3P2-NFC-00b-1in1

東日本大震災時のツイートの単語2-gramに基づくトピックの可視化

Visualization of Topics Using Word 2-grams in Tweets of Great East Japan Earthquake

久保 侑哉 *1 Yuya KUBO　風間 一洋 *1 Kazuhiro KAZAMA　鳥海 不二夫 *2 Fujio TORIUMI　斉藤 和己 *3 Kazumi SAITO

*1和歌山大学システム工学部 Fuculty of Systems Engineering, Wakayama University

*2東京大学大学院工学系研究科 School of Engineering, The University of Tokyo

*3静岡県立大学経営情報学部 School of Management and Information, University of Shizuoka

Twitter is an important tool as it can know real-world information in real time. However, it is difficult to grasp discussed topics and their translation continuously on Twitter. In this case, it is conceivable to classify various topics on Twitter by using topic models such as Latent Dirichlet Allocation (LDA). However, it was difficult to understand the contents of the topic extracted by LDA in detail. In this paper, we propose a method to detect topics for bag-of-words, which is created with word 2-grams instead of words, and visualize the results as a topic graph, which is composed of top n word 2-grams in each topic. In fact, we show topics and their topic graphs, which are extracted from tweet archive in the Great East Japan Earthquake, and evaluate them.

1. はじめに

現実世界で起こった出来事を知るための手段の中でも，Twitter は現実世界の情報をリアルタイムに知ることができることから，特に重要な手段となりつつある．ただし，Twitter のタイムラインに表示されるツイートは，自分がフォローしているユーザの投稿やリツイートに限られることから，Twitter の情報空間を俯瞰して，どのような話題が議論されているかを把握することは困難である．このような場合には，Twitter 上に存在する種々の話題を，潜在的ディリクレ配分法（LDA）のようなトピックモデルを用いて分類することが考えられる．

しかし，LDA のようなトピックモデルでは，文書を単語とその頻度の集合である BoW に変換してから扱うために，抽出したトピックの内容が単語の集合として表され，トピックの解釈において，いくつかの問題点が存在する．まず，たとえ term-score を用いてそのトピック固有の単語に絞ったとしても．出現頻度が高い単名詞が上位を占める傾向があり，トピックの内容を詳しく表す複合語が出てこない．次に，あるトピックの内容を的確に表すと思われる新語を扱うためには辞書登録が必要であるが，大量の情報に対してどのような新語が登場するかを予測するのは困難である．また，トピックの内容を単語の羅列で表されても，内容を推測しにくい．

これは，文書を独立した単語に分解した BoW（Bag of Words）として扱うことが原因であると考えられる．そこで，本稿では，単語ではなく，単語の関係性に対するトピック抽出するアプローチを用いる．まず，単語間の関係性と方向性も扱えるようにするために，文書から単語ではなく単語 2-gram で作成した BoW を作成して，単語 2-gram 単位で LDA でトピック抽出するとともに，一旦失われた 3 語以上の隣接関係や事象間の因果関係を得るために，各トピックの上位の単語 2-gram を合成して，トピックの概要を表す単語の有向グラフ構造であるトピックグラフとして可視化する手法を提案する．

連絡先: 久保 侑哉 (s181019@center.wakayama-u.ac.jp)
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

2. トピック抽出方法

北田らは，ツイートアーカイブ中の各日ツイートから LDA で抽出したトピックを，さらにコサイン類似度に基づいて接続することで，時間と共に内容が変化するトピック系列として可視化する手法を提案した [1]．本研究では北田らのシステムを拡張したので，まず北田らの手法の概要を述べる．

2.1 ツイートの BoW の作成

まず，各ツイートの本文を日本語形態素解析器の Mecab で形態素解析してから，単語とその単語の出現回数の値の組の集合である BoW 形式に変換する．ただし，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を使用した．さらに，トピック抽出する際のノイズの原因になる単語は，ストップワードリストを用いて除去する．

2.2 LDA によるトピック分類

次に，全ツイートを発言時刻に応じて複数の期間に分類し，各期間の BoW 集合から LDA を用いてトピックを抽出する．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．各文書はそれぞれトピック分布 θ を持ち，文書中の各単語について，トピック分布 θ に従ってトピック z が選択され，トピック z に対応する単語分布 ϕ に従って，単語 w が生成される．

