DEIM Forum 2012 F9-4

Twitterのタイムラインの可視化の一手法

内藤宗一朗† 太田 学†

† 岡山大学大学院自然科学研究科 〒700–8530 岡山県岡山市北区津島中 3–1–1
E-mail: †{naito,ohta}@de.cs.okayama-u.ac.jp

あらまし Twitter には，多くの人が様々な話題に関する情報をリアルタイムに投稿している．特定の話題に関する Twitter のタイムラインを閲覧することにより，その話題に関連した最新の情報を得ることができる．しかし，その話題に対して投稿されるツイートの数が多くなるほど，タイムラインの流れが速くなり，その内容の把握が困難になる．本稿では，刻々と変化する Twitter のタイムラインの内容把握を支援するために，タイムラインを可視化する手法を提案する．タイムライン上に現れる特徴語を抽出し，その特徴語の出現頻度の変化に着目して可視化する．これにより，タイムライン上で多く投稿されている内容や，その話題の変化をユーザに提示する．本稿では，Twitter のツイート検索で得られるツイートのログを実験に利用し，可視化結果を分析する．

キーワード Twitter，可視化，ユーザインタフェース

A Method for Visualizing Twitter Timelines

Soichiro NAITO† and Manabu OHTA†

† Graduate School of Natural Science and Technology, Okayama University
3–1–1 Tsushima-naka, Kita-ku, Okayama, 700–8530 Japan
E-mail: †{naito,ohta}@de.cs.okayama-u.ac.jp

Abstract Many people contribute information about various topics to Twitter in real time. We can get latest information about a specific topic we want by reading Twitter timelines. However, the more tweets about the topic are contributed, the harder for us to follow the timelines. In this paper, we propose a method to visualize Twitter timelines for helping us to understand their content. Our proposed method extracts feature terms from timelines to visualize them by using the change of their term frequencies. Thus, it shows users the content of timelines and the change of topics discussed there. In this paper, we also visualize logs of tweets we get by using the Twitter Search API and analyze the visualization results.

Key words Twitter, Visualization, User interface

1. はじめに

Twitter(注1)は，近年ユーザが増加してきたコミュニケーション・サービスである．Twitter のユーザは，ツイートと呼ばれる短い文章を投稿することで，情報の発信や他ユーザとの交流を行う．そのユーザの多さと情報発信の手軽さから，多様でリアルタイムな情報を多く得る際に Twitter は有用である．例えば，現在放送されているテレビ番組や最新のニュースを閲覧している人が，その場で発信した意見や感想，関連情報を収集することができる．

Twitter では，キーワードやハッシュタグを用いてツイートを検索することで，特定の話題に関するツイートのみを表示するタイムラインを作成することができる．しかし，その話題について大量のツイートが投稿されていると，タイムラインに新たなツイートが次々と表示され，内容を把握しきれなくなる．そこで本稿では，Twitter のタイムラインから特徴語を抽出して可視化し，概要として提示することでタイムラインの内容把握を支援する手法を提案する．タイムライン上に現れる特徴語の出現頻度の推移を計算することで，タイムライン上の話題の変化を可視化する．また提案手法のプロトタイプを実装し，提案手法の有効性を検討する．

本稿ではまず，2 節で関連研究について述べる．3 節では提案する可視化手法，4 節ではプロトタイプの実装について説明する．5 節では，可視化結果の分析と特徴語の評価を行う．6 節では，まとめと今後の課題について述べる．

(注1)：http://twitter.com/

2. 関 連 研 究

坂本ら [1] は，マイクロブログを対象としたリアルタイムな要約生成システムを提案し，Twitter を対象に実験した．彼らはバースト検出法によって Twitter からリアルタイムにイベントを検出し，イベント区間内から代表となるツイートを選択することでイベントの要約を生成した．リアルタイムに進行するタイムラインの内容把握を支援するという点では，本稿と目的が一致する．しかし，本稿のタイムライン可視化手法では，タイムライン全体を可視化して内容把握を支援するため，ユーザへの結果の提示方法が坂本らとは異なっている．

太田ら [2] は，Twitter のリツイート経路を可視化することで，自分と興味が類似するユーザの発見を支援するシステムを提案した．彼らのシステムはリツイートが伝播する様子を可視化して提示することで，自分と同じツイートに興味を持っているユーザを探すことができる．これにより，システム利用者はフォローするユーザの候補を発見できる．

