DEIM2020 P1-2 (day1 d21)

単語の分散表現を用いたLDAのトピックラベリングと時系列可視化

大町　凌弥†　　風間　一洋†　　榊　　剛史††

† 和歌山大学システム工学部　〒 640–8510 和歌山県和歌山市栄谷 930
†† （株）ホットリンク　〒 102–0071 東京都千代田区富士見一丁目 3–11
E-mail: †{s216319,kazama}@wakayama-u.ac.jp, ††t.sakaki@hottolink.co.jp

あらまし　本論文では，LDA で分類されたトピックの理解を支援するために，抽出されたトピックに既知の論点や概要を示すラベルを自動付与すると共に，同一ラベルのトピックの内容の時間に伴う変遷をわかりやすく可視化する手法を提案する．まず，トピック全体を的確に表すラベルを求めるために，最上位を除く上位 5 件の単語の平均ベクトルに最も類似するラベルを付与する．また，既知の論点文字列の分散表現ベクトルに最も類似する単語が上位に出現するトピックにラベルを付与する．さらに，一定期間ごとに LDA を適用し，共通のラベルでトピックを関連づけて，その時系列変化をアニメーションとして可視化する．

キーワード　LDA，word2vec，分散表現，トピックラベリング，時系列可視化

1 はじめに

Blei らが提案した LDA は，データセット内に存在している潜在的な複数のトピックを推定するための代表的なトピックモデルである [1]．文書の Bag of Words（BoW）から教師なし学習でトピックに分類できることから，広く使われている．さらに，LDA を改良して，実世界から取得したニュース記事や電子メール，Twitter[1]のツイートのような時系列に沿って生成される文書群のデータにおけるトピック内容の追跡や変遷をおこなう様々な手法も提案されている．しかし，トピックを単語集合として表現するだけなので内容の適切な把握が困難であることや，所与のトピックの関連づけができないなどの問題が存在するために，一般的なユーザにとっては必ずしも理解しやすいとは言えない．

本論文では，LDA で分類されたトピックの理解を支援するために，word2vec [2] で求めた単語の分散表現を用いて抽出されたトピックに既知の論点や概要を示すラベルを自動付与すると共に，同一ラベルのトピックの内容の時間に伴う変遷をわかりやすく可視化する手法を提案する．具体的には，まず，既知の論点文字列の分散表現ベクトルに最も類似する単語が上位に出現するトピックに論点ラベルを付与し，それ以外のトピックは全体を的確に表すラベルを求めるために，最上位を除く単語の平均ベクトルに最も類似する内容ラベルを付与する．さらに，一定期間ごとに LDA を適用し，共通のラベルでトピックを関連づけて，その時系列変化をアニメーションとして可視化する．

2 関連研究

所与のラベルに関係づけながら分類するトピックモデルに関する研究がおこなわれている．例えば，Ramage らは，人手で付与されたタグを文書の意味や内容を表すものとみなして，LDA におけるトピック分布を推定する Labeled LDA を提案した [3]．ただし，Labeled LDA ではタグが付与された文書に限定されることから，鈴木らは，単語共起や文書類似度などを用いて，文書集合からタグの代わりになる擬似ラベルを作成して Labeled LDA を適用可能にする手法を提案した [4]．また，Wood らは，与えられた内容が既知の情報源からディリクレ分布のハイパーパラメータを求める半教師あり学習によるトピックモデル Source LDA を提案した [5]．これらに対して，本論文では任意の文書集合に適用できるように LDA の後処理として，既知および未知のトピックへのラベリングの実現を試みる．

ニュース記事や電子メール，Twitter のツイートのように時系列に沿って生成される文書群から，時間と共に変化するトピックを追跡する研究は数多く存在する．例えば，Blei らは時系列文書集合中のトピックを追跡できるように拡張した DTM（Dynamic Topic Model）を提案した [6]．芹沢らは，新聞記事に存在するトピックを LDA を用いて抽出し，抽出された連続する 2 日間のトピックを類似度によって関連づけることで，トピックを追跡する手法を提案した [7]．藤田らは，Streaming API を使って収集したツイートを LDA で解析する際のトピック数を相関係数検定で決定し，さらに隣接する日付で強い相関がある場合のトピックの系列をグラフで図示した [8]．北田らは，東日本大震災前後のツイートを，LDA を並列分散化した PLDA [9] を用いて 1 日ごとに並列抽出したトピックをコサイン類似度に基づいて関連付けたトピック系列を作成し，選別・順位付けした後に時間的変化を重視して可視化することにより，ツイートアーカイブにおける話題とその変化の理解を容易にする手法を提案した [10]．本論文では，トピックに付与したタグで時系列を求め，内容の変化をアニメーションで可視化する．

LDA と単語の分散表現を組み合わせた研究としては，東らが LDA の分類結果に対して，word2vec で得られた単語の分散表現を用いて，単語の出現文書数と類似度によるストップワードリストの自動作成と，トピック間距離に基づいた類似トピック統合をおこなう手法を提案した [11]．本論文では，既知およ

