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
IN2018-64（2018-12）

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

塩田　茂雄†　南川　雅人†　中島　圭佑†

† 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: †shioda@faculty.chiba-u.jp, ††{adda2062,afwa5455}@chiba-u.jp

あらまし　本稿では，Twitter API のキーワード検索機能により収集したツイートデータを分析し，(1) 現実の事象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，といった幾つかの興味深い特徴が見いだされることを述べる．さらに，ツイートデータに見られる特徴のうち (2) と (3) が，人々のリツイート（情報拡散）行動に関する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説明する．

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 Collected by Keyword Search and Information Diffusion Process on Twitter

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

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

Abstract　Analyzing the large number of tweets collected by keyword search using the Twitter API, we show that Twitter has the following common features: (1) The majority of tweets collected by keywords related to actual events are retweets; (2) the number of retweets rapidly increases and, after reaching the peak, it gradually decreases with day and night fluctuation; (3) the number of retweets often follows a long-tailed distribution, regardless of the keyword for collecting the data; (4) the correlation coefficient between the number of retweets and the number of followers of the user posting the tweet is close to zero. Furthermore, we show that features (2) and (3) mentioned in the above can be reproduced by using a probabilistic model based on simple rules on people's retweet (information 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）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索する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，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

This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.
Copyright ©2018　by　IEICE

表 1: キーワード，総収集数

キーワード	総収集ツイート数
北海道 & 地震	3,536,153
台風 24 号	1,447,677
さくらももこ	676,678
樹木希林	554,052
大坂なおみ	622,121
ドラフト	642,195
牛乳プリン	70,255
区役所と仕事したまんが & 山本さほ	92,297
美しい	573,436
なう	772,841

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，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，Twitter API のキーワード検索機能により収集した Twitter データを用いて上述の特徴を詳細に説明するとともに，特に 2 番目と 3 番目の特徴が，人々のリツイート（情報拡散）行動に関する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明らかにする．

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，感染症の数理モデルに類似した単純な情報拡散モデルを用いて 2 番目の特徴が再現できる様子を示す．**4.** 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，**5.** 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 において，指定したキーワードで検索を行い，キーワードを含むツイート（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 は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と収集したツイート総数である．それぞれのキーワードについて，検索を行った日を起点として過去 10 日間のデータを収集している (注1)．ツイート総数にはリツイートも含まれる．

