DEIM Forum 2011 A10-5

投稿日時とユーザの広がりに基づくツイート分類手法

伊藤　勇也† 　浅野　泰仁†† 　吉川　正俊††

† 京都大学工学部情報学科　〒 606–8051 京都市左京区吉田本町
†† 京都大学大学院情報学研究科　〒 606–8051 京都市左京区吉田本町
E-mail: †itoh@db.soc.i.kyoto-u.ac.jp, ††{asano,yoshikawa}@i.kyoto-u.ac.jp

あらまし　2006 年のサービス開始以来，Twitter の利用者数もツイート投稿数も著しく増加している．興味のあるトピックがはっきりしている場合は，キーワード検索などで読むべきツイートを得られるが，そうでない場合は全てのツイートに目を通すしかなく，これは非常に時間のかかる作業である．そこで，本論文では，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する．本論文で提案する手法は，関連ツイートの投稿日時とそれらを投稿したユーザの分布に基づくスコアをツイートに付加することにより，ツイートを分類する．

キーワード　マイクロブログ，Twitter，ツイート分類

Tweet Classification Based on the Distribution of Post-Time and Users

Yuya ITOH†, Yasuhito ASANO††, and Masatoshi YOSHIKAWA††

† Undergraduate School of Informatics and Mathematical Science, Faculty of Engineering, Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto, 606–8501 Japan
†† Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto, 606–8501 Japan
E-mail: †itoh@db.soc.i.kyoto-u.ac.jp, ††{asano,yoshikawa}@i.kyoto-u.ac.jp

1. はじめに

近年，Twitter(注1)に代表される，マイクロブログと呼ばれる Web サービスが普及してきている．マイクロブログとは，チャットとブログの中間に位置し，短い文章のやり取りを通してユーザが情報を発信したり，コミュニケーションを取ったりすることができる新しいサービスである．Twitter では，この短い文章のことをツイート (tweet) と呼んでいる．

Twitter の特徴として，ツイートに 140 文字以内という文字数制限が設定されていることが挙げられる．これによって，ユーザは従来のブログサービスと比べて，気軽に情報を発信できるようになっている．また，個人のページ上に他のユーザのツイートを表示するためのフォローと呼ばれる機能にも特徴がある．フォローはツイートを読みたいユーザを登録する機能であり，SNS などのサービスと違い，相手に許可を取らなくてもフォローすることができる．これにより，ユーザが双方向のやり取りを強制されることなく，他のユーザを自由にフォローすることができる．さらに，各ユーザのツイートは基本的に公開されており，ユーザが特に設定しない限り，自由に閲覧できるようになっている．このような自由な閲覧形態により，Twitter では手軽で開かれた情報発信が実現されている．

Twitter は，身の回りで起こっていることを共有するサービスであるが，実際にはチャットや日記帳，情報を発信するためのメディアとしてなど，様々な形で利用されている．そのため，ツイートの内容も多種多様であり，ユーザにとって有用で興味深いものと関心のないものが混在した状態になっている．そのような中で，多くのユーザの投稿する大量のツイートを常にチェックすることはとても難しく，興味のあるツイートを見逃してしまうことがある．そこで，投稿されたツイートを後で読み返すことで読みたいツイートを探すことになる．このような場合，読みたいツイートが明確であればキーワード検索を利用することができるが，そうでないことも多い．そこで，ユーザが興味のあるツイートを効率よく発見するために，ツイートの分類が必要になる．

また，キーワード検索が利用できる場合であっても，読みたいツイートが簡単に発見できるとは限らない．その理由の一つ

(注1)：http://twitter.com/

に，ユーザが投稿するツイートには様々な意図が含まれていることが挙げられる．Twitter のユーザの多くが情報発信・受信や他のユーザとのコミュニケーション，自分の日常に関することを共有するために Twitter を利用しており [1] [2]，ツイートについてもこれらに関連する内容が多い [3] ことがわかっている．このように，同じキーワードを含むツイートの中にも，情報発信のためのツイートや日常に関するツイートなどが混在している場合が考えられ，このようなツイートはキーワードだけで分類することができない．

そこで本論文では，特定のツイートとそれに関連するツイートに関して，それらの投稿日時の広がりとそれらを投稿したユーザの幅広さの観点から分類を試みる．これにより，そのツイートが日常的な内容であることや，知り合いとの内輪話であることなど，ツイートの傾向に関する情報を付加することができ，ユーザがツイートを読む際に，読みたいツイートを発見するのに役立つと考えられる．

実際に投稿されたツイートを使った評価実験を通して，これらの広がりが指標の大小に反映されていること，及びツイートの傾向に基づいて，ツイートが分類できることを確認している．

2. Twitter

Twitter は，2006 年 7 月に開始されたマイクロブログサービスで，サービス開始以来，世界的に急成長を遂げている．2010 年 9 月現在，登録ユーザ数は 1 億 4500 万ユーザを超えている(注2)．Twitter では,「いまどうしてる？(What's Happening?)」という質問に答える形でツイートを投稿するのが，基本的な使い方である．

Twitter において，各ユーザのツイートは基本的に公開されているが，特にツイートを読みたいユーザをフォローという機能で登録することができる．利用者がフォローしているユーザはフォローイング (following)，フォローされているユーザはフォロワー (follower) と呼ばれる．フォローイングのツイートは，自分の投稿したツイートと共に，利用者専用ページのタイムラインと呼ばれる領域に，時系列順に表示される．多くのユーザは，タイムライン上のツイートを読むことで他のユーザと情報を共有している．

