情報処理学会論文誌 Vol.55 No.1 563–573 (Jan. 2014)

推薦論文

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

宮部 真衣[1,a)] 灘本 明代[2] 荒牧 英治[1,3]

受付日 2013年4月10日, 採録日 2013年10月9日

概要：近年，マイクロブログの普及にともない，マイクロブログを用いた個人での情報発信が増加している．2011 年 3 月 11 日に発生した東日本大震災においては，Twitter に多くの情報が投稿された．Twitter では，重要な情報が伝搬された一方で，様々な流言の拡散も多数行われた．特に災害時は，流言が救援活動などに悪影響を及ぼす可能性が高いため，流言の広がりにくい環境を作る必要がある．本研究では，マイクロブログ上の流言に対して，どのような対応がされているかを調査した．また，人間によって発信される訂正情報に着目し，訂正情報に基づいて流言情報を収集・提供することにより流言拡散を防止するサービスを構築した．本提案手法では，流言であることを指摘する表現（流言マーカ）を含む情報を収集し，それらが訂正情報であるかどうかを判定する．平常時のツイート，災害時のツイートを用いて，構築した訂正情報分類器の判定精度を検証した結果，平常時を含めたデータを用いることで，平常時・災害時のどちらでも高精度な判定が期待できることを示した．

キーワード：マイクロブログ，流言，災害

Development of Service for Prevention of Spreading of False Rumors based on Rumor-correction Information

Mai Miyabe[1,a)] Akiyo Nadamoto[2] Eiji Aramaki[1,3]

Received: April 10, 2013, Accepted: October 9, 2013

Abstract: Recently, conveying information from individuals has increased due to the development of microblog. After the Great Eastern Japan Earthquake in Japan 2011, numerous tweets were exchanged on Twitter. Twitter provided a lot of important information after the earthquake. On the other hand, various false rumors were spread on Twitter. False rumors negatively affect important activities such as relief activities. Therefore, it is required to structure the environment that prevents people from spreading false-rumors. In this paper, we analyze how people deal with false rumors on Twitter. Moreover, we propose rumor information cloud based on rumor-correction information. The proposed method collects information including indication of rumor, then classifies the information into correct information or other information. On the basis of these evaluation experiments, the method using data of nonemergency situation can classify both data of an emergency and a nonemergency situation with high accuracy.

Keywords: microblogging system, rumor, emergency situation

1. はじめに

近年，Twitter[*1]などのマイクロブログが急速に普及し，ユーザによるマイクロブログを用いた情報発信が活発化し

[1] 京都大学学際融合教育研究推進センターデザイン学ユニット
Unit of Design, Center for the Promotion of Interdisciplinary Education and Research, Kyoto University, Shimogyo, Kyoto 600–8815, Japan
[2] 甲南大学知能情報学部
Faculty of Intelligence and Informatics, Konan University, Kobe, Hyogo 658–8501, Japan
[3] 科学技術振興機構さきがけ
JST PRESTO, Chiyoda, Tokyo 102–0076, Japan
[a)] mai.miyabe@gmail.com

[*1] http://twitter.com/
本論文の内容は 2012 年 7 月のマルチメディア，分散，協調とモバイル（DICOMO2012）シンポジウム 2012 にて報告され，グループウェアとネットワークサービス研究会主査により情報処理学会論文誌ジャーナルへの掲載が推薦された論文である．

ている．特に Twitter は，140 文字という制限によりユーザの情報発信への敷居が大きく下がっており [1]，2011 年 3 月 11 日に発生した東日本大震災においては，リアルタイムに情報を伝える重要な情報インフラの 1 つとして活用された [2], [3], [4]．しかし，安否情報などの重要な情報の共有・伝搬が行われた一方で，多くの流言も拡散されるという問題も生じた [5]．流言は適切な情報共有を阻害する要因となり，悪影響を及ぼす．特に災害発生時には，流言が救命のための機会損失を生む場合もあるため，流言の広がりにくい環境を作る必要がある．

では，流言はどのように拡散されていくのだろうか？まず，人々がある情報を他者に伝える場合，その情報が正しいと思って伝えていることが多く，本人がでたらめだと思う話を，悪意をもって他者に伝えることは少ない [6]．つまり，流言の伝達は，主に伝達している情報が流言であることを認識していないことに起因すると考えられる．そこで，もしそうであるならば，人々に流言情報を提供することにより，流言の拡散を防止できる可能性があると考えられる．

そこで本研究では，流言拡散を防ぐための仕組みとして，流言情報クラウドを提案する．流言情報クラウドは，リアルタイムに流言情報を蓄積し，その情報を提供することにより，流言拡散を防止するサービスである．本研究では，流言情報を収集するためのリソースとして，Twitter を用いる．本研究のポイントは以下の 2 点である．

(1) 従来の流言に関する研究（現実社会の中での流言に関する研究）での知見や Twitter の特徴に基づき，マイクロブログ上での流言の特徴を明らかにする．
(2) **流言が含まれるテキストではなく，流言の不確かさについて言及しているテキスト（以下，訂正情報と呼ぶ）を抽出することにより，**自動的に流言情報を収集・提供するサービスを提案する．

本研究では訂正情報を抽出することにより，その訂正情報の中で訂正している流言情報を間接的に収集することを目指している．訂正情報に含まれる流言情報の抽出については，パターンマッチングによる抽出または人手による抽出を想定しており，本論文では訂正情報を正しく判定できるかどうか検証することを研究課題とする．

本論文では，マイクロブログ上での流言の特徴について示した後，提案するサービスおよび流言情報を自動的に収集するための訂正情報分類器について述べる．

2. 関連研究

本章では，まず，流言に関するこれまでの定義について述べた後，流言について扱ったソーシャルメディアに関する研究について述べる．

2.1 流言の定義と流言の伝達

流言については，これまでに多くの研究が多方面からなされている．流言と関連した概念として噂，風評，デマなどの研究がある．これらの定義の違いについては諸説あり，文献ごとにゆれているのが実情である．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．よって，最終的に正しい情報であっても，発言した当時に，十分な根拠がない場合は，流言と見なす．

流言の分類としては，ナップによる第 2 次世界大戦時の流言の分類がある [6]．ナップは，流言を「恐怖流言（不安や恐れの投影）」「願望流言（願望の投影）」「分裂流言（憎しみや反感の投影）」の 3 つに分類している．また，これらの流言がどの程度の割合で流通するかは社会状況によって決まると述べられている．社会状況は流言を伝達させる要因の 1 つであり，たとえば震災の直後など，社会状況が多くの人々に不安を感じさせる状況は，流言の発生や伝達に関係する．