2.2 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．

(注1)：Twitter API では，(検索時点を起点として）過去 10 日間のデータをさかのぼって収集することができる．

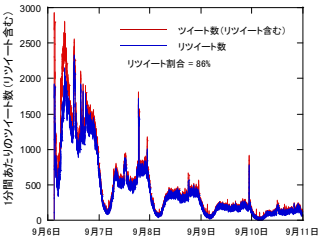

図 1: 1 分当たりのツイート数の時間変化（北海道地震）

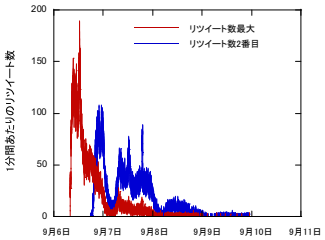

図 2: リツイート数上位 2 件の時間変化（北海道地震）

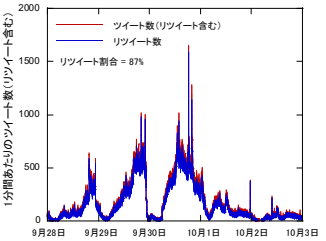

図 3: 1 分当たりのツイート数の時間変化（台風 24 号）

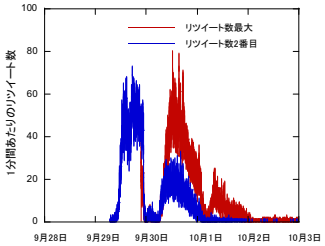

図 4: リツイート数上位 2 件の時間変化（台風 24 号）

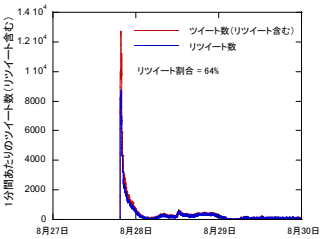

図 5: 1 分当たりのツイート数の時間変化（さくらももこ）

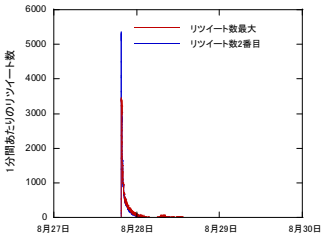

図 6: リツイート数上位 2 件の時間変化（さくらももこ）

2.2.1 災　　害

図 1 は「北海道地震」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．地震発生時（9 月 6 日深夜 3 時）に急峻なピークが見られ，その後，昼間や増加し夜間は減少するという昼夜パタンの変動を繰り返しながら，全体として減衰していく様子が見える．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，赤線と青線がほぼ重なっており，大半がリツイートである（リツイートが占める割合は約 86%）．リツイートはオリジナルツイートのコピーの拡散であることから，北海道地震の際に Twitter 上で行われた行動は情報拡散そのものであると言える．図 2 はリツイート数が上位 2 件（10 万 2 千件，9 万 3 千件）のツイートのリツイート数の変化である．北海道地震では 300 万件を超えるツイートが流れており，上位 2 件のリツイート数が全体に占めるボリュームは小さい．上位 2 件とも，急峻なピークを迎えたのち（昼夜変動を繰り返しながら）減衰している．なお，リツイート数が最大のツイートは札幌の焼き肉屋が無償で焼肉を振る舞っているという内容のツイートであった．図 3 は「台風 24 号」というキーワードで収集した場合の結果である．台風が 9 月 30 日に日本を縦断しており，30 日にツイート数のピークが見られ，やはり大半がリツイートである（リツイートが占める割合は約 80%）．図 4 はやはり上位 2 件（7 万 5 千件，7 万件）のツイートのリツイート数の変化であり，ピークののち昼夜変動を繰り返しながら減衰するというパタンに従う．リツイート数が最大のツイートは国際宇宙ステーションから撮影された台風 24 号の画像を貼り付けてコメントしたものであった．

