1H2-NFC-02a-1

情報拡散における重要人物の推定

Finding Important Users for Information Diffusion

鳥海不二夫[*1] Fujio Toriumi
榊 剛史[*1] Takeshi Sakaki
吉田 光男[*2] Mitsuo Yoshida
篠田 孝祐[*3] Kosuke Shinoda
栗原 聡[*3] Satoshi Kurihara
風間 一洋[*4] Kazuhiro Kazama
野田 五十樹[*5] Itsuki Noda

[*1]東京大学 The University of Tokyo
[*2]筑波大学 University of Tsukuba
[*3]電気通信大学 The University of Electro-Communications
[*4]和歌山大学 Wakayama University
[*5]産業技術総合研究所 The National Institude of Advanced Industrial Science and Technology

Wide-scale disasters such as earthquakes, hurricanes and so on, occur unpredictably. During a disaster, it's important to collect information appropriately to save own lives. However, it is difficult to collect information from mass media, such as TV, Newspapers, which contains information which is of use for the general public. Under the disaster situation, victims require information which shows place of shelters or danger points. Also, not only victims but also rescuers require information of victim location or that of shorted supplies. In this paper, we analyse one billion retweet data to find important user on information diffusion.

1. はじめに

大規模な災害の発生を予測することは難しいが，いつ発生してもおかしくはない．この 10 年に限定しても，スマトラ沖地震 (2004)，ハリケーンカトリーナ (2005)，四川大地震 (2008)，チリ地震 (2010)，東日本大震災 (2011) など数多くの大災害が人々を襲っている．また，マグニチュード 7.0 を超える地震だけでも 2010 年には 24 回，2011 年には 20 回観測されている[*1]．このような災害時には，情報を正確に素早く集めることが人命を守るために重要となる．しかしながら，新聞やテレビといったマスメディアは一般的な情報を提供することを目的としている．そのため，避難所の場所や被災地に必要な物資など，被災者や救助者が必要としている情報を必ずしも提供していない．

このような状況下で，ソーシャルメディアによる情報の共有が注目されている．特に，2011 年 3 月 11 日に発生した東日本大震災でソーシャルメディアがさまざまな目的で広く活用されたことは記憶に新しい．ソーシャルメディアの中でも，ツイッターによる災害時の情報共有については，多くの報告が存在する [Vieweg 10] [Heverin 10][篠田 13].

ツイッターには簡単に情報を拡散するための機能として，リツイートが存在する．リツイートはワンクリックで自分をフォローしているユーザに情報を広めることが出来るため，ツイッターが情報共有システムとして機能する上で重要な役割を担っている．

ところで，情報の共有という観点では，情報を発信するユーザと，それを広めるユーザが存在する．Twitter 上において，情報を広めるユーザは自らツイートを行うのではなく，他のユーザの有益なツイートをリツイートすることで，情報の拡散を手助けする．このようなユーザを発見しておくことで，効率よく情報を収集できるようになると期待される．

そこで，本研究ではツイッターが持つ情報拡散機能であるリツイートに着目し，情報拡散において重要な役割を担うユーザを発見することを目指す.

連絡先: 鳥海不二夫，東京大学大学院工学系研究科システム創成学専攻，東京都文京区本郷 7-3-1，tori@sys.t.u-tokyo.ac.jp

*1 http://on.doi.gov/7cqeex

2. リツイート行動の分析

2.1 利用データ

本論文では，ツイートデータの内リツイートデータを TwitterAPI を用いて収集したものを用いる．データは 2013 年 7 月〜11 月まで収集した．その結果，305,876,541 ツイートが，8,917,364 人のユーザによって 1,066,239,711 回リツイートされたデータを収集することに成功した．

本研究では，収集したリツイートデータを用いて分析を行う．

2.2 リツイート回数と被リツイート回数

まず，図 1 に，データ収集期間内の 1 ユーザ当たりのリツイート回数と，総被リツイート回数の分布を示す．リツイート回数とは各ユーザが何回リツイートしたかであり，総被リツイート回数とはあるユーザのツイートについて，リツイートされた回数の総和を取ったものである．これより，総被リツイート回数はほぼベキ分布になっていることが分かる．一方，リツイート回数は同回数であれば被リツイート回数よりもユーザ数が多く，たとえば 100 回のリツイートしたユーザとリツイートされたユーザを比較すると，13525 人と 8460 人である．ここから，ツイート行動によってのべ 100 人に情報を伝えられるユーザと比べ，リツイート行動によってのべ 100 人のツイートを集約できるユーザの方が 1.5 倍いることになる．

