Vol.2013-NL-214 No.14
2013/11/15

マイクロブログ上の告知投稿に対する非明示的な関連投稿の収集

塚本 悠馬[1,a)]　笹野 遼平[2,b)]　高村 大也[2,c)]　奥村 学[2,d)]

概要：近年，Twitter をはじめとするマイクロブログを利用した商品やイベントの告知に対し，多くのユーザが感想など告知の投稿者やその告知への感想に関心があるユーザにとって有益な関連投稿を行うようになってきている．しかし，関連投稿の多くは告知投稿とは明示的に関連付けられてはいないため，告知の投稿者がこれらの関連投稿を見つけるのは容易ではない．そこで本研究では，特に Twitter の機能であるリツイートに注目し，告知に対する非明示的な関連投稿を効率的に収集する手法を提案する．

キーワード：マイクロブログ

Collecting Microblog Posts Implicitly Related to an Announcement Post

Abstract: Events, festivals or product releases are often announced via microblogs such as Twitter. Many users post messages that are relevant to the announcements. Such relevant posts are useful for both the authors of the announcements and the users who are interested in public opinions to the announcements. However, it is not easy to collect such relevant posts because many of the relevant posts are not explicitly associated with the announcement post. In this paper, we propose a method for efficiently collecting such posts that are only implicitly relevant posts to announcements, focusing on retweets of Twitter in particular.

Keywords: Microblog Service

1. はじめに

近年，Twitter[*1] をはじめとするマイクロブログを利用した商品やイベントの告知が増えている．たとえば Twitter では図 1 に示すような告知投稿が日々行われている．また，その告知に対する反応として多くのユーザが感想などのその告知に関連した投稿を行っている．告知に対するユーザの反応には告知に注目した理由など告知の感想を知りたい者にとって有益な情報が含まれているため，告知に対する反応を知ることは告知の投稿者やその告知への感想に関心があるユーザにとって有用であると考えられる．そこで，本研究では告知に対する関連投稿を自動収集することを目指す．

関連投稿には Twitter におけるリプライ機能のように告知投稿と明示的に関連付けられているものと，非明示的なものがある．このうち，明示的なものは収集が容易であるものの，その数は限られており，告知投稿との関連が明示的なもののみを扱った場合，十分な数の関連投稿を収集できない．このため，情報をより多く収集するには非明示的な関連投稿も収集することが必要となる．告知投稿との関連が非明示的な関連投稿を収集するためには，無数の投稿の中から関連する投稿を見つけ出す必要があり，非明示的な関連投稿を網羅的に収集することは困難であると考えられる．しかし，非明示的な関連投稿の中には，共有機能を使って対象の告知投稿を共有した後に告知の関連投稿を行うものが多く存在しており，収集の対象をこのような関連投稿に限定することで効率的に関連投稿の収集を行うことが可能となると考えられる．本研究では，告知との投稿時

[1] 東京工業大学 総合理工学研究科
[2] 東京工業大学 精密工学研究所
[a)] tsukamoto@lr.pi.titech.ac.jp
[b)] sasano@pi.titech.ac.jp
[c)] takamura@pi.titech.ac.jp
[d)] oku@pi.titech.ac.jp
[*1] http://twitter.com/

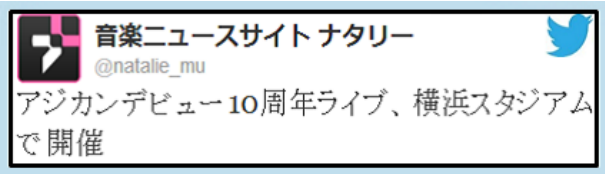

図 1　Twitter における告知投稿の例

間差や関連投稿同士の類似性等に注目することで，告知投稿を共有した直後の投稿の中から告知投稿に対する関連投稿を自動収集する．

本研究では，特に Twitter を対象に，他者のツイートを自分をフォローしているユーザと共有する機能であるリツイートに注目し，リツイート直後のツイートがそのツイートと関連したツイートかどうかを識別する技術を確立することで，告知投稿と関連する投稿の効率的な収集システムの構築を目指す．

