Twitterにおけるトピック間類似度を用いたトピック転換後の人気予測

岩永　雅史† 　田島　敬史† 　山肩　洋子††

† 京都大学大学院情報学研究科　〒606–8301 京都府京都市左京区吉田本町
†† 東京大学大学院情報理工学系研究科　〒113–8656 東京都文京区本郷 7 丁目 3 番 1 号
E-mail: †miwanaga@dl.soc.i.kyoto-u.ac.jp, †tajima@i.kyoto-u.ac.jp, ††yamakata@mi.u-tokyo.ac.jp

あらまし　Twitter において長期的にいいねやリツイートを獲得し続けるには，ツイート内容を現在人気の高いトピックに転換すべき場合がある．本研究では，ユーザーがそれまでに触れていたトピックと，トピック転換後に新しく触れるトピックとの類似度を考慮することで，トピック転換時のツイートがどの程度の人気を獲得出来るかを予測する．トピック間類似度の算出方法については, 各トピックに関連の深いアカウントを用いて算出する方法と各トピックについてツイートしたユーザーの集合を用いて算出する方法の二つを提案する．実験では，各トピックとしてビデオゲームを設定した場合とテレビ番組を設定した場合について，それぞれ実際のツイートデータを収集して予測精度の検証を行った．実験の結果として，特にトピック転換時のツイートのいいね数を予測する場合，提案手法の予測精度が高いことを示した．

キーワード　Twitter，ソーシャルネットワーク，マイクロブログ，人気予測．

1 はじめに

近年では Facebook[1]や Instagram[2]などの SNS が普及したことにより，誰もが気軽に情報を発信し，共有できるようになった．その中でも Twitter[3]は全世界で 3 億 3000 万人のアクティブユーザー [4]が存在する代表的な SNS であり，多くの企業が商品やイベントのプロモーションを目的としてアカウントを保有, 運営している．また，一個人もマイクロブログとしての利用のみならず，ファンの獲得や新たなコミュニティへの参加を求めて Twitter を利用する様子が多く見られる．Twitter において人気を獲得するためには，ユーザーの興味，関心を惹くトピックについてツイートすることが重要でり，より多くの人にとって関心のあるトピックを選択する必要がある．また，長期的に人気を獲得したい場合は，扱うトピックを新しいものに転換していく必要がある場合がある．なぜならば，一つのトピックに関してツイートし続けると，人気を獲得する速度が減少していくことが多いためである．人気の獲得速度が減少していく理由としては主に以下の二つが考えられる．

- 多くのトピックは時間経過によって関心が下がっていく．
- 一つのトピックに関心があるユーザーの数には上限があるため，実際にファンを獲得すればするほど，まだ獲得していない潜在的なファンの数は減少していく．

Twitter のユーザー全体として関心の高いトピックを見つける方法としては，公式の機能である「トレンド」のリストを利用することが考えられる．また，学術研究においても Twitter の大きなユーザー集合において関心の高いトピックを検出する手法 [1],[4],[5] は多く存在する．しかし，トピック転換時において，Twitter のユーザー全体として関心の高いトピックを扱うことで，必ずしてもリツイートやいいねといった人気を獲得出来るとは限らない．なぜならば，Twitter のユーザー全体にとって関心の高いトピックが，特定のユーザーのフォロワーにとっても関心が高いとは限らないからである．そこで，各ユーザーにパーソナライズしたトピックを検出する手法 [6],[7] を用いることも考えられる．これらの手法は，様々なトピックの中でも，各ユーザーやそのフォロワーに適したトピックを考えるため，よりトピック転換時に適している可能性が高い．しかし，これらの手法はユーザーにとってより良い情報の獲得することを目的として検索キーワードを生成する手法であり，いいね数やリツイート数といったツイートの人気を直接的に予測出来る手法ではなく，人気を獲得する目的には最適でない．

ツイートの人気を予測する手法としては，あるツイートを各ユーザーがリツイートする確率を予測する研究 [10],[11],[12] が多く存在する．これらの研究は，新トピックに関するツイートの人気を予測する手法となり得るが，予測出来るのはあくまで各ユーザーのリツイートの確率であり，総リツイート数ではない．リツイートする可能性のあるユーザー集合を定義すれば総リツイート数は算出できるが，特にリツイート数が大きくなる場合は，そのようなユーザー集合の定義や収集は非常に難しい．よって，人気を獲得する目的でツイートするユーザーの視点で考える場合，総リツイート数を直接計算出来ないという点でこれらの手法は最適とは言い難い．総リツイート数を予測する手法としては，ツイート後，初期段階のリツイート数を用いて最終的なリツイート数を予測する手法 [13],[14],[15] も多く存在す

1 : https://www.facebook.com/
2 : https://www.instagram.com
3 : https://twitter.com
4 : https://investor.twitterinc.com/home/default.aspx

