C-12

新着ツイート群からの興味をひくツイートの抽出に関する考察

A study on the extraction of interesting tweets from new tweets

辻 一明†　宝珍 輝尚†　野宮 浩揮†
Kazuaki Tsuji　Teruhisa Hochin　Hiroki Nomiya

1．はじめに

近年，Twitter[1]に代表されるマイクロブログサービスが急速に普及してきている．マイクロブログとは，ユーザが自身の今の状況（いまどうしてる？）をつぶやく（ツイートする）Web サービスのことである．投稿内容が従来のブログに比べて短文（140 字制限）であることから更新が容易で，チャットのようなリアルタイムのコミュニケーションとしても用いられることもある．したがって，記事の件数が従来のブログと比べてかなり多くなる．

Twitter には，タイムラインの急速化による新着ツイートの蓄積という問題がある．ここで，タイムラインとは，投稿されたツイートを時系列順に並べた画面で，最新の投稿が上に表示され，古い投稿が下に流れていく．フォロー（他ユーザのツイートを購読すること）数が多ければ多いほど，タイムラインが急速化し，単位時間あたりの新着記事数が増加する．また，ユーザが常に Twitter を見られる状況にあるとは言えず，退席時間中に新着ツイートが蓄積する（図 1 参照）．新着ツイートが蓄積した時，1 つ 1 つ遡って読むのが面倒になりがちである．中には，自分へのリプライ（他ユーザから自分への返信）以外は全く読まないようなユーザも実際に存在する．

そこで本論文では，蓄積した新着ツイートから読むべきツイートを自動抽出してユーザに提示することを目的として，新着ツイートからユーザの興味をひくツイートの抽出に向けた検討を行う．ここでは，200 件程度の蓄積ツイートからキーフレーズを抽出し，キーフレーズを用いて話題の手動での抽出を試みた．この結果，10～20 件の話題が抽出された．また，話題抽出の過程で注目すべき事項が明らかになった．

2．で Twitter の特徴と問題点を述べ，3．で問題点の解決法について考察し，4．で検証と分析を行っている．5．で関連研究について述べ，6．でまとめている．

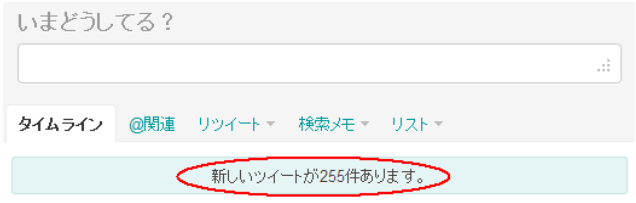

図 1　新着ツイートの蓄積

2．Twitter の特徴と問題点

2.1　Twitter について

Twitter とは，2006 年 7 月に Obvious 社（現 Twitter 社）が開始したマイクロブログサービスである．広義の SNS（Social Network Service）であるとも言える．

ユーザは，ツイートと呼ばれる 140 字以内の短文を投稿することができ，それらがホーム画面に表示される．ホーム画面に表示されているタイムラインをホームタイムラインと呼ぶ．また，特定のユーザをフォローすることで，フォローしているユーザのツイートもホームタイムラインに表示されるようになる．あるユーザをフォローする際に，そのユーザの許可が必要ない（ただし，ツイートを非公開に設定しているユーザに対しては，そのユーザの承認が必要となる）ので，必ずしも双方向にツイートを発信・受信することがない点が一般的な SNS との違いである．

また，ツイートは 140 字以内の短文であることから，Twitter はリアルタイム性の高いコミュニケーションサービスにもなり得る．文頭に“@ユーザ名”を付け加えることで，特定のユーザに宛ててツイートを投稿することができる（これをリプライと呼ぶ）．リプライを用いてチャットのように会話しているユーザも少なくない．リプライは少し特殊で，必ずしもタイムラインには現れず，リプライを行ったユーザとリプライされたユーザの双方をフォローしている場合のみ，タイムラインに現れるようになる．したがって，リプライは第三者にも見ることができるので，プライベート性はあまり高くないと言える．代わりに DM（ダイレクトメッセージ）という機能を用いれば，お互いにフォローし合っているユーザに，第三者からは見ることができないメッセージを送ることができる．

