Twitterにおけることばの流行予測

田中　勝也†　　田島　敬史††

† 京都大学工学部情報学科　〒606–8501 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科　〒606–8501 京都府京都市左京区吉田本町
E-mail: †katsuya@dl.kuis.kyoto-u.ac.jp, ††tajima@i.kyoto-u.ac.jp

あらまし　SNSが普及するにつれて，SNSユーザの話題や流行を分析，予測することは，社会分析やマーケティングにおいて重要となってきているが，急速に移り変わるSNS上の話題の流行を追うこと，また未来の流行を予測することは非常に困難である．本研究ではSNS上のユーザ間のつながりを表すグラフのデータと，ユーザの投稿に含まれるデータから得られる様々な特徴量を検討し，SNS上での話題の流行の予測に役立つものについて比較，考察する．

キーワード　SNS，Twitter，マイクロブログ，グラフデータ，流行予測

1. はじめに

インターネットが普及した現代においては，Facebook(注1)やTwitter(注2)のようなソーシャル・ネットワーキング・サービス（SNS）の普及により，様々な人々が全世界へ情報を発信できるようになった．SNSで作り出される情報量は日々増加しており，Twitterでは，2010年7月には1日平均で約5500万ツイートが投稿されていたものが，2016年では1日平均約5億ツイートとなり(注3)(注4)，日本における月間アクティブユーザ数も，2011年3月時点で670万人だったものが，2016年9月時点で4000万人(注5)(注6)となっている．

SNSが普及するにつれ，ユーザ間で話さ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．ツイート数やユーザ数の変化を見ても分かる通り，SNSはいまや人々のコミュニケーションにおいて大きな位置を占めるようになっており，またSNSの特徴として，リアルタイム性が高く，急速に情報が広がるという点があるため，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においても重要になってきている．

このような話題の流行の予測に関する研究として，Twitterのハッシュタグの使用数の変化を予測する研究[1]や，Facebookにおける画像投稿のシェアの広がり方を予測する研究[2]が知られている．これらの研究はハッシュタグやシェアといった，対象のSNSに備わっている機能を利用した手法となっているが，こうした情報拡散機能が利用されるのは短い期間に限られている[3]ため，あることばについて，どのように流行が変化し，どれくらいユーザ間に定着していくのかということを長期的に調べるには適していないと考えられる．

(注1)：https://www.facebook.com/
(注2)：https://twitter.com/
(注3)：http://www.dsayce.com/social-media/tweets-day/
(注4)：http://www.mediapost.com/publications/article/160712/n average-twitter-user-sends-half-a-tweet-per-day.html
(注5)：http://www.huffingtonpost.jp/2016/02/18/n twitter-japan_n_9260630.html
(注6)：http://gaiax-socialmedialab.jp/post-30833/

そこで，本研究では，SNSに備わっているシェア機能を利用した情報の拡散ではなく，ユーザが投稿する文章（e.g. Twitterにおけるツイート）の中でことばを使用するということに着目し，投稿文章中で対象のことばを使用したユーザ数がどのように変化するかについてHochreiterらが発表したLSTM（Long Short-Term Memory）[11]を用いた時系列データ予測を行う．LSTMはRNN（Recurrent Newral Network）の一種であり，従来のRNNが実際には長期依存を学習できない（過去の状態を覚えきれない）という問題を解決した．

従来の，ハッシュタグなどのSNSに備わっている情報拡散機能に着目した研究では，予測の際に，ハッシュタグ自体に用いられている単語の長さなど，対象となることばの属性を利用するものがよく知られているが，本研究では，ことばの属性ではなく，Twitterにおけるソーシャルグラフに注目し，対象のことば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おいて役に立つのかを検討する．

本研究に際して行った実験では，2015年1月から2015年12月までのTwitterユーザのツイートデータと，2015年時点でTwitterをクロールして収集された，ユーザのフォロー関係を表すソーシャルグラフを用いて入力とした特徴量ごとの予測精度の比較を行う．

2. 論文の構成

以下，3.では関連研究について述べ，本研究の立場を明らかにする．4.では本研究で用いる特徴量について説明する．5.では本研究において特徴量を比較するための提案手法，アルゴリズムについて述べる．6.では5.で述べる手法を用いた実験とその評価を行う．7.では本論文における結論を述べ，今後の課題について考察する．

3. 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．

ソーシャルネットワークにおける情報の拡散に関しては，すでに多くの研究がなされている．Twitter では，ユーザの投稿（ツイート）中で，ハッシュタグと呼ばれる#keyword という形式の文字列を用いることで，このツイートがどういう文脈でなされたものかを他のユーザと共有することができる．このハッシュタグについて，Tsur ら [4] は，ハッシュタグから，ハッシュタグ自体の内容（e.g. ハッシュタグの長さ，ハッシュタグを構成する単語数），そのハッシュタグを含むツイートの内容，ハッシュタグを含むツイートがリツイートされる確率などのコンテンツ的属性を抽出し，回帰分析によって，各属性とハッシュタグを含むツイート数の関係を週ベースで検討した．