利用者は，他の人が投稿したツイートに対して，返事を書くという形でツイートを投稿することができる．このようなツイートは返信 (リプライ, reply) と呼ばれ，ユーザがコミュニケーションをとるためなどに用いられる．また，リツイート (retweet, RT) と呼ばれる，他のユーザが投稿したツイートを再投稿する形で，ツイートを投稿する形式もある．リツイートは，利用者が自分のフォロワーに対して，自分のフォローイングのツイートを転送したいときに用いられ，ニュースなどに関するツイートはリツイートされやすい [2]．

3. 関 連 研 究

青島ら [4] の研究では，制約付きクラスタリングのアプローチからツイートを分類する手法を提案している．Twitter では個々のツイートの文章が非常に短くなるため，ツイートを分類する際，ツイートの特徴を単語の出現頻度で表現することが難しいという問題がある．青島らは，ツイートに対して単語の表記揺れの緩和などの前処理を行うことで，この問題に対処している．その上で，TF-IDF によって算出したコサイン類似度と，単語が属す概念の類似度などから制約付クラスタリングを適用している．

岩木ら [5] は，キーワード検索の結果に対して，ユーザの読みたいツイートを表現する感性による絞り込みと，興味の近接度による再ランキングを行うことで，ユーザの読みたい有用なツイートの分類を行っている．感性による絞り込みでは，語の共起に基づいて，単語と感性の相関ルールを抽出することで感性辞書を作成している．興味の近接度は，検索結果に表れるツイートを投稿したユーザと利用者に対して,「興味の類似度」と「ユーザの近さ」に基づいて定義している．

田中ら [6] は，広く一般の人にとって有用なツイートと，そうでないツイートを分類する手法を提案している．この研究では，ユーザが一般への情報発信を指向しているかどうかを分類することで，ツイート分類の精度の向上を図っている．ユーザが情報発信を指向しているかどうかの判定には，フォロー関係を友人関係とそれ以外に分けることによって行う手法と，お気に入りに登録されたツイートの割合による手法を用いている．

Sakaki ら [7] の研究では，Twitter のリアルタイム性に着目し，地震や台風のようなイベントの発生の検出と位置の推定をリアルタイムに行うための手法を提案している．Sakaki らの手法では，投稿されたツイートが，対象とする特定のイベント (例えば地震) の発生に関連したツイートかどうかを分類することで，イベントの発生を検出する．ツイートの分類は，予め与えられたキーワードを含むツイートに対して，SVM(Support Vector Machine) を用いて行っている．また，ツイートに関連付けられている位置情報から経度と緯度を算出し，カルマンフィルタなどの手法を用いてイベントの位置を推定している．

これらの研究では，キーワードによる分類やその絞り込みによって読みたいツイートを提示したり，ツイートのトピックを限定した中で分類を行っている．このような手法は，利用者が読みたいツイートがある程度絞れている場合には有効であるが，そうでない場合には不十分であると思われる．これに対して本研究では，キーワード検索などにより，利用者が読みたいツイートを限定することができない場合でも適用できるようなツイートの分類手法を考えている．

また，Sriram ら [8] の手法のように，機械学習のアプローチからツイートを分類する手法も提案されているが，このような手法は，新語に対応することが難しいという問題がある．

4. 提 案 手 法

本論文では，Twitter を利用している特定のユーザが，大量にあるツイートの中から読みたいツイートを探すという状況を想定している．以下，そのようなユーザを「読者」とする．

(注2)：http://blog.twitter.com/2010/09/evolving-ecosystem.html

4.1 投稿日時・ユーザの広がり

本論文では，ツイートを分類するための指標として，投稿日時の広がりとツイートを投稿するユーザの広がりを考える．以下，分類を行うツイートを「対象ツイート」，対象ツイートに関連したツイートを「関連ツイート」とする．また,「関連ツイート」は，対象ツイートに含まれるキーワードのいずれかを含むツイートと定義する．対象ツイートからキーワードを抽出する方法は，4.2 節で述べる．

対象ツイートに対して，関連ツイートがどのように分布しているかを考えることで，ツイートを読まなくてもツイートの内容を簡単に把握することができる．例えば，関連ツイートが投稿日時に関して狭い範囲に集中しているものは，ニュースやイベントなどに関連した内容であると考えられる．一方で，投稿日時に関して広く分布しているものは,「お腹空いた」や「おはよう」のような，日常的にツイートされる内容である可能性が高い．また，ユーザの広がりに関して，特定のユーザの間でツイートされている場合は，趣味の話題に関するツイートなど，グループ内でのコミュニケーションに関連したツイートであると考えられる．それとは逆に，幅広いユーザにツイートされているものは，より一般的な内容に関するツイートであると考えられる．

投稿日時の広がりには，関連ツイートとの投稿日時の差に基づく指標を用いる．関連ツイートが特定の時期に集中している場合，投稿日時の広がりは小さく，逆に，多くの時期に投稿されている場合，投稿日時の広がりが大きいとする．

