DEIM Forum 2014 E1-5

Twitterタイムラインの話題の可視化の一手法

松尾　哉太† 　新妻　弘崇†† 　太田　学 †††

† 岡山大学工学部情報工学科　〒 700–8530 岡山県岡山市北区津島中 3–1–1
†† , ††† 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: †,†††{matsuo, ohta}@de.cs.okayama-u.ac.jp, ††niitsuma@suri.cs.okayama-u.ac.jp

あらまし　Twitter では，多くのユーザにより多様な話題に関する情報が投稿され，タイムラインが時々刻々と変化するため，有用な情報を収集することが困難である．そこで本稿では，Twitter タイムラインの話題の内容理解を支援するために，その話題を可視化する一手法を提案する．提案する可視化手法は，タイムラインに現れる名詞や形容詞等の単語の出現頻度等を利用して，主要な話題や，時間経過による話題の変化を可視化する．実験では，Twitter のツイート検索で得られるツイートのログを利用し，提案手法の有効性を評価する．

キーワード　Twitter，可視化，ユーザインタフェース

A Topic Visualization Method for Twitter Timelines

Kanata MATSUO†, Hirotaka NIITSUMA††, and Manabu OHTA†††

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

Key words　Twitter, Visualization, User interface

1. はじめに

Twitter(注1) は，マイクロブログサービスとして普及し，近年ユーザが増加しているコミュニケーション・サービスの一つである．ユーザは，ツイートと呼ばれる 140 字以内の短い文章を投稿することで情報を発信して他のユーザと交流する．その情報の発信・閲覧の手軽さから，最新の多様な情報を取得する際に Twitter は有用である．

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が作成できる．しかし、ある話題に関するツイートが大量に投稿されていると，タイムラインに新たなツイートが次々と表示され，情報の内容把握が困難になる．そこで内藤ら [1] は，Twitter のタイムラインから特徴語を抽出して可視化することで，Twitter のタイムラインの内容理解を支援する手法を提案した．

本研究では，内藤らが用いた Twitter のタイムラインの可視化手法を改良する．具体的には，Twitter のタイムラインから抽出する特徴語に形容詞と形容動詞語幹を追加し，これらを可視化結果に追加することで，ユーザの意見，感想などを新たに可視化する．また，特徴語の出現頻度を利用し，話題の中心となる特徴語を可視化結果において容易に把握できるようにする．

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 節では実装したプロトタイプシステムについて説明する．5 節では内藤らの可視化結果との比較と，ツイートから抽出した特徴語の評価を行なう．6 節ではまとめと今後の課題について述べる．

2. 関連研究

内藤ら [1] は，Twitter のタイムラインの可視化手法を提案した．この研究では，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タイムラインを可視化した．彼らは特徴語をノード，特徴語間の共起関係をエッジとするグラフにより可視化をした．本研究では，彼らのシステムを拡張して，可視化結果の可読性向上を図る．

坂本ら [2] は，あるトピックにおけるトレンドの変遷を追うために，あるトピックの Twitter ストリームにおけるバーストの断続性に着目した．過去のバーストの情報を用いて，新たな

(注1) : http://twitter.com/

バーストを表すようなキーワードを発見する手法を提案した．これは，時間経過により話題が変化するタイムラインの内容を要約するための適切なキーワードを抽出する可視化手法の一つであるといえる．

加藤ら [3] は，Twitter の情報をグラフで表現する手法として，ユーザ間のリプライ数に注目し，有用でないコミュニティを判定する方法を提案した．このグラフ表現には，ユーザの所属するコミュニティを認識しやすくする特徴がある．

森田ら [4] は，タグクラウドを用いた可視化により，ウェブ上の CGM（Consumer Generated Media）サイトに投稿された評判情報を把握する負担を軽減させる手法を提案した．この可視化手法では，評価表現のみならず，評価対象となる単語も可視化している．これにより，単なる評価表現の可視化よりも，ユーザが評判情報を把握しやすくなるという特徴がある．