以上より，個々のユーザに注目すると，情報を提供する力よりも情報を集約する力の方が強いことが示唆された．

事実，7 月から 11 月までの 4ヶ月に，合計で 100 回以上リツイートされたユーザは 1,133,410 人であるのに対し，100 回以上リツイートしたユーザは 1,975,155 人おり，情報拡散を積極的に行うの方がその数が多いことが分かる．

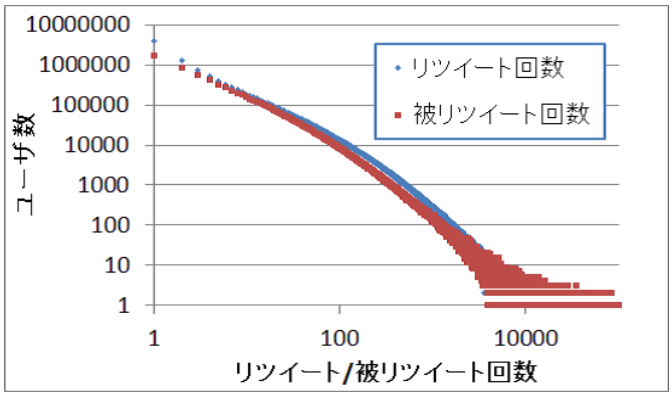

図 1: 総被リツイート数分布と総リツイート数分布

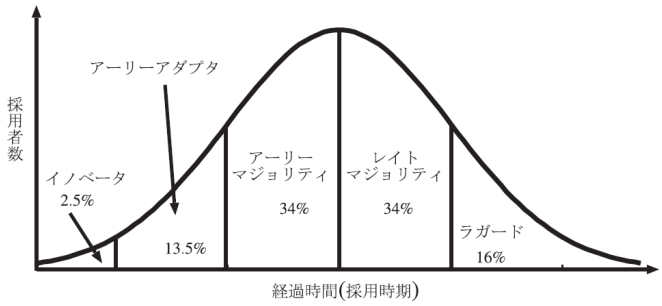

図 2: ロジャーズの普及モデルとユーザ総分類 [Rogers 10]

2.3 情報発信者の継続率

情報に対する態度をロジャーズの普及モデルに当てはめみる (図 2)．その場合，常に自ら情報を得ようと活動し，他の媒体からツイッター上に情報を流すユーザをイノベータであり，イノベータの流したツイートをいち早くリツイートし情報の拡散を行うユーザをアーリーアダプタと捉えることができる．

ロジャーズの普及モデルにおけるイノベータは最初にイノベーションを採用するユーザであるが，情報という観点から見ると，最初に情報をツイッター上に持ち込むユーザであると置き換えられる．ただし，意味の無い情報をツイートし続けるボットのような存在はイノベータとは言いがたいため，ここでは，ツイートした内容が常に多くのユーザにリツイートされているようなユーザが，重要情報を発見できるイノベータであると考える．

ここで，ある月にイノベータとして行動していたユーザが，次の月も継続してイノベータとして行動しているかどうかを確認する．もし，長期にわたって継続的にあるユーザがイノベータであれば，当該ユーザを補足することで，多くの情報を得ることが可能である．一方で，継続性がなければ，イノベータを監視することには余り意味が無い．そこで，大量にリツイートされたユーザが継続的にリツイートされるかどうかに注目する．

まず，イノベータを，一つのツイートが平均 100 回以上リツイートされたユーザであると定義する．各月に 100 回以上リツイートされたユーザの数と，前月から継続して平均 100 回以上リツイートされたユーザの数を図 3 に示す．これより，8 月に平均 100 回 RT されたユーザは 4317 人いるが，そのうち前月から継続して 100 回以上リツイートされているユーザは 637 人しかいない．これは，すなわち全体の 12.4% しか継続して大量にリツイートされることはないことを意味している．

この意味からも，大量被リツイートユーザを捉えることで情報収集を効率化することは適切ではないと考えられる．

2.4 情報拡散エージェントの発見

全節で見たとおり，積極的に自ら情報を拡散するイノベータ的ユーザは，それほど多く，継続性も高くない．

そのため，情報発信者よりも情報拡散者を見つける方がいち早く情報を捉えるには適していると考えられる．ロジャーズの普及モデルにおいても，アーリーアダプタ-が最も重要であると言われている．