また，流言の伝達には，**曖昧さ，重要さ，不安**という 3 つの要因が強く関係することが示されている [6]．オルポートとポストマンは，流言の流布量を，$R \sim i \times a$ のように定式化し，「流言の流布量（R）は，重要さ（i）と曖昧さ（a）の積に比例する」と述べている [6]．

しかし，これらの流言に関する先行研究は，現実社会の中での口伝えでの流言の伝達について行われたものであり，マイクロブログでの流言の伝達に関して，十分な分析は行われていない．本研究では，マイクロブログを対象とし，先行研究での知見とマイクロブログの特徴に基づき，マイクロブログ上の流言の特徴を分析し，流言拡散防止サービスを提案する．

2.2 異常状態の検出

流言の拡散を防止するためには，ある時点で拡散されている流言を検出する必要がある．流言が拡散されている場合，ある流言を含む情報の頻度が急激に高まる可能性があり，流言の拡散されている状態は，異常状態の 1 種と見なせる．そこで，本節では，異常状態の検出に関する研究について述べる．

Twitter をセンサとしてとらえ，災害などの異常事態の検出を試みた研究がある．Sakaki らは，Twitter を用いた地震や台風の位置の推定に関する研究を行っている [7]．Abel らは，緊急放送システムをモニタリングしておき，災害の発生を確認した後，Twitter から災害に関連するツイートを収集し，有益な情報をユーザに提供するシステムの開発を行っている [8]．Aramaki らや Paul らは，Twitter を用いてインフルエンザの把握を行っている [9], [10]．これらは，平常時からソーシャルメディアなどを監視しておくこ

とで，異常事態発生時にいち早くその情報を伝えるという警告型のサービスである．

また，流言に関する検出を試みた研究も行われている．Qazvinian らは，マイクロブログ（Twitter）における特定の流言に関する情報を網羅的に取得することを目的とし，流言に関連するツイートを識別する手法を提案している [11]．実験の結果，ある流言に関連するツイートを高精度に識別可能であることを示しているが，課題として新しく発生した流言の検出があげられている．また，Rattanaxay らは，「らしい」といった，流言に含まれる曖昧な表現に着目した流言情報の検出手法を提案している [12]．しかし，Rattanaxay らの手法では，曖昧な表現を含む情報はすべて流言と見なしており，たとえ正確な情報であっても，曖昧な表現を含むものは流言として検出してしまう．

本研究では，誤った情報および根拠がない情報を検出するために，人間によって発信される訂正情報に着目した，流言拡散防止サービスを提案する．

3. マイクロブログにおける流言の特徴

流言拡散を防止するには，ある情報が流言かどうかを判定しなければならない．流言かどうかを判定する方法としては，あらかじめ蓄積した流言情報に基づき判定する，情報に含まれる表現やユーザ属性をもとに判定するなどの方法が考えられる．本研究では，あらかじめ流言情報を蓄積することにより，流言拡散を防止する手法を検討する．

本研究では，流言情報を収集するためのリソースとして，Twitter を用いる．Twitter は，投稿する文章（以下，ツイート）が 140 字以内に制限されていることよる情報発信の敷居の低さと，リツイート（RT）という情報拡散機能により，流言が拡散されやすくなっている．実際に，東日本大震災においては，Twitter では様々な流言が拡散されていたという指摘もあることから [13]，Twitter を収集のリソースとして用いることとした．

本章では，まず，Twitter 上で拡散された流言の特徴について調査し，流言収集手法の方針を立てる．

3.1 対象データセット

本研究では，分析対象のデータとして，以下の 2 種類のデータを用いる．

平常時データ 2010 年 3 月[*2]に投稿されたツイート（一部の流言データについては，2010 年 2 月のツイートも含む）

災害時データ 2011 年 3 月 11 日～30 日までに投稿されたツイート

平常時データについては 3 種類（平常時 A～C），災害時データについては 5 種類（災害時 A～E）の流言に関するツイートを用いた．これらの 8 種類のデータの選定にあたっては，まず，平常時データ・災害時データからランダムに抽出した 1,000 件のツイートを人手で確認し，間違いを含む情報（流言）を分析候補として抽出した[*3]．その後，分析候補である流言を特定可能なキーワードをもとに，各データ全体からツイートを抽出し，流言を特定可能なキーワードを含むツイートが 100 件以上[*4]存在する流言を分析対象とした．

分析に用いた流言の内容およびツイート例を**表 1** に示す．流言データについては，それぞれの流言に関するキーワード（**表 2**）を含むツイートを抽出した後，今回取り扱う流言とは関係のないツイートを人手で確認し除外した[*5]．

3.2 ツイートの分類

本研究では，まず，抽出した流言に関するツイートを以下の 3 種類に分類する．

(1) 流言ツイート：流言に関するツイート
(2) 疑問ツイート：流言に対して，疑問を表すツイート（例：**** (流言内容) は本当なの？デマじゃないの？）
(3) 訂正ツイート：流言であることを指摘するツイート（例：このツイートはデマです。RT xxx: ****（流言内容））

情報（流言）の信頼性について疑問を表す表現を含むものは疑問ツイート，明らかに訂正であると分かるものは訂正ツイートとして扱う．

3 種類のツイートへの分類を人手で行った結果を**表 3** に示す．表 3 より，発信数に違いはあるものの，いずれの流言についても，訂正ツイートや疑問ツイートが発信されていることが分かる．

3.3 流言への対処

本節では，マイクロブログ上での流言に対してどのような対処が行われるのかについて分析する．

先行研究では，噂への対処戦略として，以下の 3 種類があると述べられている [6]．

否定戦略： 噂の内容を明確に否定する．
対抗戦略： 噂について否定しない．噂自体とは異なるイメージを流す．
無視戦略： 噂に対して反応せず，噂が流れるままにしておく．

[*2] 平常時のデータについては，暫定的に災害時データの 1 年前に設定した．
[*3] 本論文の流言の定義に基づき，「間違いを含む情報である」と十分に判断してよいことを，文献 [5]，ニュースの情報をもって確認した．
[*4] 母数の少ないものを除外するため，暫定的に閾値を 100 件とした．
[*5] たとえば，表 1 の平常時 A の流言については，「アニメ D」「声優」というキーワード（表 2）でツイートを抽出しているが，「アニメ D アプリ入れてみた．声優 が変わってから見てないけど好き．」のように，流言とは関係のないツイートも含まれる．このようなツイートについては，人手で確認作業を行い，流言データから除外した．

表 1 データセット
Table 1 Examples of rumor tweets.