さらに，“RT(QT) @ユーザ名: (そのユーザのツイート)”という形式のリツイート（ReTweet）と呼ばれる投稿形態もある．これは他のユーザのツイートの引用で，特定のツイートを自分のフォロワー（自分をフォローしているユーザのこと）に周知させることができる．また，自分のコメントを加えることで，そのツイートへコメントすると言った使い方もある．この形式の RT は一般的に非公式 RT と呼ばれる．公式 RT ではコメントを挿入することができず，発言元のツイートがそのまま表示される仕様となっている．

以上，Twitter には様々な投稿形態があり，リアルタイム性も高く，常に多くの情報で満ち溢れていると言える．

2.2　問題点

Twitter の問題点として挙げられるのがタイムラインの急速化による新着ツイートの蓄積である．ツイートは 140 字制限で短文のため「おはよう」や「おやすみ」のような従来のブログでは見られない一言だけの投稿も多数存在する．そのため，普通のブログではあまり投稿しないようなユーザでも記事を複数回投稿することが多い．したがって，それなりの人数のユーザをフォローしていると，新着ツイートがホームタイムラインに蓄積しているという状況が多く見られる．特にフォロー数が 1000 や

†京都工芸繊維大学, Kyoto Institute of Technology

10000 を超えるヘビーユーザは，この状況が顕著に現れる．ユーザが常に Twitter を見られる状況にあるとは言えず，睡眠時間や退席時間など Twitter を見られない時間は存在し，その間に新着ツイートは蓄積する．新着ツイート数が多いと，全て読むのにそこまで時間はかからないが（新着ツイート約 300 件で大体 10～20 分程度），ツイートを 1 つずつ遡って読むのが面倒になりがちである．自分へのリプライ以外は読まないようなユーザも実際に存在する．蓄積した新着ツイートを読まなかった結果，それらの中に読みたかった（興味をひくような）ツイートが埋れていた等という問題が発生する．

2.3 Twitter 標準搭載のグルーピング機能

Twitter に標準搭載されているグルーピング（ツイートの分類）機能について説明する．グルーピング機能は，サーチ，流行のトピック，ハッシュタグ，リスト，お気に入りの 5 つである（Wikipedia の Twitter の項目[2]より抜粋）．

（1）サーチ：Search

任意の検索ワードでパブリックタイムライン（ツイートを公開設定にしている全ユーザのタイムライン）からツイートを検索することができる．検索結果はタイムラインのように表示され，リアルタイムで更新される．

サーチは，知りたい検索ワードが既に分かっている場合に有効だが，ユーザにとって未知の話題には対応することはできない．また，検索対象がパブリックタイムラインなので，ホームタイムラインに対する検索は Web 画面では不可能である（ただし，使用するクライアントによっては可能である）．

（2）流行のトピック：Trends

今，数多く投稿されているワードがホーム画面に表示される．それらをクリックすることで，そのワードでサーチした結果が表示される．国や地域別にトレンドを表示させることができ，現在は対象地域として「日本」「東京」も追加されている．

トレンドを用いればユーザが知らない未知の話題も見つけることができる．しかし，対象がパブリックタイムラインなので，ホームタイムラインでのローカルな流行等は見つけることができない．

（3）ハッシュタグ：Hashtag

#hogehoge のように，#（ハッシュシンボル）の後ろに半角文字列を羅列したものをツイートの末尾に付け加えることで，ツイートを特定のトピックごとに分類することができる．ハッシュタグをツイート末尾に付けることで，ユーザはそのトピックのツイートをサーチしやすくなる．なお，ハッシュタグは日本語には対応していなかったが．2011 年 7 月 13 日から利用可能となった．

ハッシュタグは，特定のトピックについてツイートする時や TV 番組などの実況時に使われることが多い．ハッシュタグは自動で付与されるものでは無いため，ハッシュタグだけで特定のトピックに関するツイートを全て集めることはできない．また，ハッシュタグは誰でも作成できることから，同じトピックに対して複数のハッシュタグが乱立するという問題も発生している．日本語ハッシュタグが利用可能となったことから，ハッシュタグ乱立の問題はより顕著になると考えられる．