ここで，アーリーアダプタは他のユーザがリツイートしたときに有用かどうかを判断し，他のユーザに広める役割を果たす存在であると捉えられる．そのため，アーリーアダプタとはバーストするようなリツイートを初期に行っているユーザであるといえよう．そこで，ここではあるユーザがリツイートした後に，大量のユーザがリツイートすることが多いようなユーザがアーリーアダプタ的なユーザであると定義する．

あるユーザがリツイートを行った後，同じツイートをリツイートしたユーザが平均 100 以上であるユーザについて，月ごとの継続率を求めた．その結果を図 4 に示す．これより，アーリーアダプタとしての役割を果たしているユーザの多くが継続的にその役目を果たしていることが分かる．したがって，これらのユーザを追うことで継続していち早く情報を獲得できる可能性がある．

しかしながら，これらのユーザが実際にアーリーアダプタとして他のユーザに情報を拡散させる能力を有しているかどうかは不明である．特に，ツイッターでは情報がどのような経路を通って拡散されたかが不明であるため，直接データから確認することが困難である．

そこで，次章ではシミュレーションによって，早い段階でリツイートを行うユーザが高い影響力を持ったアーリーアダプタであるかどうかを確認する．

3. 情報拡散シミュレーションによる重要ユーザの発見

3.1 シミュレーションの目的

ツイッター上のユーザが持つ真の拡散能力，すなわち影響力を実データから分析することは難しい．そこで，本章ではエージェントベースシミュレーションによって早い段階でリツイートを行うアーリーアダプタが実際に影響力を持っているかを確認する．

シミュレーションでは，まず伝播経路となる仮想的なネットワークを構築し，SIR モデル [Landau 53] に基づくリツイートをモデル化した情報伝播シミュレーションを行い，その結果に基づいてアーリーアダプタの影響力を明らかにする．

本シミュレーションでは，各エージェントは一定確率でツイッターに接続し，情報伝播ネットワーク上で接続しているエージェントから情報を取得するものとする．このとき，新しい情報を受け取った場合リツイートを行うかどうか選択する．このようにして一定期間リツイート行動を繰り返した場合に，情報がどのように拡散したかを確認し，そこからアーリーアダプタが影響力を持っているかを確認する．

3.2 エージェントの設計

本シミュレーションにおける，エージェントはツイッター上の 1 ユーザを表し，対象とする情報に対して，以下の 3 状態

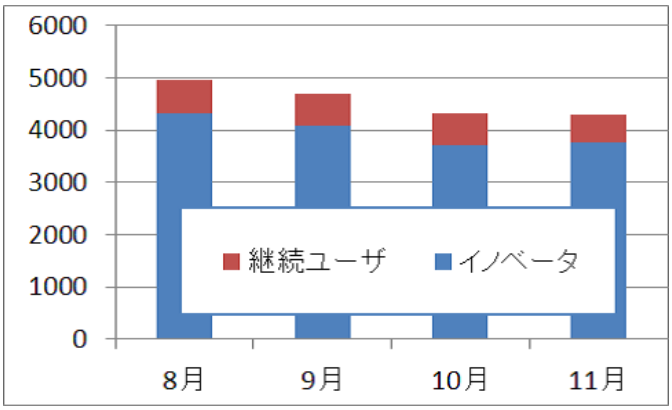

図 3: イノベータの継続率

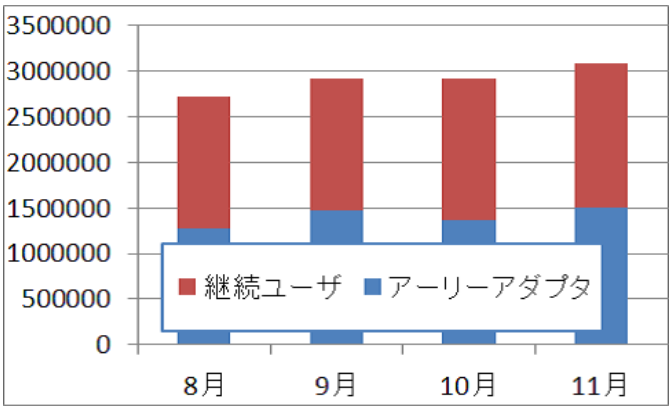

図 4: アーリーアダプタの継続率

を持つ.

