E-05

流言情報の真偽確認促進システムの評価

Evaluation of Rumors-truth Verication-facilitation System

柿本　大輔 †　　宮部　真衣 ††　荒牧　英治 †††　　吉野　孝 ††††
Daisuke Kakimoto　Mai Miyabe　Eiji Aramaki　Takashi Yoshino

1. はじめに

近年，SNS およびマイクロブログサービスが普及しており，ユーザはリアルタイムに多種多様な情報を取得・発信することが可能である．代表的なマイクロブログサービスの一つに，Twitter[*1]がある．Twitter では，情報発信時の入力可能文字数が 140 字に制限されていることなどから，ユーザの情報発信に対するハードルが大きく下がっている [1]．2011 年 3 月に発生した東日本大震災の際には，Twitter は，速報性の高さを発揮し，重要な情報インフラとして活用されていた [2][3][4][5]．

一方，情報の取得・発信時に，誰もが情報の信頼性を正しく判断することができるとは限らず，流言[*2]が伝播されるという問題も起こった [6]．東日本大震災の際に，Twitter において多く拡散された流言の一つに，「放射性物質にはうがい薬が効く」という内容のものがある．この流言に対し，後に独立行政法人放射線医学総合研究所から訂正と注意喚起がなされた [6]．このような流言は，専門知識を持たないユーザにとっては真偽の判断が難しく，その情報を鵜呑みにして行動に至ると，人間の身体に有害な影響を及ぼす可能性がある．流言は，ユーザ間の適切な情報共有を阻害し，その伝播が深刻な問題を引き起こす可能性がある．そのため，流言の拡散を防止する仕組みが必要である．

これまでに我々は，流言の拡散防止を目的とした，真偽確認促進システムの開発を行ってきた [7]．本システムは，流言である可能性が含まれる情報を閲覧している際，ユーザに気づきを与えることで，情報の真偽確認行動を促進し，流言の拡散防止を支援する．Web ページおよび Twitter タイムライン[*3]閲覧場面を対象としたシステムの評価実験の結果，本システムは，ユーザのページ閲覧を妨げることなく，流言に関する気づきを与えることが出来る可能性があることがわかった [7]．しかし，ユーザに対し，流言に関する気づきをより効果的に与えるためには，システムのインタフェースの改善が必要であることがわかった．

本稿では，真偽確認促進システムのインタフェースの改善について述べる．また，システムの一般公開による，ユーザの利用状況を分析し，システムの効果について評価および考察を行う．

† 和歌山大学大学院システム工学研究科，Graduate School of Systems Engineering, Wakayama University
†† 公立諏訪東京理科大学工学部，Faculty of Engineering, Suwa Tokyo University of Science
††† 奈良先端科学技術大学院大学研究推進機構，Center for Frontier Science and Technology, Nara Institute of Science and Technology
†††† 和歌山大学システム工学部，Faculty of Systems Engineering, Wakayama University

[*1] http://twitter.com

[*2] 本研究では，十分な根拠がなく，その真偽が人々に疑われている情報を流言と定義し，その発生過程（悪意をもった捏造か自然発生か）は問わないものとする．

[*3] フォローしているユーザの発信したツイートが一覧で表示されるページ

2. 関連研究

Twitter では情報発信の手軽さゆえに，流言が多数発信・拡散されており，Twitter における流言に関する研究は盛んである．Castillo ら [8] は，ソーシャルメディアの中には，ユーザが情報の信頼性判断をすることができるシグナルがあると考え，ソーシャルメディア上に存在する情報のみを用いて情報の信頼性を自動評価する手法を提案した．Castillo ら [8] と同様の手法を用いて，Yang ら [9] は，Sina Weibo[*4]を対象に，情報の信頼性を自動評価する手法を提案した．

また，Gupta ら [10] は，ユーザが Twitter タイムラインを閲覧している際，リアルタイムにツイートの信頼性評価を行い，その指標を提示するシステムを提案した．Lim ら [11] は，Web 検索結果を特徴量として利用し，情報の信頼性の自動評価を行うシステムを提案した．

このように，流言自体の検出および信頼性の自動推定を試みる研究は多数行われており，流言として検出あるいは信頼性が低いと推定された情報は，流言の拡散防止の一助となる可能性がある．しかし，流言の拡散防止を支援するためには，流言の検出や信頼性の推定にとどまらず，それらの結果を人々に適切かつ効果的に提供することが重要であると考えられる．そのため，情報を閲覧している段階で，情報確認行動を促すような仕組みおよびインタフェースが必要である．そこで本研究では，これまでに我々が提案した流言情報蓄積システム [12] により収集した流言情報を用いて，ユーザに対し流言に関する気づきを提供し，情報の真偽確認行動を促進するシステムを検討する．

