コンテンツ的特徴を用いたTwitterのトレンド予測
Predicting Twitter Trends Using Content-Based Features

光末　慈瑛
Jiei Mitsusue
法政大学情報科学部コンピュータ科学科
E-mail: jiei.mitsusue.9n@stu.hosei.ac.jp

Abstract

Social networking services (SNSs) occupy a major position in people's communication, and their real-time and rapid nature of spreading information makes it important to analyze and predict the trends of topics on SNSs in advance. Tanaka et al. focused on social graphs as a method for predicting word trends in social networking services. They proposed features that are thought to be indicators of the distribution of nodes in graphs and studied a method for predicting by learning features as time-series data using a model based on LSTM. However, their method was not useful for time-series data prediction, and it needs to improve the quality of features and to use content-based features. In this paper, we propose a method that solves the problems of Tanaka et al.'s method. To improve the quality of the features of the graph, we consider the appropriate pruning of the social graph data. Also, we extract content-based features from the degree of diffusion by the share function of Twitter. To evaluate our method, we compared the prediction accuracy depending on the branch pruning methods and whether the content-based features were used. The experimental results show the usefulness of the graphs and content-based features considered in this research for predicting time-series data, but the problem is that they cannot cope with extreme changes in actual measurements.

1. はじめに

インターネットが普及した現代では，Twitter や Instagram などのソーシャル・ネットワーキング・サービス（SNS)の普及により，多くの人々が全世界へ情報を発信できるようになった．SNS が普及するにつれ，ユーザ間で扱われる話題も多様化し，その流行も激しく変化するようになった．SNS はいまや人々のコミュニケーションで大きな位置を占めるようになっており，リアルタイム性が高く急速に情報が広がるという特徴のため，SNS における話題の流行を事前に分析，予測することは，社会分析や市場調査でも重要になってきている．

SNS における話題の流行に関する研究として，田中ら[1]は，ユーザが投稿する文章中で言葉を使用することに着目し，投稿文章中で対象の言葉を使用したユーザ数がどのように変化するかについて LSTM を用いた時系列データ予測を行った．Twitter におけるソーシャルグラフに着目し，対象の言葉を使用したユーザの次数のような，グラフにおける特徴量を用いた手法を比較し，どのような特徴量が予測に有効かを検討した．しかし，田中らの検討したグラフと特徴量では，予測を行う際の有用性は認められなかったとされており，特徴量の質の改善やコンテンツ的特徴量の利用が課題とされている．

本研究の目的は，田中らの研究に対して，適切なソーシャルグラフの作成によるグラフ的特徴量の質の向上と，コンテンツ的特徴量の導入の処理を加え，より効果的な言葉の流行予測の手法を提案することである．正しい予測結果に結びつかないと考えられるアカウントの削除を施し，グラフ的特徴量の質の向上と共にグラフのサイズの削減を図る．SNS での言葉の拡散において，グラフの構造だけでなくコンテンツ的特徴も大きく関係していると考えられるため，SNS のシェア機能による拡散具合といったコンテンツ的特徴からも特徴量を抽出する．グラフの作成方法の違いとコンテンツ的特徴量の有無に応じた予測精度の比較を行い，提案手法の有用性を評価する．実験の結果，本研究で検討したグラフとコンテンツ的特徴量による時系列データ予測は田中らの手法より高精度な予測結果を残し，田中らの研究の問題点を解決した．

2. 関連研究

本章では，本研究と関連する研究を紹介し，SNS における話題の流行の予測というテーマにおいて，それらの研究と比較した際の本研究の位置付けを述べる．

田中ら[1]は，Twitter の投稿文章中で対象の言葉を使用したユーザ数がどのように変化するかについて，LSTM を用いた時系列データ予測を行った．田中らは，SNS における言葉の流行予測の手法として，ソーシャルグラフに着目し，グラフにおいてノードが「散らばっている」ことの指標になると思われる特徴量を提案し，実際にこれらの特徴量から，LSTM を用いたモデルにより，時系列データとして学習することで予測を行う手法を検討した．田中らは，提案した特徴量による時系列データ予測は有効でなかったとしており，原因として，特徴量の質が十分でなかった可能性と，SNS での言葉の拡散においてはコンテンツ的特徴も大きく関わっている可能性を挙げている．本研究では，田中らの用いた特徴量の質を向上させつつ有効なコンテンツ的特徴量を導入することで田中らの研究を発展させ，時系列データ予測の精度を高める．