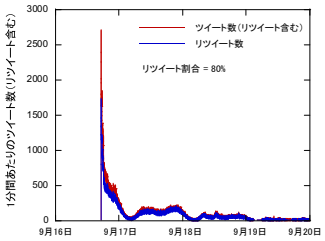

図 7: 1 分当たりのツイート数の時間変化（樹木希林）

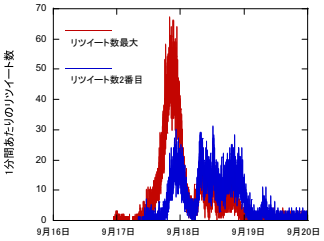

図 8: リツイート数上位 2 件の時間変化（樹木希林）

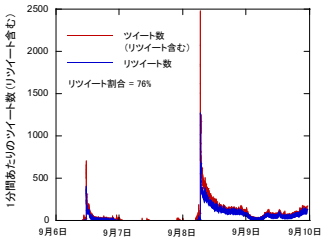

図 9: 1 分当たりのツイート数の時間変化（大坂なおみ）

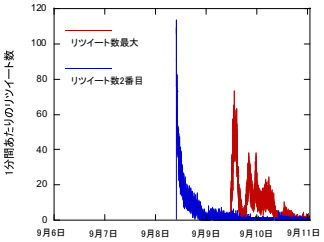

図 10: リツイート数上位 2 件の時間変化（大坂なおみ）

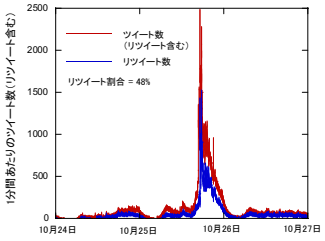

図 11: 1 分当たりのツイート数の時間変化（ドラフト）

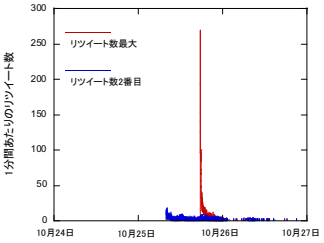

図 12: リツイート数上位 2 件の時間変化（ドラフト）

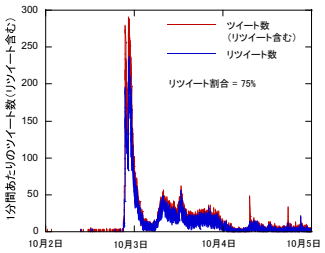

図 13: 1 分当たりのツイート数の時間変化（牛乳プリン）

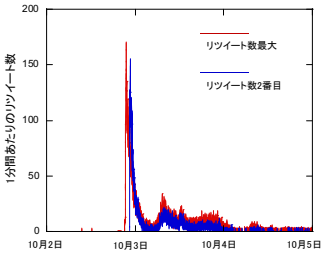

図 14: リツイート数上位 2 件の時間変化（牛乳プリン）

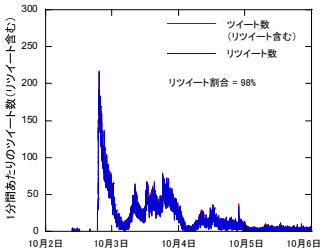

図 15: 1 分当たりのツイート数の時間変化（山本さほ）

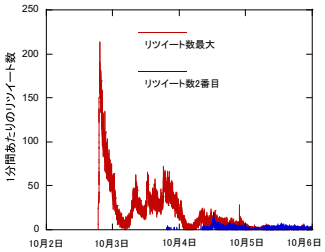

図 16: リツイート数上位 2 件の時間変化（山本さほ）

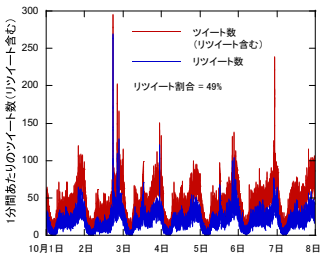

図 17: 1 分当たりのツイート数の時間変化（美しい）

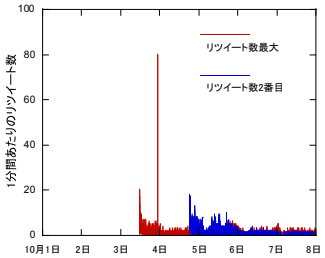

図 18: リツイート数上位 2 件の時間変化（美しい）

2.2.2 芸能，スポーツ

今年亡くなられた「ちびまる子ちゃん」の作者のさくらももこさんの名前をキーワードとして収集した結果を図 5 と図 6 に示す．さくらももこさんの訃報は講談社からツイートとして（おそらく最初に）流れ，その直後に 1 分間に 12000 件を超えるツイート（講談社からのツイートのリツイートを含む）が書き込まれている．このピーク以降は書き込み数は急速に減少した．なおリツイート数が上位 2 件のツイート（いずれもリツイート数は 20 万件超）はいずれも講談社からの訃報である．リツイートは全体の 65%を占めた．

図 7 と図 8 は，やはり今年亡くなられた女優の樹木希林さんに関するツイートである．樹木希林さんの訃報は最初に TV 等で流れ，それを見た人が書き込んだと思われるツイート数の急峻なピークが見られる．リツイート数が上位 2 件のツイートは，訃報から少し時間的に経過してから書き込まれており，リツイート数が最大（約 52000 件）のツイートは樹木希林さんの内田裕也さんに対する想いを記載した内容であった．リツイートは全体の 80%である．

図 5 と図 6 は今年の全米テニスで優勝した大坂なおみ選手の名前をキーワードとして収集した結果である．全米テニスの準決勝と決勝で勝利した直後にツイート数のピークが見られる．リツイート数が最大のツイートは決勝戦から 1 日程度経過してから書き込まれており，内容は大坂なおみ選手の出身地に関するツイートであった（約 43000 件）．リツイートは全体の 76%である．図 7 と 8 は「ドラフト」をキーワードに野球のドラフト会議の数日後に収集した結果である．ドラフト会議の日の 10 月 25 日をピークとしてツイートが書き込まれている様子が見える．リツイート数が最大のツイートは根尾昂選手の交渉権を引き当てた中日ドラゴンズの公式ツイートである（リツイート数約 6200 件）．リツイートは全体の 48%であり，他のケースとやや異なる．