ユーザの広がりには，読者と他のユーザの近接度に基づく指標を用いる．読者に近いユーザが関連ツイートを比較的多くツイートしている場合は，ユーザの広がりが小さいと定義し，読者から遠いユーザも多くツイートしている場合はユーザの広がりが大きいとする．本論文では，ユーザのフォロー関係を表すグラフ構造から読者と他のユーザとの非類似度を定義し，読者に近いユーザほど非類似度が小さくなるようにする．

ユーザのフォロー関係を表すグラフ $G = (V, E)$ は，以下のように定義する．

$$V = \{u | u \text{ は Twitter のユーザである } \}$$

$$E = \{(u_1, u_2) | u_1, u_2 \in V, u_1 \text{は } u_2 \text{をフォローしている } \}$$

また，ユーザ u, v 間の非類似度 $D(u, v)$ を，G 上で辺の向きを考慮しないときの 2 頂点間の最短経路に表れる辺の数で定義する．以下，読者 r と他のユーザ u 間の非類似度 $D(r, u)$ を，$D_r(u)$ で表す．

このようにして定義されるグラフ G の例を図 1 に示す．図 1 において，赤色は読者，青色は $D_r = 1$ のユーザ，緑色は $D_r = 2$ のユーザ，そして黄色は $D_r = 3$ のユーザを表す．

投稿日時とユーザの広がりを表す二つの指標の大小から，図 2 のようにツイートを大きく 4 つに分類することができる．次に，これら 4 つのタイプに属すツイートの特徴について説明する．

- タイプ 1

タイプ 1 に属すツイートは，投稿日時・ユーザの広がりがともに大きいツイートである．すなわち，このタイプに属すツイートは，多くの時間帯で幅広いユーザにツイートされている内容である．このようなツイートは「お腹空いた」や「眠い」など，個人の状態を表すツイートや日常的な内容に関するツイートが多いと考えられる．本論文では，個人的な内容に関するツイートが多いことから，タイプ 1 に属すツイートをプライベート型のツイートと呼ぶ．

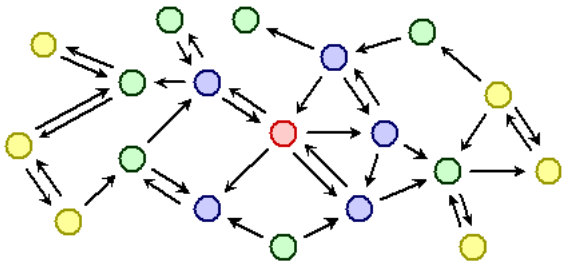

図 1　フォロー関係を表すグラフ

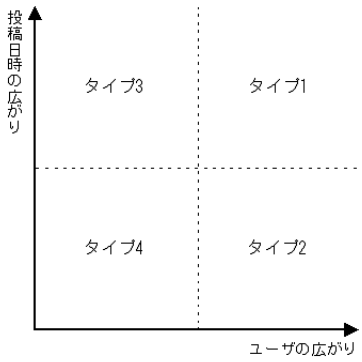

図 2　投稿日時とユーザの広がりによるツイートの分類

- タイプ 2

タイプ 2 に属すツイートは，投稿日時の広がりは小さく，ユーザの広がりは大きいようなツイートである．このようなツイートは，多くのユーザが興味を持っているニュースなどのツイートが多いと考えられる．特に，ニュースに関するツイートは，情報発信を目的とするユーザによって投稿され，それを見た他のユーザがそのニュースに関連するツイートを投稿したり，他のユーザにリツイートされることで，短時間の間に沢山のユーザが関連ツイートを投稿する．本論文では，一般のユーザに向けて情報を発信することを目的とした，ニュースなどに関するツイートが多いことから，タイプ 2 に属すツイートをパブリック型のツイートと呼ぶ．

- タイプ 3

タイプ 3 に属すツイートは，投稿日時の広がりは大きく，ユーザの広がりは小さいようなツイートである．このようなツイートは，限られたユーザの間で共通する話題に関するツイートが多いと考えられる．例えば，共通の趣味に関するツイートや，知り合いとの身内話などが当てはまる．本論文では，限られたユーザの中でコミュニケーションをとるためのツイートが多いという意味で，タイプ 3 に属すツイートをコミュニティ型のツイートと呼ぶ．

- タイプ 4

タイプ 4 に属すツイートは，投稿日時・ユーザの広がりがともに小さいツイート，すなわち，特定の時期において，読者に近いユーザによって投稿されているツイートである．このようなツイートは，地震のように地域的に偏りがあったり，コミュニティ型のツイートの中でも特に時間的に偏りがあるような内容のツイートなどが考えられる．他にも，一般的なツイートに現れないような珍しいトピックに関するツイートは投稿日時・ユーザの広がりがともに小さくなると考えられる．本論文では，タイプ 4 に属すツイートを特殊型のツイートと呼ぶ．

4.2 分類手順

ツイートの分類は以下に示す手順で行う．

- ステップ 1：分類を行うツイートの範囲を決定する．

まず，分類を行うツイートの指定を行う．ツイートの範囲は，ユーザの集合及び投稿日時の期間を選択することによって決定し，この指定は読者から与えられるものとする．

ユーザの集合としては，読みたいツイートを投稿していると考えられるユーザを読者が指定する．Twitter を利用するユーザの多くが，読みたいツイートをするユーザをフォローする形で予め選択しているということを踏まえ，本論文では，読者がフォローしているユーザ全体，もしくはその一部が指定されることを想定する．期間の指定では，読者が読みたいツイートのある期間を，例えば，昨日一日，過去一週間，特定の日の朝や夕方 (もしくは具体的な時間帯) のような形で指定する．