川井ら [5] は，テキスト中の評判情報の可視化を目的として，ユーザレビュー中の文章から評判情報を抽出し，それらを肯定，否定情報と合わせて可視化する手法を提案した．彼らは評価表現辞書を用いて評価文を分類してスコア付けし，HK Graph と呼ばれる手法を用いて評判情報可視化している．

3. Twitter のタイムライン可視化手法

3.1 可視化手法の概要

本研究では，内藤らの Twitter のタイムラインの可視化手法を改良する．内藤らは，タイムラインを一定の件数で分割し，分割された各ツイート群から特徴語を抽出し，特徴語をノード，特徴語の共起頻度をエッジとするグラフにより可視化した．その可視化結果を話題の概要としてユーザに提示することで，タイムラインの内容の把握支援を行なう手法を提案した．本研究では，以下の改良を内藤らの提案手法に加える．

（1） 特徴語に形容詞と形容動詞語幹を追加して可視化
（2） 特徴語の出現回数をノードの大きさで表示
（3） ツイート群を任意の時間間隔で分割

形容詞，形容動詞語幹を特徴語とすることで，可視化する名詞などの特徴語に対するユーザの意見や感想といった情報を新たに可視化することができる．これにより，評判情報などが把握しやすくなる．

ノードの大きさを特徴語の出現頻度で区別することで，可視化結果からどの話題が主要であるかを読み取ることができる．これにより，タイムラインに目を通すことなく主要な話題を把握することができる．

内藤らはタイムライン上のツイートを一定の件数で分割していたが，これではツイートの投稿頻度が高いタイムラインを可視化する際，タイムラインの短時間ぼ部分列しか可視化できないことがある．そこで，タイムラインを任意の時間間隔で分割することにより，特定の時刻における主要な話題や，時間経過における話題の変化を把握しやすくする．

これらの改良を加えることで，内藤らの手法を拡張し，可視化結果の可読性を向上させる．以下に本研究における Twitter タイムラインの可視化手法の大まかな流れを示す．

（1） 検索語やハッシュタグからタイムラインを生成
（2） タイムラインを任意の時間間隔でツイート群に分割
（3） ツイート群から特徴語を抽出
（4） 特徴語の出現頻度の変化や共起度を算出
（5） 出現頻度や共起関係に基づき特徴語を可視化
（6） (3) から (5) 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の特徴語の抽出は，内藤らの手法を拡張し以下のように行なう．内藤らは日本語形態素解析器を利用し，以下の形態素を連結して，特徴語とした．

（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

彼らは，形態素を連結する前後で除外語リストを用いて，ツイート中に頻出する Twitter 特有の表記などがノイズとして抽出されないように語をフィルタリングした．

本研究ではこのような形態素に加え，新たに以下の形態素を抽出する．

（1） 形容詞
（2） 形容動詞語幹の名詞

これらは他の特徴語に対するユーザの意見，感想といった評判情報として扱うため，他の形態素と連結せず，単独で特徴語として抽出する．以後，これらの特徴語を評判情報語と呼び，従来の特徴語を話題語と呼ぶ．

3.3 ユーザの意見や感想を考慮した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 節の方法でタイムラインから特徴語を抽出する．話題語は出現頻度を計算し，出現回数の多い上位 10 件までの語を可視化対象とする．また，可視化対象の話題語間の共起関係を調べる．評判情報語については，抽出した語が可視化対象の話題語と共起関係にあるかを調べる．そして，出現頻度に関わらず，話題語と共起関係にある評判情報語のみを可視化する．なお，特徴語 w_i と特徴語 w_j が同一のツイートに含まれるとき，w_i と w_j は共起関係にあるとする．特徴語 w_i と特徴語 w_j との共起関係の有無 $IsCooccur(t,i,j)$ は，式 (1) により算出する．式 (1) の T_t は t 番目のツイートを指す．