これに対し，Ma ら [1] は，ハッシュタグを含むツイート数ではなく，ハッシュタグを使用したユーザ数に関して，また，ハッシュタグ自体やハッシュタグを含むツイートの内容といった，コンテンツ的属性だけでなく，他のユーザへのメッセージとしてのツイート（メンション）を抽出し，メンションを送った関係に基づくグラフの特徴量（e.g. PageRank アルゴリズム [5] に基いたユーザのオーソリティスコア）を利用し，クラス分類問題として検討を行った．この研究は．ハッシュタグの内容のようなコンテンツの属性だけでなく，ユーザ間の関係に着目した検討を行っている点で本研究と通じる部分がある．しかし，Ma らの研究ではユーザ間のメンションに基づいたグラフを利用しているが，Twitter では，ユーザは自分宛てのメンションだけでなく，自分がフォローしているユーザ全員のツイートが表示される，タイムラインと呼ばれるツイートのストリームを参照することが多いため，メンション関係だけでは Twitter におけるユーザ同士の関係を正確に表現できていないのではないかと考えられる．そこで本研究では，メンション関係ではなく，実際のフォロー関係に基づいたグラフを用いることで，より有効な特徴量を得ることを目的とする．

また SNS のようなマイクロブログサービスに関する研究でよく見られるのが，優良な情報源を見つける能力の高い，いわゆるアーリーアダプターとよばれるユーザたちに注目し，彼らの性質を分析することで次に人気が出る情報源を見つけるというものがある．アーリーアダプターを検知するための手法としては PageRank アルゴリズム [5] や HITS アルゴリズム [6] が知られている．

Cheng ら [2] の研究では，Facebook に投稿された写真がシェア機能によって広まっていくということに焦点を当て，コンテンツの属性に加えて，最初の n 回までのシェアがグラフにおいてどのような形になっているかといったグラフ的特徴や，どれくらいの速さでシェアされるかといった時間的特徴といった特徴量を用いて，ある投稿をシェアしているユーザ数が今の 2 倍になるかという問題に対して成果を上げている．Cheng らの研究はシェア機能を用いた画像情報の伝播についてであるが，本研究でもグラフの特徴や時間的特徴を利用した比較を行っていく．

中島ら [7] の研究では，流行語を，「メディアに取り上げられたことで一気に認知度が高まり様々なコミュニティで一時的に話題になる」突発型流行語と，「小さなコミュニティから口コミなどにより徐々に様々なコミュニティで使用されるようになる」拡張型流行語があるとした上で，拡張型流行語では，ことばの使用数が増えるにつれて，そのことばを使用する発言者の年齢層が，偏りのある状態から一様に近い状態に変化していくことを指摘し，ブロガーをそれぞれの興味や趣味に基づいたコミュニティに分類して，あることばの出現数が最も多いコミュニティでの出現数が，全ての出現数に占める割合が一定の値まで低下してきたことを流行の基準として，流行語を検知するという手法が提案されている．この研究ではブログのエントリからそれぞれが属するコミュニティへ分類しているが，本研究ではグラフのクラスタリングによるクラスタを用いた手法を検討する．

グラフのクラスタリングでは，コミュニティ間のエッジが少なく，コミュニティ内のエッジが多くなるようなクラスタリング（モジュラリティが高いクラスタリング）がよいとされる．Girvan と Newman は，すべてのエッジに対して，そのエッジがグラフのすべてのノード間の最短路にどれくらい寄与するかをあらわす Betweenness Centrality を計算し，値の高いエッジから切り離していくことでモジュラリティの高いクラスタリングを実現するトップダウンクラスタリングの手法 [9] を提案した．しかし，Girvan-Newman のアルゴリズムは，コミュニティを結合するたび，各エッジの Betweenness Centrality を計算し直す必要があるため，非常に計算量が大きく，大規模なソーシャルグラフのクラスタリングには不向きである．後に Newman は，コミュニティ同士を結合したときのモジュラリティの変化量に着目し，変化量が最大となるコミュニティ同士を順に結合するボトムアップ・アプローチ（Newman Fast アルゴリズム）によって，Betweenness Centrality を計算せずにモジュラリティの高いクラスタリングが可能であることを示した [10]．Newman Fast アルゴリズムでは，コミュニティを結合するたびにモジュラリティの変化量を計算する必要があるが，結合されなかったコミュニティ間では，結合時のモジュラリティの変化量は変わらないため，結合して新たにできたコミュニティと，他のコミュニティでの変化量のみを計算しなおせばよいため，Girvan-Newman のアルゴリズムに比べ計算量の面で優れている．本研究では，Newman Fast アルゴリズムにより得られたクラスタを用いて，ことばの使用率の変化を計算する．

4. 特 徴 量

本章では本研究で検討する特徴量について説明する．

3. でも挙げた中島ら [7] の研究でも指摘されていたように，ことばが流行するとは，ことばが本来主に使われていた年齢層やコミュニティを抜け出し，普遍的に使われるようになることだと考えられる．