1 : https://twitter.com

び未知のトピックの自動ラベリングと，時系列追跡に適用する点が異なる．

3 提案手法

提案手法の手順を以下に示す．LDA に関連する処理は (1)〜(3) であり，本論文で提案するトピックラベリングと時系列可視化は，その後処理として，それぞれ (4)，(5) のように実行する．

（1） BoW の作成
（2） LDA によるトピック抽出
（3） term-score による順位付け
（4） 単語の分散表現を用いたトピックラベリング
（5） トピックの時系列変化の可視化

3.1 ツイートデータの BoW の作成

まず最初に，単語とその出現頻度の組の集合である BoW を，以下の手順で作成する．

3.1.1 記号類とストップワードの除去

一般的にツイートは口語で書かれるために正しく形態素解析できないだけでなく，Twitter 上で流行している独特の言い回しなどがあり，それらがノイズの原因になることが多い．そこで，以下のような単語や文字列をストップワードとした．

- 記号のみで構成されている単語
- 記号，アルファベットのみで構成されている単語
- ひらがなのみで構成された一般名詞
- 長音符 (ー) や「笑」などの 1 文字の単語
- URL，日本語・英語ハッシュタグ

3.1.2 日本語形態素解析による単語への分割

英語文書の場合は空白で区切れば単語単位に分解できるが，本論文では日本語のツイートデータを利用するため，Mecab [12] で日本語形態素解析して，単語に分割する．LDA では潜在的なトピックを確率的に推定するために，BoW にどのような単語を含めるかが，抽出されるトピックの質に大きく影響する．そこで，標準の IPA 辞書ではなく，新語や固有表現が強化されている mecab-ipadic-Neologd [13] を使用する．さらに，抽出された形態素のうち，一般・自立・固有名詞・サ変接続・形容動詞語幹の名詞だけを対象とする．

3.1.3 BoW の作成

最後に文書ごとに，それに含まれる単語とその出現頻度を表す BoW ベクトルを作成し，これを LDA の入力データとする．

3.2 LDA によるトピック抽出

分析対象のツイートアーカイブを一定の時間間隔で複数のツイート集合に分割し，LDA を用いて各時間区間のトピックを抽出する．

LDA は文書の生成過程を確率的にモデル化したトピックモデルのひとつであり，一つの文書中に複数のトピックが混合されていると仮定して，各単語ごとに潜在的なトピックを決定する．LDA では，文書のトピック分布を確率変数とみて生成するために，単語やトピックの多項分布に対する共役事前分布であるディリクレ分布を使用する．つまり，$Multi(\cdots)$ を多項分布，$Dir(\cdots)$ をディリクレ分布として，LDA では文書は以下のように生成される．

（1） 各トピック $k = 1, \cdots, K$ について

（a） 単語分布 ϕ_k を生成: $\phi_k \sim Dir(\beta)$

（2） 各文書 $d = 1, \cdots, D$ について

（a） トピック分布 θ_d を生成: $\theta_d \sim Dir(\alpha)$
（b） 各単語 $n = 1, \cdots, N_d$ について

i. トピックを生成: $z_{dn} \sim Multi(\theta_d)$
ii. 単語を生成: $w_{dn} \sim Multi(\phi_{z_{dn}})$

ここで，ϕ_k はトピック k の単語分布，θ_d は文書 d のトピック分布，z_{dn} は文書 d の n 番目の単語の潜在トピック，w_{dn} は文書 d の n 番目の単語を表す．K はトピック数，D は文書数，N は単語数，α はトピック分布 θ が従うディリクレ分布のハイパーパラメータ，β は単語分布 ϕ が従うディリクレ分布のハイパーパラメータである．

3.3 単語の term-score の計算

LDA では，トピックごとに各単語がそのトピックから生成された出現頻度を出力として得る．ただし，出現頻度でランキングすると，どの文書にも出現するような一般的な単語が上位にくる傾向がある．そこで，単語のスコアとして，多くのトピックに出現する単語ほど小さく，特定のトピックに出現する単語ほど大きくなる term-score [14] を用いて，トピック k における全単語を term-score で降順にソートし，その上位の単語群を用いてトピックの内容を表す．

トピック k における単語 w の term-score は式 1 によって計算される．

$$term\text{-}score_{k,w} = \hat{\beta}_{k,w} \log \frac{\hat{\beta}_{k,w}}{(\prod_{j=1}^{K} \hat{\beta}_{j,w})^{\frac{1}{K}}} \qquad (1)$$

$\hat{\beta}_{k,w}$ はトピック k における単語 w の生起確率であり，式 2 によりトピックの単語分布の推定量として求める．

$$\hat{\beta}_{k,v} = \frac{n_{kw} + \gamma}{n_{k-} + V\gamma} \qquad (2)$$

n_{kw} はトピック k における単語 w の出現確率，n_{k-} はトピック k における単語の出現確率の総和，γ はハイパーパラメータ，V は全文書における語彙数を示す．トピック k における全単語を term-score で降順にソートし，トピックの内容はその上位の単語群を用いて表す．