2.2.3 デマ，炎上

今年，京都大学の生協が流した「森永牛乳プリンが終売になる」というツイートが拡散し，しかし「終売」というのは誤報だったという事件が起きた．図 13 と図 14 は「牛乳プリン」というキーワードで検索した結果である．リツイート数の上位 2 件は京大生協からの誤報と，それを訂正する（森永乳牛からの）ツイートである（それぞれ約 41000 件と約 26000 件）．誤報の訂正のツイートは最初の誤報のあと比較的直ぐ配信されており，最初のツイートの発信者数のフォロワー数（711）よりも，訂正を流した森永乳牛のフォロワー数の方が圧倒的に多い（651715）にもかかわらず，誤報の拡散が止まっていない．

今年の 10 月に，世田谷区役所との仕事上のトラブルを記載した漫画を含むツイートを漫画家の山本さほさんが Twitter に投稿したところ，11 万件を超えるリツイートが行われ，世田谷区長が謝罪するという事件がおきた．典型的な「炎上」である．図 15 は「区役所と仕事したまんが」と「山本さほ」というキーワードで検索した結果，16 は漫画家のツイートのリツイート数の変化である．このケースでは，リツイートは全体のほぼ 100%を占め，98%を最初の山本さほさんのツイートのリツイートが占めている．炎上自体は約 2 日で収束している．