つまり，LDA では，文書は以下のように生成される．$Multi(\cdots)$ は多項分布，$Dir(\cdots)$ はディリクレ分布である．

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 は単語数である．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を表す．

この結果，各期間ごとに K 個のトピックが抽出される．

2.3 term-score による単語のランキング

LDA で，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が得られるが，これを用いて単語をランキングすると，どの文書にも多く出現する一般的な単語が各トピックの上位にくる傾向がある．そこで，トピック固有の特徴を適切に表す単語を選択するために，term-score[2] を単語のスコアとして用いる．term-score では，多くのトピックに出現する単語のスコアは小さくなり，特定のトピックにのみ出現する単語のスコアは大きくなる．term-score は次式で表される．

$$term\text{-}score_{k,v} = \hat{\beta}_{k,v} \log\left(\frac{\hat{\beta}_{k,v}}{(\prod_{j=1}^{K}\hat{\beta}_{j,v})^{\frac{1}{K}}}\right) \quad (1)$$

$\hat{\beta}_{k,v}$ はトピック k における単語 v の出現確率，K はトピック数である．term-score は情報検索における TF-IDF と同様の考えに基づいており，$\hat{\beta}_{k,v}$ が TF に，log( ) が IDF に相当する．

$\hat{\beta}_{k,v}$ は，次式によりトピックの単語分布の推定量として求めることができる．

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

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．

2.4 トピック系列の作成

次に，LDA で抽出した各期間の関連するトピックが時系列順に繋がった有向グラフ構造を作成する．これを本稿ではトピック系列と呼ぶ．

まず，各期間 t のトピック $i(1 \leq i \leq K)$ の term-score の上位 N 件の単語を抽出し，その term-score から成るベクトル $\boldsymbol{T_i^t}$ を作成する．この理由は，トピックの内容との直接的な関係が薄い一般的な単語や，ほとんど使われていない膨大な単語の影響を除去するためであり，これによりトピック間の類似性の有無がより明確に判断できるようになる．

次に，隣接する期間 t と期間 $t+1$ の間のトピックのすべての組み合わせに対して，次式に示すコサイン類似度 $csim(\boldsymbol{T_i^t}, \boldsymbol{T_j^{+1}})$ を計算し，閾値 S を超える場合に限り連続していると判定する．

$$csim(\boldsymbol{T_i^t}, \boldsymbol{T_j^{t+1}}) = \frac{\boldsymbol{T_i^t} \cdot \boldsymbol{T_j^{t+1}}}{|\boldsymbol{T_i^t}||\boldsymbol{T_j^{t+1}}|} \quad (3)$$

そして，期間を t_s から t_e，トピックをノード，トピック間の関係をエッジとして，次の手順に基づいて各区間 $[t, t+1]$ ごとのエッジリストとしてトピック系列を作成する．

1. 期間 $t = t_s$ とする．
2. 期間 t と $t+1$ の間の各エッジ e_{ij} の始点ノード n_i を終端ノードとして含むトピック系列があれば，その区間 $[t, t+1]$ にエッジ e_{ij} を追加する．なければ，新しいトピック系列を生成してエッジ e_{ij} を追加する．
3. 各エッジ e_{ej} の終点ノード n_j を含むトピック系列が複数あればマージする．
4. 期間 $t = t+1$ とし，$t = t_e$ なら終了し，そうでなければ 2. に戻る．

3. 単語 2-gram を用いたトピック抽出と可視化

提案手法は，北田らの手法におけるツイートの BoW を単語ではなく単語 2-gram で作成するように変更するとともに，新たに単語 2-gram 合成により作成したトピックグラフの可視化機能を追加した．

3.1 単語 2-gram を用いた BoW の生成

単語 2-gram として扱うことで単語の関係性や方向性が扱えるようになる反面，本来は存在しないような関係までも誤抽出される可能性がある．LDA のようなトピックモデルでは，語彙数の増加に伴う計算コストの増大に繋がるために，できる限り低減することが望ましい．そこで，文の境界を超える 2-gram を作成しないように，次の手順に従って，文単位で 2-gram の BoW を作成する．

