Vol.2019-MPS-122 No.8
2019/2/28

ツイート文中の語句に基づいたデマ状態推定モデルの提案

牛込 龍太郎[1,a)]　松田 健[2]　園田 道夫[3]　趙 晋輝[1]

概要：SNS(Social Networking Service) は誰でも情報を容易に拡散できる利点があるが，その反面誤った事実や虚偽の情報といったいわゆるデマも同時に拡散されやすい．SNS 上のデマはそのコミュニティの外にも影響を及ぼすことがあるため，近年社会問題としても認知されつつあるものの，未だにデマの発生は続いている．本研究の目的は，デマが発生するタイミングや，拡散・収束する様子を調査することで，デマの発生や拡散を抑制するための知見の獲得に繋げることである．本稿ではツイート文に含まれる語句に着目し，投稿時刻の情報と合わせた上でツイート文から推定できるデマの話題の状態を推定するモデルを提案した．

キーワード：SNS，デマ，ベイズ推定

Proposal of Model Inferring Diffusion Status of Hoaxes Based on Phrases in Tweets

Abstract: Information can be spread easily in SNS (Social) community. This also means that hoaxes like disinformation or misinformation are spread easily. Hoaxes become famous as a social problem because their information can affect not only SNS users but also public. However, they are generated continuously. Our purpose is to obtain the knowledge on reduce generating or spreading hoaxes through investigating when hoaxes are generated and how hoaxes become popular or unpopular. In this paper, we focused on phrases in tweets and proposed a model which infers diffusion status of hoaxes combining phrases with posting timestamps.

Keywords: SNS, hoax, Bayes inference

1. 序章

全世界での利用者数が 20 億人を超える [1] とされる SNS (Social Networking Service) について，近年デマやフェイクニュースといった誤りを含む情報が拡散され市民生活に混乱をきたすケースが後を絶たず，社会的な問題として認知されつつあるとともに対策がはじまっている?[2][3]．こうした社会的背景から，研究の分野においてもデマやフェイクニュースの拡散を防止することを目的とするものが活発になされており，それらの大半は SNS 投稿の文章表現やユーザアカウントが持つ統計的な情報を利用し投稿内容やアカウントの信頼性の判定することに主眼を置いている [6][12]．一方 SNS を対象とする研究には情報が SNS のコミュニティの内部で拡散される様子を，アカウントや投稿，システムに結びつけられた統計的な情報からモデリング等の手法を用いて見出すもの [5][9][10] が存在する．

そこで本研究ではベイズ推定の手法を応用し，複数の SNS 投稿の文章に含まれる語句の属性からデマが発生する兆候や拡散されている状態が表現できるかどうかについて，実データを用いた検証を通じて検討を行った．なお本研究におけるデマの定義は，「実際に発生した事象や存在する事物と矛盾する主張」とする．検証のための実データには，SNS のひとつである Twitter[4] を利用した．その結果デマの話題に言及するツイートデータから作成したモデルとデマではない通常の話題に言及するツイートデータから作成したモデルから得られたベイズ予測分布に違いがあることが確認された．

[1] 中央大学
Chuo University, Bunkyo, Tokyo 112–8551, Japan
[2] 長崎県立大学
University of Nagasaki, Nishi-Sonogi, Nagasaki 851–2195, Japan
[3] 情報通信研究機構
National Institute of Information and Communications Technology, Koganei, Tokyo 184–8795, Japan
[a)] a13.hpbb@g.chuo-u.ac.jp

2. 関連研究

SNS へ投稿されるデマやフェイクニュースの検知を試みる研究は数多く行われている．Krishman ら [12] はフェイクニュースを検知するフレームワークを提案しており，ユーザのフォロー数/フォロワー数やツイート中の文字列の記号の割合などの情報に加えて，ツイートに添付されているメディアファイルの真偽性もフェイクニュースの判定を行う特徴量として用いている．

Buntain ら [6] は，fact-check がなされた複数のデータセットに対して，異なる組み合わせの特徴量を用いて，フェイクニュースか否かについてクラス分類を実施している．そのため各データセットに対して最も有効な特徴量の組み合わせが異なり，汎用性に課題が残る．

Kumar ら [8] は，認知心理学の観点から誤って拡散された誤情報 (misinformation) の要素について考察し，誤情報の拡散の検知を試みている．考察によって，情報の信頼性は情報の出自が一つの重要な要素であることと，またユーザ間のリツイートの流れが情報伝播の把握に役立つことが知見として得られ，それらを基にリツイートの流れと情報の一貫性に着目した検知手法を提案している．