そこで本研究では，あることばが今後どれくらい広く使われるようになるかを予測する際に，単純にこれまでそのことばに言及したことのあるユーザの数だけではなく，それらのユーザノードがソーシャルグラフ上でどれくらい「散らばっている」かという情報を含めて考えることで，より正確な予測を行う手法を提案する．

Twitter におけるフォロー関係のグラフを，頂点となるユーザの集合 V と，u が v をフォローしていることを表すエッジ集合 $E = \{(u,v)|u,v \in V\}$，エッジの重み $W : E \to \mathbb{R}^+$，および区間 Δ_t においてキーワード w をツイートしたユーザの集合を $V_{\Delta_t} \subset V$ とし，重み付き有向グラフ $G = (V, E, W, V_{\Delta_t})$ とする．

それでは，「散らばっている」ということを表すための特徴量について，以下で具体的に説明する．

4.1 フォロワー数

まずもっとも単純な指標として考えられるのがフォロワー数である．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られる．$v \in V$ 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 の入次数であり，例えば，図 4.1 のグラフにおいて，follower$(F) = 4$ である．本研究においては

$$f = \frac{\sum_{v \in V_{\Delta_t}} \text{follower}(v)}{|V_{\Delta_t}|} \tag{1}$$

を用いて予測を行うこととする．

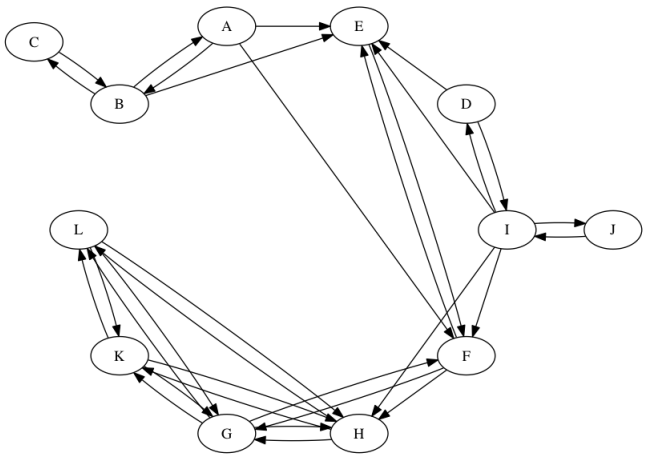

図 1　グラフの例

4.2 V_{Δ_t} までの平均距離

対象のことば w について言及していないユーザ $u \notin V_{\Delta_t}$ は，すでに言及しているユーザ $v \in V_{\Delta_t}$ のからの影響を受けてことば w を使用するようになる，ということを繰り返してことばが流行していくものだと考えられる．したがって，u と V_{Δ_t} の距離 $d(u) = \min_{v \in V_{\Delta_t}} \text{dist}(u,v)$ は，u がどの程度ことば w の流行に影響されやすいかを表す指標とできる．本実験では $d(u)$ の平均値

$$\bar{d} = \frac{\sum_{u \in V \setminus V_{\Delta_t}} d(u)}{|V \setminus V_{\Delta_t}|} \tag{2}$$

を用いた手法を検討する．例えば，図 4.2 のグラフにおいては $\bar{d} = 2.3$ となる．

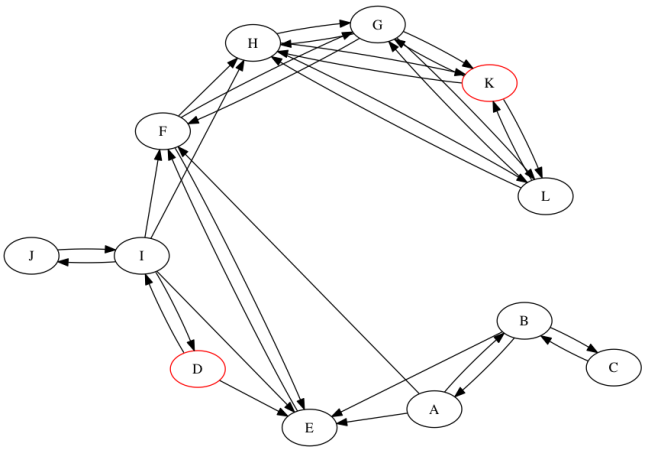

図 2　グラフの例（赤いノードが V_{Δ_t} に属しているとする）

4.3 Kermack-McKendrick の流行モデル

4.2 で述べたように，あることば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，ことばが流行するものと考えられる．これを数理モデルとして考える手法について説明する．伝染病の流行を記述するための数理モデルの研究は古くから行われてきた．Kermack と McKendrick [8] が提唱したモデル（Kermack-McKendrick モデル，Susceptibles-Infected-Removed モデル，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，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 3　SIR モデル

S に属する人は I に属する人と接触することで病気に感染し，I に属する人は一定確率で R へと変化する．これを定式化すると以下 (3)(4)(5)(6) 式のようになる．ただし β は感染率，γ は隔離率である．

