

移動エントロピーを用いた災害時における ソーシャル・メディア上の情報伝播分析

Information Spreading on Twitter Before and During the 2015 Torrential Rain in Eastern Japan

上子 優香 鳥海 不二夫 大橋 弘忠
Yuka Kamiko Fujio Toriumi Hirotada Ohashi

東京大学大学院工学系研究科システム創成学専攻
Department of Systems Innovation, School of Engineering, The University of Tokyo

Who transfers information during disaster? Assuming that users are inclined to share more trustworthy information under crisis, we can recognize trusted users by detecting users whose sharing behavior induces the sharing behavior of many others and who therefore contribute to information spreading during disaster. In this study, we use Twitter data and focus on the retweet behavior, the sharing behavior on Twitter. First, we propose the methodology, which adopted the approach of transfer entropy to track the flow of retweet. Second, we apply it to the data of the 2015 torrential rain in Eastern Japan and detect trusted users. Third, we analyze how much influence they had before the disaster. As a result, we clarified that some of widely or deeply trusted users, who highly contribute to information spreading during the disaster, do not contribute to information spreading before the disaster.

1. 序論

ソーシャルメディアが日常的に活用され、災害時におけるソーシャルメディアの有用性が明らかになると同時に、デマへの警戒が重要であると声高に叫ばれる昨今、情報の真偽を見極める力が、あらゆるユーザーに求められている。しかしながら情報の真偽を内容から瞬時に判断することは非常に難しく、災害時のような急を要する状況ではその困難さは殊更である。

では情報が信頼に値するか否かをユーザーが判断する基準は何か。社会学の分野では、山岸が安心を定義することによって信頼の概念を整理 [山岸 98] し、宮田がオンライン・コミュニティにおいては、日常生活環境以上に社会的不確実性が高い為に信頼が重要である [宮田 05] と説いている。

情報を初めに投稿したユーザー、いわゆる情報源が著名な人・企業・機関等であれば、それらに対して特定の信頼を持つ可能性も大いに考えられる。しかしながら多くの場合はそうとは限らず、情報を自身へ伝えた人への信頼が判断基準として大きな比重を占めるだろう。

本研究においては情報を自分へ伝えた人が信頼に値することが、ユーザーが伝えられた情報を新たに周囲と共有しようと行動する動機になる、つまり信頼される人が情報共有行動を起こすことが多くの他者の情報共有行動を誘発し、結果として広域に渡る情報伝播に繋がると仮定する。そして平成 27 年 9 月関東・東北豪雨を災害事象として取り上げ、災害前と災害時の Twitter における情報伝播への貢献度をユーザーごとと比較し、その関係性を分析する。

2. リツイートによる情報伝播経路の推測手法

本研究ではソーシャル・メディア Twitter のデータを用いて、災害時にあるユーザーによるリツイートを再度リツイートする行動が、そのユーザーへの信頼を反映すると仮定し、災害前の情報伝播経路の比較をする。これまでの災害時のリツイートによる情報拡散分析では、初めにツイートを作成したユーザーと

それをリツイートしたユーザーとの関係を扱うものが一般的であったが、本研究においてはユーザーがどのユーザーのリツイートを信頼し再びリツイートしたかを調査する為、情報伝播経路の推測を行うことが必要不可欠である。

よって本章では時系列データから情報伝播の大きさと向きを推測する移動エントロピーの考え方を用いた情報伝播経路の推測手法を新たに提案する。提案手法は以下の 3 段階を踏む。

1. 対象ユーザーの選定
2. ユーザー間のリツイート類似性の検証
3. 移動エントロピーによるユーザー間の情報伝播量の推定

2.1 対象ユーザーの選定

本手法においては事前に Twitter のフォロー・フォロワー関係やそれらを示唆するプライ情報等を一切用いない。従って、次段階において組み合わせ爆発が起こることを防ぐ為、リツイートの数に閾値を設けてユーザーを選定する。