2.2.4 日常的キーワード

最後に「美しい」というキーワード，および「なう」という

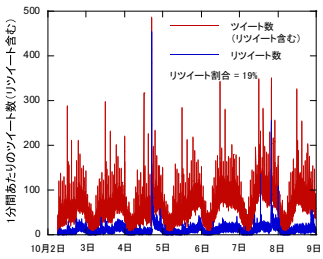

図 19: 1 分当たりのツイート数の時間変化（なう）

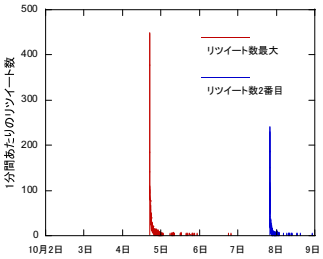

図 20: リツイート数上位 2 件の時間変化（なう）

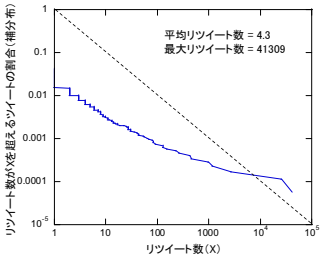

図 27: リツイート数の補分布（牛乳プリン）

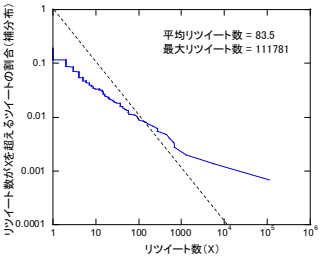

図 28: リツイート数の補分布（山本さほ）

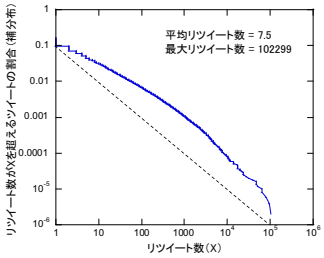

図 21: リツイート数の補分布（北海道地震）

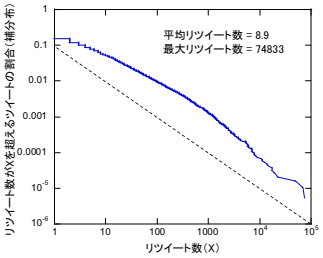

図 22: リツイート数の補分布（台風 24 号）

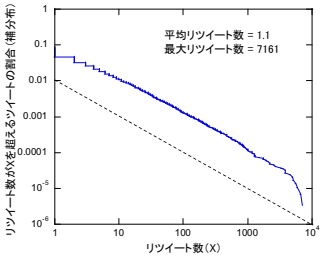

図 29: リツイート数の補分布（美しい）

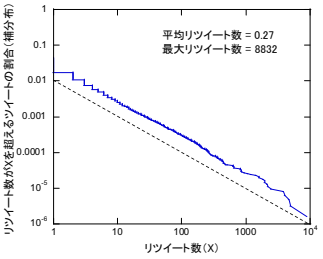

図 30: 1 分当たりのツイート数の時間変化（なう）

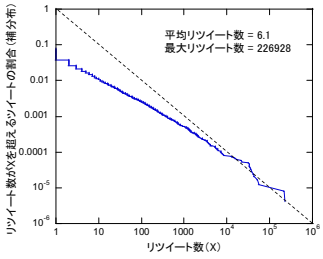

図 23: リツイート数の補分布（さくらももこ）

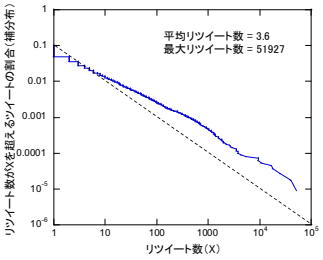

図 24: リツイート数の補分布（樹木希林）

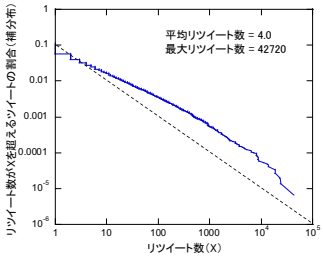

図 25: リツイート数の補分布（大坂なおみ）

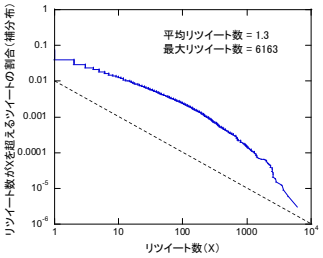

図 26: リツイート数の補分布（ドラフト）

キーワードで検索した結果を図 17 から図 20 に示す．特定の現実の事象とかかわりのない日常的なキーワードで検索を行うと，リツイートの占める割合が 5 割程度以下に減少する．例えば「なう」というキーワードで検索を行った場合，リツイートが占める割合は 20%であった．つまり，Twitter は通常時にはユーザがつぶやく（日常の感想を記載する）ツールとして機能し，実社会で重大な事象が発生すると，その情報を拡散するためのツールとして機能する二面性を持つ．なお「なう」というキーワードでのツイートのうち最も多数回リツイートされていたのは，某声優がハワイからツイートしたものであった．

2.3 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．結果を図 21 から図 30 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．

大変興味深いことに，どのキーワードの場合でも，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．実際，最大リツイート数は平均リツイート数は数千倍から数万倍に達し，一方，多くの場合，リツイート数が 1 以下のツイートが 90%以上を占める．また，両対数グラフでプロットした際の直線の傾きはおよそ 1 に近く，「牛乳プリン」,「区役所と仕事したまんが & 山本さほ」,「さくらももこ」などのキーワードで検索した場合は 1 未満である．これは，べき分布で近似した際のべき指数がおよそ 2（もしくは 2 未満）であることを意味する．これは，複雑ネットワークで見られる次数のべき指数が 2 から 3 の間にあることとは対照的な特徴である．

2.4 リツイート数と発信者のフォロワー数との相関

最後に，各ツイートの発信者のフォロワー数とリツイート数との相関を調べた．表 1 はその結果である．意外なことに，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．一般に，相関係数が 0.2 を下回ると相関はないとみなされるので，少なくとも収集したデータからは，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間に有意な因果関係は見いだされなかった．フォロワー数は Twitter 上でのそのユーザの影響度を表すと考えられており，フォロワー数が特に大きなユーザは「インフルエンサー」とみなされることが多い．従って，表 1 は（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ことを示唆している．なお，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（「北海道地震」と「さくらももこ」のキーワードで検索したデータでは，お気に入り数が入っていなかったので，表には掲載していない）．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り，人々がお気に入り数の多いツイートをリツイートする傾向があることがうかがえる．

3. リツイート数の時間変化と情報拡散モデル

2.2 節で述べたように，リツイート数の多いツイートに関す

表 2: 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
北海道 & 地震	0.08	–
台風 24 号	0.08	0.88
さくらももこ	0.04	–
樹木希林	0.08	0.97
大坂なおみ	0.07	0.95
ドラフト	0.27	0.91
牛乳プリン	0.20	0.96
世田谷区役所 &まんが	0.10	1.00
美しい	0.16	0.83
なう	0.23	0.88