2. 関連研究

本研究ではある告知投稿から，その告知に対する反応を表現した関連投稿を収集する．TDT (topic detection and tracking) タスク [1] は，時間軸上に整列した文書群から同じトピックに関する文書を見つけることを目的にしており，関連性をもつ文書を見つけるという点で本研究と関連がある．Twitter を利用した研究としては，ある組織に関する投稿からその組織に関する新しいトピックを特定する Yan ら [2] の研究がある．しかし，その目的は同じトピックに関する投稿のクラスタリングやトピックの抽出が目的であるため，関連投稿の自動収集という本研究の目的とは異なる．

また，Deepak ら [3] は Twitter 上のある投稿に関連する投稿の収集を目指した．この研究ではあるツイートを入力とし，余弦類似度などを用いてそのツイートに内容が類似したツイートをランキング形式で出力している．しかし，本研究で収集する関連投稿は告知投稿に対する反応を表現したものであり，告知を目的とする告知投稿とは性質が異なっているため，告知投稿の語句を利用した類似度では本研究で収集を目指す関連投稿の収集は難しい．

ある投稿に対する反応を表現したツイートの収集を目指した研究として Kothari ら [4] がニュース記事に対するコメントを表現した投稿の自動収集を行っている．しかし，Kothari らは，自動収集の対象をニュース記事へのリンクや記事の一部を含むツイートに限定している．そのため，分類対象は元のニュース記事との関連性が明示されているという点で関連性が非明示的な投稿を対象とする本研究とは異なっている．

本研究では告知投稿に関連する投稿の自動判定を行うが，ある文書と関連する文書の自動判定を行う研究としては，真野ら [5] が新聞記事を元にその記事で取り上げられた技術に関する特許公報の検索を行っている．真野らは検索条件として入力する文書 (新聞記事) と検索対象としている文書 (特許公報) の間で使用されている語彙が異なる点を考慮して検索に用いる語を決める手法を提案している．また，ニュースサイトの記事とブログ記事の対応付けを行った研究としては池田ら [6] の研究がある．池田らは有名なニュースに関するブログ記事は対応するニュース記事の内容を省略する傾向があることを指摘し，そのことを考慮した特徴語の重み付けを提案している．本研究で収集を目指す関連投稿も直前に告知投稿を共有しているため，告知投稿の内容を省略する傾向がある．そのため，本研究の分類には告知投稿の内容だけではなく，告知投稿の内容が省略された場合を考慮した情報も必要になる．

3. 提案手法

3.1 手法概要

提案手法の概要を図 2 に示す．提案手法では，まず告知投稿を入力とし，その投稿をリツイートしたユーザのリツイート直後のツイートを収集する．本研究では Twitter の retweeters API[*2] を使用して告知投稿をリツイートしたユーザを収集した後，その各ユーザのリツイート直後のツイートを Twitter の home_timeline API[*3] を使用して収集した．次に，二値分類器によって収集したリツイート直後のツイートそれぞれが告知投稿の関連投稿かどうかを判定する．関連投稿であると判定されたリツイート直後のツイートを収集することで告知投稿に対する関連投稿を自動収集する．[*4]

分類に利用した素性は**告知投稿と判別対象の投稿のみから得られる素性**と，判別対象の投稿の集合など**他のユーザの投稿も考慮した素性**の 2 つに大きく分けられる．まず，告知投稿と判別対象の投稿から得られる素性として告知投稿に出現する語句 [*5] や一般的に関連投稿によく見られる特徴から考案した素性を説明する．しかし，告知投稿と判別対象の投稿だけから得られる情報では分類に利用できる語句情報が限られ，また，関連投稿は告知投稿との重複内容を省略したものも多いため，十分な分類性能が得られないことが考えられる．そのため，本研究では告知投稿や判別対象の語句だけではなく，他のユーザの投稿を考慮した素性を導入することでさらなる分類性能の向上を目指す．これらの素性は，図 2 での関連投稿にはライブの 2 日間通し券に言及しているものが複数あるなど，他のユーザの投稿には関連投稿かどうかの分類に有用な情報が含まれている点に注目して考案したものである．以下では，これらの素性をそれぞれ**局所的素性**，**大域的素性**と呼ぶ．

[*2] https://api.twitter.com/1.1/statuses/retweets/:id.json
[*3] https://api.twitter.com/1.1/statuses/home_timeline.json
[*4] 本研究では分類器に SVM を用いた．
[*5] 以降，すべて語句としては名詞，形容詞，動詞を用いる．