2.2 ユーザー間のリツイート類似性の検証手法

本研究では信頼されるユーザーがリツイートした投稿を見た他ユーザーが、それを再度リツイートすることにより同一の情報が伝播していく経路を抽出することを目標とする。しかしながら短期間のデータのみを対象とするなどデータ数に限界がある場合においては、移動エントロピーで情報伝播量を計算するのみでは、コンテンツ内容を扱うことが困難である為、情報伝播が起こっていると判断されたユーザー同士でも、実は全く異なる情報をリツイートしており情報伝播が発生していないという可能性が考えられる。それ故予め各ユーザーがリツイートした内容の類似度の閾値を設定することによって、各ユーザーのリツイートした投稿の類似性を保証する。リツイートした投稿の類似性はシンプソン係数により判断する。

$$\text{シンプソン係数} : \frac{|A \cap B|}{\min(|A|, |B|)}$$

例えばあるユーザー A は非常に頻繁にリツイートを行い、ユーザー B は減多にリツイートを行わないとする。両者のリ

連絡先: 東京大学大学院工学系研究科システム創成学専攻
〒113-8656 東京都文京区本郷 7-3-1
Email: kamiko@crimson.q.t.u-tokyo.ac.jp

ツイート数の差は大きいものの、もしユーザー B のリツイートが全てユーザー A のリツイートに起因するものであれば、ユーザー A のユーザー B に対する信頼は高いと考えられる。このようなペアが排除されることを防ぐ為、リツイート数が少ないユーザーのリツイート数を分母にとるシンプソン係数を採用する。

2.3 移動エントロピーによるユーザー間の情報伝播量の推定

移動エントロピーとは離散変数で表現された 2 つの時系列データから、相互間の情報伝播を定量化する目的で Schreiber [Schreiber 00] らが提案した考え方である。

ある事象 x のもつ情報量 $I(x) = -\log P(x)$ に対して起こりうる全ての事象の情報量の和、すなわちエントロピー $H = -\sum_x p(x)I(x)$ が定義される。ここで条件付きのエントロピー $H(A|B)$ を式 (1) と表す。

$$H(A|B) = -\sum_{A,B} P(A,B) \log P(A,B) \quad (1)$$

移動エントロピーは 2 つの時系列データ X, Y 間で計算され、 Y の $t=n+1$ における振る舞い y_{n+1} を Y の過去データ $Y_n^{n-k} (= \{y_n, y_{n-1}, \dots, y_{n-k}\})$ のみから予測した場合と Y の過去データ及び X の過去データ $X_n^{n-l} (= \{x_n, x_{n-1}, \dots, x_{n-l}\})$ から予測した場合とを比較し、後者の場合にどれだけ不確実性が減少したかを計測する。式 (1) を用いると、 X から Y への情報伝播量 $T_{X \rightarrow Y}$ は式 (2) と定義でき、これを計算することにより情報伝播の方向性、すなわち因果関係を定量的に測定することが可能である。

$$T_{X \rightarrow Y} = H(y_{n+1}|Y_n^{n-k}) - H(y_{n+1}|Y_n^{n-k}, X_n^{n-l}) \quad (2)$$

移動エントロピーを用いたソーシャル・メディアの分析としては Ver Steeg らが行った Twitter における情報伝播分析 [Ver Steeg 12] があり、その研究においては Twitter 上における URL 付きのツイートを対象とし、直前期における振る舞いが与える影響を強く反映させる為に、過去データの時間区切りを不均一にするという工夫が為されている。

本研究においては、リツイートデータのみを対象として分析を行い、該当時間 $t = [i + \delta_0, i]$ にリツイート行動を起こしたか否かを振る舞いとして $y[i + \delta_0, i] = \{0, 1\}$ とバイナリで定義する。過去データとしては $Y_i^{i-\delta_3} (= \{y[i, i - \delta_1], y[i - \delta_1, i - \delta_2], y[i - \delta_2, i - \delta_3]\})$ および $X_i^{i-\delta_3} (= \{x[i, i - \delta_1], x[i - \delta_1, i - \delta_2], x[i - \delta_2, i - \delta_3]\})$ を用意し、直前の振る舞いに対して重み付けを行うことを目的として、 $\delta_0 = 1min, \delta_1 = 10min, \delta_2 = 1hour, \delta_3 = 3hours$ と設定する。