- ステップ 2：ツイートの分類に必要な情報を収集する．

Twitter では，ツイートの投稿や，投稿されたツイートを閲覧するための API が外部アプリケーション向けに提供されており，ユーザに関する情報や，投稿されたツイートに関する情報を取得することができる．本論文でも，Twitter が提供している API を通して，分類を行うために必要な情報を取得する．このステップで収集する情報は，ステップ 1 で指定されたユーザのツイートとそれに付加された投稿日時などの情報である．それに加えて，指定されたユーザ以外からもランダムにユーザを選択し，関連ツイートを投稿するユーザとしてツイートを収集する．

また，取得した各ツイートに対して，ツイートに含まれるキーワードの分析も行う．分析には，Yahoo!デベロッパーネットワーク(注3)が提供している WebAPI のキーフレーズ抽出を利用する．Yahoo!デベロッパーネットワークでは，Yahoo!Japan のサービスでも用いられている，Yahoo!JAPAN 研究所の研究成果を API として公開している．その一つであるキーフレーズ抽出 API では，文章を解析し特徴的な表現 (キーフレーズ) を抽出することができる [9].

ツイートに含まれるキーワードの分析を行うための他の方法として，Mecab(注4)などを利用した形態素解析が考えられるが，形態素解析による分析では複数の名詞から成る複合名詞が分割されてしまう場合がある．例えば,「誕生日」という単語は，Mecab の解析では「誕生」と「日」に分割されてしまうが，キーフレーズ抽出 API を利用すると「誕生日」をキーフレーズとして抽出することができる．

- ステップ 3：関連ツイートの分布からスコアを付与する．

ステップ 2 で収集した情報から，関連ツイートの分布に基づくスコアを対象ツイートに付与する．

投稿日時の広がりの指標は，対象ツイートと関連ツイートの投稿日時の差の標準偏差として定義する．対象ツイート t に対して，投稿日時の広がりを表す指標 α は，式 (1) のようになる．

$$\alpha = \sqrt{\frac{1}{|T_t|}\sum_{\tau \in T_t}(\bar{d} - d_{t,\tau})^2} \tag{1}$$

式 (1) において，T_t は t の関連ツイートの集合であり，$|T_t|$ はその要素数を表す．$d_{t,\tau}$ はツイート t とツイート τ の投稿日時の差であり，式 (2) のように表される．ここで，$PT(t)$ はツイート t の投稿日時を表す．

$$d_{t_1,t_2} = PT(t_1) - PT(t_2) \tag{2}$$

また，$\bar{d}$ は対象ツイートと関連ツイートの投稿日時の差の平均であり，式 (3) で表される．

$$\bar{d} = \frac{1}{|T_t|}\sum_{\tau \in T_t} d_{t,\tau} \tag{3}$$

次に，ユーザの広がりを表す指標を定義する．ユーザの広がりに関しては，二種類の指標を提案する．ユーザの広がりの指標 β_1, β_2 を式 (4)，式 (5) に示す．

$$\beta_1 = \sum_{i=2}^{N} i\frac{R(i)}{R(1)} \tag{4}$$

$$\beta_2 = \sum_{j=1}^{N} j\frac{U_t(j)}{\sum_{i=1}^{N} U(i)} \tag{5}$$

ここで，N は読者と他のユーザとの非類似度 D_r の最大値である．$R(x)$ は，$D_r = x$ であるユーザの中で関連ツイートを投稿しているユーザの比であり，式 (6) で表される．また，$U(x)$ は $D_r = x$ であるユーザの数であり，$U_t(x)$ はその中で関連ツイートを投稿しているユーザの数である．

$$R(x) = \frac{U_t(x)}{U(x)} \tag{6}$$

- ステップ 4：スコアによる分類を行う．

4.1 節で述べたように，ステップ 3 で算出した投稿日時・ユーザの広がりの指標の大小に基づき，ツイートを 4 つのタイプのいずれかに分類する．このとき，読者がスコアの大小の基準値を与えることができるものとする．

5. 実　　験

5.1 実験方法

まず，あるユーザを読者とし，その読者を起点としてフォロー関係を収集した．その中で，読者のフォローイング 198 人と，それ以外のユーザの中からランダムに 4000 人を選び，11 月 26 日 00:00 から 12 月 10 日 00:00 までの間に投稿されたツイートを収集した．ランダムに選んだユーザの内訳は，$U(2) = 2000$,

(注3)：http://developer.yahoo.co.jp/
(注4)：http://mecab.sourceforge.net/

$U(3) = 2000$ である．

収集したツイートの数は，合計 643,645 個であった．その内訳は $D_r = 1$ のユーザが 47,617 個，$D_r = 2$ のユーザが 444,139 個，$D_r = 3$ のユーザが 151,889 個である．$D_r = 1$ のユーザが投稿した 47,617 個のツイートについて，キーワードの分析 (サーバに問合せる時間を含む) に要した計算時間はツイート 1 個当たり 0.141[s] であった．その結果，43,253 種類，119,410 個のキーワードが抽出された．また，これらのツイートにはそれぞれ平均 10,275 個の関連ツイートが存在し，関連ツイートを求めてツイートのスコアを計算するのにツイート 1 個当たり 2.37[s] を要した．計算時間は，CPU:Intel Core i7-920 (2.67GHz)，メモリ:9GB の構成を持つ計算機上で計測した結果である．