3.4 単語の分散表現を用いたトピックラベリング

LDA では，文書を単語とその頻度の集合である BoW として扱うために，抽出したトピックの内容も単語の集合として表される．そのため，トピックの内容を単語の羅列で表しても，内容を解釈しにくいという問題点が存在する．

そこで，分類されたトピック自体にラベルを付与する手法を

提案する．トピックごとに，そのトピックの概要を表すようなラベルを付与することによって，ラベルを見ればそのトピックがだいたい何を表しているかが分かるようになると考えられる．さらに細かい内容まで把握したければ，トピック内の単語を見ればよい．

本論文では，トピック抽出の対象とする文書群に対して，目的とする分析のためにあらかじめ用意した既知の論点のラベリングと，それ以外の未知の内容のトピックに対するラベリングの 2 種類のトピックラベリングを行う．以下では，それらの手法の詳細について述べる．

3.4.1 未知の内容のトピックのラベリング

対象とする文書群において，抽出されたトピック全てが，それぞれに対応する既知の論点を持つとは限らない．そこで本節では，既知の論点を持たない場合でもトピックにラベルを付ける手法を提案する．トピックごとにラベル付けを行うにあたって，ラベルとなる単語はそのトピックの概要を表していなければならない．そのため，word2vec の分散表現ベクトルを用いて，トピック内の単語をベクトルとみなしてトピック内の単語ベクトルの平均を取ることで，トピック全体を一言で表すようなラベルを生成する．本論文では，このラベルを**内容ラベル**と呼ぶ．トピックごとに term-score による順位付けはすでに完了しているものとし，以下の手順ですべてのトピックに対して内容ラベルを生成する．

（1） トピック内の上位 n 件の単語を抽出

（2） 抽出した単語を word2vec を用いてベクトル化

（3） 最上位の単語を除いた上位 5 件のベクトルをすべて加算

（4） 加算結果のベクトルに最も類似する単語を内容ラベルとする

なお，単語群全体の分散表現ベクトルを求める手法はいろいろ考えられるが，本論文では，最上位の単語を除いた上位 5 件の単語の分散表現ベクトルの平均を求めて，これをトピックの分散表現ベクトルとする．この理由は，分散表現の再学習が不要であると共に，トピックへの寄与度が一番強い最上位の単語をあえて除いて，それより下位の単語から再推定させることで，より抽象度の高いラベルを付与することを意図しているからである．

3.4.2 既知の内容のトピックのラベリング

文書群のトピックを分析する場合には，分析したい論点や分類したいサブトピックが存在することが多い．本論文では，これらのような既知の論点を表す文字列を**論点ラベル**と呼び，抽出されたトピックのさらなる理解を支援するために付加情報として利用する．そのため，論点ラベルとの内容の類似性に基づいて，対応するトピックを探索する．そこで，あらかじめ用意した論点ラベルに最も内容的に近いトピックに，次の手順で論点ラベルを付与する．

（1） 論点ラベルを日本語形態素解析して，単語に分割する．これは，複合語の場合は word2vec では分散表現が得られないからである．

（2） 全単語の分散表現ベクトルの平均を計算し，それを論点ラベルの分散表現ベクトルとする．

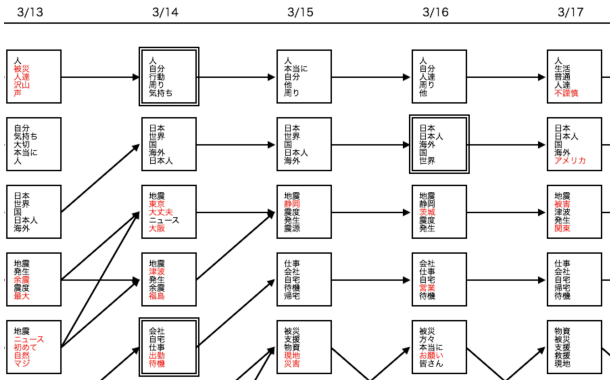

図 1　北田らの手法の出力例 [10]

（3） 論点ラベルの分散表現ベクトルとのコサイン類似度が最も高い単語を上位 n 件に含むトピックを探し，そのトピックに論点ラベルを付与する．

この方法では Labeled LDA や Source LDA のようにトピック抽出の性能を改善することはできないが，ラベル付与された文書集合である必要はなく，単にラベルとして付与したい文字列集合を用意すればよいこと，後処理として実装するだけなので Source LDA のような極端な性能低下のために現実の大規模データへの適用が困難にならないなどの利点がある．