風間ら [3] は，Twitter のツイートに出現する名詞の出現頻度の時間的変化の類似性を用い，Twitter のトピックを分析する手法を提案した．彼らは東日本大震災発生時の Twitter のログを用い，名詞の出現頻度の推移を解析した．そして，名詞間の出現頻度の推移の類似度を算出することで，特定の名詞の関連語を抽出した．彼らの提案手法では，既存の単語共起による手法とは異なる関連語を抽出することができた．

長畑ら [4] は，連続したウェブ検索における検索結果の推移に着目し，検索結果を可視化する手法を提案した．彼らはユーザの検索意図を反映した検索支援を行うために，検索結果中の特徴語の出現頻度の変化に基づき，特徴語と特徴語間の類似度をバネグラフによって可視化した．本稿では Twitter のタイムライン上の話題の推移を扱うため，検索結果の推移を扱う彼らの研究を参考にした．

3. 提 案 手 法

3.1 提案手法の概要

本節では，検索語やハッシュタグから生成した Twitter のタイムラインを，時間経過による話題の変化に着目して可視化する手法を提案する．提案手法では，タイムライン上のツイートから特徴語を抽出し，特徴語の出現頻度と特徴語間の共起頻度を基に可視化する．可視化は特徴語をノード，特徴語間の共起関係をエッジとするグラフにより行う．この可視化結果をユーザに提示することで，タイムラインの内容把握を支援する．また，タイムラインを一定件数のツイートに分割して特徴語抽出と可視化を繰り返すことで，タイムライン上の話題の時間経過による変化を知ることができる．

以下に提案手法の大まかな処理の流れを示す．

（1） 検索語やハッシュタグを用いてタイムラインを生成
（2） タイムラインを複数のツイート群に分割
（3） 可視化対象とするツイート群を選び特徴語を抽出
（4） 特徴語の出現頻度の変化や共起頻度を算出
（5） 出現頻度や共起頻度に基づき特徴語を可視化

表 1　ツイート中に含まれる除外語の例

ストップワード	意味
RT	リツイート，他ツイートの引用
QT	他ツイートの引用
via	発言元，引用元
‘ # ’から始まる文字列	ハッシュタグ
‘ @ ’から始まる半角英数文字列	宛名や発言者などを表すユーザ名

（6） 時系列に沿って可視化対象のツイート群を変更
（7） 3 から 6 を繰り返す

3.2 特徴語抽出

タイムライン上のツイートから特徴語を抽出する手法について説明する．提案手法で用いる特徴語抽出は，藤田ら [5] の手法を参考にした．まず，日本語形態素解析器を利用して，対象となるツイートを形態素解析する．そして，以下の形態素を対象として連結し，特徴語とする．

（1） 名詞（非自立名詞，代名詞等の一部名詞は除く）
（2） アルファベットから構成される未知語
（3） 漢字から構成される未知語
（4） カタカナのみから構成される形態素
（5） 数字
（6） 連体助詞の「の」
（7） 名前区切り記号

特徴語が長くなりすぎることを防ぐため，接頭辞の前と接尾辞の後ろには形態素を連結しない．また，形態素を連結する前後で除外語リストを用いて語をフィルタリングし，ノイズとなる特徴語の抽出を抑えている．ツイート中には，通常の Web 検索結果では見られないような Twitter 特有の表記が頻出するため，それらがノイズとして抽出されないようにフィルタリングする．除外語としている語の一部を表 1 に示す．

3.3 共起関係に基づくタイムラインの可視化

抽出した特徴語を基に，タイムラインを可視化する手法について述べる．まず，3.2 節の手法でタイムラインから抽出した特徴語のうち，出現回数が多い語を可視化の対象とする．出現回数が多い特徴語を可視化することで，タイムライン上の主要な話題をユーザに提示する．提案手法では，出現回数の上位 10 位までの特徴語を可視化対象とする．そして，可視化の対象とした特徴語間の共起回数を算出する．すなわち，特徴語 w_i と特徴語 w_j が同一のツイート内に含まれるとき，w_i と w_j は共起しているとし，その共起の回数を計算する．特徴語 w_i と特徴語 w_j の共起回数 $Co\text{-}occur_{i,j}$ は，式 (1) と式 (2) により算出する．式 (1) の N は可視化対象とするツイートの総数，式 (2) の T_t は t 番目のツイートを表す．

$$Co\text{-}occur_{i,j} = \sum_{t=1}^{N} IsCooccur(t,i,j) \tag{1}$$

$$IsCooccur(t,i,j) = \begin{cases} 1 & \text{if } w_i \in T_t \wedge w_j \in T_t \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