次に，収集したツイートの中からいくつかを選び，投稿日時とユーザの広がりの指標に対して適当な基準値を設定した上で，ツイートの分類がどのように行われているかを確認する．ユーザの広がりの指標に関しては，β_1 を基準に分類を行った．ユーザの広がりにおける β_1 と β_2 の比較，及び評価に関しては 5.3 節で述べる．

5.2 実験結果

実際のツイートにスコアを付け，ツイートを分類した結果の一部を，分類毎に表 1 から表 4 に示す．投稿日時の広がりの指標である α の単位は [day]，ユーザの広がりの指標である β_1 は無次元数である．分類の基準値には，分類したツイートの中での平均値 ($\alpha = 3.93, \beta_1 = 2.19$) を用いた．このとき，分類したときに抽出されたキーワードがないなどの理由で関連するツイートが一つもないようなものは除いて算出している．

5.2.1 プライベート型のツイート

プライベート型に分類されたツイートを表 1 に示す．

表 1 のように，個人の行動や状態，周囲の様子に関するツイートがプライベート型のツイートとして分類されている．例えば，一番上のツイートは，クリスマスツリーやリースを飾っている家をよく見かけたという身の回りの様子を述べたツイートであり，上から二番目のツイートは，映画を見に行こうとするユーザの行動を表している．また，一番下のツイートは自分が空腹であるという状態を表現している．このようなツイートは，あらゆるユーザが投稿する可能性のあるツイートである．実際に，上から二番目のツイートに関連したツイートを投稿したユーザの数は，それぞれ $U_t(1) = 23$，$U_t(2) = 478$，$U_t(3) = 34$ であった．これより，$R(1) = 0.116, R(2) = 0.239, R(3) = 0.017$ となり，$D_r = 1$ のユーザと比べて，$D_r = 3$ のユーザは少ないが，$D_r = 2$ のユーザでは関連ツイートを投稿したユーザの比率が高いことがわかる．すなわち，このツイートは読者に近いユーザに固有の話題ではなく，それ以外のユーザにもツイートされる一般的な内容であると考えられる．これにより，このツイートの β_1 も大きい値をとっており，ユーザの広がりの大きさが指標に表れていることを確認できる．

次に，上から二番目のツイートに含まれるキーワード「映画」に関連するツイートの数を，日付毎にまとめたグラフを図 3 に示す．図 3 では，D_r が等しいユーザの間で投稿された関連ツイートの数を棒グラフで示している．

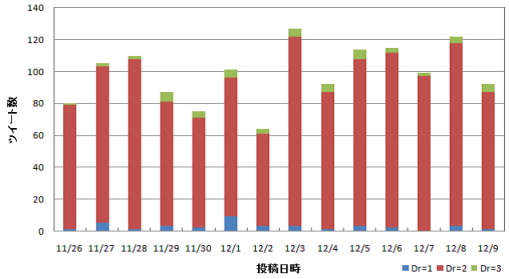

図 3 「映画」関連ツイートの分布

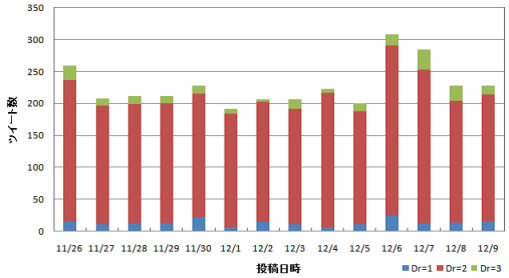

図 4 「腹」関連ツイートの分布

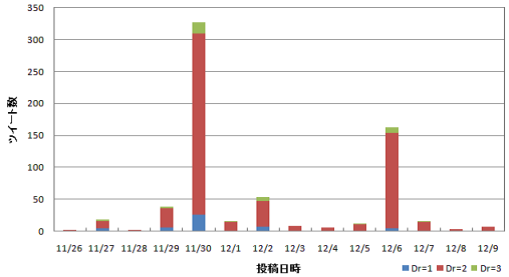

図 5 「地震」関連ツイートの分布

図 3 から，「映画」に関連したツイートは，毎日 80〜100 個程度投稿されており，時間的な偏りはほとんど見られないことが分かる．実際に 1 日当たりの平均ツイート数は 98.8 個である．また，図 4 に示すように，一番下のツイートに含まれるキーワード「腹」に関連するツイートの数にも時間的な偏りはほとんど見られない．

これらのことから，投稿日時の広がりが大きいツイートでは，α も大きな値をとることが確認できる．

5.2.2 パブリック型のツイート

パブリック型に分類されたツイートを表 2 に示す．

表 2 の一番上のツイートは，地震に関連するツイートであると考えられる．実際に，10:20 頃に京都府南部を震源とする地震が発生し，この地震によって京都府を中心に震度 2 が観測されている(注5)．次に，このツイートに含まれるキーワード「地震」に関連するツイートの数を，日付毎にまとめたグラフを図 5 に示す．

図 5 から，「地震」に関連するツイートは 11 月 30 日と 12 月 6 日に集中しており，それ以外の日ではほとんどツイートされ

(注5)：気象庁 - 震度データベース
http://www.seisvol.kishou.go.jp/eq/shindo_db/shindo_index.html