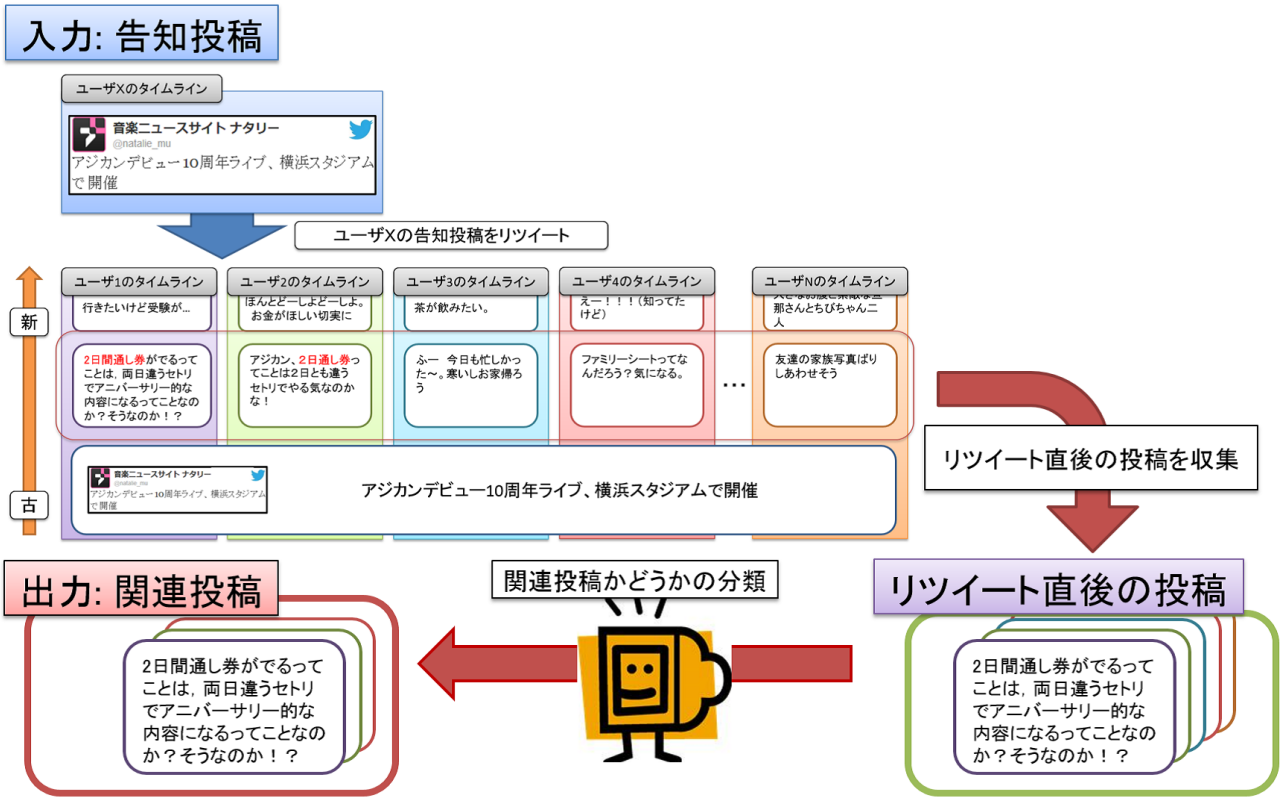

図 2　本研究の概要

3.2　局所的素性

本研究では局所的素性として以下の 3 つの素性を用いる．

告知投稿と判別対象の対から得られる語句情報

告知投稿内に出現する語句やそれらの語句と頻繁に共起する語がリツイート直後のツイートに多く含まれている場合，告知投稿とリツイート直後のツイートは関連している可能性が高いと考えられる．また，関連ツイートは告知に対する感想を含むものが多いと考えられるため，判別対象に “すごい” などの評価表現が出現したかどうかという情報も分類に役立つと考えられる．このように告知投稿の語句情報から得られる語句情報や，評価に多く使われると考えられる語が含まれるかといった判別対象の性質を考慮した語句情報は関連投稿かどうかの判定に役立つと考えられる．この点を考慮し，本研究では素性として告知投稿内に出現する語句との一致数，内容語の極性の最大値と最小値，告知投稿内の語句の共起語との一致数を利用する．各内容語の極性に関しては高村ら [7] の極性辞書を用いた．共起語に関しては，2012/2/1 から 2012/2/29 の期間に Twitter の Streaming API で収集したツイートデータ内で，告知投稿内の名詞と同じツイートに出現する回数が上位 10 位以内の名詞を用いた．