1. 未接触状態 (S:Susceptible)
2. 伝播状態 (I:Information Transmitting)
3. 取得済状態 (R:Received)

ここで，情報未接触状態 (S) とはまだ情報を獲得していない状態であり，情報伝播状態 (I) はリツイートによって周囲に情報を伝播している状態である．また，情報取得済状態 (R) はすでに情報を受け取っているが，リツイートを行っていない状態である．

各エージェントは S 状態から開始され，隣接エージェントの状態が I になった場合一定確率で状態 I または状態 R となる．なお，初期状態として一体のエージェント (初期エージェント) がシミュレーション開始時点で状態 I となるものとする．

エージェント a_i はパラメータとして，

- 活動頻度 v_i
- 情報伝播頻度 r_i

の 2 つを持つ.

活動頻度 v_i は当該ステップに活動するかどうかを決定するパラメータであり，現実社会においてはツイッターの利用頻度に当たる．各エージェントは確率 v_i で活動を行う．

情報伝播頻度 r_i は，隣接エージェントが情報伝播状態 (I) だった場合に，エージェント a_i も情報伝播状態 (I) になる確率を示す．なお，エージェントは情報伝播状態 (I) にならない場合は取得済状態 (R) となる．

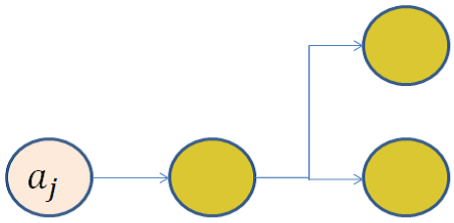

図 5: 情報伝播の例

3.3 シミュレーションの流れ

シミュレーションの手順は以下の通りである.

1. 情報を参照しあう関係をリンクとして，エージェント間にネットワークを構築する
2. 初期エージェント a_0 を決定し，エージェントの状態を情報伝播状態 (I) に変更する
3. すべてのエージェント $a_i(i = 1, \cdots, N-1)$ について以下の処理を行う
   (a) エージェント a_i が状態 (I) または (R) の場合，次のエージェントの処理に移る
   (b) 隣接エージェントに状態 (R) のエージェントがいない場合，次のエージェントの処理に移る
   (c) 確率 r_i でエージェント a_i の状態を (R) とし，そうでなければ状態を (I) にする.
4. 規定ステップに達するまで 3 を繰り返す

このようにして指定ステップが経過するまでシミュレーションを行い，リツイートが行われる様子を分析する．なお，本シミュレーションでは一つのネットワークにつき，すべてのノードが一回ずつ初期ノード a_0 となるようシミュレーションを行った．すなわち，一つのネットワークごとに，ノード数 N 回のシミュレーションが行われる．

3.4 真の影響力の定義

本シミュレーションでは，真の影響力を「当該エージェントを経由して情報を獲得したエージェントがどの程度いるか」と定義する．すなわち，情報の伝播をツリー構造と考えると，子孫ノードの数が当該エージェントの真の影響力となる．

図 5 のようにエージェント a_j から情報が広まっていったとすると，(直接・間接を含め) 情報を受け取ったエージェント数は 3 体存在することから，エージェント a_j の真の影響力は 5 となる．

本シミュレーションでは，複数回の情報伝播シミュレーションを行いそれらの合計を当該エージェントが持つ真の影響力とする．

3.5 シミュレーション結果

表 1 に示した条件でシミュレーションを行い，各指標と真の影響力との比較し，真の影響力と相関の高い指標を明らかにする．

具体的には，以下のような指標と比較を行う.

- 後発情報拡散数
  当該エージェントよりも時間的に遅れて情報拡散行動を行ったエージェント数

表 1: シミュレーション設定

Name	Value
Num of Agents	1000
Network Generate Model	CNN-Model
Simulation Step	1000
No. of Simulation	1000
First Agent Probability b_i	0-1(uniform distribution)
Active Frequency v_i	0-1(uniform distribution)
Retweet Probability r_i	0-1(uniform distribution)