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

次に，算出した特徴語の出現回数と特徴語間の共起関係を基に，可視化するグラフを生成する．本研究では，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する．ノードは，可視化対象となった特徴語である．可視化の際，話題語は丸のノード，評判情報語は正方形のノードで示される．また，特徴語間に共起関係があれば，対応する特徴語間にエッジを生成する．話題語同士の共起関係は黒

のエッジで，評判情報語と話題語の共起関係は赤のエッジで表示する．可視化結果において，赤のエッジで結ばれる評判情報語ノードは，話題語ノードに関するユーザの意見や感想といった情報を表していることが多い．

3.4 時系列に変化する話題を考慮した可視化

タイムラインにおいて，話題の内容は時間経過により変化していく．内藤らは，タイムラインを 100 件のツイート毎に分割し，それぞれの可視化結果を時系列順にユーザに提示することで，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した．また，ツイート中に現れる特徴語の出現回数を計算し，直前の 100 件のツイートでの出現回数と比較して，可視化する特徴語を以下の 3 種類に分類した．

（1） 新しく可視化対象となった語
（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

この分類に従い，(1) を黄色，(2) を赤色，(3) を青色のノードで表示した．(2) の特徴語と (3) の特徴語は，時系列に並べた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．本研究では，(1)，(2)，(3) の分類を話題語のみに適用する．本研究で提案する評判情報語は，話題語と区別するため，出現回数の変化に関わらず，すべてピンク色のノードで表示する．

更に本研究では，可視化結果として表示する特徴語を含んだツイートの出現回数を計算し，その出現回数を基に各特徴語のノードの大きさを定める．これにより，可視化結果からタイムライン上における主要な話題を詳しく読み取ることができる．

また，ツイート群の分割単位を，件数から任意の時間間隔に変更する．任意の時間間隔を指定しタイムラインを分割するので，前後の可視化結果を比較することにより，時間経過による話題の変化を読み取ることが容易になる．

本手法において作成した可視化グラフの例を図 1 に示す．これは，2014 年 1 月 20 日に，ハッシュタグ「#news」を指定して検索したツイートから作成した可視化グラフである．

4. 実　　装

提案手法を Java で実装し，ツイート検索により取得したツイートを可視化した．ツイートの取得には，Twitter API の Java ラッパである Twitter4j(注2) を用いた．形態素解析には日本語形態素解析システムの Sen を，可視化のためのグラフ生成にはフリーのグラフ描画ツールである JUNG(注3) を用いた．可視化結果のグラフのノードは JUNG により自動で生成されるが，ユーザの操作でノードの位置を変更することができる．

5. 評 価 実 験

実験として，内藤らの手法と本研究の提案手法を比較し，可視化結果の可読性等について検討する．また可視化した特徴語について，その特徴語が出現するツイートの件数を数え，タイムライン上の話題として適切な語が可視化されているか評価する．