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = -\beta S(t) I(t) \tag{3}$$

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = \beta S(t) I(t) - \gamma I(t) \tag{4}$$

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t} = \gamma I(t) \tag{5}$$

$$S + I + R = \text{const.} \tag{6}$$

しかし，話題の流行を記述するために，Kermack-McKendrick モデルをそのまま当てはめることは適切ではない．まず，話題の流行で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．よって本研究では隔離人口を考えず，感染人口は一定確率で感受性人口に戻るとする SIS モデルをベースに考えることにする．

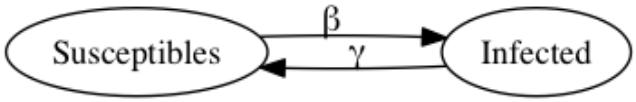

図 4　SIS モデル

SIS モデルは感染率 β と 回復率 γ を用いて以下 (7)(8)(9) のように定式化できる．

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = -\beta S(t)I(t) + \gamma I(t) \tag{7}$$
$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = \beta S(t)I(t) - \gamma I(t) \tag{8}$$
$$S + I = \text{const.} \tag{9}$$

本研究ではこの SIS モデルの感染率と回復率を利用して今後ことばがどのように広がっていくかを予測する手法を検討する．

4.4　クラスタにおけることばの使用率

中島ら [7] の研究では，ブログの投稿を分析してブロガーたちをコミュニティに分類し，最も多く対象のことばを使用しているコミュニティが全体に占める割合の低下を検知し，流行の早期検知をする方法が提案されていた．

本研究においては，投稿内容での分類ではなく，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 $C_1, \ldots, C_n$ について，グラフ全体におけることば w の使用数 w_{all} に対する，各クラスタでのことばの使用率 $o_1, \ldots, o_n$ を計算し，その最大値 $o_{\max}$ を利用する．

例えばある区間において，あることば w の使用数と，グラフのクラスタリング結果が図 4.4 のようになっているとすると，全体の w の使用数は $w_{\text{all}} = 14$ であり，

$$o_1 = \frac{1+2+1+3+2}{w_{\text{all}}} = \frac{9}{14} \tag{10}$$
$$o_2 = \frac{1+0+0+0}{w_{\text{all}}} = \frac{1}{14} \tag{11}$$
$$o_3 = \frac{1+2+1+0+0}{w_{\text{all}}} = \frac{4}{14} \tag{12}$$

となり，

$$o_{\max} = \max_i o_i = \frac{9}{14} \tag{13}$$

となる．

ことば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，ことばが広まり，多くのユーザに使わ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．本研究ではこの $o_{\max}$ を特徴量として今後ことばがどのように広がっていくかを予測する手法を検討する．

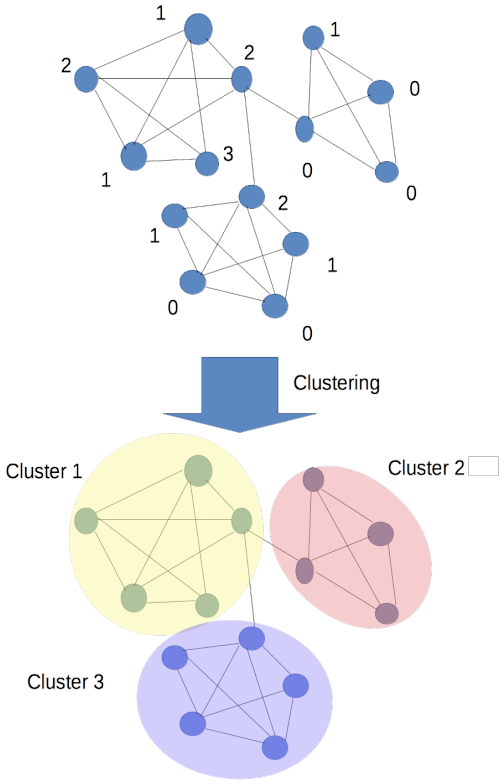

図 5　クラスタ分割から $o_{\max}$ を求める概念図

5.　提 案 手 法

本章では本研究で我々が提案する手法について述べる．

本研究の目的は，Twitter において，ことば w が含まれる過去 10 か月分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 2 か月間で w を含むツイートを行うユーザの数の変化を予測し，特徴量の有効さについて検討することである．

5.1　特徴量の抽出

各特徴量について，その抽出の方法を説明する．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 においてことば w を含むツイートをしたユーザの集合を V_{Δ_t} とする．