るが，これはツイートを行う以前に予測を行えないという点で問題がある．

以上を踏まえ，本研究ではトピックの転換時において，新たなトピックを含むツイートが最終的にどの程度人気を獲得出来るかを予測する手法を考える．具体的には，各ユーザーがそれまでに触れていたトピック（以下では，旧トピックと呼ぶ）を考慮した上で，トピック転換時に新しく触れるトピックに関するツイートが最終的にどの程度いいね数またはリツイート数を得られるかを予測することを考える．なお，以下では新しく触れるトピックのことを新トピックと呼ぶが，新トピックとはそのユーザーにとって初めて触れるトピックのことを指し，世間一般にとっての新しいトピックとは必ずしも一致しない．例えば,「スーパーマリオブラザーズ」は長年多くの人々に親しまれてきたビデオゲームである．しかし，ある少年が初めて「スーパーマリオブラザーズ」というビデオゲームの存在を知り，その面白さについてツイートし始めた場合,「スーパーマリオブラザーズ」はその少年にとって新トピックと言える．旧トピック及び新トピックを定義した上で，本稿では二つの仮説を考える．

- 旧トピックに関するツイートが人気であるユーザーは，新トピックに関するツイートも人気が得やすい．
- 旧トピックと新トピック間の類似度が高いほど，新トピックに関するツイートも人気が得やすい．

一つ目の仮説はどのトピックを選択しているかに関係なく，各ユーザーの素質に着目したものであり，直感的に予想できる．より細分化して考えると，旧トピックに関するツイートが人気であるユーザーは，人気が得られるような興味深いツイートをする技術に長けており，新トピックに関しても興味深いツイートをすることができると考えられる．また，旧トピックに関するツイートに人気であるユーザーは既にフォロワーを多く獲得しており，新トピックに関するツイートを閲覧される回数も多くなる．よって，旧トピックに関するツイートが人気であるユーザーは，新トピックに関するツイートも人気が得やすいと考えられる．

二つ目の仮説は，各ユーザーのフォロワーの性質に着目している．各ユーザーのフォロワーは，そのユーザーが触れてきた旧トピックへの関心が高いためにフォローを行っていることが多いと考えられる．そのため，旧トピックと類似性が高い新トピックにに対しても関心が高く，その結果として人気が得られることが予想できる．なお，トピックを転換する過程では，旧トピックに関するツイートと新トピックに関するツイートをそれぞれどのような頻度で行うか，また，どのような順序で行うかなども人気の獲得に関わると考えられる．例えば，ある新作ビデオゲームに関して 1 度だけツイートし，その後は他のビデオゲームに触れる場合と，新作ビデオゲームに関するツイートのみを 10 回連続して行った場合とでは，その新作ビデオゲームに関心のあるユーザーにフォローされる確率は異なることが予想される．つまり，新しいトピックに転換する行為は，新しいトピックを選出する段階と，実際にツイートする段階の二つに分けることができる．本研究では，まず前者の新しいトピックを選出する段階に注目し，新しいトピックに変更した直後にどれだけ人気が得られるかを予測する．

旧トピックと新トピック間の類似度は以下の二つの情報を用いて算出する．

- 旧トピックに関係の深いアカウントのフォロワーのうち，新トピックに関係の深いアカウントのフォロワーでもあるユーザーの割合
- 旧トピックについてツイートしているユーザーのうち，新トピックについてもツイートしているユーザーの割合

トピック間類似度は旧トピックに関心の高いユーザーのうち，新トピックに対する関心も高いユーザーの割合を表現しようとするものである．

トピック間類似度を用いてツイートの人気を予測する手法として二つの手法を提案する．いずれも目的変数は新トピックに関するツイートのいいね数またはリツイート数である．一つ目の手法は，それまでのツイートのいいね数またはリツイート数，フォロワー数，フォロー数といった説明変数に，トピック間類似度を加えて回帰予測を行う手法である．二つ目の手法は，それまでのツイートのいいね数またはリツイート数の代わりに，いいね数またはリツイート数とトピック間類似度の積を説明変数として用いて回帰予測を行う手法である．

また，提案手法を評価するために，Twitter REST API を使用して収集した実際の Twitter 上のデータを用いて二つの実験を行った．一つ目の実験では，旧トピックとして 2018 年以前にリリースされたビデオゲーム 11 個，新トピックとして 2019 年 11 月以降にリリースされたビデオゲーム 11 個を設定した．二つ目の実験では，旧トピックとして 2019 年 9 月から 12 月にかけて放送されたアニメ番組 15 個，新トピックとして 2020 年 1 月以降に放送開始されたアニメ番組 15 個を設定した．旧トピックと新トピックの全ての組み合わせに対して，トピック間類似度を算出し，二つの提案手法によって人気を予測した．

予測精度の評価基準は決定係数 R^2，Root Mean Squared Error (RMSE)，Mean Absolute Error (MAE)，Mean Absolute Percentage Error (MAPE) とし，各説明変数の寄与率は標準偏回帰係数を用いて比較した．

結果として，トピック間類似度を説明変数として加える一つ目の提案手法は，ベースライン手法より予測精度が僅かに上回ることが示された一方で，トピック間類似度は説明変数として精度への寄与率が低く，導入するコストに見合った適切な説明変数とは言い難い結果となった．旧トピックに関するツイートのいいね数またはリツイート数とトピック間類似度の積を説明変数として加える二つ目の提案手法は，多くの場合でベースライン手法の精度を上回った．特に，各トピックに関係の深いアカウントのフォロワー集合を用いて算出したトピック間類似度を用いた場合において，ベースライン手法の精度を大きく上回った．

本稿では以下の構成を取る．まず 2 章で関連研究について述べ，続いて 3 章ではトピック転換時の人気を予測するための二つの提案手法について述べる．4 章では提案手法に対する実験とその結果について説明し，結果を踏まえて考察を述べる．最後に 5 章では本論文の結論と今後の課題を述べる．