3. 災害リツイートデータに対する推測手法の適用

本章では扱う災害事象および災害データ、提案手法適用時の各種設定について述べる。

3.1 災害リツイートデータ概要

2015 年 9 月 7 日から 9 月 11 日にかけて関東地方及び東北地方で発生した豪雨災害は「平成 27 年 9 月関東・東北豪雨」と命名され、甚大災害法に基づく甚大災害指定を受けている。特に茨城県・栃木県・宮城県における被害が大きく、鬼怒川の氾濫による周囲への被害は多くの人に衝撃を与えた。

本研究では災害事象として本災害を取り上げ、2015 年 8 月から 9 月にかけて収集した Twitter の日本語リツイートデータ約 365,000,000 件を用いて、提案手法により情報伝播経路を推測する。そして推測された経路より災害前と災害時の情報の流れを比較する。収集データ数は図 1 の通りであり、災害関連語 (表 1) を含むリツイート数が急増し落ちつきを見せる、9 月 8 日から 9 月 12 日までを災害時データと定め、それ以前の 10 日間のデータを表 2 の通り 5 日間ごとに災害前データとして 2 セット用意し災害前と災害時の情報伝播比較を行う。ただし災害時期間においても災害と無関係なデータも多数含まれている為、災害時データにおいては災害関連語を含むデータのみを対象とした。

災害関連語 (38 語)
注意報, 警報, 災害, 減災, 被災, 防災, 避難, ハザードマップ, 大雨, 豪雨, ゲリラ豪雨, 集中豪雨, 台風, 寒冷前線, 洪水, 浸水, 大雪, 雪崩, 突風, 竜巻, 停電, 火山, 噴火, 噴石, 溶岩, 火砕流, 地震, 断層, 活断層, マグニチュード, 揺れ, 震度, 震源, 震災, 余震, 津波, 帰宅困難, 仮設住宅

表 1: 災害関連語一覧

	データ期間	備考
災害前データ	08/29-09/02	災害無関係データ含む
	09/03-09/07	災害無関係データ含む
災害時データ	09/08-09/12	災害関連語を含むデータのみ

表 2: データセット概要

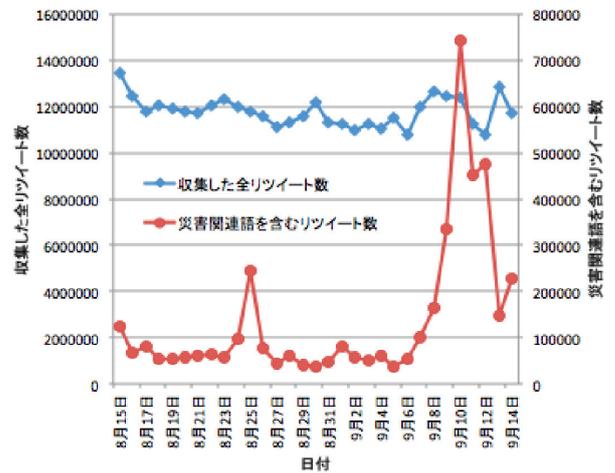


図 1: 収集データ数

3.2 情報伝播経路特定に伴う各種設定

対象ユーザーは 9 月 3 日から 9 月 7 日の期間に 25 回以上 1000 回未満リツイートを行ったユーザーと設定した。1000 回以上リツイートを行ったユーザーを排除した理由は短期間に非常に多い回数のリツイートを行っているユーザーは bot である可能性が高いと考えるからである。この条件を満たした全ユーザー