Campan ら [7] はフェイクニュースがソーシャルネットワークの中でどのように拡散されているのかについて，複数の論文の知見をまとめている．Campan らはソーシャルネットワーク内での情報の拡散には，拡散方法，ユーザに人気の話題，他のより多くのユーザへ情報を広められるユーザを把握することが重要であると述べている．その中で情報の拡散方法については SIR モデルや SIS モデルといった感染症のモデルが広く応用されている，とされている．

また SNS において情報が拡散されていく様子のモデリングを試みた研究も数多く存在する．Jin ら [5] はマイクロブログ上で話題になっているニュースの信頼性がどのように構成されているのかを示す Hierarchical Credibility Network を提案している．ネットワークは，ニュースの話題と，話題に言及しているマイクロブログへの投稿，投稿が言及しているニュースの部分的要素から構成され，構築したネットワークをグラフ最適化問題として定式化することでニュースの信頼性を形成する要素を数値化している．

三浦ら [9] は自然災害のような緊急事態において人間の情報共有行動の特徴を捉えることを目的として，自然災害発生時における実際のツイートから感情語を抽出し，その出現傾向と災害の種類の関連を分析している．その結果，ネガティブな感情語や活性度の高い感情語が多く含まれるほどユーザらに対して高い伝播性を示していることや，災害の種類に関係なく「不安」を示すものが伝播されやすいことを明らかにしている．

池田ら [10][11] は AIDM と呼ばれる情報拡散モデルを提案している．AIDM には Twitter ユーザの多様性やネットワークの持つ特徴，情報拡散メカニズムを盛り込み，マルチバースト型のデマの拡散をシミュレーションを通じて再現するとともに，拡散を抑制する手法の検証を行っている．

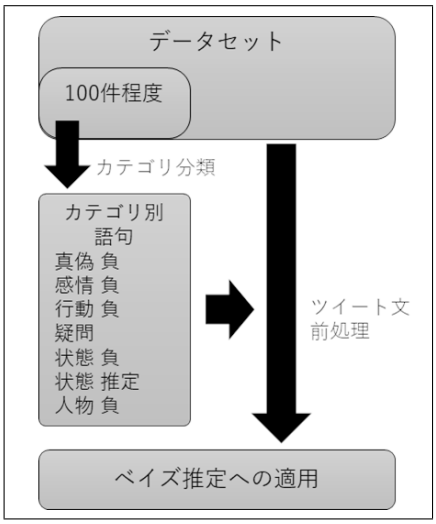


図 1: 提案手法の流れ
Fig. 1 Procedure of the proposed method

3. 提案手法

本研究において「デマ」とは，「実際に発生した事象や存在する事物と矛盾する主張」と定義し，そのため事件や事故、自然災害といった実際に発生したかどうかの確認が容易に行える話題を本研究で提案する手法によるモデル化の対象としている．デマはその内容に騙される人間が存在することによっていわば「生きている」状態にあると解釈することができ，全ての人間が虚偽であることに気づいているようなデマはデマとしての価値がほぼ皆無であるとみなせる．すなわち，SNS においてはデマに対してユーザが信頼していると考えられる表現をしている時，デマは「生きて」おり，拡散の途上にあるとみなすことができる．そこで本研究の提案手法では，ツイート文中の表現から話題に対してユーザが抱いている疑念の程度を数値化し，それらの数値を二項分布をベースとしたモデルへ適用することでベイズ推定への適用を行う．以降の節で提案手法の流れと，今回の検証で使用したデータセットの詳細について述べる．**図 1** に提案手法の流れの概要を示す．

3.1 語句のカテゴリ分類

3.3 節で述べるツイートデータセットから 100 件程度のツイートを選び，語句の持つ性質や印象の観点で，それぞれのツイート文に含まれる語句をカテゴリに手作業で分類する．分類先として設定するカテゴリは話題に対してユーザが疑念を持っている状態を示すものである．設定したカテゴリとそれらに分類した語句の例を**表 1** に示す．

表 1: 語句カテゴリと各カテゴリに属する語句の例

カテゴリ	語句例
真偽 負	架空，偽，騙され
感情 負	かわいそう，おかしい，怖い
行動 負	謝罪，悪趣味
疑問	かな，なぜ
状態 負	酷い，悪質
状態 推定	らしい，はず
人物 負	DQN