1. 文書リストから文書を取り出し，日本語形態素解析して，抽出された形態素の中から一般・自立・固有名詞・サ変接続・形容動詞語幹の名詞を取り出て，単語リストを作成する．文書がない場合は終了する．
2. 単語リストの連続する形態素の組をセパレータで結合した単一の文字列に変換して，頻度表のその項目を加算してから，次の形態素に進む．ただし，EOS あるいは句点に到達した場合には，次の単語から新たな文が始まったとみなす．
3. 単語リストをすべて処理してから，単語 2-gram の頻度表を元に文書の BoW を作成し，1. に戻る．

なお，EOS と句点の判定は，日本語形態素解析の結果に基づく．例えば，ツイートなどの文章においては，必ずしも句点が正しく用いられなかったり，顔文字を文末に用いることもあり，文の境界を越えた単語 2-gram を誤抽出する可能性がある．ただし，そのような単語 2-gram の出現頻度は低く，トピックの上位にはならないと考えられる．

また，単語 2-gram 単位で扱うために，単一の単語は対象にはならない．しかし，後述するトピックグラフでは再び分解されて単語がノードになる上に，1 つの単語しか含まない文章が多く存在したとしても，それは議論を抽出したいという目的の場合はノイズとなることから，問題にならない．

3.2 単語 2-gram 合成を用いたトピックグラフの可視化

提案手法では，各トピックの内容は，単語ではなく単語 2-gram の集合として得られる．

ただし，複合語の処理を特に行わないこともあり，形態素辞書に登録されていない複合語や新語は分割されてしまうので，一組以上の 2-gram として表される．また，文中の主語や述語，目的語などの関係も，一組以上の 2-gram として表される．さらに，単語 2-gram として扱うことで，同一トピック中に同じ単語が繰り返し登場することになる．つまり，単語 2-gram を用いてトピック抽出した結果は，必ずしもユーザにとって理解しやすいとは言えない．

そこで，各トピックの上位の単語 2-gram を合成して，ノードを単語，エッジを単語 2-gram の関係とする有向グラフ構造を作成して，それを可視化する．このトピックの内容を表すグラフ構造をトピックグラフと呼ぶ．これにより，一組以上の 2-gram で表される場合は，トピックグラフにおける連続した部分グラフに変換される．また，同一トピック内の単語の重複がなくなり，内容の視認性が向上する．

トピックグラフの作成手順は，以下の通りである．

1. トピックの上位 N 件の単語 2-gram を取り出す．
2. 各単語 2-gram(w_0, w_1) に対して，ノード w_0 を始点，ノード w_1 を終点とするエッジを作成して，有向グラフを作成する．ただし，自己ループ（$w_0 = w_1$）は除外する．
3. 作成した有向グラフを可視化する．

ただし，一度単語 2-gram に分解した結果を再合成することから，本来のデータにない単語の関係が誤抽出される可能性があるが，合成前に各トピックに分類されていることから，トピック固有の単語の関係だけが高く評価されることになり，結果には大きく影響しないと考えられる．

なお，Wang らによる，n-gram のトピックモデルを用いて，機械学習のトップカンファレンスである NIPS のデータから，1-gram，2-gram のトピックを抽出した既存研究 [3] が存在するが，本研究はさらにトピックグラフを作成する点が異なる．

4. 東日本大震災時のツイートの分析

4.1 データセットとトピック抽出

Twitter API[*1] を用いて，3 月 5 日〜24 日の間に 200 件以上日本語でツイートしたアクティブなユーザのツイートを収集し，後日収集漏れを減らすために各ユーザに対して再収集して，それをデータセットとした．200 件は，Twitter API の呼び出し 1 回で取得できる最大ツイート数である．

データセットの規模は，ツイート数が 362,435,649 件，ユーザ数が 2,711,473 人である．データセットには，ツイート ID (64 ビット整数)，ツイートしたユーザのスクリーン名，本文，ツイート元．ツイート時間，リプライ先のツイート ID，リプライ先のスクリーン名などの情報が含まれる．本稿では，ツイートの本文だけを用いて単語 2-gram の BoW を作成した．日本語形態素解析には，IPA 辞書を辞書登録せずにそのまま使用した．