次に，算出した特徴語の出現回数と特徴語間の共起回数を基

に，グラフを生成する．提案手法では，特徴語をノード，特徴語間の共起関係をエッジとするグラフによりタイムラインを可視化している．ノードは，可視化対象とした特徴語を基に生成する．また，共起回数が多い特徴語間には，対応するノード間にエッジを生成する．提案手法では，共起回数の上位 10 位までの特徴語間にエッジを生成する．この共起回数は，可視化対象とした特徴語 10 語の間でのみ計算する．同一のツイート中に含まれる回数が多い特徴語の組は，共起回数が多くなる．そのため，可視化結果中でエッジによって結ばれている特徴語ノードの組は，同一の話題に関連する語である．また，複数の特徴語ノードとの間にエッジが存在する語は，他の多くの特徴語と共起している語であり，話題の中心となる語であることが多い．

3.4 時間経過による話題の変化を考慮した可視化

ある特定のトピックに関するタイムラインにおいても，詳細な話題や内容は時間経過と共に変化していくと考えられる．そこで提案手法では，タイムラインを一定件数のツイート毎に可視化し，その可視化結果をスライドショーのように時系列でユーザに提示する．これにより，タイムライン上に現れる話題の変化をユーザが把握できるようになる．

また提案手法では，ツイート中に現れる特徴語の出現回数の変化を計算し，それにより可視化する特徴語を以下の 3 種類に分類する．

(1) 新しく可視化対象になった語
(2) 継続して可視化され，出現回数が増加した語
(3) 継続して可視化され，出現回数が減少した語

可視化する特徴語の変化とその出現回数の変化は，直前に可視化した特徴語およびその出現回数と比較する．

この分類に従い，特徴語のノードを生成する際にノードの色を変更する．可視化では，上記項目の (1) の特徴語を黄色，(2) の特徴語を赤色，(3) の特徴語を桃色のノードで表示する．(1) の特徴語は新たに可視化対象になった特徴語であり，タイムライン上に新たに現れた話題や，その時点で発生した出来事を表している．(2) の特徴語と (3) の特徴語は，連続する可視化結果に継続して出現する語であり，タイムライン上でも継続して話題となっている語である．このように特徴語を分類して可視化することで，可視化結果からタイムライン上の話題の変化を読み取ることができる．

4. 実　　装

提案手法のプロトタイプを Java で実装した．Twitter API を利用するために，Twitter API の Java ラッパである Twitter4J (注2) を用いた．グラフの生成には，自動レイアウト等の機能を備えた JUNG (注3) を利用した．グラフのノードは JUNG により自動でレイアウトされるが，ユーザの操作でノードの位置を変更することもできる．また，同じ特徴語を表すノードが複数の可視化結果に連続して出現する場合，ノードの再レイアウトの軌跡をアニメーションで可視化する．これにより，同じ特徴語が連続して可視化された場合に，その特徴語ノードを目で追いやすくなる．

(注2)：http://twitter4j.org/
(注3)：http://jung.sourceforge.net/

5. 評 価 実 験

本節では，提案手法の可視化結果の分析と評価を行う．可視化結果については，可視化対象としたタイムラインと見比べることで，タイムライン上の話題を的確に可視化できているか分析する．また可視化した特徴語については，その特徴語が出現するツイートの件数を数え，話題として適当な語が可視化されているか評価する．

実験のために，2012 年 1 月 23 日にハッシュタグ「#NHK」を指定し，1,500 件のツイートから成るタイムラインを生成した．そして，生成したタイムラインの中から時系列に連続するツイート 500 件を選び，可視化結果を分析した．なお，このタイムライン上にリツイートは含んでいない．

5.1 可視化結果の分析

実験データのツイートを時系列に従って 100 件毎に分割し，五つのツイート群を生成して可視化した．可視化結果を図 1〜図 5 に示す．また，実験に用いたツイート群の統計情報と主な話題をそれぞれ表 2 と表 3 に示す．表 2 のツイート番号は，生成したタイムラインに含まれるツイート 1500 件の中での通し番号である．