3.5 トピックの時系列変化の可視化

トピックのラベル付けに加えて，トピック理解への支援手法としてトピックの時系列変化の可視化手法を提案する．既存の時系列のトピック抽出手法においては，隣接する時刻の同一と推測されるトピック同士を関係づけ，それを表やグラフ構造で可視化していた．これらの可視化手法では，トピック間の関係を知ることはできるものの，同一トピック内の時系列変化を理解することは難しく，例えば北田らの研究では新出語を赤色で表示するなどの工夫をしていたものの，必ずしも理解しやすいものではなかった [10]．北田らの手法の出力例を，図 1 に示す．

本論文では，注目している論点でトピックラベリングできていることを前提として，注目している論点のトピック内の変化を感覚的に理解できるように，以下のように動画として可視化する．

（1） 抽出された全てのトピックのうち，内容ラベルもしくは論点ラベルが付与されたトピックを 1 つ指定する．

（2） 指定されたトピックの，term-score が上位 n 件の単語を求める．

（3） 求めた単語全てについて，各時刻における term-score を求める．

（4） 3 で求めた内容を csv 形式で出力する．

（5） 以上の処理を，抽出されたトピック全てで行う．

LDA によって抽出されたトピック数を K とすると，ここまでの手順で，K 個の csv ファイルが生成されている．また，前節までのラベリングにより，K 個のトピックそれぞれに対応する内容ラベルまたは論点ラベルが付与されている．K 個の csv

ファイルと，それに対応するラベルをもとに，動画として可視化する．

可視化には，Web サービスである Flourish[2]の Bar chart race[3]という可視化手法を用いる．このサービスに，可視化対象となる論点ラベルまたは内容ラベルと，各時刻における単語の term-score を csv 形式で与えて動画として表示する．

4 評　　価

4.1 データセット

2016 年の参議院選挙の前後の 2016 年 6 月 14 日から 7 月 11 日までの 28 日間に，選挙関連の単語をクエリとして抽出した合計 36,911,653 件の日本語ツイートを評価に使用した．評価データにはツイートされた時間とツイート本文が含まれており，トピックラベリングにはツイート本文を，時系列可視化にはツイートされた時間も使用した．

4.2 不要語の除去

LDA にこの評価データを入力するために，3.1 節の手法に従って不要語を除去した後に BoW を作成する．元のデータセットと比較して，約 17 %となる 6,259,622 件のツイートが残った．

4.3 トピック抽出における条件設定

トピック抽出には，Wang らが C++で実装した並列処理可能な LDA プログラムである PLDA を使用した [9]．ハイパーパラメータは $\alpha = 50/K$，$\beta = 0.01$，反復回数は 500 回とした．

4.4 全期間での LDA の実行結果

まず，全期間でどのような話題があったかを確かめるために，全期間の BoW を結合し，抽出トピック数 K を 20 として LDA を用いてトピック抽出した．term-score の上位 10 件の単語を表 1 に示す.「Tn」はトピック番号を表す．また，単語の左の数字はトピック内の順位を示し，列は各トピックを表す．トピック内の各単語は，term-score の降順に並んでいる．

4.5 内容ラベルの付与結果の分析

term-score 上位 10 件の単語のうち，どの順位の単語を何個用いて内容ラベルを作成すればよいのかを評価するため，以下の 4 種類の手法を試した．以降，以下の手法をそれぞれ手法 1 から手法 4 と呼ぶ．なお，提案手法は手法 4 である．

（1） 10 位までの単語から内容ラベルを付与
（2） 5 位までの単語から内容ラベルを付与
（3） 2 位から 10 位の単語で内容ラベルを付与
（4） 2 位から 6 位の単語でラベルを付与（提案手法）

手法 1 をベースラインとした．手法 2 は上位 5 件の単語のみを使うことで，手法 1 と比較してより具体的なラベル付けを狙ったものである．手法 3 は，手法 1 からあえて最上位の単語のみを除外することで，手法 1 と比較してより抽象的なラベル付けを狙ったものである．提案手法である手法 4 は，手法 2 と手法 3 の中間の特徴をもったラベル付けを狙ったものである．これら 4 つの手法の比較結果を，表 2 に示す．

結果を見ると，おおむね狙い通りのラベル付けができていると考えられる．T3 のラベルが特に顕著な例で，手法 2 では「株」，手法 3 では「市場」となっているが，手法 4 では「景気」となっており，ある程度の具体性をもっておおまかにトピックの内容を表していると言える．よって，最もよい結果を得られているのは手法 4 であると言える．

すべての手法を通して，妥当とは言えないラベルが付与されているトピックがあるが，これは LDA によるトピック抽出がうまくいっていないのが原因であると考えられる．

4.6 論点ラベルの付与結果の分析

4.1 節で述べたように，本論文では選挙に関するデータセットを用いている．このデータセットの収集期間内には，2016 年における参議院選挙が行われていた．そのため，選挙の争点や各政党の関係をまとめた記事が Web 上で多く存在する．本論文では，既知の論点として時事ドットコム [4]上の 2016 年の参院選公約をまとめた記事 [5]を用いた．本サイトによれば，2016 年の参議院公約は以下の 6 つであった．