2 関 連 研 究

本節では，本研究と関連する研究について言及し，本研究の位置づけについて述べる．

2.1 流行している関心の高いトピックを検出する手法

Twitter 上で関心の高くなっている流行トピックを検出する手法は多く提案されている．流行トピックの検出に関する研究の多くはトピックの判定にクラスタリングまたはトピックモデリングを利用しているが，Sapul ら [1] は．k-means, CLOPE クラスタリング [2]，Latent Dirichlet Allocation(LDA)[3] アルゴリズムのそれぞれによるトピック検出の比較を行った．結果として CLOPE クラスタリングが多くのトピックパターンを提示できる一方で，キーワードとハッシュタグの特徴セットを追加することで k-means と LDA はより有意義なトピックを識別できることを示した．Benhardus と Kalita[4] はストリーミングデータに対して tf-idf による語への重み付けを用いて分析を行い，特にユニグラムとバイグラムで流行トピックを検出出来ることを示した．Xie ら [5] は，リアルタイムでトピックモデリングアルゴリズムを用いる手法として，スケッチベースのトピックモデルを提案した．この手法は，1 日にツイートされる全てのツイートデータ以上の膨大なストリーミングデータに対応してトピックモデリングを行うことができ，爆発的な流行を検出できるとした．これらの手法によって，Twitter 上で一般的に関心の高いトピックを検出することが出来る．

しかし，トピック転換時において，あるユーザーが Twitter 上で一般的に関心の高いトピックを扱うことで，必ずしも人気を獲得出来るとは限らない．なぜならば，Twitte 上で一般的に関心の高いトピックが，そのユーザーのフォロワーにとっても関心が高いとは限らないからである．そこで，流行トピックを検出する手法をフォロワー集合にのみ適用することが考えられる．加えて，各ユーザーやコミュニティにパーソナライズしたトピックを検出する手法としては，Cataldi ら [6] の研究がある．この研究では，ユーザーのツイート内容を分析し，ユーザーに最も関連性のあるトピックが検索できるようなキーワードを一定期間ごとに生成していく手法を提案した．さらに Cataldi ら [7] はフォロー関係を踏まえてユーザーのコミュニティ内での影響力を考慮することで，この手法をさらに拡張した．また，新トピックに対するユーザーの意見を予測する研究として Fuji と Ye の研究 [8] がある．各ユーザーのフォロー関係から得られる社会的要素と，各ユーザーが扱っていた旧トピックの類似性から得られる文脈的要素をそれぞれ考慮することで，未知のトピックに対するユーザーの意見を予測する手法を提案した．さらに，Wu ら [9] は OpinionFrowto と呼ばれる分析システムを導入することで，意見伝搬の流れを可視化した．このシステムでは様々なレベルで関心のあるトピックを選択した上で，意見伝搬の様子を比較できるようにしている．

これらの手法は，各ユーザーやそのフォロワーに最適化した上で新トピックを扱うという点で，よりトピック転換時の予測に適していると言える．しかしこれらの手法は，より良い情報の獲得や，トピックに対する意見の予測を目的とした手法である．よって，人気の獲得を目的として，いいね数やリツイート数といったツイートの人気を直接的に予測出来る手法ではない．加えて，現在のフォロワー集合に最適化したトピックは，新たに人気を獲得するためのターゲットとなる非フォロワー集合の人気を獲得出来ない可能性もある．つまり，より人気を獲得しようとトピックを転換する際は，現在のフォロワー集合と，新たにフォロワーになり得る非フォロワー集合の両方にとって，バランス良く関心の高いトピックを選択する必要があると言える．

2.2 ユーザーがリツイートする確率を予測する手法

あるツイートをユーザーがリツイートする確率を予測する研究は多く存在する．Liu ら [10] は，ファジー理論とニューラルネットワークを用いて，流行しているトピックに関するリツイート行動を予測する手法を提案した．加えて，流行トピックの人気の時間的変化を動的に把握できることを示した．Tang ら [11] はユーザーがリツイートする際の社会的な類似性を分析し，個々のユーザーが持つ情報と組み合わせることによって，個々のユーザーのリツイート行動を予測する新たなモデルを提案した．Huang ら [12] はユーザーのツイートをベイズモデルによって様々なカテゴリに分類し，各カテゴリのツイートに対するユーザーの関心を算出することでリツイート行動を予測する手法を提案した．

これらの研究は，リツイート確率を予測する手法であり，新トピックに関するツイートの人気を予測する手法となり得る．しかし，あくあで予測出来るのは各ユーザーのリツイート確率であり，総リツイート数ではない．リツイートする可能性のあるユーザー集合を定義すれば総リツイート数は算出できるが，特にリツイート数が大きくなる場合は，そのようなユーザー集合の定義や収集は非常に難しい．なぜならば，リツイートされるに従ってツイートを閲覧する可能性のあるユーザーは爆発的に増えるからである．よって，人気を獲得する目的でツイートするユーザーの視点で考える場合，総リツイート数を直接計算出来ないという点でこれらの手法は最適とは言い難い．