流言データ		流言内容	ツイート例
平常時	A	アニメ D の声優交代	テレビ局 A は 25 日にホームページ上でアニメ「アニメ D」の声優を視聴率低迷のため変更することを発表した。O（73）さんをはじめとする旧レギュラー声優が復帰する予定。
	B	テレビ局 N での Twitter 禁止	社内ツイッター禁止情報　テレビ局 N は一昨日より twitter.com への社内からのアクセスを禁止にした模様。
	C	Twitter での歌詞のつぶやき	RT @****: Twitter で歌詞をつぶやくと組織 J の利用料が発生する
災害時	A	サーバールームでの負傷	地震が起きた時、社内サーバールームにいたのだが、ラックが倒壊した。腹部を潰され、血が流れている。
	B	命の三角形	命の三角形、ためになりました。地震では机の下にすぐ入る事を考えがちだけど、机等のすぐ横のほうがつぶされない三角形の空間になり、助かる可能性が高いんだね。
	C	K 電力による節電呼びかけ	【関西地区の皆さん】K 電力が電力の提供を始めたようなので、コチラで節電すれば立派な支援になります。出来るかぎり節電を心がけましょう。あと、電子レンジや炊飯器等を使っていないときはコンセント等を抜いておくと電気節約＆もし地震が来た際、火災防止にもなります
	D	O 氏の寄付	漫画 O の作者　O 氏、地震の被害者救済に 15 億円を寄付　「自分が幸せになったということは、世の中から受けたひとつの借りだ」
	E	S 氏のラッキー発言	「K さんはとてもラッキーな人」M 党の S 前・官房長官が東北大地震を「ラッキー」と表現。

表中の流言内容およびツイート例は，実在の名称をイニシャルで提示している．

表 2 流言キーワード
Table 2 Keywords in rumor tweets.

流言データ		流言キーワード
平常時	A	アニメ D，声優
	B	テレビ局 N，社内，Twitter，アクセス，禁止
	C	Twitter，歌詞，組織 J，利用料
災害時	A	社内サーバールーム
	B	命の三角形
	C	K 電力，（節電，給電，送電，供給，提供，依頼，願い）
	D	O 氏，（寄付，15 億）
	E	S 氏，ラッキー

表中の括弧で括られたキーワードは，いずれか 1 つを含むものを抽出した（たとえば，災害時 D の場合は，「O 氏 AND 寄付」または「O 氏 AND 15 億」を含むツイートを抽出した）．
表中のキーワードは，実在の名称をイニシャルで提示している．

これらの対処戦略のうち，無視戦略が行われたかどうかを判断することは難しいが，否定戦略と対抗戦略が行われたかどうかは調査可能である．

ユーザは，噂の内容を否定する際に，「デマ」「間違い」「嘘」のような流言であることを示す表現（以下，流言マーカと呼ぶ）を用いる可能性がある．そこで，Twitter の特徴であるリツイートをふまえ，この流言マーカに注目して，訂正ツイートを以下の 3 種類に分類する．

(1) 否定戦略 1　流言マーカを用いて流言を否定しているツイート
（例：****（流言内容）は デマ です）

(2) 否定戦略 2　流言マーカを含まないが，流言を否定しているツイート
（例：****（流言内容）について，ネットではソースが見つからない）

(3) 対抗戦略　正確な情報を含むツイート

なお，本調査では，17 語句[*6]を流言マーカとし，流言

表 3 流言ツイート数，疑問ツイート数，訂正ツイート数
Table 3 Number of rumor tweets, question tweets and correction tweets.

		流言ツイート数	疑問ツイート数	訂正ツイート数	合計
平常時	A	589 (57.0%)	119 (11.5%)	326 (31.5%)	1,034 (100%)
	B	1,350 (87.9%)	60 (3.9%)	125 (8.1%)	1,535 (100%)
	C	3,934 (84.3%)	196 (4.2%)	533 (11.4%)	4,663 (100%)
災害時	A	2,822 (73.4%)	38 (1.0%)	984 (25.6%)	3,844 (100%)
	B	2,836 (87.4%)	34 (1.0%)	374 (11.5%)	3,244 (100%)
	C	15,883 (46.1%)	808 (2.3%)	17,791 (51.6%)	34,482 (100%)
	D	1,886 (27.1%)	169 (2.4%)	4,906 (70.5%)	6,961 (100%)
	E	267 (61.2%)	32 (7.3%)	137 (31.4%)	436 (100%)

表中の数値は各流言に関して抽出したツイートを 3 種類（流言，疑問，訂正）に分類した結果であり，それぞれの中にリツイート（非公式リツイートおよび公式リツイート）やリプライを含む．

[*6] 本調査で用いた流言マーカは，「デマ」「嘘」「ウソ」「釣り」「ツリ」「偽情報」「都市伝説」「ガセ」「ネタ」「狂言」「迷信」「誤報」「間違い」「いたずら」「騙され」「釣られ」「チェーンメール」である．

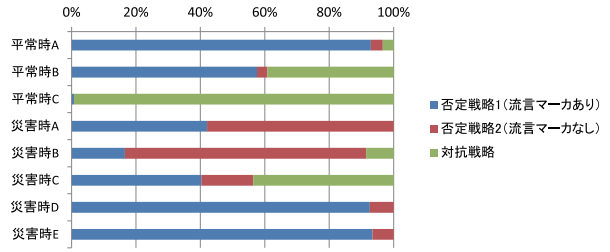

図 1 訂正ツイートの分類結果
Fig. 1 Classification result of correction tweets.