今回の実験で使用したタイムラインは，テレビ局の「NHK」に関するツイートに付けられるハッシュタグを用いて生成して

表 2　実験データに用いたツイート群

ツイート群	ツイート番号	ツイート投稿時刻
ツイート群 1	ツイート No.101 - 200	19:39:56 - 19:48:20
ツイート群 2	ツイート No.201 - 300	19:48:40 - 20:00:36
ツイート群 3	ツイート No.301 - 400	20:00:43 - 20:18:49
ツイート群 4	ツイート No.401 - 500	20:19:06 - 20:36:10
ツイート群 5	ツイート No.501 - 600	20:36:33 - 20:49:24

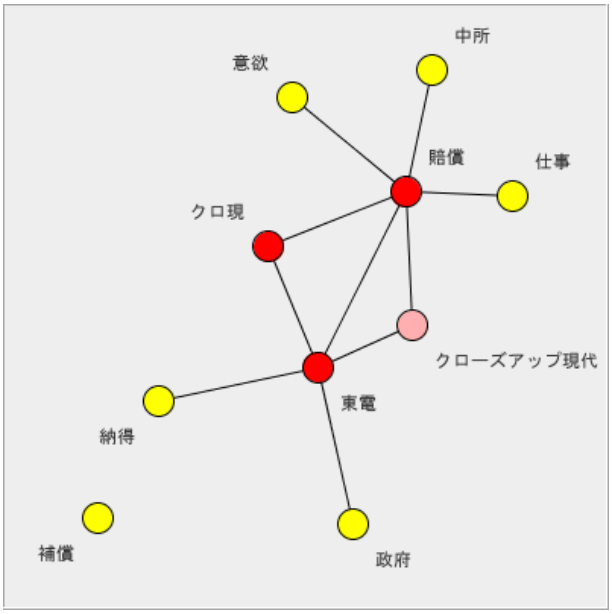

図 1　ツイート群 1 の可視化結果

表 3　実験データに用いたツイート群の主な話題

ツイート群	主な話題
ツイート群 1	TV 番組「クローズアップ現代」，番組のテーマ「原発事故の賠償」
ツイート群 2	TV 番組「クローズアップ現代」，番組のテーマ「原発事故の賠償」
ツイート群 3	TV 番組「鶴瓶の家族に乾杯」，TV 番組「クローズアップ現代」
ツイート群 4	TV 番組「鶴瓶の家族に乾杯」，TV 番組「クローズアップ現代」
ツイート群 5	TV 番組「鶴瓶の家族に乾杯」，福島県の地震