2.3 ツイートの総リツイート数を予測する手法

総リツイート数を予測する手法としては，ツイート直後のリツイート数の時間変化を用いて，最終的な総リツイート数を予測する研究 [13],[14],[15] が多く存在する．しかし，人気を獲得する目的でツイートするユーザーの視点で考える場合，ツイートする前に人気を予測出来る必要がある．ツイート後のデータを用いずに総リツイート数を予測する手法を示した Can らの研究 [16] がある．この研究では，画像付きのツイートに関して，フォロワー数やフォロー数といった基本的な特徴に加え，画像の特徴を算出して拡張することで総リツイート数をより良い精度で予測する手法を示した．

2.4 本研究の位置づけ

以上を踏まえ，本研究ではユーザーがツイートしてきた旧ト

ピックを考慮し，各ユーザーやそのフォロワーに最適化した上で，トピック転換時に新しいトピックに関するツイートがどれだけ人気を獲得できるかを予測する手法を提案する．

3 提案手法

本節では，提案する手法について詳細を述べる．本研究の目的は，ツイートするか検討している新トピックを与えた上で，ユーザーがツイートしてきた旧トピックを考慮することで，トピック転換時の新トピックに関するツイートの人気を予測する精度を向上させることである．

以下の手順により，各ユーザーに対して，旧トピックと新トピックのそれぞれに関するツイートデータを収集する．

1). 旧トピックに関するツイート及び新トピックに関するツイートを検出するためのクエリをそれぞれ与える

2). クエリを用いて新トピックに関するツイートデータを収集する

3). 得られたツイートデータから新トピックについてツイートしたユーザーの集合を求める

4). 各ユーザーのツイートを最新のものから順に取得する

5). 各ユーザーのツイートデータから旧トピック，新トピックに関するツイートデータをそれぞれ抽出する．

なお，本研究はユーザーがまだ触れたことのない新トピックに関するツイートの人気を予測することが目的であるため，旧トピックに関するツイート以前に新トピックに関するツイートを投稿しているユーザーのデータは除く必要がある．

3.1 トピック間類似度の算出

本節ではトピック間の類似度を算出する手法について述べる．本研究の目的は，旧トピックから新トピックへ転換した際の人気を予測することである．つまり，本研究におけるトピック間類似度とは，旧トピックから新トピックへ転換する際に適用できるものでなければならない．よって，同じ二つのトピックの組み合わせを仮定したとしても，どちらのトピックを新トピックとして与えるかによってトピック間類似度は異なる場合があることに注意が必要である．旧トピックに対する関心の高さを基準とした上で，どれだけ新トピックに対する関心が高いかを表現することを試みる．本研究ではトピック間類似度を以下の二つの情報を用いて算出する．どちらのトピック間類似度も旧トピックに関心の高いユーザーのうち，新トピックに対する関心も高いユーザーの割合を表現しようとするものである．

3.1.1 トピック間類似度 S_1

旧トピックに最も関係の深いアカウント及び新トピックに最も関係の深いアカウントをそれぞれ設定する．旧トピックに関係の深いアカウントのフォロワーの集合を F_o , 新トピックに関連の深いアカウントのフォロワーの集合を F_n とする．トピック間類似度 S_1 は以下のように求める．

$$S_1 = \frac{|F_o \cap F_n|}{|F_o|}$$

トピック間類似度 S_1 は，各トピックに関係の高いアカウントをフォローしているユーザーはそのトピックに対する関心が高いということを仮定している．この提案手法では，旧トピックに関係の深いアカウントは既知のものであり，与えられるものとして考えるが，実際はトピックに関係の深いアカウントを発見することが難しい場合や，与えられたアカウントでは予測の精度が優れない可能性がある．そのような場合，Zengpin と Gndz [17] の研究によって提案された Personalized PageRank を用いて各トピックに関係の深いアカウントを発見することが考えられる．この研究では，Personalized PageRank を用いることで，特定のトピックに関するインフルエンサーを検出することができるとしている．

3.1.2 トピック間類似度 S_2

旧トピックに関するツイートを検索するクエリ，新トピックに関するツイートを検索するクエリをそれぞれ与える．それぞれのクエリを用いて，ツイートデータから旧トピックについてツイートしたユーザーの集合 U_o, 旧トピックについてツイートしたユーザーの集合 U_n を収集する．トピック間類似度 S_2 は以下のように求める．

$$S_2 = \frac{|U_o \cap U_n|}{|U_o|}$$

トピック間類似度 S_2 は，各トピックに関してツイートをしたことのあるユーザーは，そのトピックに対して関心が高いということを仮定している．トピック間類似度 S_2 は，ツイートが各トピックに関係したものであるかを判定することができれば，算出することができる．よって，トピック間類似度 S_1 に比べて汎用性が高いと言える．さらに，計算対象とするツイートデータのツイート時期を制限することで，各トピックへの関心の時間変化に対応することが期待できる．

3.2 線形重回帰モデルによる予測

トピック間類似度を用いてツイートの人気を予測する手法として二つの提案手法を示す．いずれも線形重回帰モデルを用いた手法であり，目的変数は新トピックに関するツイートのいいね数またはリツイート数である．一つ目の手法は，それまでのツイートのいいね数またはリツイート数，フォロワー数，フォロー数といった説明変数に，トピック間類似度を加えて回帰予測を行う手法である．二つ目の手法は，それまでのツイートのいいね数またはリツイート数の代わりに，いいね数またはリツイート数とトピック間類似度の積を説明変数として用いて回帰予測を行う手法である．