マーカの有無を確認し，ツイートを分類する．訂正ツイートの分類結果を**図 1** に示す．

図 1 より，流言によって投稿される訂正ツイートの傾向は異なるものの，流言マーカを含む訂正ツイート（否定戦略 1）については，発信数に違いはあるものの，いずれの流言においても発信されている．一方，対抗戦略についてはまったく発信されない場合（災害時 A，災害時 D，災害時 E）がある．否定戦略 2，対抗戦略については，「否定戦略 2 である」「対抗戦略である」と判断するための，それぞれの戦略のみに共通する明確な特徴が見られない．しかし，否定戦略 1 については，明確な特徴（流言マーカ）があるため，これらと比較して流言の訂正情報として検出しやすい可能性がある．そこで，本研究では訂正情報として否定戦略 1 に着目し，流言収集手法を検討することとした．以降の章においては，「否定戦略 1」のことを「訂正情報」と呼ぶ．

なお，今回は分析したデータが 8 種類と限定的であるため，否定戦略 1 の発信されていない流言も存在する可能性がある．否定戦略 1 を用いて十分に流言情報を収集できるかどうかは，今後検証する必要がある．

4. 流言情報クラウド：訂正情報に基づく流言拡散防止サービス

1 章で述べたように，情報が誤っていることをユーザに提示できれば，誤った情報の拡散を防ぐことができる可能性がある．2011 年の東日本大震災発生後には，Web 上で広がった流言について，ブログなどでまとめ記事が作成されるなど，流言の拡散を防止するための活動が行われた[*7,*8]．しかし，これらのブログ上の情報は人手によりまとめられており，発生した流言をリアルタイムに反映することは容易ではない．

そこで，本研究では流言拡散を防止するサービスとして，流言情報クラウドを提案する．流言情報クラウドの機能は大別して 2 つある．まず，リアルタイムに流言情報を蓄積する（**情報蓄積フェーズ**）．次に，蓄積した情報をユーザに提供し，流言の拡散を防ぐ（**拡散防止フェーズ**）．情報蓄積フェーズにおける情報の収集リソースとしては，Twitter を想定する．拡散防止フェーズにおける提供先アプリケーションとしては，Twitter だけでなく，メールクライアントなどのテキストベースのアプリケーションを想定する．

以降の節において，本研究における流言情報収集のアプローチ，および流言情報クラウドのシステム構成について述べる．

4.1 本研究における流言情報収集のアプローチ

本節では，流言情報を収集するためのアプローチについて述べる．

流言情報を蓄積するためには，ある情報に流言が含まれているかを判定する必要がある．しかし，人間が信じてしまうような流言を自動的に流言だと判定することはきわめて難しい．また，その時点では情報の真偽を判断できず，後になって真偽が分かることも多い．さらに，流言の内容は多様であり，既知の流言情報を用いて判定しても，正しく抽出することは容易ではないと考えられる．

3 章における分析により，マイクロブログ上では，話題によって発信数は異なるものの，流言に対して訂正ツイートや疑問ツイートが発信されており，さらに，流言への対処として，流言マーカを含む形で訂正情報が発信されていることを示した．つまり，内容が多様であり，情報の形式に統一性のない流言自体を特定するよりも，流言を否定している訂正情報を特定する方が容易である可能性がある．

そこで，本研究では，ある情報が流言かどうかを判定し，流言情報を直接収集するのではなく，**訂正情報を抽出することにより，間接的に流言情報を収集する．**本研究では，3.3 節で述べた噂に対する 3 つの対処戦略のうち，否定戦略を訂正情報として扱うこととする．

4.2 流言情報クラウドの構成

流言情報クラウドのサービス構成を**図 2** に示す．提案サービスは，以下の 3 つの機能により構成される．

(1) クローリング機能 Twitter からテキストを収集する．
(2) 訂正情報判定機能 我々の提案する訂正情報分類器を用いて，テキストが訂正情報かどうかを判定する．
(3) 流言情報管理機能 流言情報データベースの管理（検索，登録，修正）を行う．

情報蓄積フェーズでは機能 (1)〜(3) を，拡散防止フェーズでは機能 (2)〜(3) を利用する．

次項において，各フェーズの流れについて詳細に説明する．

4.2.1 情報蓄積フェーズ

基本的には自動で流言情報を蓄積し，人手を介することなく情報提供を可能にする．また，人手による精査も可能とすることにより，提供する情報の信頼性を向上させるこ

*7 荻上式 BLOG「東北地方太平洋沖地震，ネット上でのデマまとめ」：http://d.hatena.ne.jp/seijotcp/20110312/p1
*8 ついのすみか「東北地方太平洋沖地震のデマ情報まとめ」：http://tsuinosumika.iku4.com/Entry/67/

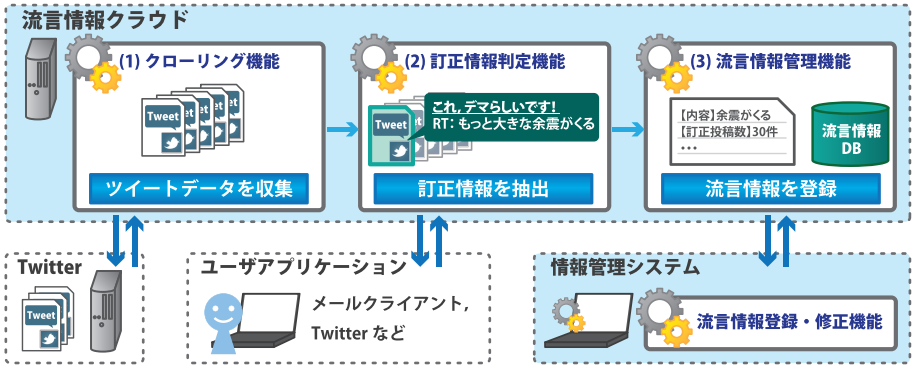

図 2 流言情報クラウドのサービス構成

Fig. 2 Service configuration of rumor information cloud.

表 4 訂正情報の定義とその例

Table 4 Examples of rumor-correction information.

判定条件	該当例
a) ある情報に関する不正確さの記述が主題である	このツイートはデマです。RT xxx: ○○○ （確信度：高） ○○○は本当なの？デマじゃないの？ （確信度：中） ○○○が、デマだとしても、備えあれば憂いなし。（確信度：低）
b) ある情報に関する不正確さの記述が含まれるが，主題ではない	○○○というデマを広げた人間がいるみたいだね。
c) 流言に関してまとめたサイトを紹介している	地震に関するデマ http://・・・

とができるようにする．

情報蓄積フェーズでは，流言マーカをもとに訂正情報を抽出することにより，間接的に流言情報を収集する．本研究における訂正情報の定義を**表 4** に示す．不正確さを含む記述が含まれていた場合，不正確さに関する確信度の高さにかかわらず，訂正情報と判定することとする．

一方，流言マーカを含むものが，必ずしも訂正情報であるとは限らない．たとえば，「デマゴギーって何？デマの省略前の言葉？」というツイートは，「デマ」という流言マーカが含まれるが，流言の訂正情報ではない．