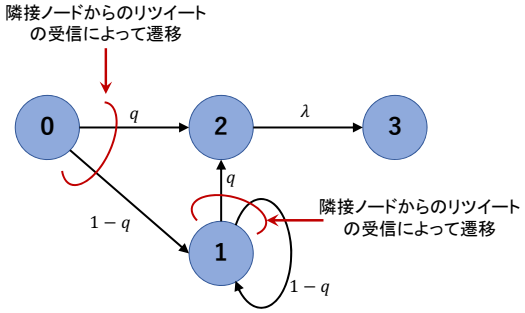

図 31: 情報拡散モデルによる状態遷移

る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．ソーシャルネットワーク上の情報拡散のメカニズムについては様々なモデルが提案されているが [1, 2]，本章では，我々が提案した情報拡散モデル [3–6] を用いて，その定型パタンが再現できるかを検証する．

3.1 情報拡散モデル

Twitter を有向グラフで表現し，1 つないしは複数のノードを起点として，情報（リツイート）が有向リンクを経由して有向グラフ内の各ノードに拡散していく現象をモデル化する．ノードは以下のいずれかの状態を取る（図 31）．

状態 0: ツイート（リツイート）を受信していない

状態 1: ツイート（リツイート）を受信したが，リツイートしない

状態 2: ツイート（リツイート）を受信し，リツイート予定

状態 3: ツイート（リツイート）を受信し，リツイート済み

初期状態ではツイート発信元のノードが状態 2 に，それ以外のノードが状態 0 にあり，ツイート発信元ノードが状態 2 から状態 3 に遷移することで，情報拡散がスタートする．各ノードは（隣接ノードから）ツイート（リツイート）を受信することで，状態 0 から確率 $1-q$ で状態 1 に，確率 q で状態 2 に遷移する．状態 2 に遷移後は，平均 $1/\lambda$ の指数分布に従う時間経過後に，全ての隣接ノードに同時にリツイートし，状態 3 に遷移する．状態 3 に遷移後は，そのまま状態 3 に留まる．状態 0 から状態 1 に遷移した場合は，状態 1 に留まるが，別の隣接ノードからツイート（リツイート）を受信すると確率 q で状態 2 に遷移する．全てのノードが状態 0，状態 2，状態 3 のいずれかになると拡散は終了する．このモデルにおいて，確率 q はノードがツイートの内容に興味を持つ確率に相当する．

3.2 現実の現象の再現性

3.1 節のモデルは，ネットワーク構造，さらには λ と q という二つのパラメタを含む．現実の Twitter のネットワーク構造を完全に知ることは困難であるため，ネットワーク構造については仮定を置くとして，λ と q を調整することにより，現実の現象をある程度再現できると考えられる．また，3.1 節のモデルには，いわゆる昼夜変動（リツイートは昼間は行うが，夜間が行わない）が考慮されていないが，λ を時間の関数とすることにより，昼夜変動を取り入れることも可能である．

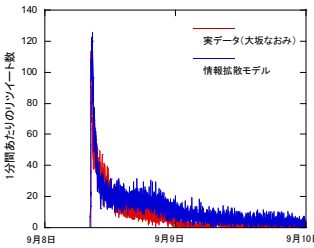

図 32: リツイート数の時間変化のモデルによる再現性（大阪なおみ）

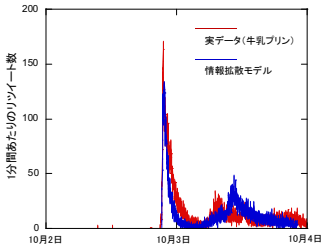

図 33: リツイート数の時間変化のモデルによる再現性（牛乳プリン）

図 32 は，2.2.2 項で述べた大坂なおみ選手の（2 番目にリツイートが多かった）例を再現することを目的として，3.1 節の拡散モデルによりシミュレーションを行った例である．9 月 8 日のおよそ 11 時に最初のツイートが書き込まれたと仮定し，$1/\lambda$（状態 2→3 の遷移にかかる平均時間）は通常は 8 時間とし，夜間（23 時半〜6 時）は $1/\lambda$ =80 時間，また，最初のツイートが書き込まれてから 10 分間は $1/\lambda$ =20 分，次の 10 分間は $1/\lambda$ =1 時間とした．また $q = 0.1$ とした．ネットワーク構造は，インターネット上に公開されている Twitter の部分トポロジーデータ（ノード数 81306，リンク数 1768149）[7] を用いた．最初のツイートは出次数 1111 のノードから書き込まれたとした．図 33 は，2.2.3 項で述べた森永牛乳プリンの例を再現することを目的として，同様のシミュレーションを行った例である．10 月 2 日の 23 時に最初のツイートが書き込まれたと仮定し，$1/\lambda$（状態 2→3 の遷移にかかる平均時間）は通常は 1 時間とし，夜間（23 時半〜6 時）は $1/\lambda$ =100 時間，また，最初のツイートが書き込まれてから 10 分間は $1/\lambda$ =20 分とした．他の条件は先ほどと同じである．いずれのケースにおいても，実データを比較的良く再現できている．