（4）リスト：List

ユーザを名前をつけたリストで分類する機能である．最大 20 個まで作成可能で，1 つのリストあたり 500 人までフォローすることができる．

（5）お気に入り：Favorites

自分が気に入ったツイートをお気に入りとして登録し，後に一覧として見ることができる．

3. 解決へのアプローチ

3.1 手動による問題点の解決

本研究の問題点は「蓄積した新着ツイートを全て読んで自分にとって興味をひくツイートを探すのは面倒」という点である．この問題の解決法として「興味をひくツイートの抽出手法」を考える前に，手動によるこれらのツイートの発見方法について考える．手動での方法（第一著者の場合）を以下に示す．

（1）頻出語を探す
（2）ユーザにとって興味があるワードを探す
（3）興味のあるユーザのツイートを内容問わずに読む

以上の 3 つである．これら 3 つを自動化できれば問題は解決すると言える．

3.2 解決法の自動化について

（1）頻出語を探す

頻出しているワードがあれば，何か同じ話題で盛り上がっているということであり，それは興味をひきそうなものであると考えられる．Twitter 標準搭載のトレンド機能を用いれば，パブリックタイムラインにおける頻出語は抽出できる．しかし，2.3 で述べたようにホームタイムラインにおける頻出語は抽出できない．本論文では，ホームタイムラインにおける頻出語抽出に関する検証と分析を 4. で行う．

(2) ユーザにとって興味があるワードを探す

ユーザにとって興味があると思われるワードを探す．このワードは単独で現れ，（1）のように複数回現れているわけではないので探すのが難しいと考えられる．ユーザによる任意の入力を必要とせずに，すなわちシステム側で自動で収集できる情報のみを用いて，自動で抽出するのが理想である．

そこで，ユーザの興味を測る指標として利用する情報について熟考しなければならない．Twitter から得られる情報に加え，それ以外から得られる情報も利用する必要がある．本論文では，ユーザにとって興味があるワードを抽出する手法までは検討できておらず，今後の課題である．

(3) 興味のあるユーザのツイートを内容問わずに読む

Twitter 標準搭載のリスト機能を用いて，興味のあるユーザを集めたリストを作れば問題は解決する．リストを自動で作成することはできないが，リストを作るのにさほど手間はかからない．したがって，本研究では取り扱わないものとする．

4. 頻出語抽出実験

4.1 検証の対象とするツイート群

検証の対象とするツイート群は，第一著者のホームタイムライン内から取得した．以下の 3 つの時間帯においてツイートを取得した．

・DataSet 1

2011 年 6 月 20 日 (月) 9 時 56 分 2 秒～12 時 30 分 57 秒

・DataSet 2
2011 年 6 月 21 日 (火) 5 時 44 分 13 秒～10 時 3 分 26 秒
・DataSet 3
2011 年 6 月 22 日 (水) 4 時 44 分 26 秒～10 時 10 分 7 秒

時間帯の選択に意図は無く，無作為に決定した．Twitter API を用いてそれぞれ 200 件のツイートを取得した．200 件は Twitter API のメソッド 1 回の呼び出しで取得できる上限である．次に，これらのツイートに含まれる以下の文字列を排除する．これは，ユーザ名等の頻出語として適さないような語が抽出されるのを防ぐためである．

・リプライの宛先ユーザ名（@username）
・URL（http(s)://...）
・笑い（wの複数個連結）

上記の文字列を排除したものに対して，Yahoo Japan API のキーフレーズ抽出[3]を行っている．これは，Yahoo デベロッパーネットワークが提供している日本語文を解析して特徴的な表現をキーフレーズとして抽出するための API である．例えば，「東京ミッドタウンから国立新美術館まで歩いて 5 分で着きます。」という文章に対して，キーフレーズ抽出を行うと「東京ミッドタウン」「国立新美術館」「5 分」がキーフレーズとして抽出される．今回の検証では，取得したツイート群の中で複数回発生しているキーフレーズを頻出語として取り扱っている．