そこで，以下の手順により流言情報を収集する．

(1) 流言マーカを含むツイート群を収集する．

(2) 5 章で述べる訂正情報分類器を用いて，収集したツイート群から訂正情報を抽出する．

(3) “「～」というデマ” のような，流言部分が明示的な訂正情報については，パターンマッチング[*9]により訂正情報に含まれる流言部分を抽出し，「登録日時」「ID」「流言訂正情報」「流言内容」「訂正数」を流言情報管理機能により蓄積する．

(4) 手順 (3) で流言部分を特定できなかった場合は，抽出した訂正情報をクラスタリングツール[*10]を用いて分類し，各分類結果について，最も所属度の高い訂正情報を代表として「登録日時」「ID」「流言訂正情報」「訂正数」を流言情報管理機能により蓄積する．なお，手順 (4) では「流言内容」は登録しない．

一方，上記の流れで自動的に蓄積された流言訂正情報には，誤って訂正情報と判定されたものや，正しく訂正情報と判定されていないものが含まれる可能性もある．そこで，人手による流言情報の精査を可能にし，提供する情報の信頼性を向上させることができるようにする．人手で登録または修正が行われた流言情報については，流言情報としての信頼度を高く設定することにより，精査が行われていない情報との区別ができるようにする．

4.2.2 拡散防止フェーズ

蓄積した流言情報を用いて，Twitter や電子メールなどでの誤った情報の拡散を防止できるようにする．流言情報クラウドは，ユーザアプリケーションから受け取った入力テキストに関連する流言情報を提供する．

情報提供の流れを以下に示す．

(1) ユーザアプリケーションからテキストを受け取る．

(2) 訂正情報判定機能により，受け取ったテキストが訂正情報かどうかを判定する．

(3) 流言情報管理機能により，テキストに流言情報が含まれるかどうかを調べる．

(4) 訂正情報ではなく，流言情報が含まれる場合，ユーザアプリケーション上でユーザに警告を行う（**図 3**）．

5. 訂正情報分類器の構築

本章では，流言情報クラウドの構成機能の 1 つである訂正情報分類器について述べる．

[*9] パターンマッチングでは，“「～」というデマ”，“『～』っていうデマ” など，かぎ括弧や特定するためのフレーズを組み合わせた 27 種類のパターンを利用している．

[*10] http://code.google.com/p/bayon/

表 5 訂正情報コーパス
Table 5 Examples of rumor-correction information corpus.

正例（+1）／負例（−1）	ツイート
+1	千葉の C 石油、有害な雨が…の件、デマ確定です。拡散しないようにご注意。→【東北地方太平洋沖地震】C 石油、「有害物質が降る」メールに注意呼びかけ
+1	近畿の地震デマだったんだ～複雑だけどよかった
+1	千葉の有害雨もプレート型による深夜の地震もデマか
−1	デマゴギーって何？デマの省略前の言葉？
−1	なにかデマ騒動があったのかな？
−1	明らかなデマであったなら、論外だけど、そうでないんだから、頭使えよ！、ってかんじだよね。

表中のツイート例は，実在の名称をイニシャルで提示している．

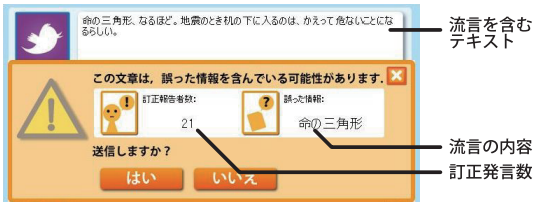

図 3 流言拡散防止のイメージ
Fig. 3 Image of prevention of rumor-spreading.

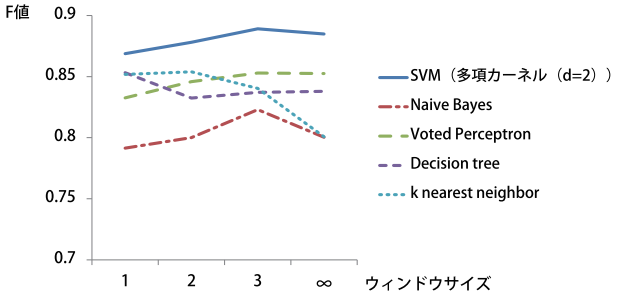

図 4 各学習アルゴリズムによる判定精度
Fig. 4 Accuracy of each learning algorithm.

5.1 コーパス

今回は，以下の 2 種類のデータを用いてコーパスを構築した．

平常時データ：2010 年 3 月のツイート

災害時データ：2011 年 3 月のツイート

今回は，流言マーカを「デマ」とし，訂正情報の抽出を試みる．まず，各データから「デマ」という表現を含むツイートを無作為にそれぞれ 1,000 件抽出した．次に，表 4 に示した判定条件に基づき訂正情報かどうかを人手で判定し，コーパスとした．コーパスの一部を**表 5** に示す．ツイートが訂正情報である場合は正例，そうでない場合は負例とした．平常時データにおける正例の数は 1,000 件中 187 件，災害時データにおける正例の数は 1,000 件中 602 件である．

5.2 訂正情報分類器と学習アルゴリズム

5.1 節で述べたコーパスを用いて，テキストの内容が訂正情報であるかを判定する分類器を構築した．

5.2.1 素性

今回は，素性として以下の項目を用いた．

- 流言マーカ「デマ」の周辺文脈[*11]
- 形態素数
- URL の有無
- 引用（RT @）の有無

[*11] 予備実験により検証した結果，周辺文脈として用いる形態素の品詞の違いを考慮した場合とそうでない場合との精度に大きな違いは見られなかったため，本手法においては品詞の違いを考慮していない．

素性とする流言マーカの周辺文脈の適切な大きさを調査するために予備実験を行ったところ，流言マーカの両側の周辺文脈のウィンドウサイズを 1～3（形態素数[*12]）とした場合[*13]に，比較的精度の良い結果が得られた．そこで，これ以降はウィンドウサイズを 1～3 とし，精度検証を行う．

5.2.2 学習アルゴリズム

分類器構築に用いる学習アルゴリズムを選択するために，学習アルゴリズムの精度を比較した．今回は，**図 4** に示す 5 種類のアルゴリズムを比較した[*14]．