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

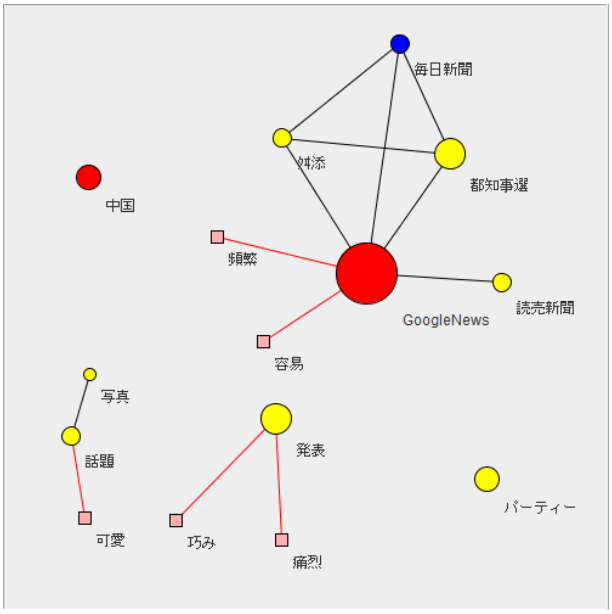

図 1　本手法における可視化グラフ例

5.1 可視化結果の比較

提案手法における可視化結果の例を図 2〜 図 5 に示す．また，内藤らの手法における可視化結果の例を図 6〜 図 9 に示す．これらは，2014 年 2 月 3 日にハッシュタグ「#政治」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，それを時系列順に可視化したものである．そのため生成したタイムラインでは，政治に関する話題が可視化されている．なお，取得したツイートの中にリツイートは含まない．可視化に用いたツイート群の説明を表 1 と表 2 に示す．表 1 と表 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 1,500 件の中での通し番号である．本研究では内藤らの可視化結果との比較のため，表 2 のツイート群 N-1 の時間間隔を分割単位に使用した．

表 1　提案手法で用いたツイート群

ツイート群	ツイート番号	ツイート投稿時刻
M-1	101 - 200	18:04:09 - 19:42:23
M-2	201 - 328	19:42:56 - 21:19:12
M-3	329 - 417	21:21:53 - 23:00:42
M-4	418 - 529	23:01:06 - 00:40:12

表 2　内藤らの手法で用いたツイート群

ツイート群	ツイート番号	ツイート投稿時刻
N-1	101 - 200	18:04:09 - 19:42:23
N-2	201 - 300	19:42:56 - 21:04:09
N-3	301 - 400	21:04:13 - 22:37:53
N-4	401 - 500	22:38:09 - 00:07:50