- 経済
- 憲法改正
- TPP[6]
- 原発
- 安全保障法制
- 子育て支援

なお，論点ラベルとの近似性を内容ラベルとの類似性で求める以下の手法を，比較手法として用いる．

（1） 論点ラベルを日本語形態素解析する．
（2） 論点ラベルを構成する単語の分散表現ベクトルの平均を計算し，それを論点ラベルの分散表現ベクトルとする．
（3） 論点ラベルの分散表現ベクトルとのコサイン類似度が最も高い内容ラベルが付与されたトピックを探し，そのトピックに論点ラベルを付与する．

提案手法と比較手法で論点ラベルを付与した結果を，それぞれ表 3 と表 4 に示す．

末尾に*が付加されているラベルが論点ラベルである．比較手法では，内容ラベルをもとに論点ラベルを付与するため，4 通りの内容ラベルの付け方のそれぞれに対して比較手法を適用した．この結果から，妥当な論点ラベルが付与されていると共に，付与された論点ラベルと一緒に見ることで，該当するトピックが把握しやすくなっていると考えられる．

比較手法と提案手法を比較すると，比較手法ではひとつのトピックに対して論点ラベルが複数付いているが，提案手法ではそれがない．これは，また，比較手法では「子育て支援」のラベルが「年金」を表すと思われるトピックに対して付与されて

2 : https://flourish.studio/features/
3 : https://app.flourish.studio/@flourish/bar-chart-race/9
4 : https://www.jiji.com/
5 : https://www.jiji.com/jc/graphics?p=ve_pol_election-sangiin20160616j-10-w680
6 : https://www.jftc.or.jp/kids/kids_news/japan/kyotei04.html

表 1　全期間での LDA の実行結果

	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9
1	保育園	情報	投票	円	政治	賃金	人	日本	民進党	選挙
2	生活	ブログ	自民	月	公明党	社会	自民党	経済	批判	議員
3	奨学	動画	野党	株	韓国	若者	自分	国	代表	太郎
4	子供	更新	参院	景気	日本	労働	話	世界	民	維新
5	保護	希望	選挙	海外	資金	仕事	人間	離脱	報道	演説
6	教育	猪木	候補	影響	日本人	女性	意味	国民	進	応援
7	保育	拡散	議席	可能	団体	地方	ダメ	イギリス	岡田	山本
8	先生	サイト	与党	市場	組織	企業	理解	政策	声	大阪
9	学校	紹介	共闘	平均	自民党	格差	意見	状況	党首	比例
10	登校	本	公明	稼働	外国	マイ	頭	失敗	自民党	参議院
	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19
1	共産党	舛添	速報	年金	防衛	憲法	主義	原発	安倍	沖縄
2	党	東京	第三者	消費	中国	改憲	民主	福島	政権	熊本
3	支持	都知事	朝日新聞	税	日本	国民	自由	事故	首相	地震
4	政党	蓮	大臣	財政	戦争	改正	国家	放射能	民主党	支援
5	共産	知事	保守	増税	政府	反対	社会	放射線	総理	月
6	民進党	舛	総理	介護	日	自民党	否定	汚染	晋	被災
7	発言	出馬	デジタル	保険	事件	平和	思想	東電	会議	基地
8	人	小池	青山	税金	米	草案	革命	報告	安保	参加
9	政策	五輪	麻生	金融	テロ	緊急	独裁	調査	責任	米
10	予算	オリンピック	拉致	運用	日本人	争点	立憲	原子力	安全	オバマ

表 2　付与された内容ラベルの比較

	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9
手法 1	学校	ブログ記事	与党	水準	国	労働	それ	国	民進党	議員
手法 2	子ども	ブログ	選挙	株	国	労働	人	国	批判	議員
手法 3	学校	ブログ記事	与党	市場	国	社会	それ	国	議員	議員
手法 4	教育	ブログ	自民党	景気	国	社会	それ	国	議員	議員
1	保育園	情報	投票	円	政治	賃金	人	日本	民進党	選挙
2	生活	ブログ	自民	月	公明党	社会	自民党	経済	批判	議員
3	奨学	動画	野党	株	韓国	若者	自分	国	代表	太郎
4	子供	更新	参院	景気	日本	労働	話	世界	民	維新
5	保護	希望	選挙	海外	資金	仕事	人間	離脱	報道	演説
6	教育	猪木	候補	影響	日本人	女性	意味	国民	進	応援
7	保育	拡散	議席	可能	団体	地方	ダメ	イギリス	岡田	山本
8	先生	サイト	与党	市場	組織	企業	理解	政策	声	大阪
9	学校	紹介	共闘	平均	自民党	格差	意見	状況	党首	比例
10	登校	本	公明	稼働	外国	マイ	頭	失敗	自民党	参議院
	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19
手法 1	政党	都知事	総理	年金	北朝鮮	改憲	民主主義	原発事故	政権	被災
手法 2	共産党	都知事	朝日新聞	増税	わが国	改憲	国家	放射能	首相	熊本
手法 3	政党	都知事	総理	税金	テロ	改憲	民主主義	放射能	政権	被災
手法 4	政党	知事	総理	増税	中国	改憲	民主主義	放射能	首相	被災
1	共産党	舛添	速報	年金	防衛	憲法	主義	原発	安倍	沖縄
2	党	東京	第三者	消費	中国	改憲	民主	福島	政権	熊本
3	支持	都知事	朝日新聞	税	日本	国民	自由	事故	首相	地震
4	政党	蓮	大臣	財政	戦争	改正	国家	放射能	民主党	支援
5	共産	知事	保守	増税	政府	反対	社会	放射線	総理	月
6	民進党	舛	総理	介護	日	自民党	否定	汚染	晋	被災
7	発言	出馬	デジタル	保険	事件	平和	思想	東電	会議	基地
8	人	小池	青山	税金	米	草案	革命	報告	安保	参加
9	政策	五輪	麻生	金融	テロ	緊急	独裁	調査	責任	米
10	予算	オリンピック	拉致	運用	日本人	争点	立憲	原子力	安全	オバマ