4. リツイート数分布のべき則性の出現モデル

2.3 節で述べたように，検索するキーワードによらず，リツイート数は共通にべき分布のような裾の長い分布に従う．本章では，その理由について考察するために，Barabási-Albert モデル [8] の優先的選択ルールにヒントを得た，次のシンプルなリツイートモデルを考える．以下，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

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

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 $\gamma \geqq 1$ である．

なお，このモデルでは，a（a_i）が 0 以下とすると，最初に書き込まれたツイートしかリツイートされなくなるが，リツイートモデルの 3 番目の項目を以下に変えると a（a_i）が負の場合の扱いが可能となり，γ が 1 未満（べき指数が 2 未満）の分布が再現できる．

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

2.3 節の結果を見ると，リツイート数の補分布は $\gamma \leqq 1$（べき指数 2 以下）であるが，これは $a \leqq 0$ であること，すなわち，人々はツイートの内容ではなく，主としてリツイートされている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 (注2)．

図 34 は $a=1$，$\lambda_0=1$，$\lambda_1=10$ の設定（べき指数 2.1）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，「美しい」というキーワードで取得したツイッターデータのリツイート数の補分布も示す．分布の傾きはほぼ等しく，その意味で再現性が得られていることが確認できる．一方，図 35 は $a=-5$，$\lambda_0=1$，$\lambda_1=10$ の設定（べき指数 1.5）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，「牛乳プリン」というキーワードで取得したツイッター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，分布の裾付近の傾きはほぼ等しく，その意味での再現性は実現できている．なお，例えば図 34 では実際の分布はモデルによる分布を下に平行移動したものとなっているが，これは平均リツイート数が異なることを意味する．実データでは平均リツイート数は 1.07 であり，一方，モデルでは平均リツイート数は $10(=\lambda_1/\lambda_0)$ である．その意味では，現実の現象の再現性はできておらず，モデルの改良が必要である．

(注2)：$a<0$ は，リツイート数が $|a|$ を超えるツイートのみリツイートの対象とすることを意味する．

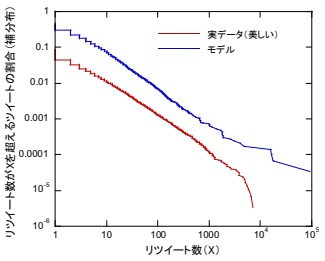

図 34: 提案モデルによるリツイート数の補分布と実データとの比較（美しい）

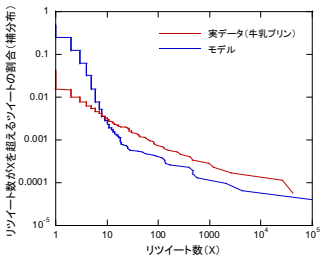

図 35: 提案モデルによるリツイート数の補分布と実データとの比較（牛乳プリン）

5. 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興味深い特徴が見いだされることを示すとともに，単位時間あたりのリツイート数の変化や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] 高野知佐, 会田雅樹, “Scaled Laplacian 行列に基づいた固有ベクトル中心性の考察,” 電子情報通信学会複雑コミュニケーションサイエンス研究会, CCS2017-12, pp.13–18, 2017.

[2] 久保尊広, 高野知佐, 会田雅樹, “縮退した振動モードから生じる新しいネット炎上モデル,” 電子情報通信学会ネットワークシステム研究会, NS2017-80, pp.55–60, 2017.

[3] 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 コミュニケーションクオリティ研究会, CQ2017-58, pp.43–58, 2017.

[4] 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 コミュニケーションクオリティ研究会, CQ2017-84, pp.79–84, 2017.

[5] 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.

[6] 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.

[7] “Stanford large network dataset collection,” http://snap.stanford.edu/data/.

[8] A.-L. Barábasi and R. Albert, “Emergence of scaling in random networks,” Science, vol.286, pp.509–512, 1999.