5.1.1　フォロワー数

(1) 式により，各区間 $\Delta_{t_1} \ldots \Delta_{t_n}$ での，V_{Δ_t} に含まれるユーザの平均フォロワー数 $f_{\Delta_{t_1}} \ldots f_{\Delta_{t_n}}$ を求めたのち，その平均

$$\bar{f}_{\Delta_t} = \frac{\sum_{i=1}^{n} f_{\Delta_{t_i}}}{n} \tag{14}$$

と標準偏差

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n}} \tag{15}$$

を求め，平均 0 となるようセンタリングしたものを，フォロワー数の入力データとして用いる．

5.1.2 V_{Δ_t} までの平均距離

V_{Δ_t} までの平均距離は以下の手順により近似値を求める．

まず，V_{Δ_t} を，V_{Δ_t} からの距離 0 の頂点集合とし，距離 0 の集合に含まれる頂点から，エッジを逆向きに 1 回だけ辿ってたどり着ける頂点の集合を求め，そのうち，距離 0 の集合には含まれない頂点の集合を距離 1 の集合とする．以下同様に，距離 i の集合からエッジを逆向きに 1 回だけ辿ることでたどり着ける距離 $i+1$ の集合を求める．そして，距離 $0 \ldots n$ の集合に含まれる頂点数と，グラフに含まれるすべての頂点の数の比が予め決められた値を越えたとき，そこで作業を終了し，平均距離を計算する．終了条件となる頂点数の比を p として，この手順のアルゴリズムを 1 に示す．

Algorithm 1 V_{Δ_t} までの平均距離を近似するアルゴリズム

Require: 各頂点から V_{Δ_t} までの平均距離 $\bar{d}$ を近似する．

Ensure: 終了条件となる頂点数の比 p，グラフの全ての頂点の集合 V，グラフのエッジ集合 E

1: $V_0 \leftarrow V_{\Delta_t}$

2: $n \leftarrow 0$

3: **while** $V_n \neq \emptyset$ **and** $\dfrac{\sum_{i=0}^{n} |V_i|}{|V|} < p$ **do**

4: $\quad V_{n+1} \leftarrow \{u \in V \setminus (V_0 \cup \cdots \cup V_n) | {}^{\exists} v \in V_n \text{ s.t. } (u,v) \in E\}$

5: $\quad n \leftarrow n+1$

6: **end while**

7: $\bar{d} \leftarrow \dfrac{\sum_{i=0}^{n} n|V_n|}{\sum_{i=0}^{n} |V_n|}$

8: **return** $\bar{d}$

このアルゴリズムにより，各区間における $\bar{d}$ の近似値を求めた後，5.1.1 と同様にセンタリングを行ったデータを V_{Δ_t} までの平均距離の入力データとして用いる．実験においては，$p = 0.7$ として近似値を求めた．

5.1.3 SIS モデル

区間 Δ_{t_i} における SIS モデルの感染率 β_i と回復率 γ_i は以下のように計算する．まず，回復率 γ_i は，区間 Δ_{t_i} で対象のことば w を投稿したユーザが，区間 $\Delta_{t_{i+1}}, \ldots, \Delta_{t_{i+j}}$ のいずれにおいても w を投稿しない確率として計算できると考えた．本研究においては $j = 7$ とし，区間ごとに回復率を求めた．こうして求めた回復率と，Δ_{t_i} から $\Delta_{t_{i+1}}$ での感染人口の変化から区間ごとに感染率 β を求める．

5.1.4 クラスタにおけることばの使用率

本研究においては，ユーザのクラスタリングに際して，ユーザの投稿内容，自己紹介文，位置情報などを参照したユーザプロファイルの手法ではなく，ソーシャルグラフをモジュラリティが高くなるようなクラスタに分けるグラフクラスタリングの手法により行う．クラスタリングアルゴリズムには，Newman Fast アルゴリズム [10] を利用した．Newman Fast アルゴリズムは，クラスタ同士を結合したときのモジュラリティの変化量が最大となるようなクラスタ同士を貪欲に結合していくボトムアップクラスタリングの手法である．Newman Fast アルゴリズムにより得られたクラスタ $C_1, \ldots, C_n$ の中で，区間 Δ_{t_i} における対象のことば w の使用数が最も多いクラスタが，全体の使用数に占める割合 o_max を計算する．

5.2 ニューラルネットワークによる学習・予測

本研究においては，ことばを使用するユーザの数の時間的特徴を学習するため，RNN の一種である LSTM を用いて学習を行う．RNN はステップ t におけるインプット i_t が入力される際，ステップ $t-1$ での自身の状態 w_{t-1} も同時に考慮され，活性化関数 f を適用した $f(i_t, w_{t-1})$ を入力とすることで，過去の入力を記憶する能力を持つモデルである．

RNN は過去の状態を入力として受け取り続けることで時系列データを扱うことができるが，実際には過去の入力をすぐに忘れてしまうことが知られている．そこで，RNN の中間層を入出力に加え，メモリと，入力，出力，忘却を判断する 3 つのゲートを持つ LSTM ブロックに置き換えることで，長期依存を学習できないという問題を解決したのが LSTM である．

6. 実験と評価

本章では本研究に際して行った実験とその評価について述べる．

6.1 実験に用いたデータ

実験では，2015 年 1 月から 2015 年 12 月までの Twitter ユーザのツイートデータおよび，2015 年時点で Twitter をクロールして収集された，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を説明する．