3.2 予測モデルの生成

はじめに各ツイート文の文字列が 3.1 節で設定したそれぞれのカテゴリの語句を含むか否かを 0/1 値で表し，その総和をツイート文が持つカテゴリ数として算出を行う．文字列データであるツイート文を $S_i(i = 1, 2, ...)$ とおき，S_i がカテゴリ $C_j(j = 1, 2, ..., n)$ に属する語句を含む場合を $c_{i,j} = 1$，含まない場合を $c_{i,j} = 0$ とする．なお，表 1 より今回の検証において $n = 7$ である．このときそれぞれのツイート文 S_i に

$$(c_{i,1}, c_{i,2}, ..., c_{i,n})$$

が作成され，S_i のカテゴリ数を x_i は

$$x_i = \sum_{k=1}^{n} c_{i,k}$$

で求まる．またツイート文へのカテゴリ数の算出と並行して，ツイートデータセットを一定の時間間隔ごとに分割する．今回の検証においては 5 分間隔でデータを分割した．続いてベイズ推定へ適用するため，二項分布を基にしたモデルを導入する．

$$f(x_i|p) = {}_nC_{x_i} p^x (1-p)^{n-x_i} \quad (1)$$

(1) 式において p はユーザが話題に対して懐疑的である確率を表す．生成したモデルを分割した時間帯ごとに積をとり，ここに事前分布としてベータ分布を導入すると事後分布がベータ分布となる性質を利用することで，ベイズ予測分布を計算しデマ拡散モデルとして利用する．

3.3 データセット

モデルの検証に用いたツイートデータはデマの話題に言及するツイート (以下，デマツイート) とそれ以外のツイート（以下，通常ツイート）の 2 種類である．全てのツイートデータには，ツイート本文のほかに投稿時刻の情報が付されている．**表 2** にツイートデータの詳細を記す．デマツイートの話題は，日本国内のある飲食店を名乗るアカウントが近隣の大学の職員団体からの貸し切りの予約を当日に無断でキャンセルされたと主張するものである．しかし実際は飲食店も職員らが所属する大学も架空のものであった．デマツイートデータは，話題の元となるツイートの投稿時刻から 1 時間以内に投稿されたツイートとそのツイートに対するリプライのツイートを Twitter の Web ページの検索機能を用いて収集した．

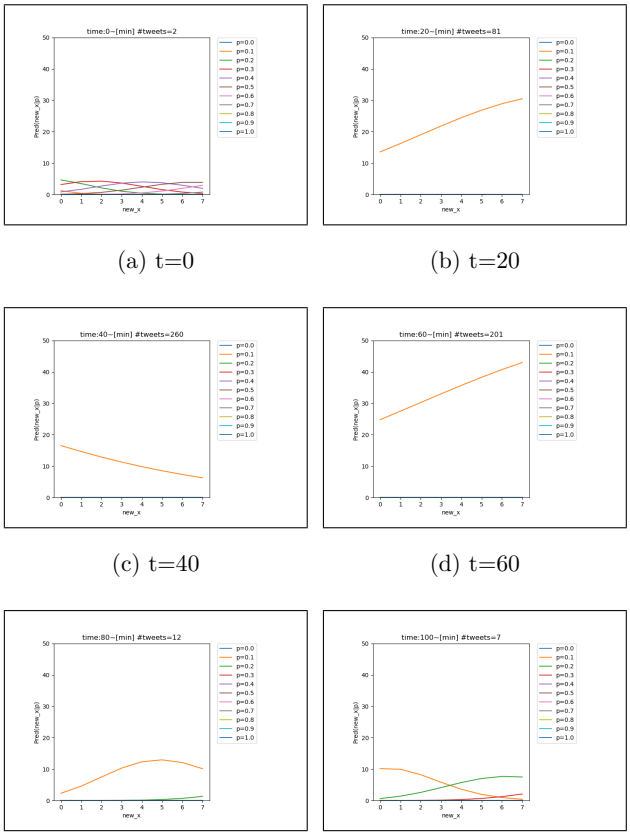
表 2: ツイートデータの詳細

属性	データ名	期間	件数	検索キーワード
デマ	F1	2018/5/13〜16	2149	–
通常	N1	2017/6/1〜2	3776	ゲリラ豪雨
通常	N2	2018/2/7〜8	4398	大阪 駅

