} \approx x^{-1/2}.$$

つまり，べき指数を 1.5 まで小さくする（補分布の傾きを-0.5 まで大きくする）ことが可能となる．

図 29 は $\lambda_0 = 1$，$\lambda_1 = 1 + 1000t$，$a = 0.02$ の設定で，後者の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，結果を図示したものである．比較として「紅白」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も示す．補分布は比較的一致しており，べき指数が 2 未満の分布が再現されている．

4. むすび

本稿では，Twitter API を通して種々のキーワードで検索を掛けて収集したツイートデータを分析し，そこに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，リツイート数の分布の特徴を，人々のリツイート行動に関してシンプルな仮定をおいた確率モデルにより再現できることを示した．本稿で用いた確率モデルはまだ単純であり，現実の現象の再現性を十分に確保するものではないが，今後モデルの精緻化を進めるとともに，パラメタフィッティングの手法や 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．

文　献

[1] 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.
[2] 南川雅人, 塩田茂雄, “ネットワーク上の情報拡散過程におけるノード相関の影響,” 電子情報通信学会 コミニュケーションクオリティ研究会, CQ2017-58, pp.43–58, 2017.
[3] 中島圭佑, 南川雅人, 塩田茂雄, “SNS における投稿件数推移分析のための情報拡散モデル,” 電子情報通信学会 コミニュケーションクオリティ研究会, CQ2017-84, pp.79–84, 2017.
[4] 南川雅人, 中島圭佑, 塩田茂雄, “強相関近似による複雑ネットワーク上の情報拡散過程の解析,” 2017 年度待ち行列シンポジウム「確率モデルとその応用」, pp.194–195, 2018.
[5] 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, 2017.
[6] A.-L. Barábasi and R. Albert, “Emergence of scaling in random networks,” Science, vol.286, pp.509–512, 1999.