6.1.1 ツイートデータ

ツイートデータは，調べたいことば w をクエリとして，Twitter の WebAPI のひとつである search API(注7) を用いて，2015 年 1 月から 2015 年 12 月までの 12 か月分のツイートを収集したのち，API のオプションで除外することのできない以下のようなツイートを削除した．

- ツイート内容にクエリが含まれない場合．対象のクエリがユーザ名に含まれているだけの投稿も取得されてしまうため，ユーザ名にクエリが含まれているツイートを削除した．

収集したツイートデータは，ツイート 1 件ごとに表 1 のようになっている．

表 1 収集したツイートデータの概要

データ	説明
ts	ツイートが投稿された時間のタイムスタンプ
uid	ツイートを投稿したユーザの内部 ID
text	対象のツイートの投稿内容
query	API での検索に使用したクエリ w

ツイートデータ収集に際して使用したクエリ，そのクエリで取得したツイートの数（#tweets），ツイートを投稿したユーザ数（#users）を表 2 に示す．

(注7)：https://dev.twitter.com/rest/reference/get/search/tweets

表 2　ツイートデータ収集に使用したクエリ

query	#tweets	#users
爆買い	232719	128176
ドローン	421763	160516
トリプルスリー	132725	76104
大阪都構想	379022	100645

表 3　実際に収集したツイートデータの例

ts	uid	text	query
1422619875000	2232923797	この前爆買いし過ぎた	爆買い
1422613688000	2836712954	765 円に窓埋めあるからそこまで下落したら爆買いだけどね。	爆買い
1422611440000	2432043643	えまちゃんのエンゼル体操がくそかわいかったから、バイト前にラノベ爆買いしてくる	爆買い
1422641991000	6525302	都市部でのドローンは危険だから Amazon は地下にエアシューター設置してそれを使って配送してくるようになるのでは？ #ねーよ	ドローン
1422640634000	429632033	お掃除ドローンいたらいいのに。	ドローン
1422639559000	2986441345	ドローンは早すぎた 逆か 対策が遅すぎた まだ無人機をノーリスクで飛ばせる時代が来るとは思われてなかった	ドローン
1425047772000	59961735	柳田トリプルスリーありそうですな、ことし	トリプルスリー
1425039137000	439656024	ヤフオクドームランすごいなこれ。東京ドームのカラクリたる所以である左中間右中間の膨らみの小ささも謎のリスペクトしてるようだし、柳田のトリプルスリー待ったなしよ	トリプルスリー
1425035797000	248792018	そろそろプロ野球でトリプルスリー達成者が出て欲しいな。一番最後が 2002 年の松井稼頭央さんだったよな。	トリプルスリー

6.1.2　フォロー関係グラフ

フォロー関係グラフは，Twitter の WebAPI のひとつである followers/ids API [注8] を用いて，言語設定が日本語になっているユーザを幅優先探索により収集した．フォロー関係グラフはフォロワーとフォロイー（フォロワーにフォローされているユーザ）の内部 ID のペアの集合として収集した．収集したグラフは，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，このグラフを直接扱うことは困難であったため，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．

6.2　実験の手順

今回行った実験は特徴量の抽出段階と，ニューラルネットワークによる学習・予測段階の 2 つに分けられる．2 つの段階の詳細についてそれぞれ説明する．

6.2.1　特徴量の抽出

4 つのクエリにより収集した 2015 年 1 月から 2015 年 12 月までのツイートデータのうち，モデルの学習のために，300 日分のツイートデータに対して特徴量を抽出した．区間は $|\Delta_{t_1}| = \cdots = |\Delta_{t_300}| = 24(\text{hours})$ とした．