Supervisor: Prof. Hiroshi Hosobe

松浦ら [2]は，一定の制約がある中でも個人による大量ツイートデータの収集・分析を実現し得る手法を提案した．社会現象などの膨大な量のツイートを発生させる話題を取り扱おうとした場合，費用的，技術的な問題から，それが可能なのは十分な予算や設備を持った大企業や組織に限られていた．そこで松浦らは，Twitter の無料 API である Standard search API を利用しつつ，レートリミットがある中で大量ツイートの収集を行うことを目標とした．本研究では，松浦らが作成したツイート収集プログラムを利用することでツイートの収集を行い，ソーシャルグラフの作成，特徴量の抽出に役立てる．

3. 準備

本章では，田中らの検討した特徴量について説明する．本研究では，この田中らの特徴量を用いてツイート数の時系列データ予測を行う．

以下では，Twitter におけるフォロー関係のグラフを重み付き有向グラフ$G = (V, E, W, V_{\Delta_t})$で表す．ただし，$V$は頂点となるユーザの集合，$E = \{(u, v) | u, v \in V\}$はユーザ$u$が$v$をフォローしていることを表すエッジ集合，$W : E \to \mathbf{R}^+$はエッジの重み，$V_{\Delta_t} \subset V$は区間$\Delta_t$においてキーワード$w$をツイートしたユーザの集合である．

3.1. フォロワー数

田中らは，より多くのフォロワーにフォローされているユーザほど，話題の流行への影響力が大きいと考えた．$u \in V$に対して，uがvのフォロワーであるとは，$e = (u, v) \in E$なるeが存在することであり，vのフォロワー数 follower(v)とは，Gにおけるvの入次数である．本研究では次式を用いて予測を行う．

$$f = \frac{\sum_{v \in V_{\Delta_t}} \text{follower}(v)}{|V_{\Delta_t}|}$$

3.2. 言及者までの平均距離

田中らは，対象の言葉wについて言及していないユーザ$v \notin V_{\Delta_t}$は，すでに言及しているユーザ$v \in V_{\Delta_t}$からの影響を受けて言葉を使用する，ということを繰り返して言葉が流行していくものだと考えた．そこで，uとV_{Δ_t}の距離$d(u) = \min_{v \in V_{\Delta_t}} \text{dist}(u, v)$を，$u$がどの程度言葉$w$の流行に影響されやすいかを表す指標とした．本研究では$d(u)$の平均値である次式を用いて予測を行う．

$$\bar{d} = \frac{\sum_{u \in V \setminus V_{\Delta_t}} d(u)}{|V \setminus V_{\Delta_t}|}$$

3.3. Kermack-McKendrick の流行モデル

3.1 節で述べたように，ある言葉についてまだ言及していないユーザは，すでに言及しているユーザから影響を受け言及するようになる，ということを繰り返して言葉が流行するものと考えられている．田中らはこれを数理モデルとして考えた．Kermack と McKendrick が提唱した SIR モデルは，ある集団に属する人々を，感染症に対する免疫を持たず病気にかかりうる感受性人口（Susceptibles; S），すでに病気に感染している感染人口（Infected; I），病気に感染後，治癒したことで病気に対する免疫を獲得する，または病気によって死亡するなどして二度とその病気にかからない隔離人口（Removed; R）の 3 種類に分類する（図 1(a)）．しかし，話題の流行を記述するために SIS モデルをそのまま適用することは適切ではないとされた．話題の流行には免疫に相当するものがないと考えられるため，隔離人口の変化はアクティブユーザの減少のみであるが，Twitter のユーザ数は年々増加しており，アクティブユーザの減少による効果は外部からの人口増加の影響に比べると小さい．よって，田中らの研究では隔離人口を考えず，感染人口は一定確率で感染人口に戻るとする SIS モデルをベースに考えた（図 1(b)）．

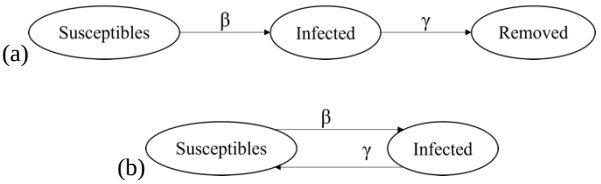

図 1　(a) SIR モデルと(b) SIS モデル

SIS モデルは感染率βと回復率γを用いて以下のように定式化できる．

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = -\beta S(t) I(t) + \gamma I(t)$$
$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = \beta S(t) I(t) - \gamma I(t)$$
$$S + I = \text{const.}$$