検証結果を図 4 に示す．図 4 より，SVM が最も高い精度を示した．そこで，以降の検証においては，SVM を用いることとする．

6. 実験

6.1 概要

構築した訂正情報分類器の精度を検証するため，5.1 節で述べたデータを用いて 2 種類の実験を行う．それぞれの実験において，以下の内容を検証する．

実験 A：訂正情報分類器によって，訂正情報を判定できるか？

実験 B：平常時のデータを用いた分類器は，災害時にも性

[*12] 形態素解析には JUMAN を用いた．
[*13] ツイート中に 2 回以上「デマ」が出現する場合，2 回目以降の周辺文脈は素性として用いないこととした．
[*14] 検証においては，TinySVM およびデータマイニングツールである WEKA を利用した．それぞれ，パラメータはデフォルト値を用いた．

能を発揮できるか？

流言の発生には，災害などの社会状況も関わっており，特に災害時には流言が発生しやすいと考えられるが，平常時でも流言情報は発信されている．そこで，上記 2 種類の実験を行うことにより，提案手法が平常時・災害時のいずれの場合にも対応可能であるかを検証する．

6.2 実験結果と考察

6.2.1 分類器による判定精度

訂正情報分類器によって，訂正情報を判定できるかどうか（実験 A）を検証するために，各データを用いた 10 分割交差検定を行った．結果を**図 5** に示す．図 5 より，各データにおいておおむね良好な結果が得られた．

次に，平常時のデータを用いた分類器が，災害時にも性能を発揮できるかどうか（実験 B）を検証するために，一方のデータをトレーニングデータにした場合の精度を確認した．結果を**図 6** に示す．まず，平常時データをトレーニングデータとした災害時の訂正情報の判定精度を見ると，ウィンドウサイズ 1 の場合に最も精度の良い F 値 0.841（適合率 0.819，再現率 0.864）が得られ，比較的高精度に判定できることが分かる．一方，災害時のデータに基づく平常時の訂正情報の判定精度を見ると，最も精度の良い F 値がウィンドウサイズ 1 の場合の 0.667（適合率 0.608，再現率 0.738）であり，平常時のデータに基づく判定精度を下回った．

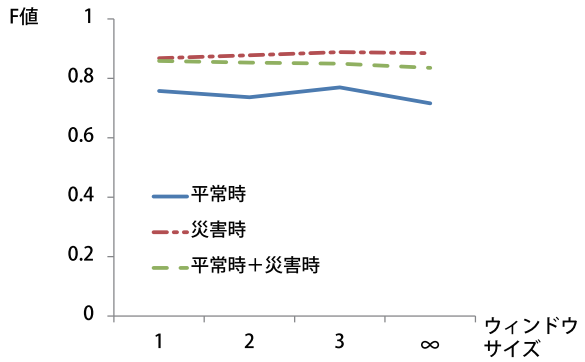

図 5 各データの 10 分割交差検定結果

Fig. 5 Result of 10-fold cross-validation.

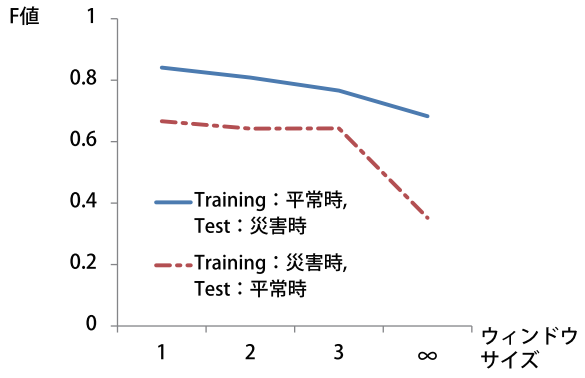

図 6 各データによる判定精度

Fig. 6 Result of validation using other data.

どちらの場合でも最も良い精度が得られたウィンドウサイズ 1 における判定結果を**表 6** に示す．表 6 より，正しく訂正情報と判定されなかったもの（未検出）は，平常時データがトレーニングデータの場合に 13.6%（正例 602 件中 82 件），災害時データがトレーニングデータの場合に 26.2%（正例 187 件中 49 件）であった．また，誤って訂正情報と判定されたもの（誤検出）は，平常時データがトレーニングデータの場合に 28.9%（負例 398 件中 115 件），災害時データがトレーニングデータの場合に 10.9%（負例 813 件中 89 件）であった．判定に失敗したツイートの例を**表 7** に示す．判定に失敗した原因としては，(1) トレーニングデータ中にその周辺文脈素性が存在しない，(2) トレーニングデータ中にその周辺文脈素性は存在するが，テストデータとは正負が逆である，ということが考えられる．表 7 に示した事例のうち，平常時 (a)，(c)，災害時 (a)，(c) はいずれもトレーニングデータ中にその周辺文脈素性が存在しておらず，正しく判定できなかった可能性がある．また，表 7 の，平常時 (b)，災害時 (b) は訂正情報であると判定すべきツイートであるが，トレーニングデータ中では同じ周辺文脈素性が負例として存在したため，訂正情報と判定されなかったと考えられる．同様に，平常時 (d)，災害時 (d) は訂正情報ではないと判定すべきだが，トレーニングデータ中では同じ周辺文脈素性が正例として存在したため，訂正情報として誤検出されたと考えられる．上述したように，災害時のデータに基づく平常時の訂正情報の判定精度は平常時のデータに基づく判定精度を下回った．ウィンドウサイズ 1 における周辺文脈の異なりパターン数を確認したところ，平常時データは 514 パターン，災害時データは 369 パターンであった．つまり，平常時のデータは災害時のデータと比較して出現する周辺文脈が多様であり，トレーニングデータとした災害時データには出現しない表現が多いため，判定精度が低下した可能性があると考えられる．

表 6 ウィンドウサイズ 1 での判定結果

Table 6 Results of judgment in window size 1.

実験条件	判定成功		未検出	誤検出
	p/p	n/n	p/n	n/p
Training：平常時 Test：災害時	520	283	82	115
Training：災害時 Test：平常時	138	724	49	89

表中の p は正例を，n は負例を表し，それぞれ以下を意味する．

- p/p：正例を正例と判断
- n/n：負例を負例と判断
- p/n：正例を負例と判断（未検出）
- n/p：負例を正例と判断（誤検出）

表 7 ウィンドウサイズ 1 の場合の判定失敗例
Table 7 Examples of failure in window size 1.