6.2.2　ニューラルネットワークによる学習・予測

本節では得られた特徴量から，目標となる，対象のことばを用いるユーザ数の予測値を算出する方法を示す．時系列データの学習を行うために，LSTM を用いて，

m_1　LSTM 層のみの 1 層モデル

m_2　LSTM 層+LSTM 層の 2 層モデル

m_3　LSTM 層+LSTM 層+LSTM 層の 3 層モデル

の 3 種類のニューラルネットワークモデルを構築し，それぞれのモデルについて 300 エポックの学習を行い，得られたモデルにより予測を行った．

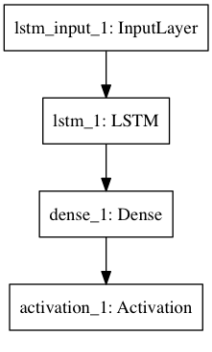

図 6　m_1

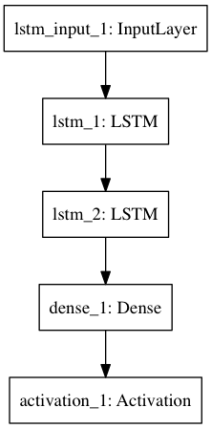

図 7　m_2

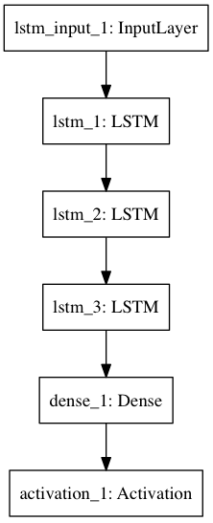

図 8　m_3

(注8)：https://dev.twitter.com/rest/reference/get/followers/ids

表 4 実験結果

入力に使用した特徴量	m_1(MAPE)	m_2(MAPE)	m_3(MAPE)
f	104.7193	104.0217	100.2460
$f, \bar{d}$	99.1705	107.3842	102.8128
f, r_i, r_r	104.9314	105.3243	118.2380
$f, o_{\max}$	99.6183	108.2882	110.9932
$\bar{d}$	105.5918	116.6669	102.4934
$\bar{d}, r_i, r_r$	113.9591	114.1747	101.5679
$\bar{d}, o_{\max}$	103.4175	112.2400	106.4599
r_i, r_r	104.5680	104.8184	105.7658
$r_i, r_r, o_{\max}$	102.2065	106.4259	119.4674

それぞれのモデルには，連続した 10 区間の特徴量の値の組を学習データ，その次の区間での目標値（ことばを使用したユーザ数）を教師データとして学習させた．最適化アルゴリズムには Adam [12] を用いた．

また今回の実装において，ニューラルネットワークモデルの学習には，オープンソースの機械学習ライブラリである TensorFlow(注9) をバックエンドとして，Keras(注10) を用いた．

6.3 評価手法

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）

$$\mathrm{mape} = \frac{100}{n}\sum_{i}^{n}\left|\frac{u^i_{\mathrm{true}} - u^i_{\mathrm{predict}}}{u^i_{\mathrm{true}}}\right| \tag{16}$$

を使用し，4 つのクエリに対しての MAPE の平均値を評価値とした．ただし u^i_{true} は区間 Δ_{t_i} に対象のことばを含むツイートをしたユーザの数，u^i_{predict} はモデルによるその予測値である．

6.4 実験結果

表 4 に実験結果を示す．

今回の実験において考慮したすべての特徴量の組み合わせで，モデルにかかわらず，MAPE の平均値は 100 に近いスコアとなり，有効な予測ができていないことがわかった．

6.5 考察

一連の実験を踏まえ，今後の課題について述べる．

今回，我々の提案する特徴量による時系列データ分析は有効ではなかった．原因として，まず特徴量の質が十分でなかったという可能性が考えられる．実験において特徴量を抽出する際，巨大なグラフデータを扱うための前処理として枝刈りを行い，5.1.2 節，5.1.3 節のように様々な仮定を置き，近似を用いた．これらの手順の妥当性，またより適当な手法について検討し直すことが必要である．

グラフデータに対し，クラスタに含まれるノードの逐次集約などの手法を用いることで，Louvain アルゴリズムによるクラスタリングを高速化した塩川ら [13] のように，グラフ処理を高速化し，グラフ全体を扱うことが可能であれば，より質の良いデータを得られると考えられるが，塩川らの研究で扱われているグラフデータは最大でもエッジ数 367,662 本のグラフであり，今回扱った Twitter のフォロー関係グラフのような巨大なグラフに対してクラスタリングを行うことは計算量の観点から非常に難しい．よって，

- フォロー関係グラフを用いるのではなく，代わりとなるユーザ同士の関係を表すようなグラフを用いる
- フォロー関係グラフを適切な方法で枝刈りする

という 2 つのアプローチが考えられる．

まず，フォロー関係グラフの代わりとなるようなグラフに関して，Ma ら [1] は，フォロー関係グラフではなく，ユーザのメンション関係に着目することで予測を行い，Cheng ら [2] は，投稿をシェア機能により再投稿したユーザのグラフに着目することで予測を行った．これらのユーザ関係グラフは本研究で使用した Twitter のフォロー関係グラフよりもずっと小さいものである．これらの研究のように，ユーザのフォロー関係以外の関係に着目することにより，より，小さいグラフから効果的な特徴量を得るということも考えられる．

またフォロー関係グラフに行う枝刈りに関しても，今回のようにランダムなものではなく，ノードとなる個々のユーザの投稿内容について調べ，対象期間の間に投稿をしていない，ユーザが流行に与える影響はないとし，グラフから削除するという方法でグラフの規模を縮小するといった，より元のデータに含まれる属性を損なわないような方法も考えられる．

7. 結論

本研究では，SNS におけることばの流行予測の手法として，SNS のシェア機能や，投稿内容に着目，分析し，予測を行う従来の手法とは違った手法として，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」という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 を利用したモデルにより，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．しかし，今回検討した手法と特徴量では，時系列データ予測を行う際の有用性は認められなかった．その理由として，