本研究では，この SIS モデルの感染率と回復率を用いて予測を行う．

3.4. クラスタにおける言葉の使用率

田中らは，Twitter のフォロー関係グラフGを Newman Fast アルゴリズムによりクラスタリングし，得られたクラスタ$C_1, \ldots, C_n$ついて，グラフ全体における言葉wの使用数w_{all}に対する，各クラスタでの言葉の使用率$o_1, \ldots, o_n$を計算し，その最大値$o_{\max}$を利用した．言葉がある特定のユーザ間だけで話されるにとどまっている状態では$o_{\max}$の値は高く，言葉が広まり多くのユーザに使用されるようになればなるほど$o_{\max}$の値は低下すると考えられる．本研究では，$o_{\max}$を用いて予測を行う．

4. 提案手法

本章では，本研究で提案する手法について説明する．本研究の目的は，田中らの研究で用いられた特徴量の質の向上とコンテンツ的特徴量の導入を行い，特定の言葉を含むツイート数の変化の予測精度を高め，特徴量の有効さについて検討することである．言葉wが含まれる過去 1 か月間のツイートデータと Twitter のソーシャルグラフから各種特徴量を計算し，特徴量をもとに今後 1 か月間のwを含むツイート数の変化を予測し，特徴量の有効さについて検討する．

4.1. ソーシャルグラフ作成

本研究では，特徴量の質の向上とソーシャルグラフのサイズの削減のために，田中らのものとは異なる方法でソーシャルグラフの作成を行った．田中らは，Twitter の followers/ids API を用いて，ユーザを幅優先探索により収集してフォロー関係を表すソーシャルグラフを 1 つ作成した．グラフは 5000 万ノード，40 億本以上のエッジを含む巨大なグラフになり，エッジを無作為に削除することで 800 万ノード，3600 万本のエッジを含むグラフに縮小した．しかし，まだグラフのサイズが大きく，無作為な枝刈りの結果元のデータに含まれる属性が損なわれるという問題点があった．本研究では，収集したツイートの投稿ユーザのみがノードとなるソーシャルグラフの作成を各クエリについて行った．対象期間中に対象の言葉wについてツイートを行っていないユーザとフォロワー数の少ないユーザが言葉の流行に与える影響はないとして除外することで，有効なデータを保持させつつ規模の小さく扱いが容易なグラフを作成する．今回，フォロワー数が 50 人未満のユーザを除外した．

4.2. コンテンツ的特徴量抽出

コンテンツ的特徴とは，投稿文章の内容やサービス特有の機能などのことを指す．田中らは自身らの研究結果から，グラフ的特徴量のみを用いた時系列データ予測の有効さが認められなかったとした．そこで，本研究では時系列データ予測の精度を高めるためにコンテンツ的特徴量を導入する．グラフ的特徴だけでなくコンテンツの面からも予測に役立つ特徴量について検討することが重要だと考え，リツイート数といいね数の 2 つのコンテンツ的特徴量の有効さについて検討を行う．

以下で各コンテンツ的特徴量の抽出方法について説明する．すべての特徴量において，単位となる区間をΔ_tとし，区間Δ_tにおける言葉wを含むツイートの集合をT_{Δ_t}とし，以下の方法で特徴量を抽出する．

4.2.1. リツイート数

Twitter には，他のユーザのツイートを自身のフォロワーのタイムラインに表示させて拡散するリツイート機能が存在する．対象の言葉wについて言及していないユーザ$u \notin V_{\Delta_t}$は，自身のフォローするユーザもまた言葉wについて言及しておらずとも，言葉wを文章中に含むツイートがリツイートにより自身のタイムラインに届く可能性はある．したがって，言葉wを含むツイートtがリツイートされ，それを見たユーザuは影響されて自身も言葉 w を使用するようになり，さらにリツイート数$\mathrm{retweet}(t)$が多ければ多いほどより多くのユーザuが影響されて言葉 w を使用するようになると考えられる．本研究では，区間Δ_tにおけるツイートの集合T_{Δ_t}のリツイート数の平均値である次式を用いて予測を行う．

$$r = \frac{\sum_{t \in T_{\Delta_t}} \mathrm{retweet}(t)}{\left|T_{\Delta_t}\right|}$$