よって，図 2〜 図 5 はタイムライン中の連続するツイートを約 1 時間 40 分ごとに分割して可視化した結果である．一方，

図 6〜 図 9 は，タイムライン中の連続するツイートを 100 件ずつ可視化した結果である．

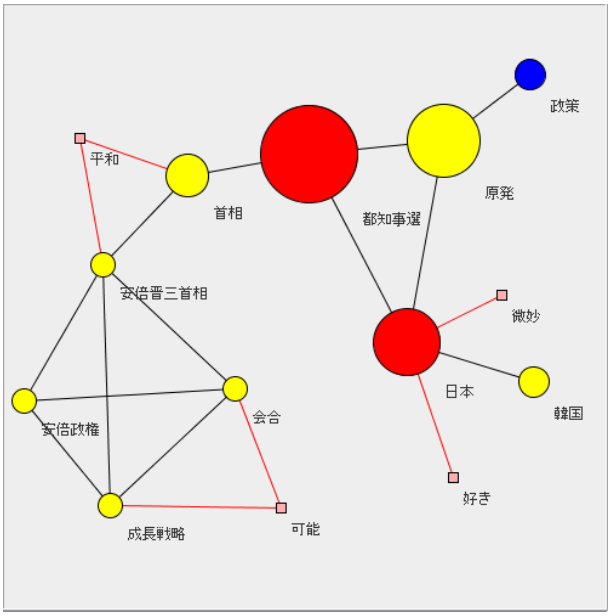

図 2　提案手法によるツイート群 M-1 の可視化結果

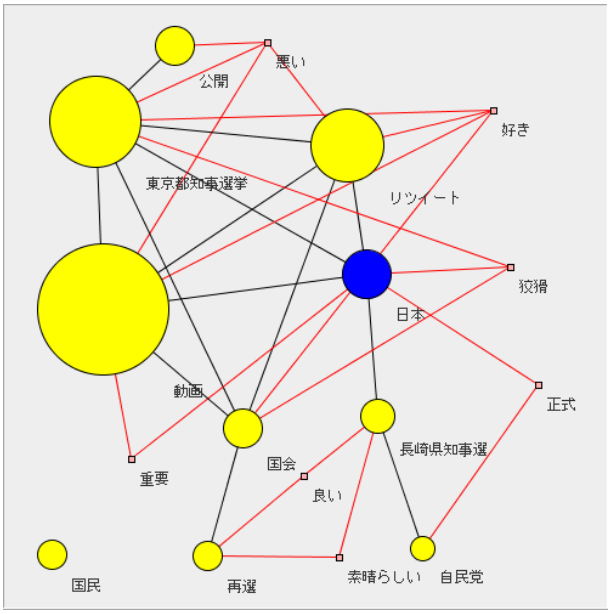

図 3　提案手法によるツイート群 M-2 の可視化結果

表 1 を見ると，タイムラインの分割単位を時間間隔に変更すると，各時間帯における投稿されたツイート数の変化が見て取れる．

図 2 を見ると，「都知事選」や「原発」といった話題語を表すノードが他のノードに比べて大きくなっており，これらの語がタイムライン上で主要な話題であることが把握できる．また，「日本」や「首相」といった話題語に対して「平和」や「好き」といった評判情報語がエッジで繋がっており，それぞれの話題に対するユーザの意見や感想が可視化されている．同様に，図 3 では「動画」や「東京都知事選」，図 4 では「都知事選」と

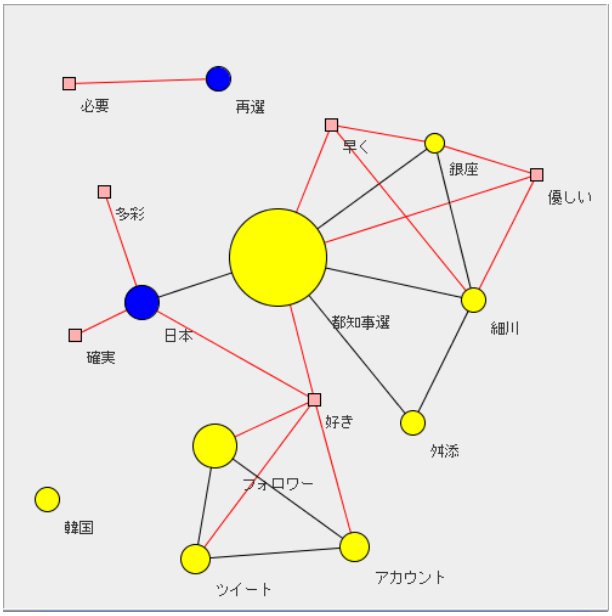

図 4　提案手法によるツイート群 M-3 の可視化結果

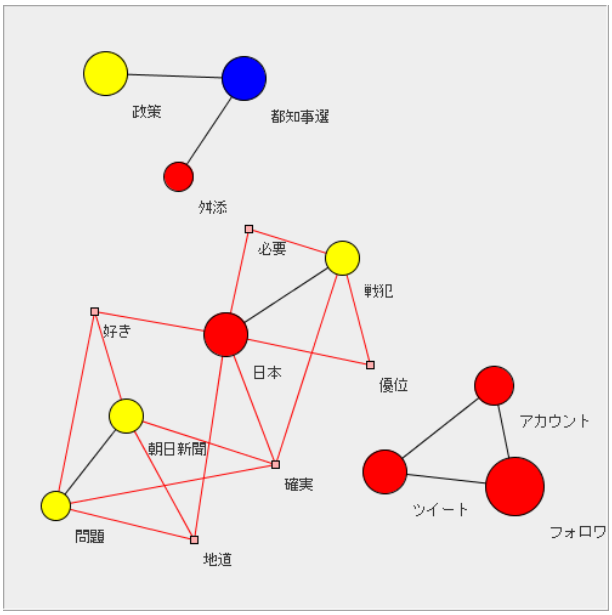

図 5　提案手法によるツイート群 M-4 の可視化結果

いった話題語を表すノードが大きくなっており，タイムライン上では，東京都知事選挙に関する話題が引き続き主要であることが分かる．また，各話題語に対する評判情報語が可視化されている．図 5 では，他のノードと比べて特に大きくなっているノードが無かった．これは複数の話題がタイムライン中に存在し，出現ツイート数が極端に多い特徴語が無かったためである．