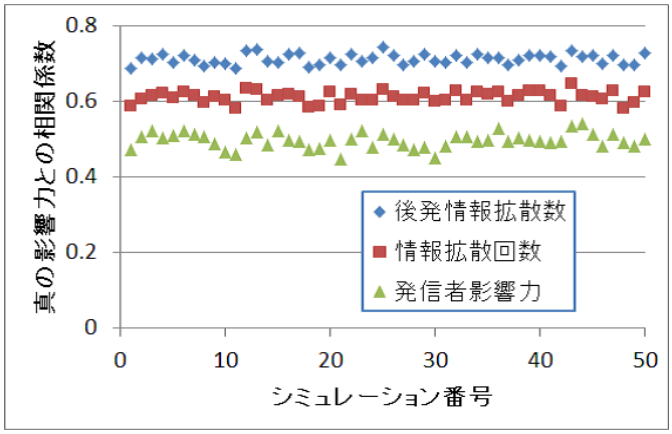

図 6: Correlatoin with True Diffusion Ability

- 情報拡散回数
  当該エージェントが情報拡散行動を行った回数
- 発信者影響力
  当該エージェントが発信者となった場合に，情報が伝わるエージェント数

シミュレーションは，異なる 50 のネットワークで試行した．なお，ネットワークの構築には大規模な SNS を表現するのに適している CNN モデル [Yuta 07][Vázquez 03] を用いた．それぞれの指標と，3.4 節で定義した真の影響力との相関を求めた結果を図 6 に示す．横軸はシミュレーション番号，縦軸は相関係数である．この図より，後発情報拡散数が最も真の影響力と相関が高いことが分かる．したがって，後発情報拡散数が大きいノードは高い影響力を持っているといえる．アーリーアダプタはその定義上後発情報拡散数が大きいノードであることから，アーリーアダプタは高い影響力を持っていることが示された．

以上より，他のユーザよりもリツイートを行うアーリーアダプタを捉えておくことは，高い影響力を持つユーザを捉えておくことに相当し，いち早く情報を獲得するために有効であることが示唆された．

4. 結言

ツイッター上の情報の伝播について，実際にどのような経路をたどって伝播したのかは分析できないため，イノベータ以外の真に影響力の高いユーザを把握することは難しい．しかしながら，イノベータは継続性が少ないため，アーリーアダプタを発見しておき，それらのユーザがリツイートを行った情報を把握することで素早く情報を獲得できるようになると期待される．

本研究では，アーリーアダプタは継続性が高く，またシミュレーションによってアーリーアダプタは実際に高い影響力を持っている可能性が高いことを示した．特に単にリツイートが多いユーザや，情報発信力が高いユーザよりも，情報拡散に寄与したユーザを推定できることは，震災時などでいち早く情報を獲得する上で有用であると考えられる．たとえば，デマのような不正確な情報が伝播しようとしたとき，影響力の高いユーザに先に注意喚起を行っておくことで，そのような情報の拡散を防ぐことが出来，また逆に重要な情報をそれらのユーザに優先的に知らせることで，より早い拡散が実現できるのではないかと期待される．

5. 謝辞

本研究は科研費 (24300064) の助成を受けて行われたものである．

参考文献

[Heverin 10] Heverin, T. and Zach, L.: Microblogging for Crisis Communication: Examination of Twitter Use in Response to a 2009 Violent Crisis in Seattle-Tacoma, Washington Area, in *Proceedings of the 7th International ISCRAM Conference* (2010)

[Landau 53] Landau, H. and Rapoport, A.: Contribution to the mathematical theory of contagion and spread of information: I. Spread through a thoroughly mixed population, *The bulletin of mathematical biophysics*, Vol. 15, pp. 173–183 (1953)

[Rogers 10] Rogers, E. M.: *Diffusion of innovations*, Simon and Schuster (2010)

[Vázquez 03] Vázquez, A.: Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations, *Physical Review E*, Vol. 67, No. 5, p. 56104 (2003)

[Vieweg 10] Vieweg, S.: Microblogged Contributions to the Emergency Arena: Discovery, Interpretation and Implications, in *Computer Supported Collaborative Work* (2010)

[Yuta 07] Yuta, K., Ono, N., and Fujiwara, Y.: A Gap in the Community-Size Distribution of a Large-Scale Social Networking Site, *Arxiv preprint physics/0701168* (2007)

[篠田 13] 篠田 孝祐, 榊 剛史, 鳥海 不二夫, 風間 一洋, 栗原 聡, 野田 五十樹, 松尾 豊: 東日本大震災時における Twitter の活用状況とコミュニケーション構造の分析, 知能と情報, Vol. 25, No. 1, pp. 598–608 (2013)