通常ツイートデータは表 2 の「検索キーワード」列の各行に記されているキーワードを Twitter API によって検索し収集した．データセット N2 の検索に用いた語句は，当時発生していた通り魔事件の発生場所に関連するものである．事件発生直後のツイートの大半は「通り魔」という表現を含んでおらず，事件に言及するツイートを効率よく収集するには場所の名前を用いるのが適切である判断したため，検索にこれらの語句を使用した．また N2 が話題としている通り魔事件は，データセットに含まれる最も古いツイートが投稿されてから約 155 分後に発生している．

4. 結果

データセット F1，N1，N2 に対して提案手法を適用し，

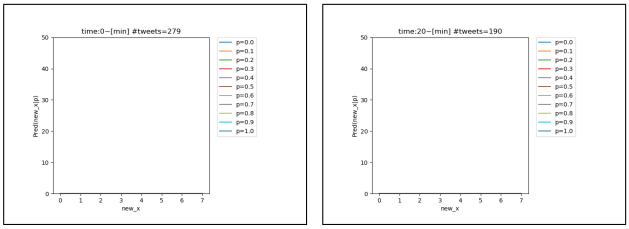


(a) t=0　(b) t=20

(c) t=40　(d) t=60

(e) t=80　(f) t=100

図 2: データセット F1 の予測分布の一部
Fig. 2 Partial prediction distributions of F1 dataset

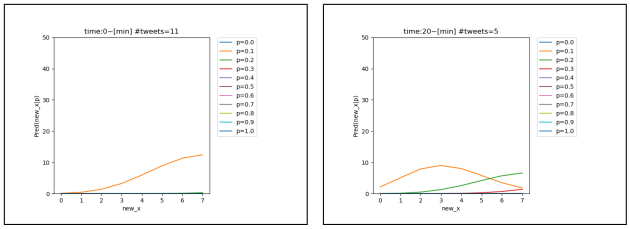


(a) t=0　(b) t=20

(c) t=100　(d) t=120

(e) t=140　(f) t=160

図 3: データセット N1 の予測分布の一部
Fig. 3 Partial prediction distributions of N1 dataset



(a) t=0　(b) t=20

(c) t=40　(d) t=155

(e) t=175　(f) t=195

図 4: データセット N2 の予測分布の一部
Fig. 4 Partial prediction distributions of N2 dataset

作成した予測分布の一部を図 **2**, 3, 4 にそれぞれ示す．各グラフの横軸の new_x は新たなデータが含む語句のカテゴリ数を示し，縦軸は予測分布の確率密度 $Pred(new\_x|p)$ を表す．また各グラフにおいて，話題に対して懐疑的な確率の値 p は凡例に，データセット内の最初のツイートの投稿時刻からの経過時間及び経過時間内に投稿されたツイートの数をタイトルにそれぞれ記載した．

F1 から生成されたグラフ同士を比較すると，$Pred(new\_x|p = 0.1)$ の頂点が**図 2c** を除いて時間の経過につれて右方向，すなわちカテゴリ数の多い方向へスライドしているのが確認できる．またデータセット N1 中の古いデータの大半は語句のカテゴリ数が 0 であったため，**図 3a**, 3b のように予測分布の確率密度の値が全ての p に対して 0 になっている．データセット N2 から生成した予測分布は，夜遅い時間帯に投稿されたデータであったため時間帯ごとのツイート数が少なく，$Pred(new\_x|p)$ の値は 0 より大きくとるものが見られた．

デマツイートの予測分布を示している図 2 と通常ツイートの予測分布を示している**図 3** 及び**図 4** を比較すると，デマの話題において一定の時間間隔内に約 50 件以上のツイートが存在する際に $Pred(new\_x|p)$ の値が大きいものが出現しているのに対して，通常の話題では全ての p において $Pred(new\_x|p) = 0$ となった．

5. 考察

デマの話題に対しては情報が誤りであることを指摘する投稿が増えるにつれて話題に対してネガティブな表現が増加するため，図 2 のような予測分布の時系列的な変化が発生したと考えられる．図 2c においてグラフの頂点が一旦左方向へスライドした要因としては，今回使用したデマの話題に「架空の大学」と「架空の飲食店」という 2 つの虚偽の情報が含まれており，一方の偽の情報に気づいたのちもう一方の偽情報に気づいたユーザが複数存在したため，と推測される．

デマの話題と通常の話題の予測分布の形状の変化についてみると，通常の話題の予測分布は投稿数が 50 件程度より多い場合は $Pred(new\_x|p)$ の値がほぼ 0 になり，投稿数が少ない場合においては 0 より大きい値をとるものが現れる．しかしデマの話題に対して否定的な投稿が盛んになされていた時間帯の予測分布を示す**図 2b**, 2d と比較すると，$Pred(new\_x|p)$ の値は小さいことから，この分布の形状の違いをデマが拡散されている状態の検知へ生かすことが期待できる．