いる場合があるが，提案手法ではどの論点ラベルも妥当と思われるトピックに付与されている．ただし，用意した論点ラベルの中で,「TPP」だけが付与されていない．これは，TPP の交渉の前に保秘契約書への署名を要求されるように徹底的な秘密主義が貫かれており，与党の国会議員でさえ協定文案を見ることができないので，マスコミの報道がほとんどなく，Twitter 上で議論されることも少なかったからだと考えられる．

以上から，提案手法のほうがよりよい付与手法であると言える．

4.7　トピックの時系列可視化

3.5 節で述べた手法を適用するためには，各トピックごとに，そのトピック内の全単語の日ごとの term-score が必要となる．そのため，追加で各日ごとに LDA を実行し，抽出結果を term-score の降順にソートした．

トピックの時系列可視化では，動画を用いるため，本論文に掲載することができない．そのため，本節では，一部のスク

表 3　比較手法により付与された論点ラベルの比較

	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9
手法 1	学校	ブログ記事	与党	水準	国	経済＊労働	それ	国	民進党	議員
手法 2	子育て支援＊子ども	ブログ	選挙	株	国	労働	人	国	批判	議員
手法 3	学校	ブログ記事	与党	市場	国	経済＊社会	それ	国	議員	議員
手法 4	子育て支援＊教育	ブログ	自民党	経済＊景気	国	社会	それ	国	議員	議員
1	保育園	情報	投票	円	政治	賃金	人	日本	民進党	選挙
2	生活	ブログ	自民	月	公明党	社会	自民党	経済	批判	議員
3	奨学	動画	野党	株	韓国	若者	自分	国	代表	太郎
4	子供	更新	参院	景気	日本	労働	話	世界	民	維新
5	保護	希望	選挙	海外	資金	仕事	人間	離脱	報道	演説
6	教育	猪木	候補	影響	日本人	女性	意味	国民	進	応援
7	保育	拡散	議席	可能	団体	地方	ダメ	イギリス	岡田	山本
8	先生	サイト	与党	市場	組織	企業	理解	政策	声	大阪
9	学校	紹介	共闘	平均	自民党	格差	意見	状況	党首	比例
10	登校	本	公明	稼働	外国	マイ	頭	失敗	自民党	参議院
	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19
手法 1	政党	都知事	総理	子育て支援＊年金	北朝鮮	憲法改正＊安全保障法制＊TPP ＊改憲	民主主義	原発＊原発事故	政権	被災
手法 2	共産党	都知事	朝日新聞	増税	わが国	憲法改正＊安全保障法制＊TPP ＊改憲	経済＊国家	原発＊放射能	首相	熊本
手法 3	政党	都知事	総理	税金	テロ	憲法改正＊安全保障法制＊TPP ＊改憲	民主主義	原発＊放射能	政権	被災
手法 4	政党	知事	総理	増税	中国	憲法改正＊安全保障法制＊TPP ＊改憲	民主主義	原発＊放射能	首相	被災
1	共産党	舛添	速報	年金	防衛	憲法	主義	原発	安倍	沖縄
2	党	東京	第三者	消費	中国	改憲	民主	福島	政権	熊本
3	支持	都知事	朝日新聞	税	日本	国民	自由	事故	首相	地震
4	政党	蓮	大臣	財政	戦争	改正	国家	放射能	民主党	支援
5	共産	知事	保守	増税	政府	反対	社会	放射線	総理	月
6	民進党	舛	総理	介護	日	自民党	否定	汚染	晋	被災
7	発言	出馬	デジタル	保険	事件	平和	思想	東電	会議	基地
8	人	小池	青山	税金	米	草案	革命	報告	安保	参加
9	政策	五輪	麻生	金融	テロ	緊急	独裁	調査	責任	米
10	予算	オリンピック	拉致	運用	日本人	争点	立憲	原子力	安全	オバマ