図 6 を見ると，ノードの色の変化で特徴語の出現回数の増減が把握できる．しかし，可視化結果からどの話題が多く呟かれているかを把握することは難しい．これは図 7〜 図 9 にも同じことがいえる．

このことから，提案手法における可視化結果は，内藤らの手法と比較して，ノードの大きさで話題語の出現頻度を可視化したことにより，タイムライン上の主要な話題を把握しやすく

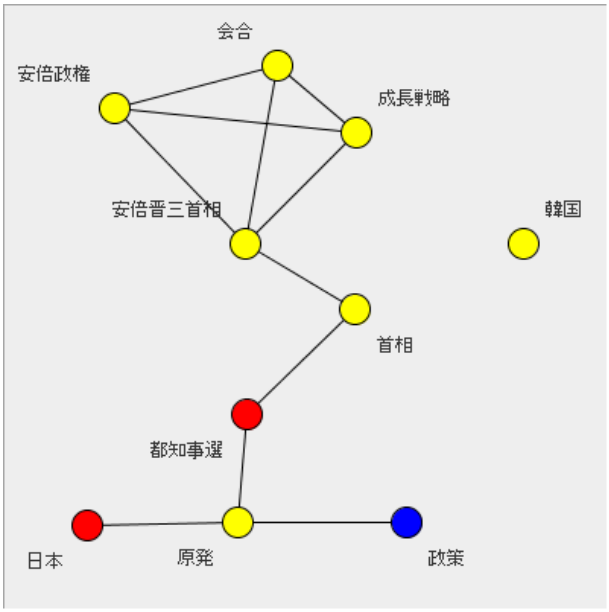

図 6　内藤らの可視化手法によるツイート群 N-1 の可視化結果

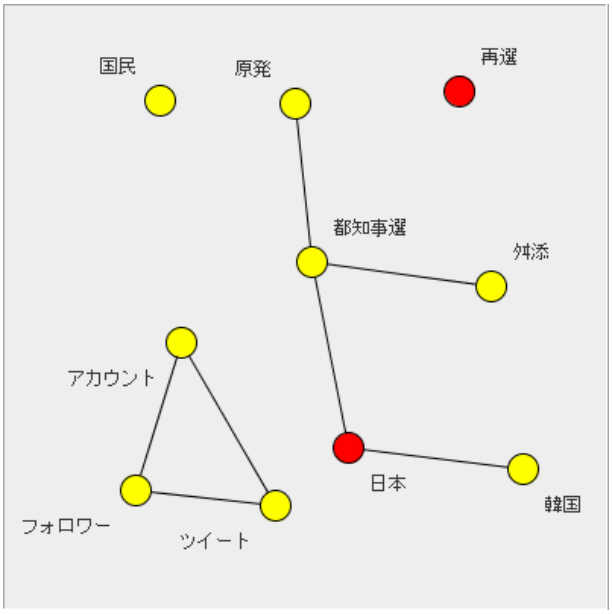

図 8　内藤らの可視化手法によるツイート群 N-3 の可視化結果

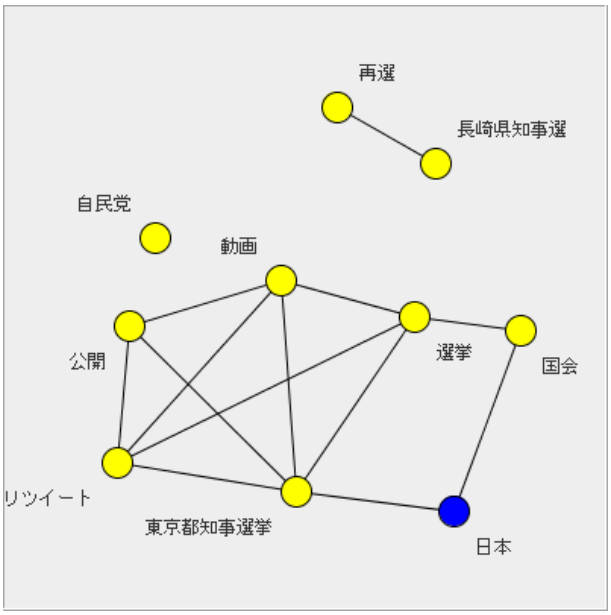

図 7　内藤らの可視化手法によるツイート群 N-2 の可視化結果

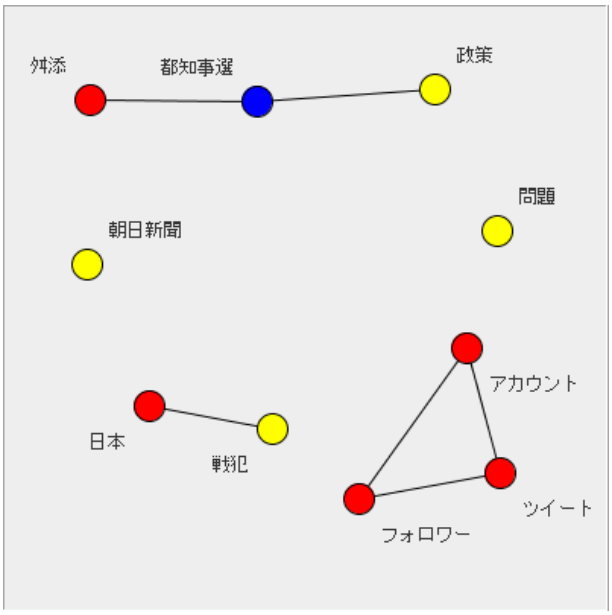

図 9　内藤らの可視化手法によるツイート群 N-4 の可視化結果

なったといえる．また，その話題語に対するユーザの意見や感想といった情報も可視化したことで，話題の詳細な内容理解が容易になったといえる．

5.2　特徴語の評価