またゲリラ豪雨の話題であるデータセット N1 の結果に

注目すると，自然災害の話題のデータに対して分析を行った三浦ら [9] の知見と異なる部分があり，この要因は本研究では感情語を一括して一つのカテゴリで取り扱ったためと，ゲリラ豪雨に対してユーザが驚きや感嘆を表現するものの不安を表す語句をさほど使っていなかったためと考えられる．

なお，本研究では比較対象に適した既存手法を発見することができなかったため，提案手法を実装したシステムを下記の URL にて公開し，今後システムから得られるフィードバックを基に手法の改良に努めていく．

https://script.google.com/macros/s/AKfycbzn_ItDd2ypdwBiZVbhIvSaa8rBl9KcFjV5gF2RpFhf_e24Qb4/exec

6. まとめ

本研究では話題に騙されているユーザが存在することでデマが「生きている」状態にあるという仮定を設定し，SNS ユーザが話題に懐疑的な度合いを投稿中に含まれる語句の属性からモデル化を行いベイズ予測分布を作成するとともに，実データを用いてモデルの精度の検証を行った．検証の結果，投稿数が少ない時間帯に投稿されたツイートから作成した予測分布同士では，デマツイートから生成されたものと通常ツイートから生成されたものとの間に，確率密度関数の最大値に違いがみられ，この値の大小の違いを生かすことで SNS コミュニティ内でデマが拡散されている状態にあるかどうかを判別できる可能性を見出した．
今後の課題としては，語句のカテゴリ分類を複数人で行う，あるいは機械的な分類手法を導入することで分類の客観性を高めることや，カテゴリに属する語句を含むものの話題との関連性は低いノイズを含むツイートデータへの対応が挙げられる．

参考文献

[1] statista: 入手先 ⟨Global social media ranking 2018 Statistic⟩, （参照 2019-01-29）
[2] WIRED: *In a Fake Fact Era, Schools Teach the ABCs of News Literacy*, 入手先 ⟨https://www.wired.com/2017/06/fake-fact-era-schools-teach-abcs-news-literacy/⟩ （参照 2019-01-29）
[3] Facebook Newsroom: *Authenticity Matters The IRA Has No Place on Facebook*, 入手先 ⟨https://newsroom.fb.com/news/2018/04/authenticity-matters/⟩ （参照 2019-01-29）
[4] Twtter，入手先 ⟨https://twitter.com/⟩（参照 2018-12-17）
[5] Zhiwei Jin, Juan Cao, Yu-Gang Jiang, Yongdong Zhang: *News Credibility Evaluation on Microblog with a Hierarchical Propagation Model*, IEEE International Conference on Data Mining (2014)
[6] Cody Buntain, Jennifer Golbeck: *Automatically Identifying Fake News in Popular Twitter Threads*, IEEE International Conference on Smart Cloud (2017).
[7] Alina Campan, Alfredo Cuzzocrea, Traian Marius Truta: *Fighting Fake News Spread in Online Social Networks: Actual Treds and Future Research Directions*, IEEE International Conference on Big Data (BIGDATA) (2017)
[8] KP Krishna Kumar, G Geethakumari: *Detecting misinformation in online social networks using cognitive psychology*, 13th International Conference on Semantics, Knowledge and Grids (2017)
[9] 三浦 麻子，鳥海 不二夫，小森 政嗣，松村 真宏，平石 界：ソーシャルメディアにおける災害情報の伝播と感情：東日本大震災に際する事例，人工知能学会論文誌 31 巻 1 号 (2016)
[10] 池田 圭佑，榊 剛史，鳥海 不二夫，栗原 聡：口コミに着目した情報拡散モデルの提案及びデマ情報拡散抑制手法の検証，情報処理学会研究報告 (2017)
[11] 池田 圭佑，榊 剛史，鳥海 不二夫，風間 一洋，野田 五十樹，諏訪 博彦，篠田 孝祐，栗原 聡：マルチエージェント型情報拡散モデルの提案，人工知能学会論文誌 Vol. 31, No.1(2016)
[12] Saranya Krishman, Min Chen: *Identifying Tweets with Fake News*, IEEE International Conference on Information Reuse and Integration for Data Science (2018)
[13] Christopher M. Bishop：パターン認識と機械学習 上，丸善出版 (2012)