3. 真偽確認促進システム

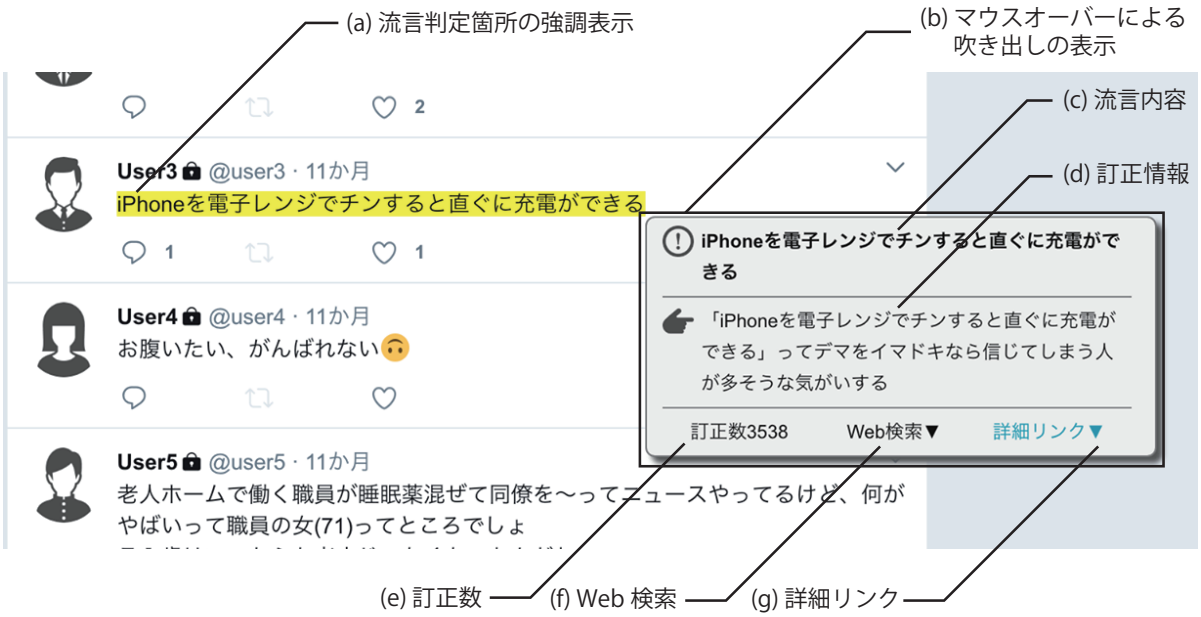
3.1 概要

流言が拡散する要因の一つとして，人間による情報の真偽確認不足が考えられる．これまでに我々は，Twitter における情報確認行動に関する調査を行った．その結果，リツイート[*5]対象とするツイート内容の真偽を意識せずにリツイート機能を利用している Twitter 利用者が多く，流言が拡散されやすい可能性があることがわかった [7]．これまでに，Twitter 上で流言の拡散防止を支援する手法およびシステムは提案されている [8][9][10][11][12] が，これらは，ユーザの能動的なシステムの利用が前提になっている．情報の真偽に関する関心の低さを鑑みると，流言拡散防止支援システムを能動的に利用してまで，ツイート内容の真偽を確認しようとする人々は少ないと考えられる．流言の拡散を防止するためには，このような人々に情報の真偽確認行動を促すことが重要であると考えられる．

そこで我々は，流言である可能性が含まれる情報を閲覧している際，ユーザに気づきを与えることで，情報の真偽

[*4] http://weibo.com

[*5] 他のユーザが発信したツイートを再発信することができる機能．



（a）流言判定箇所の強調表示：流言が含まれる可能性のある部分を判定し，強調表示．
（b）マウスオーバーによる吹き出しの表示：強調表示部分のマウスオーバーで，吹き出しを表示．
（c）流言内容：判定された流言．
（d）訂正情報：流言情報に対する訂正ツイート．
（e）訂正数：訂正情報の数．
（f）Web 検索結果へのリンク：判定された流言情報における名詞をクエリとした Web 検索ページへのリンク．
（g）詳細リンク：これまでの訂正数および直近一週間の訂正数の推移を表すグラフを確認できるページへのリンク．