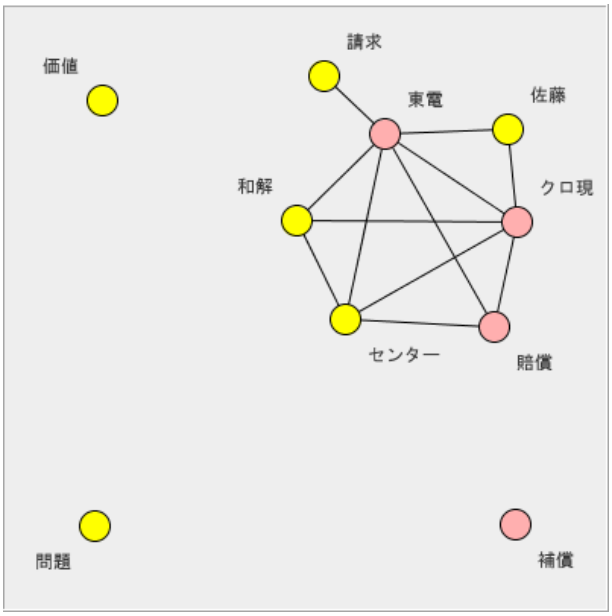

図 2　ツイート群 2 の可視化結果

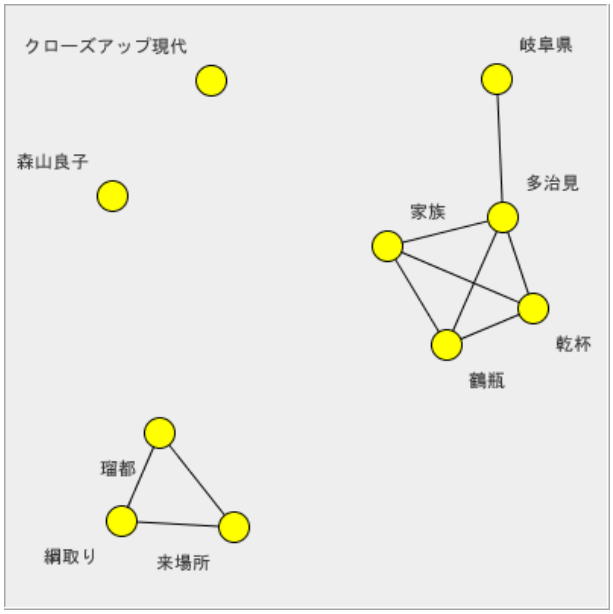

図 3　ツイート群 3 の可視化結果

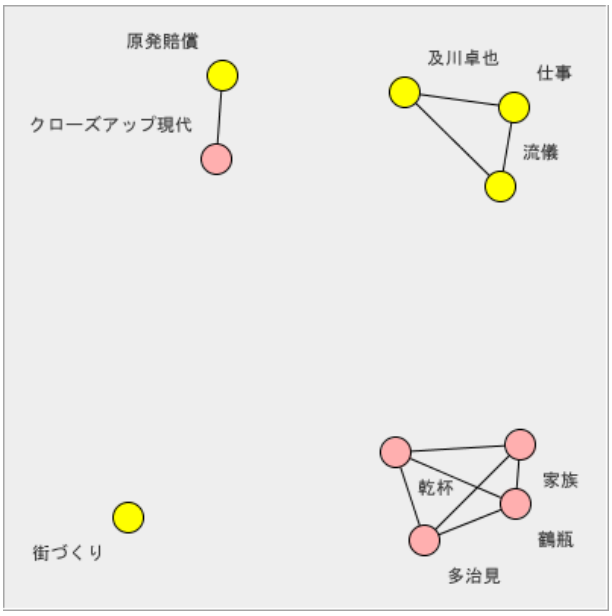

図 4　ツイート群 4 の可視化結果

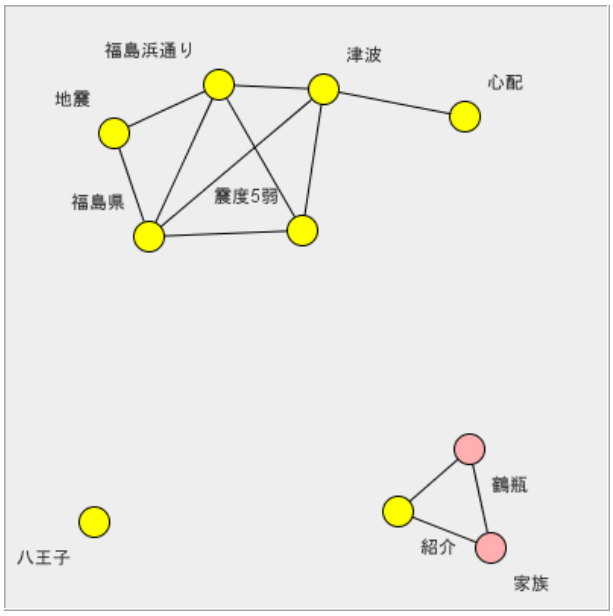

図 5　ツイート群 5 の可視化結果

いる．そのため，その時間帯に放送しているテレビ番組が主な話題となっている．この日は 19 時 58 分までは番組「クローズアップ現代」が放送されており，20 時から「鶴瓶の家族に乾杯」が放送されていた．可視化結果を見ると，図 2 と図 3 の間を境に可視化される特徴語が大きく変化している．これは，放送中の番組が変わることでタイムライン上の話題も変化したことを表している．それとは逆に，番組放送中の可視化結果には番組に関する特徴語が連続して出現している．例えば図 1〜図 2 のように，「クローズアップ現代」が放送されている最中には，番組のテーマとなっていた「賠償」や「東電」といった特徴語が連続して可視化されている．同様に図 3〜図 4 でも，「鶴瓶の家族に乾杯」が放送されている最中には，番組内容に関連