コーパス		正解/判定結果	ツイート
平常時	(a)	p/n	苦しい…甘いものが別腹 なんてデマ流し たの誰だよ！！
	(b)	p/n	テレビ局 N ツイッター禁止 、デマだった ?広めてすいません。陳謝。・・・でもありえる話だなあと思った。どの会社も上の人は何するかわからんし。
	(c)	n/p	RT @xxxx 昨日の K 氏引退報道で思ったけど、ツイッター ってデマに 弱いメディアだと感じた。2ちゃんねるよりもソースの確認がしづらいので、ツイッターしかやってない方は特に気をつけた方がいいでしょう。
	(d)	n/p	@xxxx 危篤っつーのは聞いてたからさ。状況変わってねーんだな。それにしても、twitterのデマのバズりやすさは異常。バズってる話題には乗らないことを改めて胸に誓いたい。
災害時	(a)	p/n	アメブロで例の「自衛隊が支援物資の募集してます」デマが まだ流れてます。10 代～20 代の利用者が多いブログサービスなので広がり方がハンパないです；； もし見かけたら総務省＜ URL ＞東北地方太平洋沖地震に関するチェーンメール等にご注意ください。を！
	(b)	p/n	吉野家あいてた。xxx 情報サンクス。ボンド前で大将と遭遇。パチンコは出たのだろうか…ww それにしても、関東の地震情報 がデマで 良かった…。こんな状況だから情報の取捨選択が難しいね。
	(c)	n/p	RT @xxxx: #j_j_helpme ←地震被害で身動きできない人のタグです！ ！デマ・ ガセは絶対に流さないで下さい！！
	(d)	n/p	地震で怖いの はデマな んだよね…

表中の p は正例を，n は負例を表し，それぞれ以下を意味する．
下線部は，ウィンドウサイズ 1 での周辺文脈である．
表中のツイート例は，実在の名称をイニシャルで提示している．

これらの結果から，判定の失敗は発生しうるものの，災害時の訂正情報は平常時のデータをトレーニングデータとすることでおおむね高精度に判定可能であり，平常時を含めたデータを用いることで，平常時・災害時のどちらでも高精度な判定が期待できる．

6.2.2 流言情報自体の判定との比較

本研究では，訂正情報の検出により間接的に流言を収集するアプローチを採用した．本項では，流言情報を直接検出する場合との違いについて述べる．

まず，単純に既知の流言情報を用いて，異なる流言情報が検出できるのかを調査する．流言情報として，東日本大震災発生後に Twitter 上で拡散された 6 種類の流言情報（**表 8**）を用いて，流言情報の分類器を構築した．なお，それぞれの流言情報に含まれる内容は異なるため，学習時の素性として，ツイートに含まれるすべての形態素および 5.2.1 項で述べた素性（流言マーカの周辺文脈は除く）を用いた．

6 種類の流言情報のうち，5 種類を学習データとし，残りの流言を判定できるかどうかを検証した結果，F 値の平均は 0.283 となった．表 8 に示した流言情報を見ると，すべての流言情報に共通するような，特徴的な表現は見られない．また，流言によって拡散の規模は異なり，各流言に関して十分な量のツイートが集まっているとはいえない．そのため，異なる流言情報を用いても，流言情報かどうかを精度良く判定することができなかったと考えられる．

次に，流言情報自体の検出を行っている先行研究について述べる．Rattanaxay らは，流言情報自体の検出を試みている [12]．この研究では，「らしい」といった，流言に含まれる曖昧な表現に着目し，分類器の構築を行っている．Rattanaxay らの実験結果によると，流言の検出精度（F 値）は 0.49 [*15] となっている．また，この手法では，曖昧な表現を含む情報は，たとえ正しくても流言と見なしている．そのため，ユーザに流言として提示すべきでない情報も，流言として判定してしまうという問題がある．また，上述したように，流言には必ずしも曖昧な表現が含まれるわけではない．たとえば，表 8 における流言 A，D，E は曖昧な表現が含まれておらず，Rattanaxay らの手法によって正しく判定できるかどうかは分からない．

このように，流言情報を直接検出することは容易ではない．流言情報自体を検出するタスクと比較して，本提案手法による訂正情報を検出するというタスクは比較的高精度に判定が可能なタスクであると考えられる．ただし，本提案手法と流言情報の直接検出とでは，検出対象が異なり，本提案手法では訂正情報の発信されていない流言情報の抽出はできない．流言情報の直接検出手法と，本提案手法を併用することによって，より広範囲な流言情報の収集ができる可能性がある．

7. 本研究の拡張性および限定性

本論文では，流言情報収集の手法として，訂正情報に着目し，流言マーカを含む訂正情報を分類することで，間接的に流言情報を収集する手法を提案した．

4 章では，流言拡散防止のための提供先アプリケーショ

[*15] 文献 [12] 中の実験結果をもとに算出した．

表 8 流言情報の例
Table 8 Example of rumor information.

種類	ツイート例	ツイート数
流言 A	・地震が起きた時、社内サーバールームにいたのだが、ラックが倒壊した。腹部を潰され、血が流れている。	21
流言 B	・命の三角形、ためになりました。地震では机の下にすぐ入る事を考えがちだけど、机等のすぐ横のほうがつぶされない三角形の空間になり、助かる可能性が高いんだね。 ・「ほとんどの人は、地震のとき、四つんばいになったりして机やテーブルの下にもぐりこむが、これでは、崩落したときには、助からない」命の三角形: Voila la vie en rose!	84
流言 C	・【関西地区の皆さん】K 電力が電力の提供を始めたようなので、コチラで節電すれば立派な支援になります。出来るかぎり節電を心がけましょう。 ・K 電力の友人からのお願いです。「東北地震へ電力提供を始めました。少しの節電でも立派な支援になります。電子レンジや炊飯器など、普段さしっぱなしのコンセントを今日だけでも抜いて、節電のご協力をお願いします」	86
流言 D	・漫画 O の作者　O 氏、地震の被害者救済に 15 億円を寄付　「自分が幸せになったということは、世の中から受けたひとつの借りだ」	5
流言 E	・「K さんはとてもラッキーな人」M 党の S 前・官房長官が東北大地震を「ラッキー」と表現。	15
流言 F	・近畿のプレートが小さくなっている模様。これが元に戻ろうとすれば次は近畿に大きな地震が起きる可能性が非常に大きいので明日、明後日は注意してください。	5

表中には，テキスト中の表現の差分が大きいものの一部を提示している．なお，表中の流言に関するツイートを抽出する際，投稿者のコメントが追加されていないリツイート（「RT @」から始まるツイート）は，内容が他のツイートと重複するため除外した．
表中のツイート例は，実在の名称をイニシャルで提示している．