表 1　プライベート型のツイート

投稿日時	内容	α	β_1
2010/12/01 00:12:37	もうリースとかツリーとか飾ってるお家も多いですね。	3.99	4.37
2010/12/01 11:06:00	映画見に行くよ！	4.00	4.55
2010/12/01 17:42:28	やけに腹が減った	4.11	2.94

表 2　パブリック型のツイート

投稿日時	内容	α	β_1
2010/11/30 10:21:15	地震？	2.85	2.62
2010/12/01 00:04:34	こんにちは１２月	3.51	3.16
2010/12/01 09:36:27	「小沢部隊」カネで形成、現金入り封筒手渡し http://www.yomiuri.co.jp/politics/news/20101201-OYT1T00267.htm	3.60	14.75

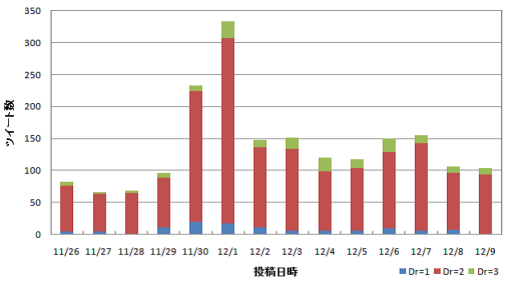

図 6　「12 月」関連ツイートの分布

ていないことがわかる．表 1 に示したツイートと比べても，このツイートの α が小さい値をとっていることから，関連ツイートの投稿日時の広がりが指標に反映されていることを確認できる．

上から二番目のツイートは，12 月が始まったことを表すツイートであると考えられ，図 6 のように 11 月 30 日から 12 月 1 日にかけて，多くのツイートが投稿されていることが分かる．また，図 5 に示した「地震」に関連するツイートと比べると，「12 月」に関連するツイートの方が投稿日時の広がりは大きい．α の値も，それぞれ 2.85 と 3.51 であり，投稿日時の広がりの大小が指標に表れている．

また，一番下のツイートのように，ニュースに関連するツイートもパブリック型のツイートに分類されていることが確認できる．

5.2.3　コミュニティ型のツイート

コミュニティ型に分類されたツイートを表 3 に示す．

表 3 の一番上のツイートのキーワードは「天一」(注6)であり，このツイートは，コミュニティ型のツイートに分類されたツイートの中でも特に β_1 の値が小さくなっている．キーワードに関連したユーザの数は，それぞれ $U_t(1) = 14$，$U_t(2) = 15$，$U_t(3) = 2$ であった．この中で，$D_r = 1$ であるユーザの中には一人で 10 個の関連ツイートを投稿しているユーザもいたが，$D_r = 2$ や $D_r = 3$ のユーザの中には，多くても一人当たり 3 個以下，すなわち 4 個以上の関連ツイートを投稿しているユーザはいなかった．すなわち，このツイートに関連するツイートは，読者に近い一部のユーザが中心となってツイートし，そのユーザに近いユーザがそれに反応して関連ツイートを投稿していると考えられる．β_1 には，このようなユーザの分布の特徴が表れていると言える．

他にも，二番目のツイートには，「授業」というキーワードが含まれており，学生であるユーザを中心にツイートされやすい内容であると考えられる．また，一番下のツイートは「プログラミング言語」に関連したツイートであり，プログラミング言語に興味があるユーザにツイートされやすい内容である．

以上のように，一部のユーザが中心となり多くのツイートを投稿しているような場合や，特定のコミュニティでツイートされやすい固有の話題などが，コミュニティ型のツイートに分類されている．

5.2.4　特殊型のツイート

特殊型に分類されたツイートを表 4 に示す．

表 4 に示したツイートは，いずれもコミュニティ型のツイートと同じように，関連ツイートを多く投稿するユーザが一人もしくは極少数存在し，さらに，時間的にも集中している，もしくは他の時間帯ではほとんどツイートされないような珍しい話題に関するツイートである．例えば，一番上のツイートのキーワードは，「ハル研」(注7)，「プロコン」(注8)であるが，図 7 に示すように，12 月には関連ツイートがほとんど投稿されていない．また，上から二番目のツイートには，駅名がキーワードとして含まれており，その駅がある地域に縁のあるユーザが関連ツイートを投稿していると考えられる．

このように，投稿日時・ユーザの広がり共に偏りがあったり，地域的に偏りがあるようなツイートが，特殊型のツイートに分類されている．

5.3　考　　察

5.3.1　ツイートの分類

5.2 節で示したツイートに関して，ユーザの広がりの指標 β_1, β_2 を比較する．表 1 から表 4 で示したツイートの一部について，β_1, β_2，及び $U_t(1), U_t(2), U_t(3)$ を表 5 に示す．ただし，ツイートに含まれる URL は省略した．また，β_2 の基準値も β_1 と同様に，分類したツイートの中での平均値 $\beta_2 = 0.388$ を用いる．

(注6)：中華そば専門店「天下一品」

(注7)：株式会社ハル研究所

(注8)：プログラミングコンテスト

表 3　コミュニティ型のツイート

投稿日時	内容	α	β_1
2010/11/30 02:46:30	いやさすがに。 しかし天一こってり食べたい ‥	4.33	0.25
2010/12/08 12:48:38	授業いこう … [はこ]	4.12	1.42
2010/12/08 20:05:06	プログラミング言語 ruby を図書館で借りてきた	4.08	1.37