3.2.1 提案手法 1

新トピックに関してツイートした各ユーザーのツイートデータのうち，旧トピックに関するツイートに対するいいねの数の平均を F_{old}, リツイートの数の平均を R_{old} とする．また，新トピックに関するツイートに対するいいねの数を F_{new}, リツイートの数を R_{new} とする．

まず，新トピックに関するツイートに対するいいねの数を予測する場合について述べる．目的変数を F_{new} とし，説明変数として以下の九つを設定した線形重回帰モデルを用いる．各ユーザーに関する説明変数として以下の四つを設定する．

- 各ユーザーのフォロー数

- 各ユーザーのフォロワー数
- 各ユーザーのフォロワー数 / フォロー数
- 各ユーザーの総ツイート数
- 各ユーザーの F_{old}

また，新トピックに関するツイートの説明変数として以下の四つを設定する．

- ツイートが含むハッシュタグの数
- ツイートが含む URL の数
- ツイートによるメンションの数
- ツイートの文字数

加えて，トピック転換に関する説明変数として以下を設定する．

- トピック間類似度 S_1 または S_2

最小二乗法を用いて，F_{new} の残差二乗和が最小となるような各偏回帰係数及び切片を求めることで，回帰直線を定める．

新トピックに関するツイートに対するリツイートの数を予測する場合についても同様にしてモデルを生成する．つまり，目的変数を新トピックに関するツイートに対するリツイート数 R_{new} とし，説明変数 F_{old} の代わりに，旧トピックに関するツイートに対するリツイート数の平均 R_{old} を設定する．その他の八つの説明変数はいいね数を予測する場合と共通のものを用いて，回帰直線を定める．

3.2.2 提案手法 2

まず，新トピックに関するツイートに対するいいねの数を予測する場合について述べる．目的変数を F_{new} とし，提案手法 1 でも設定したものと共通の説明変数として，各ユーザーのフォロー数，フォロワー数，総ツイート数，新トピックに関するツイートが含むハッシュタグの数，URL の数，メンションの数，文字数を設定する．加えて，旧トピックに関するツイートに対するいいね数の平均 F_{old} と，トピック間類似度の積，つまり $F_{old}S_1$ または $F_{old}S_1$ を説明変数に設定する．その上で最小二乗法を用いて，F_{new} の残差二乗和が最小となるような各偏回帰係数及び切片を求めることで，回帰直線を定める．

新トピックに関するツイートに対するリツイートの数を予測する場合についても同様の手順でモデルを生成する．つまり，目的変数を新トピックに関するツイートに対するリツイート数 R_{new} とし，説明変数 $F_{old}S_1$ の代わりに $R_{old}S_1$，$F_{old}S_2$ の代わりに $R_{old}S_2$ を設定する．

4 実　　験

本節では，各提案手法を用いて行った実験とその評価について述べ，結果に対する考察を行う．本実験の目的は提案手法の有効性について検証することである．実験は，各トピックとしてビデオゲームを設定した場合と，アニメ番組を設定した場合の二つの実験を行った．線形重回帰モデルによる予測には，scikit-learn[5]の LinearRegression ライブラリを使用した．

4.1 データセット

一つ目の実験では，旧トピックとして 2018 年以前にリリースされたビデオゲーム 11 個，新トピックとして 2019 年 11 月以降にリリースされたビデオゲーム 11 個を設定した．旧トピックから新トピックへ転換したユーザーとして，10176 アカウントを収集した．二つ目の実験では，旧トピックとして 2019 年 9 月から 12 月にかけて放送されたアニメ番組 15 個，新トピックとして 2020 年 1 月以降に放送開始されたアニメ番組 15 個を設定した．旧トピックから新トピックへ転換したユーザーとして，9028 アカウントを収集した．なお，各トピックとしてビデオゲームやアニメ番組を設定した主な理由は，制作会社公式のハッシュタグを用いて Twitter 上のトピック転換時のデータを効率よく収集できると考えたためである．

トピック間類似度 S_2 を算出する際に用いるトピックに関係の深いアカウントは，それぞれのビデオゲームやアニメ番組の制作会社が運営する公式アカウントとした．一つのビデオゲームやアニメ番組に対して公式アカウントが国や地域別に複数存在する場合は，日本向けのアカウントを選択した．各トピックに関するツイートを検索する際のクエリは，各公式アカウントが利用しているハッシュタグと同一の文字列とした．

1 章で定義したように，新トピックは必ずしも世間一般として新しいトピックである必要はないが，今回の実験では世間一般としても新しいトピックを新トピックとして設定している．その理由は主に二つある．一つ目は，実際にはユーザーが初めて触れるトピックは，世間一般としても新しいトピックである可能性が高いためである．二つ目は，世間一般に古いトピックを新トピックとした場合，取得できなかった古いツイートデータの中に新トピックに関するツイートデータが存在し，実際にはトピック転換した場合に当てはまらない場合を計算に含んでしまう可能性があるからである．

データを収集する際には Twitter REST API[6]を利用し，実際の Twitter 上のデータを収集した．

4.2 評 価 指 標

実験では，提案手法 1，提案手法 2, ベースライン手法のそれぞれについて決定係数 R^2 ，Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) を算出し，比較することで予測精度の比較を行う．なお，MAPE は以下の式で算出する．

$$MAPE = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{\hat{y_i} - y_i}{y_i}\right|$$