ンとしては，Twitter だけでなく，メールクライアントなどのテキストベースのアプリケーションを想定していることを述べた．流言情報クラウドにおける訂正情報分類器は，5 章で述べたように，Twitter から抽出したツイートを用いて構築しているが，表 5 に示したように，流言を訂正するツイートの場合，それほど文法的に崩れた文ではない．また，6.2.1 項で示したように，流言マーカ「デマ」の周辺文脈のウィンドウサイズを 1 とした場合に最も良い精度が得られており，判定において重要な素性は，流言マーカ「デマ」の直近のわずかな形態素パターンであると考えられる．そのため，用いられるテキストの傾向が大きく異ならない場合は，メールクライアントなど，テキストベースのアプリケーションにおけるテキストに対しても一定の精度で適用できると考えられる．

ただし，本手法で検出対象としている訂正情報は，3.3 節で述べた否定戦略 1 の訂正情報のみである．そのため，否定戦略 1 については比較的高精度に判定・収集可能であると考えられるが，流言マーカを含まない訂正情報や，訂正情報が発信されていない流言情報を本手法により収集することは困難である．

また，本論文で得られた結果は，否定戦略 1 を訂正情報とした場合の結果であるが，3 章で述べたように，方針の検討にあたり分析したデータは 8 種類と限定的であるため，否定戦略 1 の発信されていない流言も存在する可能性がある．否定戦略 1 を用いて十分に流言情報を収集できるかどうかは，今後検証する必要がある．

8. おわりに

本研究では，流言拡散を防ぐための仕組みとして，訂正情報に基づいた流言情報クラウドを提案した．流言情報の収集リソースとして Twitter を用いることとし，平常時および災害時のデータを用いて，マイクロブログ上での流言への対処方法について調査した．また，調査結果に基づき，流言情報クラウドを構成する機能の 1 つである，訂正情報分類器を構築した．本研究では以下の点を明らかにした．

(1) 流言の内容によって，投稿される訂正情報の傾向は異なるが，流言マーカ（「デマ」「間違い」「嘘」のような流言であることを示す表現）を含む形での訂正は，多くの流言への対処方法として用いられる可能性がある．
(2) 災害時の訂正情報は平常時のデータをもとに判定可能であり，平常時を含めたデータを用いることで，平常時・災害時のどちらでも高精度な判定が期待できる．
(3) 流言情報自体を判定することと比較して，訂正情報の判定は容易である可能性がある．

今後は，流言情報クラウドを実際に運用し，蓄積される流言情報の検証や流言情報の提供によるユーザへの影響，課題などを明らかにする．また，本論文では，訂正情報に含まれる流言情報の抽出については，パターンマッチングによる抽出または人手による抽出を想定しているが，流言情報の抽出手法については，より高精度な抽出技術の検討が必要である．

謝辞 本研究のデータ分析において，貴重かつ膨大なデータ（平常時）を提供していただいた兼山元太氏（クックパッド）に深く感謝する．本研究の一部は，JST 戦略的

創造研究推進事業および JSPS 科研費 24500134 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] 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).
[3] 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 (2011).
[4] 宮部真衣，荒牧英治，三浦麻子：東日本大震災における Twitter の利用傾向の分析，情報処理学会研究報告，グループウェアとネットワークサービス研究会，Vol.2011-GN-81, No.17, pp.1–7 (2011).
[5] 荻上チキ：検証　東日本大震災の流言・デマ，光文社新書 (2011).
[6] 川上善郎：うわさが走る—情報伝搬の社会心理，サイエンス社 (1997).
[7] Sakaki, T., Okazaki, M. and Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors, *Proc. 19th International Conference on World Wide Web* (*WWW '10*), pp.851–860 (2010).
[8] Abel, F., Hauff, C., Houben, G.J., et al.: Twitcident: Fighting Fire with Information from Social Web Stream, *Proc. International Conference on Hypertext and Social Media*, pp.305–308 (2012).
[9] Aramaki, E., Maskawa, S. and Morita, M.: Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter, *Proc. 2011 Conference on Empirical Methods in Natural Language Processing* (*EMNLP2011*), pp.1568–1576 (2011).
[10] Paul, M.J. and Dredze, M.: You Are What You Tweet: Analyzing Twitter for Public Health, *Proc. 5th International AAAI Conference on Weblogs and Social Media* (*ICWSM 2011*), pp.265–272 (2011).
[11] Qazvinian, V., Rosengren, E., Radev, D.R., et al.: Rumor has it: Identifying Misinformation in Microblogs, *Proc. 2011 Conference on Empirical Methods in Natural Language Processing* (*EMNLP2011*), pp.1589–1599 (2011).
[12] Rattanaxay, K.，相田　慎，青野雅樹：ツィッターのデマ率の推定，情報処理学会第 74 回全国大会，第 2 分冊，pp.523–524 (2011).
[13] 小林啓倫：災害とソーシャルメディア—混乱，そして再生へと導く人々の「つながり」，毎日コミュニケーションズ (2011).

推薦文

流言（裏付けを持たない自然発生的な情報）の拡散防止のために，流言を含む可能性のある情報を自動的に検出し，その旨をユーザに警告する仕組みを提案し，そのための流言の分類器を実現・評価した興味深い論文であり，推薦論文に値する．

（グループウェアとネットワークサービス研究会主査　小林　稔）

宮部 真衣　（正会員）

1984 年生．2006 年和歌山大学システム工学部デザイン情報学科中退．2008 年同大学大学院システム工学研究科システム工学専攻博士前期課程修了．2011 年同大学院システム工学研究科システム工学専攻博士後期課程修了．博士（工学）．2011 年東京大学知の構造化センター特任研究員．現在，京都大学学際融合教育研究推進センターデザイン学ユニット特定研究員．コミュニケーション支援に関する研究に従事．

灘本 明代　（正会員）

東京理科大学理工学部電気工学科卒業．2002 年神戸大学大学院自然科学研究科後期博士課程修了．博士（工学）．現在，甲南大学知能情報学部教授．Web コンピューティング，データ工学の研究に従事．ACM, IEEE 各会員．

荒牧 英治　（正会員）

1974 年生．2000 年京都大学総合人間学部卒業．2002 年同大学大学院情報学研究科修了．2005 年東京大学大学院情報理工系研究科修了（博士；情報理工）．2006 年東京大学医学部附属病院特任助教，2009 年東京大学知の構造化センター特任講師を経て，現在，京都大学学際融合教育研究推進センターデザイン学ユニット特定准教授．医療分野の言語処理研究に従事．