また，告知投稿は主に企業の公式アカウントが行うため，文体が比較的整っているのに対し，その関連投稿は投稿者が一般ユーザであるためにマイクロブログ特有のくだけた表現を含むことが多い．そのため，告知投稿は形態素解析が比較的成功しやすく，リツイート直後のツイートは形態素解析が失敗しやすい傾向がある．この告知投稿と判別対象間の性質の違いを考慮し，告知投稿から得られた形態素と分類対象ツイートの形態素の類似度が高い場合，それらを同じ形態素であるとみなし，類似度が高い形態素を含むかどうかという素性として関連投稿を判定する素性に利用する．

告知投稿と判別対象の投稿との投稿時間差

告知投稿とリツイート直後のツイートの投稿時間差も関連投稿であるかの判別に有用である．告知投稿に対する関連投稿は，告知投稿と連続して投稿されるものも多く，告知投稿のリツイートと関連投稿との投稿時間差は小さいことが多い．そのため，投稿時間差が小さい場合はリツイート直後のツイートが関連投稿である可能性が高く，投稿時間差が大きい場合は関連投稿である可能性が低い．本研究ではこのことを考慮し，告知投稿と判別対象の投稿との投稿時間差も関連投稿を判定する素性として利用する．

判別対象から得られる言語的特徴

(1)　うおおお土日行けるのかこれはあああああ

(2)　うおおおおお！横スタ行きたい！！

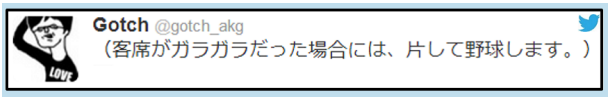

図 3 図 1 の投稿に対する共起リツイート

関連投稿の中には上記の例のように語尾や“！”などを重ねて興奮や衝撃などを表現しているツイートが見られる．このような関連投稿を判定する素性として，最後の文字と同じ数が 3 以上，連続した母音の数が 3 以上，ツイートの長さが 5 以下などの条件と“！”の出現数を利用する．

3.3 大域的素性

本研究では大域的素性として以下の 3 つの素性を用いる．

関連投稿間の類似性

同一の告知投稿に対する関連投稿の中には類似した反応を表現した関連投稿が存在する場合が多い．

(3) 2 日間通し券がでるってことは，両日違うセトリでアニバーサリー的な内容になるってことなのか？そうなのか！？

実際に，図 1 に示したライブの告知投稿の関連投稿には 2 日間通し券に言及している上記の例の投稿が存在し，この投稿以外にも 2 日間通し券に言及している関連投稿が複数存在した．このように，異なる関連投稿が同じ事柄に対して言及することは多く，その事柄は告知投稿には含まれていないことも多い．この点を考慮し，本研究ではリツイート直後のツイート集合内の頻出語 [*6] との一致数，一致した語の頻度順位を素性として利用する．

共起リツイートの語句情報

告知投稿のリツイートと共起して多くリツイートされるツイートはその内容が告知投稿に関連していると考えられる．たとえば，あるイベントの告知投稿では一緒にそのイベントの参加者の投稿がよくリツイートされる．実際に，図 1 にに示すライブの告知投稿の関連投稿の共起リツイートには図 3 の投稿が存在する．この共起リツイートの投稿者は告知されたライブのバンドメンバーであり，告知投稿と関連している．本研究ではこれらを**共起リツイート**と呼び，その語句情報を利用する．本研究では，各ユーザが告知投稿の投稿時間の前後 6 時間以内に投稿したリツイートを収集し，その頻度上位 5 位以内かつ 3 回以上のリツイートを共起リツイートとする．関連投稿かを分類する素性として，共起リツイート内の頻出語との一致数を利用する．

“＞ RT” を伴うツイートの語句情報

(4) 21 世紀に出てきたバンドでは一番好きかも ＞ RT

[*6] 以降，すべて頻出語としては頻度上位 10 語を用いる．

図 4 図 1 の投稿に対する非公式 RT

関連ツイートの中にはその前のリツイートに対する言及であることを明示するために“＞ RT”といった表記を伴っているものがある．このため，“＞ RT”を含むツイートは関連投稿である可能性が高いことから“＞ RT”を含むツイートの語句情報が有用となる場合があると考えられる．そこで本研究では，“＞ RT”を含むツイートの頻出語との一致数とその頻度順位を関連投稿を判定する素性として利用する．