図 1: システムの動作画面

確認行動を促進し，流言の拡散防止を支援するシステムを開発した．システムの動作画面を図 1 に示す．システムは，ユーザが閲覧中の Web ページ内に，流言である可能性のある情報が含まれる場合，該当テキストの強調表示を行う（図 1（a））．また，強調表示箇所をマウスオーバーすることで，吹き出しにより情報を提示する（図 1（b））．吹き出しには，流言に関する情報（判定された流言内容（図 1（c）），訂正情報（流言情報に対する訂正ツイート（図 1（d）），訂正数（訂正情報の数（図 1（e）），Web 検索結果へのリンク（判定された流言に含まれる名詞をクエリとした Web 検索ページへのリンク（図 1（f）），および詳細リンク（これまでの訂正数および直近一週間の訂正数の推移を表すグラフを確認できるページへのリンク（図 1（g））が含まれる．

これらの機能により，流言への気づきをユーザへ提供可能にする．本システムは Google Chrome[*1] のアドオンとして動作するように実装した．

3.2 インタフェースの改善

1 章で述べたように，これまでの評価実験の結果，システムは，ユーザのページ閲覧を妨げることなく，流言に関する気づきを与えることが出来る可能性があることがわかっている [7]．しかし，ユーザに対し，流言に関する気づきをより効果的に与えるためには，「Twitter タイムライン閲覧場面のような様々な情報が混在する環境においても，ユーザが流言に関する情報を見落とさないようにする」「システムがユーザの情報閲覧の妨げにならないよう，ユーザ側の設定に応じて情報提供方法を変更可能にする」などの，システムのインタフェースの改善が必要であることがわかった．そこで我々は，新たに以下の機能を本システムに追加した．

機能 1．トースト通知機能

Twitter のタイムラインのような環境では，様々なユーザの発信した情報が混在しているため，流言に関する情報が提示されても，見落としてしまう可能性がある．そこで，このような環境でもユーザが流言に関する情報を見落とさないよう，トースト通知[*2]機能を追加した．トースト通知機能の動作例を，図 2 に示す．この機能は，ページを読み込んだ際に，テキストを強調表示するだけでなく，Web ブラウザの画面右上に，検出した流言情報リストを提示する機能である．

機能 2．カスタマイズ機能

普段 Web を閲覧している際と同様に，ユーザにシステムを利用してもらうためには，強調表示箇所などに応じて，流言に関する気づきの提供方法を変更できることが望ましい．そこで，カスタマイズ機能を追加した．カスタマイズ機能の動作例を，図 3 に示す．カスタマイズ機能により，Web ページ閲覧時，本システムにおけるテキストの強調表示，トースト通知および吹き出し表示の各機能を動作させるかどうかを，ユーザは自由に切り替えることができる．また，

[*1] http://www.google.co.jp/chrome/browser/desktop/index.html

[*2] トースト通知とは，画面端に数秒表示される通知である．ユーザに操作を要求する通知ではないため，操作を妨げることがないという特徴がある．

図 2: トースト通知機能

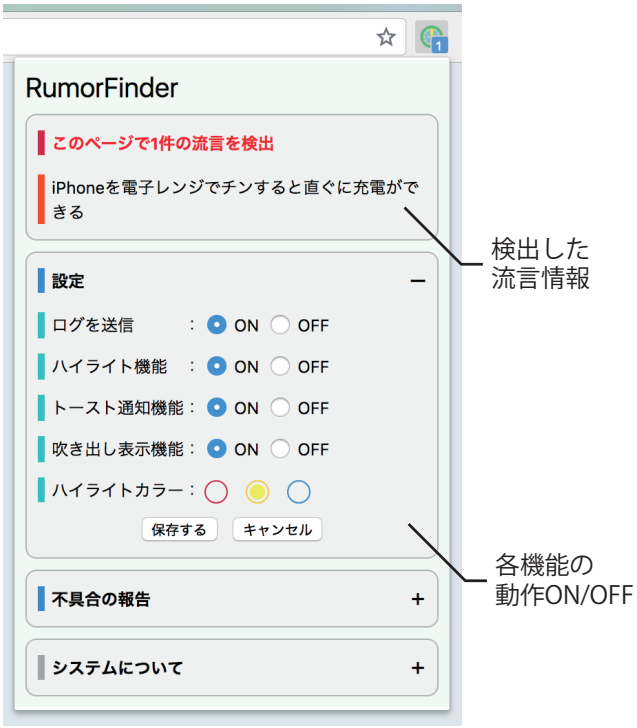


図 3: カスタマイズ機能

カスタマイズ機能は Web ブラウザの画面右上から起動することができ，図 3 上部に示すように，検出した流言情報のリストアップも行う．

4. 利用状況と評価

4.1 利用状況

我々は，Chrome ウェブストア[*1] を介して，2018 年 7 月 4 日より本システムを一般公開している (システム名は「Rumor Finder」としている)．また，本システムの利用状況の分析および評価のために，19 人の大学生および大学院生にシステムを利用してもらうよう呼びかけた．呼びかけの際，システムの概要を紹介する Web ページ[*2] の URL を伝え，利用ログの送信を許可してもらうよう伝えた．その他の指示は与えず，自由に使ってもらうこととした．