提案手法による可視化結果において，各ツイート群から抽出され可視化した話題語と，その話題語が出現したツイートの件数を表 3〜 表 6 にまとめる．また，可視化した評判情報語の各ツイート群における数を表 7 に示す．なお，ツイート群 M-1 は 100 件，ツイート群 M-2 は 129 件，ツイート群 M-3 は 89 件，ツイート群 M-4 は 112 件のツイートから構成されている．そのため表 3〜 表 6 を見ることで，可視化した特徴語がどのくらいの割合でツイートに出現したかが分かる．

表 3〜 表 6 を見ると，どのツイート群においても，話題語の出現ツイート数の上位は平均して 10 件程度のツイートに出現しており，タイムライン上の主要な話題といって問題ない．特に，ツイート群 M-2 では「動画」という話題語が 36 件，ツイート群 M-3 では「都知事選」という話題語が 20 件も出現した．このことから，実際のタイムラインでも，多くのユーザがこれらの語を使ってツイートを投稿したことを示している．表 7 を見ると，平均して 5 語程度の評判情報語が可視化された．特にツイート群 M-2 において，可視化対象となった 7 語の評判情報語が複数の話題語と共起しており，このタイムライン上の話題に対する意見や感想が多くツイートされていることが分かる．

5.3　考　　察

5.1 節では，可視化結果の可読性について，内藤らの手法と本稿の提案手法を比較した．本稿で，可視化結果に評判情報語を新たに追加したことで，タイムライン上の話題に対する評判情報が把握しやすくなったといえる．また，特徴語の出現頻度