明示的な関連投稿の語句情報

関連投稿の中には Twitter におけるリプライのように告知投稿との関連性が明示的なものが存在する．明示的な関連投稿は非明示的なものに比べて収集が容易である場合が多く，含まれる語は告知投稿に関連する語句である可能性が高いため分類性能の向上に有用な情報になり得る．そのため，これら明示的な関連投稿に含まれる語句情報を素性として利用する．本研究では明示的な関連投稿としてリプライと非公式 RT を使用する．リプライはある記事に対する Twitter 公式の返信機能であるため，Twitter 内部でその記事との関連性が保持されている．非公式 RT とは，図 4 のように元の投稿全体，RT という文字列および引用元を投稿したユーザ名とともに引用されている投稿のことである．リプライとは違い，Twitter 内部ではその記事との関連性は保持されていないが，引用部分があるために元の投稿との関連性は明示的であり，また，告知投稿を検索クエリとすることで容易に収集可能である．各告知投稿に対するリプライと非公式 RT はそれぞれ Twitter 提供の API を利用して集めた．その収集したリプライと非公式 RT は共に告知からの引用部分以外の内容語を利用した．収集したリプライと非公式 RT の集合それぞれの頻出語との一致数とその語の該当順位を素性として利用する．

4. 実験

4.1 実験設定

提案手法の分類結果を評価するため，15 個の告知投稿のリツイート直後のツイートを収集し，これらのツイートが関連投稿であるかどうかを人手で分類し，正解データを作成した．評価対象のリツイート直後のツイート 5,736 個のうち，関連投稿は 1,512 個，関連投稿でない投稿は 4,224 個であった．表 1 に実験で使用した告知投稿と，そのリツ

表 1 実験に使用した告知投稿とそれぞれのリツイート直後の投稿に占める関連投稿の割合

投稿者	ツイート	関連投稿の割合	リツイート数
音楽ニュースサイト ナタリー	アジカンデビュー 10 周年ライブ、横浜スタジアムで開催	0.415	948
NHK 紅白歌合戦	【第６３回紅白・出場歌手決定！紅組その１】 aiko、絢香、いきものがかり、石川さゆり、ＡＫＢ４８、 ＳＫＥ４８、きゃりーぱみゅぱみゅ、香西かおり、倖田來未、 伍代夏子、坂本冬美、天童よしみ (続く） ♯ NHK 紅白	0.346	8609
CNET Japan	iOS 版の Google Map 正式版、リリースされました。 記事はもうすぐ。	0.364	164
マクドナルド	本日 2/4（月）6:00〜9:00 の朝マック 「フリーマンデー」キャンペーンは、 人気の「ハッシュポテト」を無料でお試しできちゃう♪ お気に入りの朝マックメニューと一緒に、できたての アツアツをハフハフどうぞ☆ ♯ 朝マック	0.263	262
映画『レ・ミゼラブル』公式アカウント	映画『レ・ミゼラブル』本日 12 月 21 日より公開！ 劇場で待ってまーす！ ♯ レミゼラブル	0.312	409
ビクターエンタテインメント	【サカナクション】3/13 発売のニューアルバムのタイトルが 「sakanaction」に決定♪ジャケット写真も遂に公開だワン！ ♯sakanaction 初回盤・通常盤 J 写はこちら→ http://twitpic.com/c23omx http://twitpic.com/c23one	0.215	340
ソニー・ミュージックレコーズ	ナタリー - YUI が活動休止へ 「ポジティブな想いで決めたこと」 http://natalie.mu/music/news/80191 …	0.409	322
スターバックス コーヒー	本日から、スターバックス コーヒー 目黒店がオープンします。 限定のビバレッジ、マグをご用意して、みなさまのご来店を お待ちしています。 http://sbux.jp/15eAPxO pic.twitter.com/nifVFWwDPO	0.193	780
スターバックス コーヒー	本日から、『コーヒー ティラミス フラペチーノ』 発売です。贅沢なデザート感溢れるフラペチーノ を公式ブログでご紹介します。 http://sbux.jp/12KjjRF	0.251	6579
スターバックス コーヒー	いままで沢山の限定フラペチーノが販売されてきましたが、 みなさんの「お気に入り」、「復活希望」、「思い出の」 フラペチーノを教えてください。 http://sbux.jp/100rPoj	0.179	372
FOOTBALL-STATION.net	ベンゲルｖｓピクシー、師弟対決実現へ - http://bit.ly/11uBdqk 名古屋が７月に豊田スタジアムで名将 ベンゲル監督率いるアーセナルと親善試合を行う方向で 最終調整に入っていることが分かった。 開催日は日本代表の東アジア選手権のため Ｊ１が中断中の７月２３日が有力。ベン	0.285	158
famima_now	大人気ＧＥＬＡＴＯの新作『ピンクグレープフルーツ』 が登場☆果汁７５％を使用し、本物の果実を味わって いるようなフレッシュ感をお楽しみいただけます！ よろしかったらお試しくださいね♪ http://fm.eng.mg/aa686 pic.twitter.com/Vag1jafn18	0.186	66
ワールドサッカーキング	【速報】 マンチェスター・U のファーガソン監督が今季で退任!! http://www.soccer-king.jp/ news/world/eng/20130508/108741.html …	0.176	2920
セブン-イレブン・ジャパン	「濃厚ベルギーチョコまん〜生チョコ仕立て〜（100 円）」 が登場！ 濃厚な生チョコがたっぷり入ったスイーツ系中華まんです♪ http://bit.ly/W47PTG pic.twitter.com/FX5G4Ts7	0.258	220
ケンタッキーフライドチキン	新商品サンド「ケンタッキーチキンライス」！！ 食べごたえたっぷりのボリュームサンドですが… カロリーはたったの 585kcal ！！ お試しくださいね〜(*・ω・*) ノ ♯ 肉米肉 http://twitpic.com/c1qrgq	0.247	739