する「多治見」という特徴語が抽出されている．また，図 5 で「地震」や「津波」,「震度 5 弱」など地震に関連する特徴語が多く出現している．これは，20 時 45 分ごろに地震速報が流れ，それに関するツイートが多く投稿されたためである．このことから，可視化した特徴語は，タイムライン上でその時に盛り上がっている話題を示すことができていると言える．

連続する可視化結果にまたがって出現する特徴語は，タイムライン上に継続して頻出している語であり，その時のタイムライン上の主要な話題といえる．また可視化結果に新たに出現する特徴語に注目することで，タイムライン上の話題の変化や事象の発生等に気付くことができる．

さらに特徴語間の共起関係を表すエッジを見ることで，特徴語同士の関連性を把握することができる．例えば，図 3 や図 4，図 5 の可視化結果では，特徴語を結ぶエッジによって，話題が複数のグループに分割されていることがわかる．同グループに属する特徴語同士は共起頻度が高いため，共通の話題に含まれる特徴語であるということが予想される．このように，可視化結果を見ることで，タイムラインに目を通すことなく，タイムライン上の話題の大まかな内容を知ることができる．

5.2 特徴語の評価

各ツイート群から抽出された特徴語と，その特徴語が出現したツイートの件数を表 4〜表 8 にまとめる．各ツイート群はそれぞれ 100 件のツイートから構成されているため，表 4〜表 8 を見ることで，100 件のツイートのうち何件に特徴語が出現していたかがわかる．また，特徴語の出現ツイート数の平均を，各ツイート群での特徴語の出現ツイート数の順位別にまとめた結果を表 9 に示す．

表 9 を見ると，特徴語の出現ツイート数の上位 3 位までは平均して 10 件程度かそれ以上のツイート中に出現しており，ツイート群の話題を十分に表していると言える．特に表 4 では，「賠償」と「東電」という二つの特徴語が 41 回も出現しており，実際のタイムライン上でも多くの人がこれらの特徴語を使ってツイートを投稿していたことを示している．逆に表 7 では，最も多く出現した特徴語「鶴瓶」でも出現ツイート数が 9 件と少なくなっている．この時のタイムライン上では，放送中の「鶴瓶の家族に乾杯」の話題以外にも,「クローズアップ現代」や「仕事の流儀」などの他番組に関する内容がツイートされていた．このように複数の話題がタイムライン上に存在していると，出現ツイート数が極端に多い特徴語は無くなる．

5.3 考　　察

5.1 節の実験では，可視化結果と実際のタイムライン上の話題を比較し，可視化結果がタイムライン上の話題を反映できていることを確認した．提案手法は，テレビ番組の移り変わりや地震の発生など，タイムライン上の話題の変化を反映した可視化結果を生成できていると言える．また，特徴語ノード間のエッジにより，特徴語を話題別にグループ分けすることもできていた．これらの実験結果から，提案する可視化手法は Twitter のタイムライン上の話題や，話題の変化を把握することに役立つと考えられる．しかし，特徴語の出現頻度は可視化結果から知ることができないため，どの特徴語が話題の中心なのかが理解しづらくなっている．特徴語ノードの大きさや形を変えるなどして，特徴語の重要度を可視化できれば可読性の改善につながると考えられる．また特徴語だけを見ても，特徴語を含む言及の内容を把握することは難しい．特徴語を含んでいる元のツイートを表示することができれば，特徴語に関してより詳しく理解することができると考えられる．

5.2 節の実験では，可視化した特徴語がタイムライン上の話題を表す語として適切かを評価した．特徴語を出現ツイート数の順位によってソートし，各順位の特徴語について出現ツイート数の平均を計算した．出現ツイート数の上位 3 位までの特徴語は，平均して約 10 %以上のツイートに含まれており，タイムライン上の話題を表す語として十分に適切であると判断できる．4 位から 10 位の特徴語は，平均すると 3 %以上のツイートに含まれていた．出現頻度を見ると話題の中心と言える特徴語ではないが，それぞれの特徴語を見ると主要な話題に関連する語が多く含まれていた．これらの実験結果から，提案手法で可視化する特徴語は，タイムライン上の話題を表す語として一定の有効性があるといえる．可視化する特徴語の選択に，語出現頻度だけでなく文書出現頻度なども利用すれば，より有用な特徴語を選択できると考えられる．