2018 年 7 月 19 日時点のシステムのインストール総数は 28 回，同様にユーザ数は 28 人であり，2018 年 7 月 4 日～2018 年 7 月 19 日の 15 日間に，15 人の利用ログが得られた[*3]．ただし，利用ログは端末ごとに取得しているため，同一ユーザが複数端末でシステムを利用している可能性がある．

4.2 評価と考察

2018 年 7 月 4 日～2018 年 7 月 19 日の 15 日間における，ユーザごとのシステムの利用状況 (閲覧ページ数，流言検出ページ数，検出割合および吹き出し表示回数) を表 1 に示す．表 1 では，システムの概要を紹介する Web ページ[*3] および流言情報クラウドの Web ページ[*4] における動作ログは除外している．また，同期間に検出および吹き出し表示された流言情報を表 2 に示す．表 1 より，(C) 検出割合については，ユーザ 15(24.7%) を除く全てのユーザで 6%未満となった．(D) 吹き出し表示回数については，吹き出しの表示を複数回行うユーザがいるものの，全く行わないユーザがいることもわかった．これらの結果より，ユーザが Web ページ閲覧時，本システムにより流言の検出が通知される頻度は低いことがわかった．この理由としては，(1) そもそも，災害時など以外の平常時では，ユーザが流言情報を目にする機会がさほど多くないこと，(2) システムが流言として検出すべき情報を，正しく検出できていない可能性があることが考えられる．また，吹き出し表示を全く行わないユーザもいることから，流言に関する情報の提供方法の追加・変更が必要である可能性がある．そのため，トースト通知機能およびカスタマイズ機能のような，テキストの強調表示および吹き出し表示以外の情報の提供方法について，今後さらに検討する必要があると考えられる．

また，本システムは「不具合の報告」として，本システムの問題点および要望などを自由に記述し，送信できる機能を備えている．本稿で述べた 15 人のユーザのうち 1 人のユーザから「あるページの閲覧時，流言情報が含まれていないにもかかわらず，システムが特定の流言情報を検出した」「日付の情報が記載されている様々なページにおいて，特定の流言情報が検出される」というコメントが得られた．実際には Web ページの中に含まれていなかったにもかかわらず，システムが誤検出したと報告された流言情報は，表 2 における流言情報 (2)「10 月 10 日が晴れの日の特異日だから」であった．システムの流言判別手法には，流言情報の形態素解析結果における名詞が，判別対象とするテキスト内に全て含まれるかどうか，という基準が含まれる．「10 月 10 日が晴れの日の特異日だから」という流言情報の形態素解析の結果，名詞は“月”および“日”のみとなった．そのため，ユーザが閲覧している Web ページ内に，“月”および“日”の両方が存在している場合，その箇所を流言情報として判別する場合がある．これらの結果より「流言として判別すべきでない情報を，システムが流言として判別することがある」という問題があることがわかった．流言情報とは関係のない情報が強調表示されると，ユーザの Web ページ閲覧を妨げ，システムが利用されなくなる可能性が

[*1] https://chrome.google.com/webstore/category/extensions
[*2] http://www2.yoslab.net/ kakimoto/addon_intro/index.html
[*3] 利用ログを送信するかどうかは，ユーザが任意で変更することができる．そのため，呼びかけた 19 人全員の利用ログを全て取得することはできなかった．また，システムは一般公開されているため，その他のユーザの利用ログを取得している可能性がある．
[*4] http://mednlp.jp/ miyabe/rumorCloud/rumorlist.cgi

表 1: システムの利用状況

	ユーザ														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
(A) 閲覧ページ数 (件)	1373	2184	1155	586	94	130	35	9	1485	2461	667	262	4012	1099	85
(B) 流言検出ページ数 (件)	13	102	7	10	0	3	2	0	24	28	11	0	33	12	21
(C) 検出割合 (%)	0.9	4.7	0.6	1.7	0.0	2.3	5.7	0.0	1.6	1.1	1.6	0.0	0.8	1.1	24.7
(D) 吹き出し表示回数 (回)	33	65	2	12	0	3	0	0	71	3	2	0	0	1	21