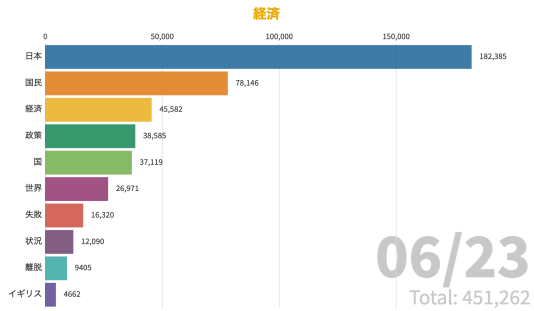

図 2　トピック 7 の時系列可視化結果のスクリーンショット（変動前）

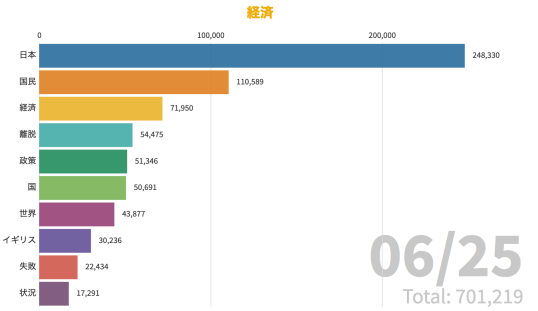

図 3　トピック 7 の時系列可視化結果のスクリーンショット（変動後）

リーンショットを掲載する．図 2 から図 5 に，短期間での順位の変動が激しい単語が確認されたトピックの，ツイートされた回数の変動の様子を示す．

図 2 と図 3 に示した 2 日間で，「離脱」という単語のツイートされた回数が大きく増加しているのがわかる．これは，イギリスの欧州連合（EU）離脱の是非を問う国民投票が 6 月 23 日に実施された影響である．6 月 24 日に結果発表され，その結果に関するツイートが増加している様子が可視化されている．図 4 と図 5 に示した 1 日間で，「円」という単語のツイートされた回数が大きく増加しているのがわかる．これは，先述と同様に EU 離脱に関する投票結果が影響している．事前の予想では残留派が優勢とされていたが，結果発表前に離脱派の優勢へと状況が傾くと，外国為替市場で急激な円高を招いた．そのため，円に関するツイートが増加し，その様子が可視化されている．

表 4　提案手法により付与された論点ラベル

	T0	T1	T2	T3	T4	T5	T6	T7	T8	T9
手法 1	子育て支援*	ブログ記事	与党	水準	国	労働	それ	経済*	民進党	議員
手法 2	子育て支援*	ブログ	選挙	株	国	労働	人	経済*	批判	議員
手法 3	子育て支援*	ブログ記事	与党	市場	国	社会	それ	経済*	議員	議員
手法 4	子育て支援*	ブログ	自民党	景気	国	社会	それ	経済*	議員	議員
1	保育園	情報	投票	円	政治	賃金	人	日本	民進党	選挙
2	生活	ブログ	自民	月	公明党	社会	自民党	経済	批判	議員
3	奨学	動画	野党	株	韓国	若者	自分	国	代表	太郎
4	子供	更新	参院	景気	日本	労働	話	世界	民	維新
5	保護	希望	選挙	海外	資金	仕事	人間	離脱	報道	演説
6	教育	猪木	候補	影響	日本人	女性	意味	国民	進	応援
7	保育	拡散	議席	可能	団体	地方	ダメ	イギリス	岡田	山本
8	先生	サイト	与党	市場	組織	企業	理解	政策	声	大阪
9	学校	紹介	共闘	平均	自民党	格差	意見	状況	党首	比例
10	登校	本	公明	稼働	外国	マイ	頭	失敗	自民党	参議院
	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19
手法 1	政党	都知事	総理	年金	北朝鮮	憲法改正*	民主主義	原発*	安全保障法制*	被災
手法 2	共産党	都知事	朝日新聞	増税	わが国	憲法改正*	国家	原発*	安全保障法制*	熊本
手法 3	政党	都知事	総理	税金	テロ	憲法改正*	民主主義	原発*	安全保障法制*	被災
手法 4	政党	知事	総理	増税	中国	憲法改正*	民主主義	原発*	安全保障法制*	被災
1	共産党	舛添	速報	年金	防衛	憲法	主義	原発	安倍	沖縄
2	党	東京	第三者	消費	中国	改憲	民主	福島	政権	熊本
3	支持	都知事	朝日新聞	税	日本	国民	自由	事故	首相	地震
4	政党	蓮	大臣	財政	戦争	改正	国家	放射能	民主党	支援
5	共産	知事	保守	増税	政府	反対	社会	放射線	総理	月
6	民進党	舫	総理	介護	日	自民党	否定	汚染	晋	被災
7	発言	出馬	デジタル	保険	事件	平和	思想	東電	会議	基地
8	人	小池	青山	税金	米	草案	革命	報告	安保	参加
9	政策	五輪	麻生	金融	テロ	緊急	独裁	調査	責任	米
10	予算	オリンピック	拉致	運用	日本人	争点	立憲	原子力	安全	オバマ