6. ま と め

本稿では，特徴語の出現頻度の変化に着目し，話題の時系列変化を考慮して Twitter のタイムラインを可視化する手法を提案した．ユーザにタイムライン上の主要な話題を提示するために，タイムラインを特徴語と特徴語間の共起頻度に基づき可視化した．また，タイムラインを時系列に従って分割することで，特徴語の出現頻度の時系列変化を計算した．出現頻度の時系列変化によって特徴語を分類することで，タイムライン上の話題の時系列変化を可視化した．

評価実験では，可視化結果とタイムライン上の話題を比較して分析した．ハッシュタグ「#NHK」によりタイムラインを生成し，そのタイムライン上のツイートを実験に用いた．可視化結果は，タイムライン上の主要な話題や話題の変化を示すことができており，タイムラインの内容把握に役立つと考えられる．また，可視化された特徴語の出現ツイート数を調査し評価した．出現ツイート数の上位 3 位までの特徴語は，タイムライン上の主要な話題を表す語として適切であった．4 位以下の特徴語にも主要な話題に関する語が多く含まれており，提案手法が提示する特徴語はタイムラインの内容把握に役立つ語であると考えられる．

今後の課題としては，可視化の可読性の向上が挙げられる．可視化に利用したグラフでは，ノードの色の他にノードの大きさや形状，エッジの方向などを変更することができる．これらの属性を有効に利用できれば，より多くの情報が視覚的に提示できると考えられる．

表 4　ツイート群 1 の特徴語と出現ツイート数

特徴語	出現ツイート数
賠償	41
東電	41
クロ現	10
クローズアップ現代	8
意欲	6
納得	6
中所	6
政府	6
仕事	5
補償	4

表 5　ツイート群 2 の特徴語と出現ツイート数

特徴語	出現ツイート数
東電	23
賠償	15
センター	7
和解	7
クロ現	7
問題	5
佐藤	4
補償	4
価値	4
請求	4

表 6　ツイート群 3 の特徴語と出現ツイート数

特徴語	出現ツイート数
多治見	22
家族	12
乾杯	11
鶴瓶	11
岐阜県	6
クローズアップ現代	6
森山良子	5
瑠都	3
来場所	3
綱取り	3

表 7　ツイート群 4 の特徴語と出現ツイート数

特徴語	出現ツイート数
鶴瓶	9
多治見	8
乾杯	7
家族	6
仕事	4
流儀	4
街づくり	4
及川卓也	3
クローズアップ現代	3
原発賠償	3

表 8　ツイート群 5 の特徴語と出現ツイート数

特徴語	出現ツイート数
震度 5 弱	16
福島浜通り	15
津波	14
心配	11
八王子	8
地震	7
福島県	5
鶴瓶	4
家族	4
紹介	2

表 9　全ツイート群の特徴語ランク毎の平均出現ツイート数

特徴語の出現ツイート数の順位	平均出現ツイート数
1 位	22.2
2 位	18.2
3 位	9.8
4 位	8.6
5 位	6.2
6 位	5.4
7 位	4.8
8 位	4.2
9 位	3.6
10 位	3.4

文　　献

[1] 坂本翼，横山昌平，福田直樹，石川博：マイクロブログを対象としたリアルタイムな要約生成システムの試作，第 3 回データ工学と情報マネジメントに関するフォーラム (DEIM 2011)，F5-5 (2011)

[2] 太田侑介，寺田実，丸山一貴：Twitter におけるリツイート経路の重ね合わせによるユーザ発見支援，第 10 回情報科学技術フォーラム (FIT2011)，RM-008 (2011)

[3] 風間一洋，鳥海不二夫，篠田孝祐，榊剛史，栗原 聡，野田五十樹：名詞出現頻度の時間的変化に着目した東日本大震災時の Twitter のトピックの分析，Web とデータベースに関するフォーラム (WebDB Forum 2011)，1G-1-2 (2011)

[4] 長畑洋臣，太田学：検索結果の推移の可視化による検索支援，Web とデータベースに関するフォーラム (WebDB Forum 2008)，5A-3 (2008)

[5] 藤田遼治，太田学：文書クラスタリングによる話題の絞込みを利用した先読み検索，情報処理学会研究報告 データベース・システム研究会報告，2010-DBS-151(2) (2010)