表 4　特殊型のツイート

投稿日時	内容	α	β_1
2010/12/01 01:18:35	ハル研のプロコンを布教しよう。 http://www.hallab.co.jp/progcon/2010/question/	2.82	0.20
2010/12/01 10:15:38	出町柳なぅ	3.52	0.12
2010/12/01 21:55:44	さすがフェラーリ、乗りやすいぜ …· なんてことはない、ただのじゃじゃ馬になっている	2.71	1.39

表 5　β_1, β_2 の比較

内容	$U_t(1)$	$U_t(2)$	$U_t(3)$	β_1	β_2
もうリースとかツリーとか飾ってるお家も多いですね。	19	362	38	4.37	0.204
やけに腹が減った	63	802	89	2.94	0.461
地震？	27	320	25	2.62	0.177
「小沢部隊」カネで形成、現金入り封筒手渡し	1	70	3	14.75	0.036
プログラミング言語 ruby を図書館で借りてきた	33	193	24	1.37	0.117
出町柳なぅ	10	6	0	0.12	0.005

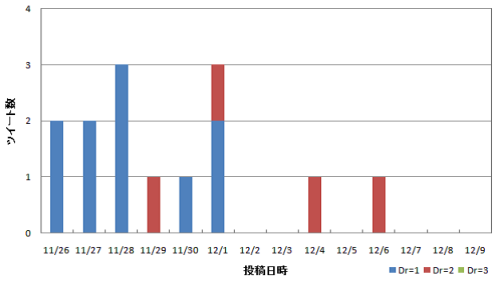

図 7　「ハル研」及び「プロコン」関連ツイートの分布

表 5 に示したツイートは，上からプライベート型のツイートとパブリック型のツイートを各 2 個と，コミュニティ型と特殊型のツイートから各 1 個を選んだものある．表 5 より，β_1 が小さいツイートでは β_2 も小さい値をとっているが，β_1 が大きいツイートについては，β_2 も大きい値をとるとは限らないことがわかる．特に，上から四番目のツイートではその差が顕著に表れている．β_1 では，$R(1)$ が 0 に近い場合，β_1 が非常に大きな値になりやすいと考えられる．すなわち，ユーザ全体で関連ツイートの数が少ない場合，ユーザの広がりをうまく評価できない可能性がある．一方で，β_2 はユーザの広がりよりも，ユーザ全体で関連ツイートを投稿しているユーザの数が，指標に大きく影響していると考えられる．

これより，β_2 よりも β_1 の方がユーザの広がりをよりよく表していると考えられるが，関連ツイートの数が少ない場合，特に $R(1)$ が非常に小さくなる場合を考慮する必要がある．

5.3.2　提 案 手 法

本論文では，投稿日時の差の標準偏差を投稿日時の広がりの指標として用いた．投稿日時の分布に関しては，特定の時間に集中しているかどうかだけでなく，ツイートが集中している時期がどのように分布しているかを考えることも重要である．例えば，表 2，及び図 5 で示した「地震」に関するツイートでは，対象ツイートの投稿日時に近い 11 月 30 日と，少し離れた 12 月 6 日に多くのツイートが投稿されていることがわかる．しかし，提案手法では，このような情報を指標に反映することができず，特定の時間に集中しているという情報しか得ることができない．このような分布の特徴は，混合ガウス分布のような混合分布モデルを考えることで，指標に反映することができると考えられる．混合分布モデルに基づく指標を考えることで，定期的にツイートが多く投稿される時間帯が表れるといった周期性を捉えることもできる．

ユーザの広がりに関しては，読者に近いユーザがそれ以外のユーザに比べて関連ツイートをどれだけ多く投稿しているかを指標として用いた．読者と他のユーザとの近接度を考える際に，フォロー関係を表すグラフ構造だけでなく，ユーザが投稿するツイートの中で出現頻度が高いキーワードや，プロフィールの類似度なども考慮することで，更に他のユーザとの関係を分類に反映することができると考えられる．

本論文で提案したユーザの広がりの指標は，読者と他のユーザの関係を利用したものであり，読者に依存した指標である．すなわち，この指標に基づく分類も読者に依存している．一般に，ツイートには様々な解釈が考えられ，ツイートの解釈は読者によって変化し得ると考えられるが，読者に依存した分類がふさわしくない場合も考えられ，別の指標を検討する際には，その指標が読者に依存したものであるかどうかを考慮することも必要である．

本手法では，ツイートの分類のために関連ツイートの分布を利用しており，関連ツイートを求める方法が分類の精度にも大きく影響する．本論文では，対象ツイートのキーワードを含むツイートを関連ツイートと定義した．この他に関連ツイートを求める方法として，TF-IDF とコサイン類似度を利用して