RMSE や MAE で評価した場合，いいね数やリツイート数が大きいデータにおいて誤差が出やすい．よって，MAPE と比較して相対的にいいね数やリツイート数が大きいツイートを重視した評価指標だと言える．一方で MAPE で評価した場合は，相対的にいいね数やリツイート数が小さい場合を重視した評価となる．本実験では，いいね数やリツイート数が小さいツイートを重視した場合と，いいね数やリツイート数が大きいツイート場合の両方を重視した場合のそれぞれの視点で考えるため，

5 : https://scikit-learn.org/stable/

6 : https://developer.twitter.com/en/docs

RMSE, MAE, MAPE の全てで評価することとした．また，各説明変数の標準偏回帰係数を比較することによって，各説明変数の寄与率を比較する．偏回帰係数によって説明変数の寄与率を比較するには，各説明変数及び目的変数を標準化する必要がある．よって用いるデータの任意の値 x は以下の式に従って標準化の処理を行う．

$$x \longmapsto \frac{x - \overline{x}}{s}$$

なお，以下で示す実験結果は，R^2，RMSE, MAE，MAPE の全てにおいて，標準化の処理を行った上で算出したものであり，MAPE が 100 を上回る場合もある．

4.3 実験結果

以下では，トピック間類似度を説明変数に加えて回帰予測を行う提案手法 1 を GS, いいね数またはリツイート数とトピック間類似度の積を説明変数として用いて回帰予測を行う提案手法 2 を GM と示す．その中でも，トピック間類似度 S_1 を用いていいねの数を予測する手法をを GSF_1,GMF_1 と示し，トピック間類似度 S_2 を用いていいねの数を予測する手法を GSF_2，GMF_2 と示す．また同様に，トピック間類似度 S_1 を用いてリツイートの数を予測する手法を GSR_1，GMR_1 と示し，トピック間類似度 S_2 を用いてリツイートの数を予測する手法を GSR_2，GMR_2 と示す．

ベースライン手法は，提案手法 1 の説明変数からトピック間類似度を除いた重回帰モデルによる予測とした．なお，以下ではベースライン手法を B と示す．その中でも，いいねの数を予測する手法をを BF, リツイートの数を予測する手法を BR と示す．

4.3.1 ビデオゲームを対象とした実験結果

各トピックとしてビデオゲームを設定した実験の結果を以下の表 1，表 2 に示す．表 1 は各手法における R^2，RMSE, MAE, MAPE の値を示すものである．表 2 は各手法における説明変数に対応する標準偏回帰係数を示すものである．S_n は，トピック間類似度 S_1 または S_2 のいずれかを指す．各手法に特有な説明変数及び，各手法に共通する説明変数のうち最も標準偏回帰係数の大きかったユーザーのフォロワー数 (Followers) について記載している．なお，この重回帰予測においては，説明変数が互いに影響を及ぼし合いながら目的変数に影響を与えていることが予想できるため，標準偏回帰係数による比較は絶対的な信頼がある訳ではないことに注意する必要がある．

トピック間類似度を直接説明変数として加える一つ目の提案手法 GS については，各予測精度においてベースライン手法 BF, BR と大きな差がない．また，標準偏回帰係数においても，他の説明変数に比較してトピック間類似度 S_n に対する標準偏回帰係数の絶対値は小さく，相対的に寄与率は低いと考えられる．よって GS においてトピック間類似度は新トピックに関する人気を予測する説明関数として優れた特徴量だとは言い難い．

二つ目の提案手法である GM においては，特にトピック間類似度 S_1 を用いた GMF_1,GMR_1 がベースライン手法の精度を比較的大きく上回り，トピック間類似度 S_2 を用いた GMF_2,GMR_2 においてもベースライン手法を上回っていることがわかる．

表 1　ビデオゲームを対象した際の予測精度

	R^2	MAE	RMSE	MAPE
BF	0.1410	0.002187	0.02111	173.5
GSF_1	0.1427	0.002334	0.02108	188.6
GSF_2	0.1412	0.002225	0.02110	189.8
GMF_1	0.2155	0.001986	0.02017	157.1
GMF_2	0.1542	0.002146	0.02094	169.6
BR	0.1309	0.002892	0.02496	99.8
GSR_1	0.1321	0.002949	0.02494	108.1
GSR_2	0.1309	0.002897	0.02496	100.2
GMR_1	0.1864	0.002730	0.02415	86.4
GMR_2	0.1346	0.002898	0.02490	96.4

表 2　ビデオゲームを対象した際の標準偏回帰係数

	F_{old}	$F_{old}S_n$	S_n	Followers
BF	0.504074	-	-	0.039071
GSF_1	0.500953	-	0.002809	0.038978
GSF_2	0.503743		0.002904	0.039197
GMF_1	-	0.497145	-	0.034257
GMF_2		0.481772	-	0.037343
	R_{old}	$R_{old}S_n$	S_n	Followers
BR	0.458020	-	-	0.052689
GSR_1	0.455297	-	0.002717	0.052659
GSR_2	0.457878	-	0.001199	0.052745
GMR_1	-	0.535426	-	0.055546
GMR_2	-	0.518203	-	0.058916