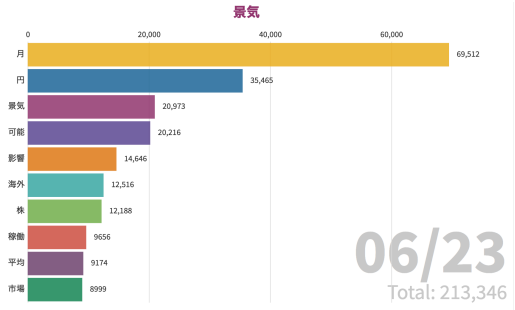

図 4　トピック 3 の時系列可視化結果のスクリーンショット（変動前）

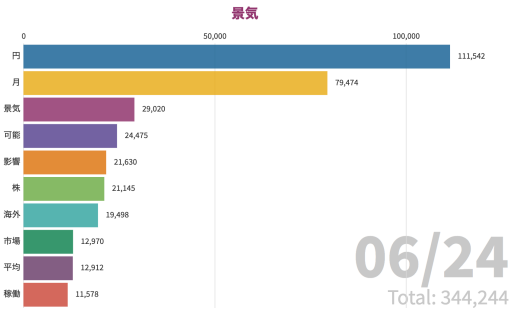

図 5　トピック 3 の時系列可視化結果のスクリーンショット（変動後）

以上より，動画によって期間中のツイート数の変動の様子を可視化することができているのがわかる．特に，ある時事の発生に関連してツイートされることが急激に増える単語に関しては，うまく補足できているといえる．

5　おわりに

本論文では，LDA で分類されたトピックの理解を支援するために，抽出されたトピックに既知の論点を示す論点ラベルやトピックの内容を示す内容ラベルを自動付与する方法と，トピックの内容の時間に伴う変遷を動画でわかりやすく可視化する手法を提案した．さらに，実際に 2016 年の参議院選挙の前後のツイートデータに提案手法を適用し，トピックラベリングとトピックの内容の時系列変化の可視化の妥当性を分析した．

謝　辞

本論文は JSPS 科研費 17H01826 の助成を受けた．

文　献

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pp. 3111–3119, 2013.

[3] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP '09)*, pp. 248–256, 2009.

[4] 鈴木聡子, 小林一郎. 疑似ラベルを用いた潜在的ディリクレ配分法の提案. 第 3 回インタラクティブ情報アクセスと可視化マイ

ニング研究会 SIG-AM-03-04, pp. 20–25. 人工知能学会, 2013.
[5] Justin Wood, Patrick Tan, Wei Wang, and Corey Arnold. Source-lda: Enhancing probabilistic topic models using prior knowledge sources. In *Proceedings of IEEE 33rd International Conference on Data Engineering (ICDE 2017)*, pp. 411–422, 2017.
[6] David M. Blei and John D. Lafferty. Dynamic topic model. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113–120, 2006.
[7] 芹澤翠, 小林一郎. 潜在的ディリクレ配分法に基づくトピック類似度を考慮したトピック追跡. *DEIM Forum 2011 F4-1*, 2011.
[8] 藤野巌, 星野祐子. Twitter におけるトピックの同定手法の提案とそれを用いたトピックの変遷解析. *DEIM Forum 2014 C4-2*, 2014.
[9] Yi Wang, Hongjie Bai, Matt Stanton, Wen-Yen Chen, and Edward Y. Chang. PLDA: Parallel latent dirichlet allocation for large-scale applications. In *Proceedings of the 5th International Conference on Algorithmic Aspects in Information and Management*, pp. 301–314, 2009.
[10] 北田剛士, 風間一洋, 榊剛史, 鳥海不二夫, 栗原聡, 篠田孝祐, 野田五十樹, 斉藤和己. Twitter のトピック変遷の可視化法の提案. *DEIM 2015 E2-6*, 2015.
[11] 東和幸, 高橋仁, 中川博之, 土屋達弘. 単語の出現頻度と類似性に基づいたトピックモデル洗練化手法. コンピュータ ソフトウェア, Vol. 36, No. 4, pp. 4_25–4_31, 2019.
[12] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pp. 230–237, 2004.
[13] 佐藤敏紀, 橋本泰一, 奥村学. 単語分かち書き辞書 mecab-ipadic-neologd の実装と情報検索における効果的な使用方法の検討. 言語処理学会第 23 回年次大会 (NLP2017). 言語処理学会, 2017.
[14] A. Srivastava and M. Sahami, editors. *Text Mining: Theory and Applications*, chapter TOPIC MODELS. Taylor and Francis, 2009.