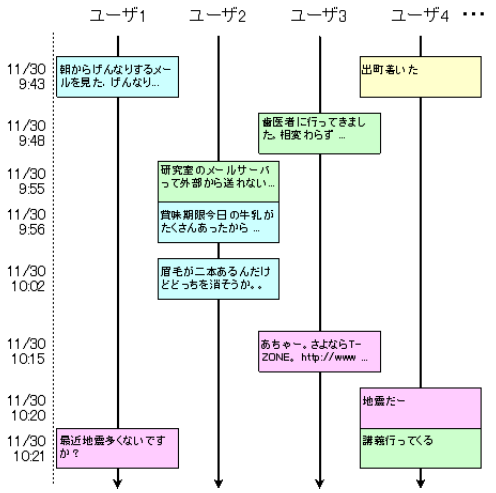

図 8　本手法を応用したシステムの例

ツイートの類似度を計算したり，キーワード間の類似度として Normalized Google Distance(NGD) [10] を利用することが挙げられる．NGD は，コルモゴロフ複雑性に基づく単語間の類似度の近似値を Google の検索結果の件数から求める手法である．また，Ramage ら [11] の手法のように，LDA(Latent Dirichlet Allocation) を利用することも考えられる．

本手法の応用として，図 8 に示すように，各ユーザのツイートを色分けして表示するようなシステムが考えられる．図 8 では，プライベート型のツイートは青，パブリック型のツイートは赤，コミュニティ型のツイートは緑，特殊型は黄色として色分けし，時系列に沿ってツイートを表示している．このようにすることで，読者はツイートのタイプを簡単に知ることができ，読みたいツイートを見つけやすくするのに役立つと考えられる．例えば，ニュースに関連していると考えられるパブリック型のツイートだけを読んだり，日常的な内容であることが多いプライベート型のツイート以外を読むなどで，読者が読みたいツイートを効率的に発見できる．他にも，α と β の値に従って色の濃さを変化させることで，投稿日時とユーザの広がりの程度を視覚的に表現することも考えられる．

このようなシステムを考える場合，ツイートを分析するために必要な計算時間も重要である．Twitter では大量のツイートがリアルタイムに投稿されるため，計算量の大きな分類手法では，実用性が損なわれてしまう．本手法では，各ツイートに対して，多くの関連ツイートの情報を基にスコアを算出しているため，計算量が非常に大きくなる可能性がある．計算量を削減するための方法として，関連ツイートの集合やそれらを投稿したユーザの集合を，キーワード毎に予め計算しておくことなどが挙げられる．

6. おわりに

本論文では，多くのツイートの中からユーザが読みたいツイートを発見しやすくするために，それぞれのツイートに対して，関連ツイートの投稿日時とそれらを投稿したユーザの広がりに基づくスコアを付加し，ツイートを分類するための手法を提案した．

投稿日時の広がりは，対象ツイートと関連ツイートの投稿日時の差の分布に着目し，対象ツイートと関連ツイートの投稿日時の差に基づく指標を用いた．また，ユーザの広がりでは，ユーザのフォロー関係を表すグラフ構造から読者と他のユーザとの非類似度を定義し，読者と他のユーザの近接度と関連ツイート投稿したユーザの数に基づく指標を用いた．これら二つの指標の大小から，ツイートを大きく 4 つのタイプに分類することができる．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た．

今後の課題としては，投稿日時の広がりやユーザの広がりに関する別の指標を検討することや，大規模なデータを用いて分類精度の評価実験を行うこと，そして，本手法を応用して，読みたいツイートを発見しやすくするためのシステムを実際に構築し，その効果を確認することが挙げられる．また，今後は Sriram ら [8] の手法との比較実験を行う予定である．

文　献

[1] Akshay Java, Xiaodan Song, Tim Finin, Belle Tseng:“ Why We Twitter: Understanding Microblogging Usage and Communities ”, Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pp. 56-65 (2007).

[2] Haewoon Kwak, Changhyun Lee, Hosung Park, Sue Moon:“ What is Twitter, a Social Network or a News Media? ”, WWW2010, pp. 591–600 (2010).

[3] Mor Naaman, Jeffrey Boase, Chih-Hui Lai :“ Is it Really About Me? Message Content in Social Awareness Streams ”, CSCW2010, pp. 189–192 (2010).

[4] 青島傳隼，福田直樹，横山昌平，石川博：“ マイクロブログを対象とした制約付きクラスタリングの実現 ”, DEIM2010, 2010.

[5] 岩木祐輔，アダム ヤトフト，田中克己：“ マイクロブログにおける有用な記事の発見支援 ”, DEIM2009, 2009.

[6] 田中淳史，田島敬史：“ twitter のツイートに関する分類手法の提案 ”, DEIM2010, 2010.

[7] Takeshi Sakaki, Makoto Okazaki, Yutaka Matsuo:“ Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors ”, WWW2010, pp. 851–860 (2010).

[8] Bharath Sriram, David Fuhry, Engin Demir, Hakan Ferhatosmanoglu, Murat Demirbas:“ Short Text Classification in Twitter to Improve Information Filtering ”, SIGIR2010, pp. 841-842 (2010).

[9] 服部峻：“ Web 知識を用いた時空間依存な対話システムの試作 ”, 電子情報通信学会技術研究報告, Vol. 110, No. 105, AI2010-3, pp. 13–18 (2010).

[10] Rudi L. Cilibrasi, Paul M.B. Vitanyi:“ The Google Similarity Distance ”, IEEE Transactions on Knowledge and Data Engineering, , Vol. 19, No. 3, pp. 370–383 (2007).

[11] Daniel Ramage, Susan Dumais, Dan Liebling:“ Characterizing Microblogs with Topic Models ”, ICWSM2010, pp. 130-137 (2010).