表 3　アニメ番組を対象した際の予測精度

	R^2	MAE	RMSE	MAPE
BF	0.7645	0.000699	0.00540	222.7
GSF_1	0.7647	0.000721	0.00540	200.1
GSF_2	0.7645	0.000699	0.00540	222.0
GMF_1	0.9164	0.000593	0.00322	171.9
GMF_2	0.7701	0.000691	0.00534	214.8
BR	0.5065	0.003277	0.01998	513.7
GSR_1	0.5065	0.003282	0.01998	515.8
GSR_2	0.5065	0.003284	0.01998	505.8
GMR_1	0.4361	0.003390	0.02135	608.0
GMR_2	0.5133	0.003240	0.01984	626.6

4.3.2 アニメ番組を対象とした実験結果

各トピックとしてビデオゲームを設定した実験の結果を以下の表 3，表 4 に示す．表 1 と同様，表 3 は各手法における R^2，RMSE, MAE，MAPE の値を示すものである．表 2 と同様，表 4 は各手法における説明変数に対応する標準偏回帰係数を示すものである．

まず，トピック間類似度を直接説明変数として加える一つ目の提案手法 GS について考える．ビデオゲームを対象とした実験と同様，各予測精度においてベースライン手法 BF, BR と大きな差がない．また，標準偏回帰係数で比較しても，相対的にトピック間類似度 S_n の寄与率は低いと考えられる．よって，アニメ番組を対象とした場合においても，GS は優れた手法だ

表 4　アニメ番組を対象した際の標準偏回帰係数

	F_{old}	$F_{old}S_n$	S_n	Followers
BF	0.575577	-	-	0.134755
GSF_1	0.575695	-	0.000756	0.134941
GSF_2	0.575571	-	0.000118	0.134751
GMF_1	-	0.891235	-	0.067850
GMF_2	-	0.584735	-	0.133284
	R_{old}	$R_{old}S_n$	S_n	Followers
BR	0.566084	-	-	0.587786
GSR_1	0.565857	-	-0.000866	0.587591
GSR_2	0.565994	-	0.002321	0.587692
GMR_1	-	0.580132	-	0.577704
GMR_2		0.778612		0.499741

とは言い難い．

二つ目の提案手法である GM においては，ビデオゲームを対象とした場合と異なる点がある．いいね数を予測した GMF_1，GMF_2 においては，ビデオゲームの場合と同様，特に GMF_1 がベースライン手法 BF の精度を比較的大きく上回っている．一方で，トピック間類似度 S_1 を用いてリツイート数を予測した GMR_1 では，各評価指標でベースライン手法 BR を下回っている．トピック間類似度 S_2 を用いた GMR_2 においては，R^2，MAE はベースライン手法 BR を上回っているものの，RSME, MAPE においてベースライン手法 BR を下回っている．このことから，特に値が小さいデータに関して予測値が大きく外れていることが予想される．

一つ目の提案手法 GS は，二回の実験，二つのトピック間類似度それぞれの組み合わせにおいて，ベースライン手法と比較して精度の差が小さく，説明変数としてのトピック間類似度の寄与率は低かった．よって，トピック間類似度を算出し，説明変数として加えるコストに見合う優れた手法とは言い難い．

二つ目の提案手法 GM のうち，トピック類似度 S_1 を用いた場合は，多くの場合，最も予測精度が高かった．一方で，アニメ番組のリツイート数を予測するにあたっては，ベースライン手法を下回った．よって，精度の高い予測を出来る場合がある一方で，汎用性の高い手法であるかどうかに関しては疑問が残った．トピック間類似度 S_2 を用いた場合においては，ほとんどの評価指標で予測精度がベースライン手法を上回ったが，多くの場合で．トピック間類似度 S_1 を用いた場合より相対的に予測精度が低かった．以上を踏まえると，今回の実験で扱ったビデオゲームやアニメ番組以外のトピックにおいても実験を行い，幅広いトピックに対して適用できる手法であるかを検証する必要がある．

5　まとめと今後の課題

本研究では，ユーザーがツイートしたことのある旧トピックの存在を考慮した上で，新しいトピックに関するツイートがどの程度人気であるかを予測する手法を示した．旧トピックと新トピックのトピック間類似度を算出し，重回帰予測の説明変数やその一部として加えることで，より高い精度の回帰予測を試みた．

二つの実験の結果として，トピック間類似度を説明変数として加える一つ目の提案手法は，ベースライン手法より予測精度が僅かに上回る一方で，トピック間類似度は説明変数として精度への寄与率が低く，導入するコストに見合った適切な説明変数とは言い難い結果となった．トピック転換前のツイートのいいね数またはリツイート数とトピック間類似度の積を説明変数として加える二つ目の提案手法は，多くの場合でベースライン手法の精度を上回った．特に，各トピックに関係の深いアカウントのフォロワー集合を用いて算出したトピック間類似度を用いた場合において，ベースライン手法の精度を大きく上回った．

本研究における今後の課題を以下に示す．

- 旧トピックの判定の改善
- 複数の旧トピックの考慮
- トピック転換の過程の考慮

まず一つ目に関してだが，今回の手法では旧トピックをハッシュタグによってのみ判定した．よって，ユーザーのツイートデータのうち，ハッシュタグが含まれるものしか活用出来ていないと言える．つまり，実際には触れたことのあるトピックが他に存在しても設定した旧トピックに含まれていない場合や，想定したトピックに関するツイートもハッシュタグによって検索出来ていない可能性が考えられる．考慮できていないトピックが存在すれば，そのトピックに触れたことによって獲得している新トピックに関心のあるユーザーの存在も考慮できていないことになる．出来るだけ多くのツイートを各トピックに分類するために，クラスタリングやトピックモデリングを用いてトピックを定義，判定する手法も考えられる．クラスタリングやトピックモデリングを用いた Twitter 上でのトピック関する研究は多く行われているため，それらを応用することを検討したい．一方その場合，多くのツイートデータを何らかのトピックに分類することができるようになる反面，今回提案したトピック間類似度の算出方法を適用することが難しくなると予想できる．