上述の式により，各区間$\Delta_{t_1}, \ldots, \Delta_{t_n}$での，$T_{\Delta_t}$に含まれるツイートの平均リツイート数$r_{\Delta_{t_1}}, \ldots, r_{\Delta_{t_n}}$を求めたのち，その平均を求め，平均 0 となるようセンタリングしたものを入力データとして用いる．

4.2.2. いいね数

Twitter には，他のユーザのツイートに「いいね」を付与できる，いいね機能が存在する．いいねされたツイートも，すべてではないが自身のフォロワーのタイムラインに表示され拡散される．したがって，リツイート同様，いいね数$\mathrm{like}(t)$が多ければ多いほど多くのユーザuが影響されて言葉wを使用するようになると考えられる．本研究では，区間Δ_tにおけるツイートの集合T_{Δ_t}のいいね数の平均値である次式を用いて予測を行う．

$$l = \frac{\sum_{t \in T_{\Delta_t}} \mathrm{like}(t)}{\left|T_{\Delta_t}\right|}$$

上述の式により，各区間$\Delta_{t_1}, \ldots, \Delta_{t_n}$での，$T_{\Delta_t}$に含まれるツイートの平均いいね数$l_{\Delta_{t_1}}, \ldots, l_{\Delta_{t_n}}$を求めたのち，その平均を求め，平均 0 となるようセンタリングしたものを入力データとして用いる．

5. 実験

本章では，本研究で行った実験とその結果について述べる．

5.1. データ

実験では，2021 年 8 月から 2021 年 12 月までの Twitter ユーザのツイートデータ，および，収集したツイートの投稿ユーザのフォロー関係を表すソーシャルグラフを用いた．それぞれのデータの詳細は以下の通りである．

- ツイートデータは，調べたい言葉wをクエリとして，松浦らが作成したツイート収集プログラム「小鳥男」を利用して，2021 年 8 月から 2021 年 12 月までの 5 か月分のツイートを収集した．
- ソーシャルグラフは，各クエリについて，収集したツイートの投稿ユーザのみがノードとなるように作成した．

使用したクエリと，各クエリで取得したツイートの数，各クエリについてのグラフのノード数とエッジ数を表 1 に示す．

表 1　クエリと取得したツイート数，
ソーシャルグラフのノード数とエッジ数

クエリ	ツイート数	ノード数	エッジ数
ウマ娘	1473491	211685	780866
うんこちゃん OR 加藤純一	160690	30972	164438
感染者	2017438	178997	818683
呪術廻戦	770564	36495	149634

5.2. 評価指数

予測結果の評価には，平均絶対パーセント誤差（Mean Absolute Percentage Error, MAPE）を使用した．MAPE は次式で定義される．

$$\mathrm{MAPE} = \frac{100}{n} \sum_{i}^{n} \left| \frac{u^i_{\mathrm{true}} - u^i_{\mathrm{predict}}}{u^i_{\mathrm{true}}} \right|$$

ただし，u^i_{true}は区間Δ_{t_i}に対象の言葉を含むツイートをしたユーザ数，u^i_{predict}はその予測値である．

5.3. 実験結果

田中らの実験で考慮された特徴量の組み合わせと，それにコンテンツ的特徴量を追加した組み合わせで実験を行った．結果を表 2 に示す．表の左側が田中らの実験で考慮された特徴量の組み合わせ，右側がそれにコンテンツ的特徴量を追加した組み合わせである．田中らの実験結果 [1]では，表 3 に示すように，すべての特徴量の組み合わせで MAPE の平均値が 100 に近いスコアとなった．

表 2　本研究の実験結果

入力した特徴量	MAPE	入力した特徴量	MAPE
f	204.90	f,r,l	81.18
$f,\bar{d}$	71.62	$f,\bar{d},r,l$	51.87
f,β,γ	56.49	f,β,γ,r,l	35.43
$f,o_{\max}$	69.06	$f,o_{\max},r,l$	51.70
$\bar{d}$	87.09	$\bar{d},r,l$	52.63
$\bar{d},\beta,\gamma$	37.56	$\bar{d},\beta,\gamma,r,l$	41.89
$\bar{d},o_{\max}$	62.15	$\bar{d},o_{\max},r,l$	44.74
β,γ	32.45	β,γ,r,l	42.54
$\beta,\gamma,o_{\max}$	50.14	$\beta,\gamma,o_{\max},r,l$	26.37

表 3　田中らの実験結果

入力した特徴量	MAPE
f	104.71
$f,\bar{d}$	99.17
f,β,γ	104.93
$f,o_{\max}$	99.61
$\bar{d}$	105.59
$\bar{d},\beta,\gamma$	113.95
$\bar{d},o_{\max}$	103.41
β,γ	104.56
$\beta,\gamma,o_{\max}$	102.20