4.2 検証結果と分析

それぞれのデータセットに対して，総ツイート数，総キーフレーズ数（種類数），出現回数 2 回以上のキーフレーズ数，同一話題と判定されたキーフレーズ数，話題数，興味をひく話題数を手動で調べた結果を表 1 に示す．なお，同一話題の判別は第一著者が手動で行った．

表 1 検証結果

	DataSet1	DataSet2	DataSet3
総ツイート数	200	199	200
総キーフレーズ数	534	506	654
出現回数2回以上のキーフレーズ数 [割合]	37 [6.92%]	49 [9.68%]	51 [7.80%]
同一話題と判定されたキーフレーズ数 [割合]	16 [2.30%]	29 [5.73%]	26 [3.98%]
話題数	9	16	18
興味をひく話題数(主観)	9	16	18

まず，9 割以上のキーフレーズが 1 回しか出現していないことが分かる．また，2 回以上出現しているキーフレーズの中でも約半数前後しか同一話題と判定されていない．そこで，同一話題と判定されるキーフレーズのみを上手く抽出するための考察を行う．

4.3 考察

(1) キーフレーズの内容による同一話題性の判別

キーフレーズが一般名詞か固有名詞かどうかで結果が変わってくる．固有名詞である場合は，その固有名詞の名前が 1 つの話題名として見なせるので，ほぼ確実に同一話題となりうる．図 2 にキーフレーズが固有名詞である場合の具体例を示す．

また，キーフレーズが一般名詞の場合は，複数回出現していても同一話題とは見なせないことが多い．図 3 にキーフレーズが一般名詞である場合の具体例を示す．

以上より，上手く固有名詞のみを抽出することが精度向上に繋がると考えられる．

キーフレーズ:アシモフ　出現回数:3
Tweet1 ユーザA
ぱって　目に入ったところにあったのは　ネメシス　って小説。(アシモフ
Mon Jun 20 11:46:48 JST 2011
Tweet2 ユーザA
アシモフの本小さい頃から部屋にあるけど　一度も読んだことがない
Mon Jun 20 11:39:06 JST 2011
Tweet3 ユーザA
アシモフあってた(ぐぐってもうた
Mon Jun 20 11:38:03 JST 2011

図 2 キーフレーズが固有名詞（アシモフ）の例

キーフレーズ:朝　出現回数:3
Tweet1 ユーザA
朝、クロとスパーリングしたけど、あいつ急所を狙う技ばっか撃ってくる。トロも急所をガードするとうまくいく。
Mon Jun 20 12:03:42 JST 2011
Tweet2 ユーザB
朝の講義おわたなーう 相変わらず両手両足はいたい
Mon Jun 20 11:28:02 JST 2011
Tweet3 ユーザC
社内席替え大会実施中。引越席替え大好き！朝から変に活気付いてるw
Mon Jun 20 10:48:17 JST 2011

図 3 キーフレーズが一般名詞（朝）の例

(2) 投稿ユーザによる同一話題性の判別

キーフレーズが一般名詞の場合でも同一話題になっている場合は存在する．それを判別するためには投稿ユーザにも着目する．今回の検証では，同じユーザによる投稿の場合，キーフレーズが一般名詞でも同一話題となることが多いという結果が出た．表 2 に同じユーザによる投稿と異なるユーザによる投稿ごとの同一話題と判定された一般名詞のキーフレーズ数を示す．

表 2 投稿ユーザ別のキーフレーズ数（一般名詞）

	DataSet1	DataSet2	DataSet3
同じユーザ	11	13	11
異なるユーザ	4	10	3

基本的に同じユーザによる投稿が多いことが分かる．しかし，異なるユーザによる投稿の場合でも同一話題である一般名詞のキーフレーズが発生しているので両方に対処する必要がある．