- 入力とした特徴量の質が十分でなかった
- SNS でのことばの拡散においては，ソーシャルグラフの構造だけでなく，コンテンツ的特徴も大きく関わっている

ということが考えられる．

まず入力データの質に関して，今回は収集された巨大なソーシャルグラフデータを扱うため，無作為にエッジの削除を行い，また，ノードから V_{Δ_t} への平均距離といった計算を行う際に近似を用いた．これらの前処理や近似により予測に用いた特徴量の質が低くなってしまった可能性がある．この問題を解決するために，適切な枝刈りアルゴリズムについて検討する，フォロー関係グラフではなく，メンション関係のグラフなど，小規模で有効なグラフについて検討するといった方法が考えられる．

次にグラフ構造以外のコンテンツ的特徴に関して，対象となることば自体や，ユーザの投稿内容，SNS のシェア機能による拡散具合のようなコンテンツ的特徴量が，ハッシュタグやシェ

(注9)：https://www.tensorflow.org/
(注10)：https://keras.io/

ア機能による投稿の拡散の予測において有効であるということは参考文献 [1] [2] [4] などにより指摘されている．本研究ではソーシャルグラフに着目し，グラフ的特徴量による予測を試みたが，最終的に予測精度を上げるためには，ユーザ関係のグラフ的特徴のみではなく，コンテンツの面からも予測に役立つ特徴量について考えることが必要だと考えられる．例えば使用率最大のクラスタにおける使用率を調べる際，グラフのクラスタリングを行うのではなく，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でコミュニティに分割する，ユーザの位置情報ツイートをもとに地域のコミュニティに分割する，などといった方法でもコミュニティ分割を行える．これらのコンテンツ的特徴を用いた手法も検討する必要がある．

また，今回は抽出した特徴量を入力データとしてニューラルネットワークモデルに学習させることで予測を行った．しかし今回提案した SIS モデルの感染率や治癒率といったパラメータは，特徴量として入力するだけでなく，実際にフォロー関係グラフ上でシミュレーションを行うことで感染しているノード数の予測値を出す，というアプローチも考えられる．実際に，SIR モデルを拡張したモデルを用いて，Twitter におけるデマツイートとデマ訂正ツイートの広がり方をシミュレートする研究 [14] や，複雑ネットワーク上に拡張した SIS モデルの解析 [15] なども行われている．

以上を考慮した上で，今後の課題として，

- メンション関係やシェア機能を利用したユーザのグラフなど，フォロー関係よりも小規模なユーザ関係グラフに関して特徴量の抽出を行い，実験を行い評価する
- グラフの特徴量だけでなく，コンテンツ的特徴も含めた特徴量で，予測において有用なものがないかを検討し，実験を行い評価する
- 特徴量を時系列データとして学習するだけでなく，実際のユーザ関係グラフ上でシミュレーションを行うなど，別の予測方法を検討する

ことが挙げられる．

謝　　辞

本研究は JSPS 科研費 16K12430 の助成を受けたものです．

文　　献

[1] Ma, Zongyang, Aixin Sun, and Gao Cong. "Will this# hashtag be popular tomorrow?." Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012.

[2] Cheng, Justin, et al. "Can cascades be predicted?." Proceedings of the 23rd international conference on World wide web. ACM, 2014.

[3] Lehmann, Janette, et al. "Dynamical classes of collective attention in twitter." Proceedings of the 21st international conference on World Wide Web. ACM, 2012.

[4] Tsur, Oren, and Ari Rappoport. "What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities." Proceedings of the fifth ACM international conference on Web search and data mining. ACM, 2012.

[5] Page, Lawrence, et al. "The PageRank citation ranking: bringing order to the web." (1999).

[6] Kleinberg, Jon M. "Authoritative sources in a hyperlinked environment." Journal of the ACM (JACM) 46.5 (1999): 604-632.

[7] 中島伸介, et al. "大規模なブログ記事時系列分析に基づく流行語候補の早期発見手法." 情報処理学会論文誌データベース (TOD) 6.1 (2013): 1-15.

[8] Kermack, William O., and Anderson G. McKendrick. "A contribution to the mathematical theory of epidemics." Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences. Vol. 115. No. 772. The Royal Society, 1927.

[9] Girvan, Michelle, and Mark EJ Newman. "Community structure in social and biological networks." Proceedings of the national academy of sciences 99.12 (2002): 7821-7826.

[10] Newman, Mark EJ. "Fast algorithm for detecting community structure in networks." Physical review E 69.6 (2004): 066133.

[11] Hochreiter, Sepp, and Jrgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

[12] Kingma, Diederik, and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

[13] 塩川浩昭, 藤原靖宏, and 鬼塚真. "ノードの逐次集約による大規模グラフクラスタリングの高速化." 第 4 回データ工学と情報マネジメントに関するフォーラム (DEIM2012), B61 (2012).

[14] 白井嵩士, et al. "Twitter におけるデマツイートの拡散モデルの構築とデマ拡散防止モデルの推定." 人工知能学会全国大会予稿集, IC3-OS-12-1 (2012).

[15] 守田　智. "複雑ネットワークと感染症モデル" 第 18 回交通流数理研究会 (2012)