本研究では，特徴量がfのみの場合はスコアが 204.9 と田中らのものを上回り誤差が大きくなったが，他の場合では誤差が小さくなった．最も誤差が小さかったのはβ,γの 32.45 である．コンテンツ的特徴量を追加した組み合わせでは，$\bar{d},\beta,\gamma,r,l$とβ,γ,r,l以外で元の組み合わせより誤差が小さくなった．先述のfの 204.9 はコンテンツ的特徴量が追加されることで 81.18 となり 100 を下回った．最も誤差が小さかったのは$\beta,\gamma,o_{\max},r,l$の 26.37 である．

6. 議論

実験結果を踏まえ，本研究で提案したソーシャルグラフとコンテンツ的特徴量について考察を行う．結論として，本研究で提案したグラフとコンテンツ的特徴量による時系列データ予測は，田中らのものより有効であった．表 2 の左半分と表 3 より，入力した特徴量が同一であるにもかかわらずfを除いて本手法による予測の誤差が小さい結果となった．したがって，本研究で提案した，対象の言葉についての言及者のみがノードとなるグラフは予測に有効であるといえる．さらに，田中らのものと比較して小さいサイズのグラフで良好な結果が得られたため，田中らの研究の，サイズが大きく，無作為な枝刈りの結果元の属性が損なわれる問題点を解決したといえる．

表 2 より，田中らの実験で考慮された特徴量の組み合わせにコンテンツ的特徴量を追加した場合に，誤差が小さくなったものがほとんどであるが大きくなったものもある．したがって，本研究でコンテンツ的特徴量として検討したリツイート数といいね数は，特徴量の組み合わせ次第で予測に有効である．Twitter のトレンドの時系列データ予測ではコンテンツ的特徴も考慮すべきである．

7. おわりに

田中らは，SNS における言葉の流行予測の手法として，ソーシャルグラフに着目し，グラフにおいてノードが散布していることの指標になると考えられる特徴量を提案し，LSTM を利用したモデルにより特徴量を時系列データとして学習することで予測を行う手法を検討した．しかし，田中らの検討したグラフと特徴量では，予測を行う際の有用性は認められなかった．本研究では，田中らの研究に対して，適切なグラフの作成によるグラフ的特徴量の質の向上や，コンテンツ的特徴量の導入などの処理を加え，より効果的な Twitter におけるトレンド予測を行う手法を検討した．結果として，本研究で検討したグラフとコンテンツ的特徴量に時系列データ予測を行う際の有用性が認められたが，実測値の極端な変化に対応できなかった．

解決策として，他のコンテンツ的特徴量の導入と，十分な期間の学習データの用意が挙げられる．まず，他のコンテンツ的特徴量として，例えばクラスタにおける言葉の使用率を調べる際，グラフのクラスタリングではなく，投稿内容からユーザがよく言及するトピックを調べることによるコミュニティ分割を行える．つぎに，十分な期間の学習データの確保に関して，本研究の手法では，Twitter API のアクセス回数制限が原因でグラフの作成と特徴量の抽出に時間を要したため，実験時間との兼ね合いから 30 日分の学習データしか用意することができなかった．十分な期間の学習データを用意するために，グラフ作成と特徴量抽出をツイートデータ取得と同時に行ったり，1 つのクエリに対して Twitter API のアプリケーションを複数用意して，API にアクセスする必要がある処理を分担させたりすることが考えられる．

また，本研究では，田中らの手法でグラフを作成しなかったため，田中らの検討したグラフに対するコンテンツ的特徴量の有無の比較実験を行うことができなかった．今後，田中らの検討したグラフに対して実験を行い，コンテンツ的特徴量の有用性を改めて確認する必要がある．フォロワー数 50 人が少ないとしてそれ未満のユーザをグラフから除外したが，フォロワー数の条件を変えて比較実験を行い，適切なフォロワー数を求める必要がある．

文　献

[1] 田中勝也, 田島敬史, "Twitter におけることばの流行予測," DEIM Forum 論文集, no. D6-2, pp. 1-8, 2017.

[2] 松浦智之, 當仲寛哲, 大野浩之, "大量ツイートの収集・分析を個人で手軽に実現可能にする方法の提案," デジタルプラクティス, vol. 11, no. 1, pp. 173-190, 2020.