まず，同じユーザによる投稿の場合は，基本的に同一話題と考えてもいいが，話題内容が異なる場合も存在する．その例を図 4 に示す．

キーフレーズ:名前　出現回数:2
Tweet1 ユーザA
アニメの名前ど忘れなう
Tue Jun 21 09:33:25 JST 2011
Tweet2 ユーザA
名前が力を持つ世界に一日だけ行きたい
Tue Jun 21 07:57:47 JST 2011

図 4 キーフレーズが一般名詞，同じユーザによる投稿でも話題内容が異なる場合

このような場合は投稿時間にも着目する必要がある．図 4 で投稿時間を見ると約 1 時間半離れていることが分かる．一般的に投稿時間の間隔が短いと話題内容が同じであり，間隔が長いと話題内容が異なることが多いが，必ずしもそうであるとはかぎらない．新谷ら[4]の研究では，ユーザの平均投稿間隔にも注目することで精度が向上することが報告されている．

次に，異なるユーザによる投稿の場合を考える．この場合は何通りかのパターンに分けられると考えている．今回の検証で現れたパターンを以下に示す（今後，これら以外のパターンが発生する可能性はある）．

1. リプライによる会話

2. 非公式 RT による会話やコメント
3. 同じ所属のユーザによる投稿（例：同じ大学）
4. 異なる所属・全く関係のないユーザによる投稿

1 と 2 の抽出については容易である．3 は手動で判別したからこそ分かる情報だが，自動で判別するためにはユーザに関する情報も収集しなければならない．また，4 に関しては，ツイートの内容等を十分に解析して判別する必要がある（具体的な手法には思い至っていない）．

（3）関連ツイートの発見

あるキーフレーズを含むツイートがある時，そのキーフレーズ自体はツイートの中に現れていないが内容は関連している別のツイート（投稿ユーザは同じ）が存在する．そのようなツイートを発見し，まとめて提示することで利便性は向上する．新谷ら[4]の研究では，単一ユーザの同一話題ツイートの集約が行われている．

（4）キーフレーズごとの同一話題性の判別

異なるキーフレーズ間で同一話題となっている場合もある．この時，同一話題となる複数のキーフレーズを集約して提示する必要がある．今回の検証では，以下のような場合に異なるキーフレーズ間で同一話題となった．

1. 表記ゆれが発生している時（例：「じぶん」「自分」）
2. 類義語が発生している時（例：「私」「自分」）
3. 関連語が発生している時（例：「みかん」「りんご」「果物」）
4. 別の同じキーフレーズを含む時（例：キーフレーズ A と B を含むツイート群 α とキーフレーズ B と C を含むツイート群 β がある時）

（5）興味をひいた話題

抽出された話題に対して，実際に興味をひくかどうかの判定を第一著者の主観で行ったところ全て興味をひくという結果になった．主観による判定なので，断定はできないが頻出語抽出により抽出された話題はある程度興味をひくものであると考えられる．

（6）まとめ

頻出語の抽出という形でユーザの興味をひくツイートを抽出するためには，上記の全てを実現しなければならない．また，今回の検証では現れなかったパターンが存在する可能性もあるので，さらなる検証を進めていく必要がある．

5． 関連研究

5.1 投稿間隔に基づくマイクロブログからの話題チャンク抽出に関する一検討 [4]

ある話題について記述された複数の記事の塊を話題チャンクとして抽出し，ユーザに提示することで話題を理解する支援を行っている．ここで抽出する話題チャンクは単一ユーザがある話題について複数の記事に分けて投稿することで形成されるものである．

共起語に基づく抽出と投稿間隔に基づく抽出の 2 つを提案し，それぞれ単体・2 つ合わせたものの精度を Jaccard 係で評価している．

この結果，共起語による抽出手法で平均約 40%の精度を示したのに対して，投稿間隔に基づく抽出手法は，6 名中 3 名のユーザが共起語に基づく抽出精度を上回る結果を示した．また，平均投稿間隔の約 0.1 倍の時間間隔で抽出したときに，最も精度が高くなる傾向があった．さらに，共起語と投稿間隔を併用した抽出手法では，3 名のユーザにおいて，それぞれを個別に適用した際の抽出精度を上回る結果を示した．これらから投稿間隔を指標とした話題チャンク抽出の有効性を確認している．