ザーの組み合わせに対して、2種類の災害前データおよび災害時データに提案手法を適用し、情報伝播経路の推測を行った。

途中リツイート類似性が0.2を下回ったペアに関しては情報が伝播した可能性が低いとして、この閾値を下回ったペアは移動エントロピーの計算対象から排除した。災害時データにおいては、災害関連語を含んだデータのみではデータ数の少なさ故、相互間影響が存在する可能性のあるペアの抽出が十分ではないと判断した。その為災害関連語を含まないデータも対象として類似性が閾値を超えるかを計測し予めペアを絞った上で、再度災害関連語を含むデータのみで類似性、移動エントロピーの計算を行い情報伝播経路推測を行った。

4. 災害前と災害時における情報伝播者の比較

4.1 災害時における情報伝播経路

提案手法により推測された災害時における情報伝播経路を有向グラフを用いて可視化した結果(図2)及び考察をまず述べる。グラフの可視化およびコミュニティ抽出にはオープンソースソフトウェア Gephi*¹ を使用し、ノードの色は所属コミュニティを表現している。コミュニティ内でのみ強い繋がりをもった集合はスパムアカウントの集合であり、これは Ver Steeg らの研究結果 [Ver Steeg 12] を支持するものである。また複数の一連のスパムアカウントを一斉に抽出できるという点で本手法が有効であると言える。

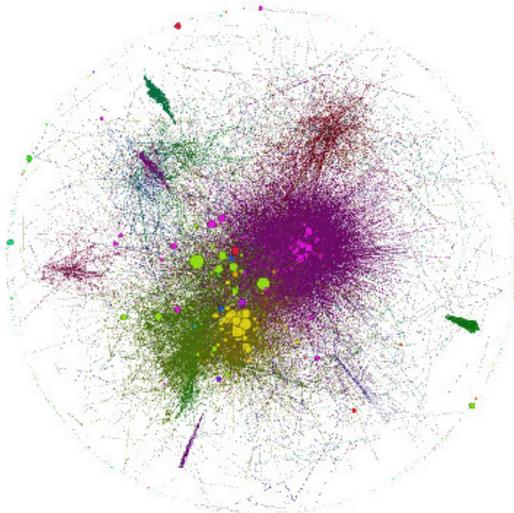


図 2: 災害時における情報伝播経路グラフ

4.2 情報伝播に対する貢献度の測定尺度

情報伝播者と一口に言っても、一人ひとりへの情報伝播量はさほど多くはないものの多数の人へ情報を伝播している人も居れば、情報を伝播させている人数はさほど多くはないものの、一人ひとりへの情報伝播量が多い人も居る。情報伝播者として前者は幅広く信頼され広域へ情報を拡散させるという観点から重要で、後者は深く信頼され高確率で他のユーザーへ情報を伝播させるという観点から重要である。前者は多数の人への情報伝播が行われている為、情報伝播量の総和が大きく、後者は情報伝播量が多いつながりがある為、情報伝播量が多い外向きエッジ数が多いと考えられる。

*1 <https://gephi.org>

よって本研究においてはユーザーの情報伝播への貢献度を測る指標として情報伝播量の総和、情報伝播量が上位5%以上の外向きエッジ数、の2つを用意し、二節に分けてそれぞれ分析を行った結果および考察を記す。

4.3 情報伝播量の総和による比較

本節では各ユーザーの情報伝播への貢献度を情報伝播量の総和、すなわち外向きエッジの重み和と定義し、災害時に情報伝播に貢献したユーザーが災害前においても情報伝播に貢献していたかを、災害前と災害時データの計算結果から比較する。図3は災害時に情報伝播を行ったユーザーを対象として、各データセットにおいて計算された貢献度を両対数グラフにプロットしたものである。災害時の情報伝播貢献度が低いユーザーは青色で、中程度のユーザーは緑色で、高いユーザーはオレンジ色でプロットされているが、左下から右上にかけてプロットの色が青からオレンジへと移り変わっていく傾向から、災害時に情報伝播量の総和が大きいユーザーほど災害前から情報伝播量の総和が大きかった可能性が高いと示唆される。しかしながら、両対数グラフである為、災害前のいずれかのデータセットにおいて情報伝播量の総和が0であったものは描画されていない。