表 2　局所的素性の各素性における関連投稿の分類結果

	精度	再現率	F 値
告知投稿と判別対象の語句情報	0.652	0.477	0.552
判別対象から得られる言語的特徴	0.651	0.608	0.629
告知投稿と判別対象との時間差	0.558	0.812	0.661
すべての局所的素性を使用	0.797	0.650	0.716

イート直後のツイートにおける関連投稿の割合を示す．音楽に関する告知は，比較的関連投稿が占める割合が高いものが多かった．なお，15 個の告知投稿は投稿者がニュースサイトや企業の公式アカウントであり，かつリツイート数が 50 以上の投稿から人手で選んだ．また，リツイート直後のツイートが別の投稿のリツイートであった場合は評価対象から除いた．これらのデータを用いて，告知投稿 15 個の交差検定を行った．すなわち，15 個の告知投稿のうち 14 個を学習に用いて残り 1 個の告知投稿のリツイート直後のツイートを分類することを 15 回繰り返した．SVM の実装は LIBLINEAR[*7] を用い，パラメータ C は 14 個の学習データのうち 1 個を無作為に選択し、それを開発データとして用いることで決定した．

4.2　局所的素性の実験結果

局所的素性の各素性を利用したときの実験結果を表 2 に示す．すべての素性を使うことで分類性能が向上することが確認できる．すべての素性を用いた場合の分類結果は，精度，再現率， F 値がそれぞれ 0.797， 0.650， 0.716 であった．以下では，各素性の特徴の考察を行う．

告知投稿と判別対象の対から得られる語句情報

関連投稿は告知投稿との重複内容を省略したものも多いため，告知投稿に出現する語句を含まないものも多い．また，Twitter には文字制限があり，告知投稿の内容だけから得られる語句情報には限界がある．そのため，告知投稿と判別対象の対から得られる語句情報だけでは分類に失敗することが多かった．また，映画の題名や紅白歌合戦の出場歌手名など，告知投稿では伝えたい事柄が正式名で記載されていることが多いのに対し，その告知に対する関連投稿では正式名よりも略称がよく使われる傾向があった．このため，正式名は語句情報として得ることが出来たとしても，その略称や愛称を得ることが出来ないために分類に失敗することが多く見られた．たとえば，図 1 で使用されている “横浜スタジアム” という語句は正式名であるが，関連投稿ではその略称である “横スタ” や “浜スタ” が使われることが多かった．