4.3 の(3)で述べたように，単一ユーザによる関連のある複数の記事を一つにまとめることは本研究の目的を実現するために必要なものの 1 つである．

5.2 データ圧縮による Twitter のツイート話題分類 [5]

Twitter において，日々生成される膨大な情報（口語的で短く，リアルタイム性の高いテキスト）の中から，利用者にとって有益な情報のみを収集するために，形態素解析に依存せず（新語や口語の出現に影響されない），学習対象の変化に素早く追従可能なアルゴリズムとして，ツイートの圧縮されやすさを応用した手法を提案している．

データ圧縮を利用した分類手段の基本的概念は，あるデータ x が，情報源となる他のデータ A を基に十分「圧縮」できる場合，2 つのデータ x, A は類似しているというものである．新しいツイートを，着目する話題に関するツイートの集合（話題モデル）と，それ以外のツイートの集合（比較モデル）の両方を用いて圧縮する．そして，新しいツイートが，話題モデルを基に圧縮した方が圧縮され易い場合に，着目する話題に関連する可能性が高いと見なす．話題モデルは，指定した文字列（キーワード，ハッシュタグ，URL など）が含まれるテキストを時間順に連結したものであり，比較モデルは，それ以外のテキストを時間順に連結したものである．

評価実験では，オンライン学習器の中で最も性能の良いものの一つである confidence-weighted linear classification について性能を比較している．この結果，提案手法が優れた識別率と再現率を実現することを示している．なお，提案手法は，ハッシュタグ分類に限らず，汎用的な話題分類に使用可能となっている．

この研究は本研究と目的が類似している．知りたい話題が既に分かっている時（話題モデルが生成できる時）には，この手法は有効であると考えられる．

5.3 投稿日時とユーザの広がりに基づくツイート分類手法 [6]

読むべきツイートを見つけやすくするために，興味のあるトピックが明確でない場合でも適用可能なツイート分類手法を提案している．

ここでは，投稿日時とユーザの広がりの大小の組み合わせ 4 パターンでツイートを分類している．誰でもどんな時でもツイートされるプライベート型のツイート，一時的に幅広くの人にツイートされるパブリック型のツイート，一部の人に日常的にツイートされるコミュニティ型のツイート，特定の時期において一部の人にツイートされる特殊型のツイートに分けられる．このようにツイートを分類することで，利用者がツイートを探しやすくなる．

実際に Twitter に投稿されたツイートのデータを用いた実験を通して，関連ツイートの時間的な広がりやユーザの幅広さに基づいてツイートが分類できることを確かめている．

この研究は，ツイートの大まかな分類を行う研究である．ツイートの特性を捉えて分類することは本研究にも応用できるかもしれないと考えている．

6． おわりに

本論文では，新着ツイート群からのユーザの興味をひくツイートの抽出に関する考察を行った．その結果，頻出語抽出により得られる話題はユーザの興味をひくものであるが，それを実現するためには様々な課題を解決しなければならないことが分かった．

今後の課題としては，4.3 で述べた様々な問題や課題を解決するための具体的な手法について 1 つ 1 つ検討していきたいと考えている．

7． 参考文献

[1] Twitter：https://twitter.com/

[2] Twitter - Wikipedia：http://ja.wikipedia.org/wiki/Twitter

[3] Yahoo Japan API キーフレーズ抽出：http://developer.yahoo.co.jp/webapi/jlp/keyphrase/v1/extract.html

[4] 新谷歩生，関洋平，佐藤哲司：”投稿間隔に基づくマイクロブログからの話題チャンク抽出に関する一検討”，Proc. of DEIM Forum 2011，A1-2（2011）

[5] 西田京介，坂野遼平，藤村孝，星出高秀：”データ圧縮による Twitter のツイート話題分類”，Proc. of DEIM Forum 2011，A1-6（2011）

[6] 伊藤勇也，浅野秦仁，吉川正俊：”投稿日時とユーザの広がりに基づくツイート分類手法”，Proc. of DEIM Forum 2011，A10-5（2011）