(A) 閲覧ページ数において，同一の Web ページの重複に関する考慮は行っていない．
(B) 流言検出ページ数は，1 件以上の流言情報が検出された Web ページの数である．
(C) 検出割合は，((B) 流言検出ページ数/(A) 閲覧ページ数)*100 の値である．

表 2: 検出した流言情報

(1)	一般の人でもフォロワーが 6%減る
(2)	10 月 10 日が晴れの特異日だから
(3)	窃盗グループが被災地に入っている
(4)	レスキューの服を着た泥棒
(5)	朝鮮人が井戸に毒を入れた
(6)	東日本には人が住めなくなる!!
(7)	ジャー・ジャー・ビンクスが登場する
(8)	北海道の水源が中国に買われる
(9)	狙われる日本の水源地
(10)	被災地に窃盗団が...
(11)	精神鑑定受けないまま執行された
(12)	ハリウッドで映画化!

ある．そのため，システムによる流言情報の判別手法について，再検討する必要がある．

5. おわりに

我々はこれまでに，ユーザが Web ページを閲覧している際，流言が含まれる可能性のある部分を強調表示し，吹き出しにより流言に関する情報を提示する真偽確認促進システムを開発してきた．情報の真偽に対して関心の低いユーザにシステムを利用してもらうためには，ユーザに対し，流言に関する気づきをより効果的に与えることが重要である．

本稿では，システムのインタフェースに関する改善を行った．Twitter のタイムラインのような，様々なユーザの発信した情報が混在している環境でも，ユーザが流言に関する情報を見落とさないよう，トースト通知機能を追加した．また，普段 Web を閲覧している際と同様に，ユーザにシステムを利用してもらうために，カスタマイズ機能を追加し，流言に関する気づきの提供方法を変更可能にした．

また，システムの利用状況の分析および評価の結果，本来流言情報と判別すべき部分を，システムが流言情報として判別しない場合があること，流言情報と判別すべきでない部分を，システムが流言情報として判別する場合があることがわかった．そのため，システムによる流言情報の判別手法を再検討する必要がある．

今後は，システムによる流言情報の判別手法を再検討する．また，ユーザの利用ログをさらに収集し，システムがユーザの流言拡散防止を支援することができるかどうかを検証する．

謝辞

本研究は，JSPS 科研費 15H05317 の助成による．

参考文献

[1] 垂水浩幸: 実世界インタフェースの新たな展開:4. ソーシャルメディアと実世界, 情報処理学会誌, Vol.51, No.7, pp.782-788 (2010).

[2] 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, Vol.31, No.1, p. NFC-A_1-9 (2016).

[3] インプレス R&D インターネットメディア総合研究所: インターネット白書 2011, インプレスジャパン, pp.44-47 (2011).

[4] 西谷智広: I 見聞録:Twitter 研究会, 情報処理学会誌, Vol.51, No.6, pp.719-724 (2010).

[5] 立入勝義: 検証 東日本大震災 そのときソーシャルメディアは何を伝えたか?, ディスカヴァー・トゥエンティワン, pp.20-26 (2011).

[6] 荻上チキ: 検証 東日本大震災の流言・デマ, 光文社, pp. 27-28 (2011).

[7] 柿本大輔，荒牧英治，宮部真衣 : 流言拡散防止のための真偽確認促進システムの構築，ヒューマンインタフェース学会論文誌，Vol.20，No.1，pp.1-11 (2018).

[8] C. Castillo, M. Mendoza, and B. Poblete.: Information Credibility on Twitter. In WWW, pp.675-684 (2011).

[9] F. Yang, Y. Liu, X. Yu, and M. Yang.: Automatic Detection of Rumor on Sina Weibo. In Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, No.13, pp.1-7 (2012).

[10] A. Gupta, P. Kumaraguru, C. Castillo and P. Meier : TweetCred: Real-time Credibility Assessment of Content on Twitter，SocInfo'14，pp.228-243 (2014).

[11] W. Y. Lim, M. L. Lee and W. Hsu : iFact: An Interractive Framework to Assess Claims from Tweets，CIKM'17，pp.787-796 (2017).

[12] 宮部真衣，灘本明代，荒牧英治 : 人間による訂正情報に着目した流言拡散防止サービスの構築，情報処理学会論文誌，Vol. 55，No. 1，pp. 563-573 (2014).