[*7] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

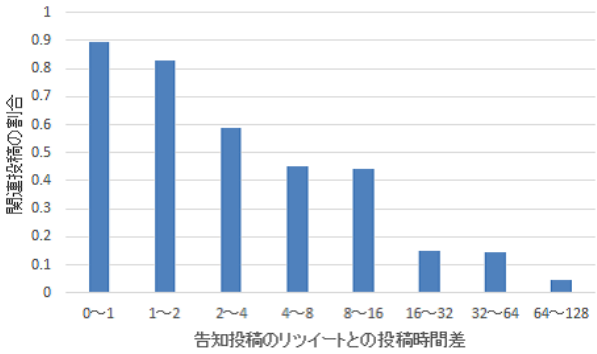

図 5　図 1 の投稿に対する関連投稿の各投稿時間差における割合

表 3　大域的素性における関連投稿の分類結果

	精度	再現率	F 値
局所的素性 (L)	0.797	0.650	0.716
(L) + 関連投稿の類似性	0.855	0.731	0.788
(L) + 共起リツイートの語句情報	0.827	0.689	0.752
(L) + “> RT” の語句情報	0.842	0.729	0.781
(L) + 明示的な関連投稿の語句情報	0.824	0.633	0.716
すべての素性を使用	0.868	0.798	0.832

判別対象から得られる言語的特徴

(5)　うおおおおおまじかあああああああ行く絶対！！てか映像集ほしいよー！

告知投稿と判別対象の対から得られる語句情報のみを用いた場合より，判別対象から得られる言語的特徴のみを用いた場合の方が良い分類結果が得られた．関連投稿には告知に対する反応を表現することが多く，上の例のように告知投稿に出現する語句を含まなくとも母音の重なりや “！” の数が多いものがある．この素性ではそれらの情報を利用することで，このような関連投稿を正しく分類できていた．

告知投稿と判別対象の投稿との投稿時間差

図 1 の投稿に対する正例と負例それぞれについて，各投稿時間差における投稿数を調べた．結果を図 5 に示す．多くの関連投稿は告知投稿と小さい投稿時間差で行われていることが確認できる．この結果が示すように，投稿時間差は関連投稿かどうかの有益な情報となっており，表 2 に示す局所的素性を 1 つずつ用いた実験の中でもっとも良い分類結果になったと考えられる．

4.3　大域的素性の実験結果

結果を表 3 に示す．大域的素性を導入することで分類性能が向上することが確認できる．すべての素性を用いた場合の分類性能は，精度，再現率，F 値がそれぞれ 0.868，0.798， 0.832 であった．以下では，各素性の特徴の考察を行う．

関連投稿の類似性

この素性によって告知投稿には含まれないが，関連投稿ではよく言及されている語句を含む関連投稿の分類が出来るようになった．図 1 に示す例では，二日通し券が関連投稿でよく言及されていたため，例 (3) のような告知投稿に出現する語句を含まないものも関連投稿に正しく分類できるようになった．

共起リツイートの利用

(6) 邦楽ニュース アジカン、デビュー 10 周年記念ライヴを横浜スタジアム 2days で開催

(7) 2days 行きたいな

共起リツイートは実際には関連したニュースやイベント参加者の発言などであることが多かった．そのため，共起リツイートの情報を考慮することは，複数の情報源から告知に対する情報を得ることになり，元の告知投稿だけを利用するよりも多くの語句情報を得られる可能性が高い．たとえば図 1 に示した告知投稿の共起リツイートには例（6）に示すような，この告知投稿と同じイベントに冠する他のニュースサイトのツイートがあったため，例 (7) のように “2days” といった告知投稿には含まれない情報を含む関連投稿を正しく分類できるようになった

“＞ RT” を伴うツイートの利用

(8) まじかたぶん 9 月なら行けるよね!?チケットさえ取れれば無理にでも行く！！＞ RT

(9) アジカンの浜スタライブいきたいけどなあって思ってたらスタッフで興味ある人いたからチケットとれますよーに！（´▽｀）

“＞ RT” を伴うツイートは今回利用した周辺投稿の中で一番関連投稿である可能性が高いため，“＞ RT” を伴うツイートに含まれる語句は告知に対する関連語句である可能性が高い．このことが”＞ RT”を伴うツイートを考慮したすることで分類結果が大幅に向上した要因である．たとえば図 1 に示した投稿に対しては例 (8) に示すような”＞ RT”を含む関連投稿が存在したため，“チケット” や “バンド” など告知投稿と関連性が高いと考えられる語句が分類の手掛かりとして利用され，結果として例 (9) に示す投稿が関連投稿として正しく分類できるようになったと考えられる．