表 3　ツイート群 M-1 の話題語と出現ツイート数

話題語	出現ツイート数
都知事選	13
日本	11
原発	11
首相	7
韓国	5
政策	5
安倍晋三首相	4
会合	4
成長戦略	4
安倍政権	4

表 4　ツイート群 M-2 の話題語と出現ツイート数

話題語	出現ツイート数
動画	36
リツイート	15
東京都知事選挙	14
日本	10
公開	8
国会	8
長崎県知事選	7
再選	6
国民	6
自民党	5

表 5　ツイート群 M-3 の話題語と出現ツイート数

話題語	出現ツイート数
都知事選	20
フォロワー	9
日本	7
アカウント	6
ツイート	6
韓国	6
再選	5
舛添	5
細川	5
銀座	3

表 6　ツイート群 M-4 の話題語と出現ツイート数

話題語	出現ツイート数
フォロワー	11
ツイート	9
都知事選	9
政策	9
日本	9
アカウント	8
戦犯	7
朝日新聞	7
舛添	6
問題	6

に応じてノードの大きさを変更し，タイムラインの分割単位を時間にしたことにより，時系列に変化するタイムライン上の主要な話題を把握しやすくなったと考えられる．

表 7　各ツイート群における可視化された評判情報語の数

ツイート群	評判情報語数
M-1	4
M-2	7
M-3	6
M-4	5

5.2 節では，可視化した話題語を出現ツイート数の順位によってソートした．出現ツイート数が上位の話題語は，平均して 10 件以上のツイートに含まれていることから，タイムライン上の主要な話題を表していると判断できる．また，各ツイート群ごとに評判情報語の出現語数を数えた．各ツイート群で平均して 5 語程度の評判情報語が可視化されており，図 2〜 図 5 から，話題語に対するユーザの意見や感想が抽出できていることも分かる．

6. まとめと今後の課題

本研究では，ユーザが Twitter タイムライン上の主な話題を把握しやすいように，内藤らが実装した Twitter のタイムラインの可視化システムを改良した．可視化する特徴語に形容詞，形容動詞語幹も加えることで，名詞などの話題を表す特徴語に対するユーザの意見，感想といった情報を新たに可視化した．また，特徴語を任意の時間間隔で分割したツイート群から抽出し，特徴語の出現ツイート数に基づいて可視化結果のノードの大きさを変更した．これにより，時間とともに変化するタイムラインにおける主要な話題を把握しやすくした．

今後の課題として，可視化する特徴語の抽出方法の改善が挙げられる．本研究では，形容詞と形容動詞語幹を評判情報語として扱ったが，抽出した語の中には，検索結果の理解にあまり役に立たない語もいくらか含まれていた．そこで，評判情報語に肯定，否定等のスコアを付与し，極性値の高いスコアを持った評判情報語のみを可視化結果に表示させる方法を検討している．これにより，話題語に対する意見や感想を分類することができるようになり，有用な情報を含む評判情報語を多く抽出できると考えられる．

他の課題として，可視化結果を表示する際に，時系列変化の様子をなめらかに表示させることが挙げられる．提案手法では，可視化結果を時系列に連続表示しても，前後の可視化結果に外観上の連続性がない．そこで，既に表示されているノードの位置を次の可視化結果でも引き継げば，時間変化になめらかに対応した可視化が可能になると考えられる．

文　献

[1] 内藤宗一郎，太田学，Twitter のタイムラインの可視化の一手法，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 (DEIM 2012)，F9-4(2012).

[2] 坂本翼，廣田雅春，横山昌平，福田直樹，石川博，Twitter ストリームのバーストの断続性に着目したキーワード抽出，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 (DEIM 2012)，C7-3(2012).
[3] 加藤翔子，斉藤和巳，風間一洋，佐藤哲司，MDSR 法を用いた reply ツイートネットワークの特性分析，Web とデータベースに関するフォーラム (WebDB Forum 2013)，A3-2(2013).
[4] 森田雄介，藤本悠，大原剛三，タグクラウドを用いた評判情報可視化システム，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 (DEIM 2012)，C9-2(2012).
[5] 川井康示，吉川大弘，古橋武，可視化によるユーザレビューの評判情報解析に関する研究，Human-Agent Interaction Symposium 2011(HAI シンポジウム 2011)，Ⅲ-2B-2(2011).