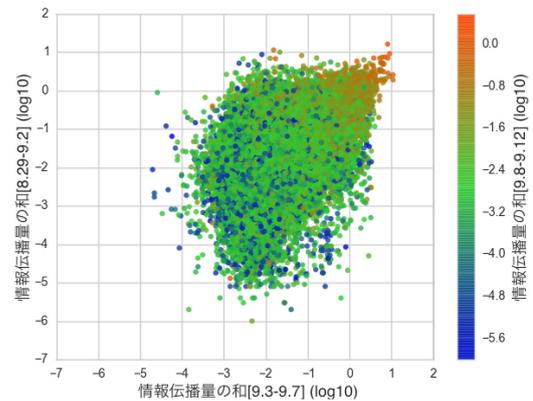


図 3: 災害前と災害時における各ユーザーの情報伝播量の総和

そこで災害時の情報伝播量の総和とユーザー数との関係をヒストグラムで表し、各階級ごとにユーザーを災害前の2つのデータセットにおいて、どちらにおいても情報伝播に貢献していた (Neither Zero)、どちらか一方のみでは情報伝播に貢献していた (Either Zero)、全く情報伝播に貢献していない (Both Zero) の3つのグループに分け、各グループの割合をプロットした(図4)。ここで特筆すべきは、災害時の情報伝播への貢献度が高くなるにつれて青線で表された Neither Zero の割合が増えるものの、災害時の情報伝播への貢献度が非常に高いユーザーになると緑線の Either Zero および赤線の Both Zero の割合が再度高まることである。災害時の情報伝播貢献度が非常に高いユーザーとなるとユーザー数は限られるものの、図4のヒストグラムの度数は災害時の情報伝播貢献度が-1を超える範囲で低い方から順に [934,278,100,65,20] であり、災害時の貢献度が高いユーザーにおいて Either Zero および Both Zero の割合が高まるという結果は、無視できないであろう。

4.4 情報伝播量上位5%以上のエッジ数による比較

本節では各ユーザーの情報伝播への貢献度を各データセットにおける情報伝播量上位5%以内の外向きエッジ数と定義し、災害前と災害時データにおけるユーザーごとの貢献度を比較

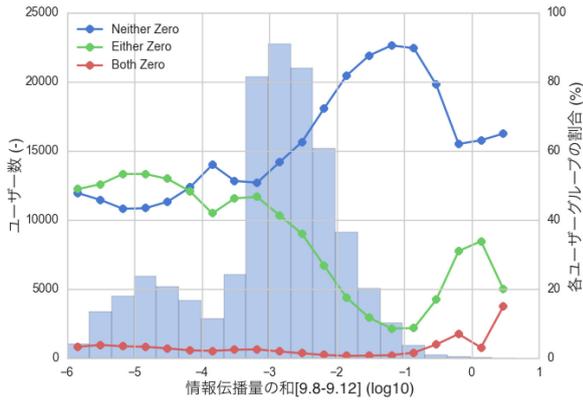


図 4: 災害時の情報伝播貢献度毎の災害前に貢献がないユーザーの割合 (情報量の総和)

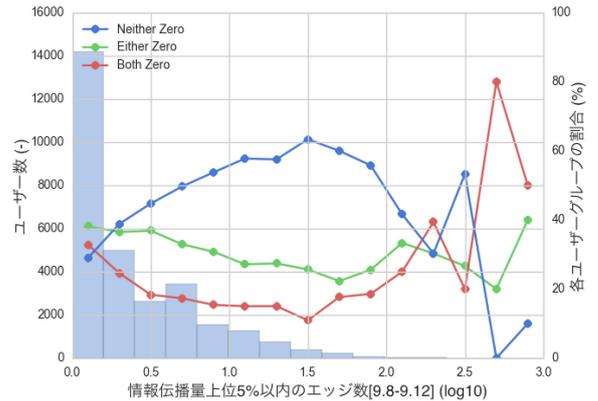


図 6: 災害時の情報伝播貢献度毎の災害前に貢献がないユーザーの割合 (情報伝播量上位 5% 以内のエッジ数)