二つ目に関して，今回の実験では 1 人のユーザーが複数の旧トピックについて触れたことがある場合を考慮しなかった．トピックの定義にもよるが，実際には多くの場合，1 人のユーザーが複数のトピックを同時に扱っている．旧トピック間の類似度も定義し，考慮する等の手法を取ることでさらに予測精度を上げることができる可能性がある．同様に，同時に複数の新トピックについてツイートし始める場合も検討したい．

三つ目に関して，1 章で触れたように，実際に新たなトピックに触れる場合，トピックを完全には転換せず，旧トピックと新トピックの両方についてツイートすることが考えられる．この時，旧トピックと新トピックのそれぞれに触れる割合や頻度によって，獲得できる人気も異なることが予想できる．今回の手法では時系列について，新トピックに関するツイートの方が旧トピックに関するツイートより新しいという条件しか考慮しなかったが，時系列データとしてツイートデータを扱うことによって，フォロワー数も含めたより長期的な人気を予測するための手法を提案したい．

謝　　辞

本研究は，JST CREST (JPMJCR16E3)，JSPS 科研費 18H03245，JSPS 科研費 18K11425 の支援を受けたものである．

文　　献

[1] Ma Shiela C Sapul, Than Htike Aung, Rachsuda Jiamthapthaksin, "Trending topic discovery of Twitter Tweets using clustering and topic modeling algorithms", 2017 14th International Joint onference on Computer Science and Software Engineering (JCSSE), pp. 1-6, 2017.

[2] Yiling Yang, Xudong Guan, Jinyuan You, "CLOPE: A fast and effective clustering algorithm for transactional data", Proceedings of KDD ' 02. , 2002(pg. 682-7)

[3] David M. Blei, Andrew Y. Ng, Michael I. Jordan, "Latent dirichlet allocation", Journal of Machine Learning Research, vol. 3, pp. 993-1022, Jan 2003.

[4] James Benhardus, Jugal Kalita,"Streaming trend detection in Twitter", International Journal on Web Based Communities, 9 (1) 2013, pp. 122-139

[5] Wei Xie, Feida Zhu, Jing Jiang, Ee-Peng Lim, Ke Wang, "Topicsketch: Realtime bursty topic detection from twitter", ICDM, pp. 837-846, 2013.

[6] Mario Cataldi, Luigi Di Caro, Claudio Schifanella, "Personalized emerging topic detection based on a term aging model", ACM Trans. Intell. Syst. Technol., vol. 5, no. 1, 2013.

[7] Mario Cataldi, Luigi Di Caro, Claudio Schifanella, "Emerging topic detection on Twitter based on temporal and social terms evaluation", Proceedings of the Tenth International Workshop on Multimedia Data Mining, p.1-10, July 25-25, 2010, Washington, D.C.

[8] Ren Fuji, Wu Ye, "Predicting user-topic opinions in twitter with social and topical context", IEEE Trans. Affect. Comput., 4 (4) 2013, pp. 412-424

[9] Yingcai Wu, Shixia Liu, Kai Yan, Mengchen Liu, Fangzhao Wu, "OpinionFlow: Visual analysis of opinion diffusion on social media", To appear in IEEE Transactions on Visualization and Computer Graphics, vol. 20, no. 12, 2014.

[10] Yanbing Liu, Jinzhe Zhao, Yunpeng Xiao, "C-RBFNN: A user retweet behavior prediction method for hotspot topics based on improved RBF neural network", Neurocomputing, v.275 n.C, p.733-746, January 2018

[11] Xing Tang, Qiguang Miao, Yining Quan, Jie Tang Kai Deng, "Predicting individual retweet behavior by user similarity: a multi-task learning approach", Knowledge-Based Systems, Vol. 89, pp. 681-688.

[12] Dongxu Huang, Jing Zhou, Dejun Mu, Feisheng Yang, "Retweet behavior prediction in twitter." In 2014 IEEE Seventh international symposium computational intelligence and design (ISCID), vol 2, pp 3033

[13] A. Kupavskii, L. Ostroumova, A. Umnov, S. Usachev, P. Serdyukov, G. Gusev, and A. Kustarev. Prediction of retweet cascade size over time. In Proc. CIKM, 2012.

[14] J. Cheng, L. Adamic, D. Alex, J. Kleinberg, and J. Leskovec. Can cascades be predicted? In WWW, 2014.

[15] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec. Seismic: A self-exciting point process model for predicting tweet popularity. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '15, pages 1513–1522, 2015.

[16] E. F. Can, H. Oktay, and R. Manmatha. Predicting retweet count using visual cues. In Proceedings of the 22nd ACM international conference on information & knowledge management, pages 1481–1484. ACM, 2013.

[17] Zeynep Zengin Alp, ule Gndz dc "Identifying topical influencers on twitter based on user behavior and network topology" Knowledge-Based Systems, 141 2018, pp. 211-221