明示的な関連投稿の利用

明示的な関連投稿は先に述べたとおり，明示的な関連投稿に比べて少数であることが多い．今回は非公式 RT と告知投稿へのリプライを明示的な関連投稿として利用したが，どちらもその傾向が見られたため，分類に必要な語句情報を安定して得ることが難しかった．リプライは特に数

図 6 図 1 の投稿に対する連続した関連投稿

が少なく，本実験では 1 つの告知に対して最大でも 14 個のリプライしか得られなかった．これは告知投稿者に企業などの公式アカウントが多いことが原因だと考えられる．公式アカウントの場合，リプライされたとしてもそれに反応することは少ない．そのため，ユーザがリプライする利点を感じず，それがリプライが少ない結果を生んでいると考えられる．また，非公式 RT はリツイート直後の関連投稿よりも数が多いこともあったが，短い感想であることが多かった．非公式 RT は元の告知投稿の引用部分も含み，文字制限が通常のツイート以上に厳しい．そのため，長いツイートを投稿するには適しておらず，その結果として短いツイートが多くなっていることが考えられる．

4.4 明示的な関連投稿と非明示的な関連投稿の収集数

本実験で収集した関連投稿の数は，明示的な関連投稿は 877，非明示的な関連投稿は 1,512 であった．告知投稿 1 個あたりの数は明示的な関連投稿と非明示的な関連投稿がそれぞれ 100.8 個，62 個であり，非明示的な関連投稿の方が多いことが確認できた．提案手法により収集非明示的な関連投稿は 1207 個が収集できており，提案手法で収集した関連投稿だけでも明示的投稿の数を上回った．このため，本研究の提案手法は関連投稿の収集に有用であることがわかる．

5. まとめと今後の方針

本稿では，Twitter における告知投稿対象に，対象の告知投稿のリツイート直後のツイートから関連投稿を収集することで，告知投稿に対して，非明示的な関連投稿を自動収集する手法を提案した．実験の結果，告知投稿と判別対象だけでなく，他ユーザの投稿も考慮することで高い精度で関連投稿を収集できることを確認した．関連投稿の収集数に関しても提案手法により収集した非明示的な関連投稿は明示的な関連投稿を上回った．今後は，今回考慮していな

かった部分からの関連投稿の収集方法の考案につとめる．具体的には，提案手法の分類対象を告知投稿のリツイート直後のツイートだけではなく，リツイート後に連続する複数のツイートにまで拡張することを考えている．たとえば図 6 に示す例ではリツイート直後だけではなく，その後の投稿も告知に対する関連投稿となっている．そのため，リツイート直後のツイートだけでなく，それに続くツイートも関連投稿の候補とし，そこから関連投稿を収集することができれば，さらに多くの非明示的な関連投稿を収集することが可能になると考えられる．

参考文献

[1] James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, Yiming Yang, Umass Amherst, James Allan Umass, “Topic detection and tracking pilot study”, In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pp. 194–218, 1998.

[2] Yan Chen, Hadi Amiri, Zhoujun Li, Tat-Seng Chua, “Emerging topic detection for organizations from microblogs”, In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 43–52, 2013

[3] Padmanabhan Deepak, Sutanu Chakraborti, “Finding Relevant Tweets”, In *Web-Age Information Management Lecture Notes in Computer Science Volume 7418*, pp. 228–240, 2012

[4] Alok Kothari, Walid Magdy, Kareem Darwish, Ahmed Mourad, Ahmed Taei, “Detecting Comments on News Articles in Microblogs”, In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pp. 293–302, 2013

[5] 真野博子，伊東秀夫，小川泰嗣，“文書検索におけるランキング検索技術”，リコーテクニカルレポート，No. 29，pp. 21–30，2003

[6] 池田大介, 藤木稔明, 奥村学, “blog とニュース記事の自動対応付け”，言語処理学会第 11 回年次大会，pp. 1030–1033，2005

[7] Hiroya Takamura, Takashi Inui, Manabu Okumura, “Extracting Semantic Orientations of Words using Spin Mode”, In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pp.133–140, 2005.