する。図 5 は災害時に情報伝播量上位 5% 以内の外向きエッジが 1 つ以上あったユーザーを対象として、各データセットにおけるエッジ数を両対数グラフに表したものである。確かに 8 月 29 日から 9 月 2 日までの情報伝播への貢献度が高い範囲において淡いオレンジ色のプロットがいくつかみられるが、災害時の貢献度が非常に高い濃いオレンジ色のプロットは災害前の貢献度が高い範囲には見られない。

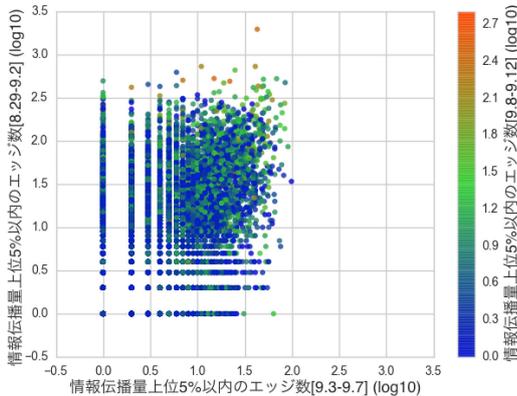


図 5: 災害前と災害時における各ユーザーの上位 5% 以内のエッジ数

以上の事実を踏まえ、前節同様災害時の情報伝播への貢献度ごとに、災害前の貢献度が 2 種類のデータセットどちらも 0 ではない (Neither Zero)、どちらか一方で 0 である (Either Zero)、どちらも 0 である (Both Zero) 割合を調べた結果が図 6 である。情報伝播への貢献度が 2.0 を超える階級でヒストグラムの度数が低い順に [86, 69, 33, 15, 5, 10] であり貢献度が 2.6 を超える階級ではユーザー数が限られているが、災害時に非常に貢献度が高いユーザーになるにつれ Neither Zero の割合が小さくなり、Both Zero の割合が高くなることは十分に推察され、これは前節と同様の傾向である。すなわち災害時に特に多くの貢献をした情報伝播者であるほど、災害前には情報伝播をしていなかった人の割合が増えており、災害時には災害前とは異なる情報の流れが生まれている。

5. 結論

本研究では、災害時において多くの人から信頼される人によるリツイート行動が、他ユーザーの新たなリツイート行動を誘発し、結果として広域にわたる情報伝播に繋がると仮定し、移動エントロピーを用いたリツイートの伝播経路推測手法を提案した。さらに提案手法を用いて平成 27 年 9 月関東・東北豪雨を含む期間のデータを分析し、情報伝播への貢献度を情報伝播量の和および情報伝播量が上位 5% 以内の外向きエッジ数の 2 種類で測ることで、災害時に情報伝播に貢献したユーザーが災害前にも同様の特徴を示していたかを比較し考察した。

結果として災害時の情報伝播への貢献度が比較的高いユーザーでは、災害前から情報伝播への貢献度が高いユーザーの割合が大きいと見えるが、災害時の貢献度が非常に高いユーザーに限ると、逆に災害前に情報伝播への貢献度が低いユーザーの割合が増えるということが判明した。これは仮定を用いて言い換えれば、災害時にのみ突如として情報伝播に貢献するようになる信頼された人が存在し、災害時には通常時とは異なる情報の流れが発生するということである。それらのユーザーの持つ特徴について調査することは今後の課題である。

6. 謝辞

本研究は日本学術振興会 課題設定による先導的人文・社会科学推進事業「リスク社会におけるメディアの発達と公共性の構造転換」プロジェクトの一部として行われた。

参考文献

- [山岸 98] 山岸俊男. 信頼の構造: こころと社会の進化ゲーム. 東京大学出版会, 1998.
- [宮田 05] 宮田加久子. きずなをつなぐメディア: ネット時代の社会関係資本. NTT 出版, 2005.
- [Schreiber 00] Schreiber, Thomas. "Measuring information transfer." *Physical review letters* 85.2 (2000): 461.
- [Ver Steeg 12] Ver Steeg, Greg, and Aram Galstyan. "Information transfer in social media." *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012.