トピック抽出には，Wang らの C++で実装された並列処理可能な LDA プログラムである PLDA を使用した [4]．各パラメータは，トピック数は $K = 50$，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．

4.2 複数トピックの可視化結果の分析

3 月 10 日から 3 月 20 日までのトピック系列の可視化結果の一部を図 1 に示す．コサイン類似度の閾値は 0.7 とした．

この結果から，通常の複合語に加えて,「東日本→大震災」のように辞書登録しても相対的な頻度が低く上位にならなかった複合語や,「太平洋→沖」のような状況依存の複合語が，単語 2-gram として抽出できていることがわかる．

さらに，出現頻度が高い「停電」に注目すると,「停電→中止」,「停電→予定」,「停電→実施」,「停電→詳細」のように複数の単語との関係が抽出され，例えば「停電する予定だ」のような元の文を推測できるような単語 2-gram も抽出されていた．ただし,「停電」のように，トピック中に特定の単語が多数出現する場合は，必ずしも内容を把握しやすいとは言えない．

4.3 トピックグラフの可視化結果の分析

東日本大震災が発生した翌日の 3 月 12 日の「津波」に関するトピックの term-score の上位 50 件の単語 2-gram からトピックグラフを作成し，その最大連結成分を Cytoscape 3.4.0 で可視化して図 2 に示す．レイアウトアルゴリズムには Force-Directed Layout を用いた

この結果から,「津波」と「警報」の次数が高いことに加えて,「津波」は入次数が,「警報」は出次数が高いなど，異なる傾向を示していた.「津波」は,「静岡」,「徳島」,「宮城」,「福島」のような津波の到達が予測されていた太平洋側の都道府県名からエッジが多く張られ,「警戒」,「注意」,「観測」,「到達」など，津波に対する対処を意味する単語にエッジが張られていた.「警報」は,「発令」,「避難」,「解除」などの単語にエッジが張られていた．また,「大津」,「波」,「警報」という順序でエッジが張られた,「大津波警報」という複合語に該当する部分グラフも存在した．

この結果から，トピックグラフ化することで,「警報発令」や「大津波警報」のような複合語や新語が辞書を強化しなくても，トピックに特有な表現として抽出できることや,「津波が到達する」,「警報を発令する」のような文も簡略な形式で表現できていることがわかる．

5. おわりに

本稿では．単語間の関係性や方向性を扱えるようにするために，文書から単語 2-gram で作成した BoW を作成して LDA でトピック抽出するとともに，各トピックの上位の単語 2-gram を合成してトピックグラフとして可視化する手法を提案した．さらに，東日本大震災前後の Twitter のデータセットに適用した結果を分析した．この結果，トピックグラフでは，トピックに依存した新語や複合語に加えて，主語と述語，動詞と目的語のような関係が抽出されるだけでなく，複数の単語間の関係をグラフとして圧縮表現できることから，トピックの内容を簡潔にわかりやすく表示する手法であることを確認した．

今後の課題として，トピックグラフの可視化機能をシステムに組み込むと共に，より多くの情報を簡潔に表現するためのグラフ作成方法と，トピック間の違いや変遷がより理解しやすいような可視化手法を検討する予定である．

謝辞

本研究を行なうにあたり，ツイートデータの収集に協力して頂いたクックパッド株式会社の兼山元太氏に感謝する．また，本研究は JSPS 科研費 26330345 の助成を受けた．

参考文献

[1] 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.

[2] A. Srivastava and M. Sahami, editors. *Text Mining: Theory and Applications*, chapter TOPIC MODELS. Taylor and Francis, 2009.

[3] Xing Wei Xuerui Wang, Andrew MSCallum. Topical n-grams: Phase and topic discovery, with an application to information retrieval. *In Proceedings of the 7th IEEE International Conference on Data Mining (ICDM '07)*, pp. 697–702, 2007.

[4] 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.

*1 http://apiwiki.twitter.com/

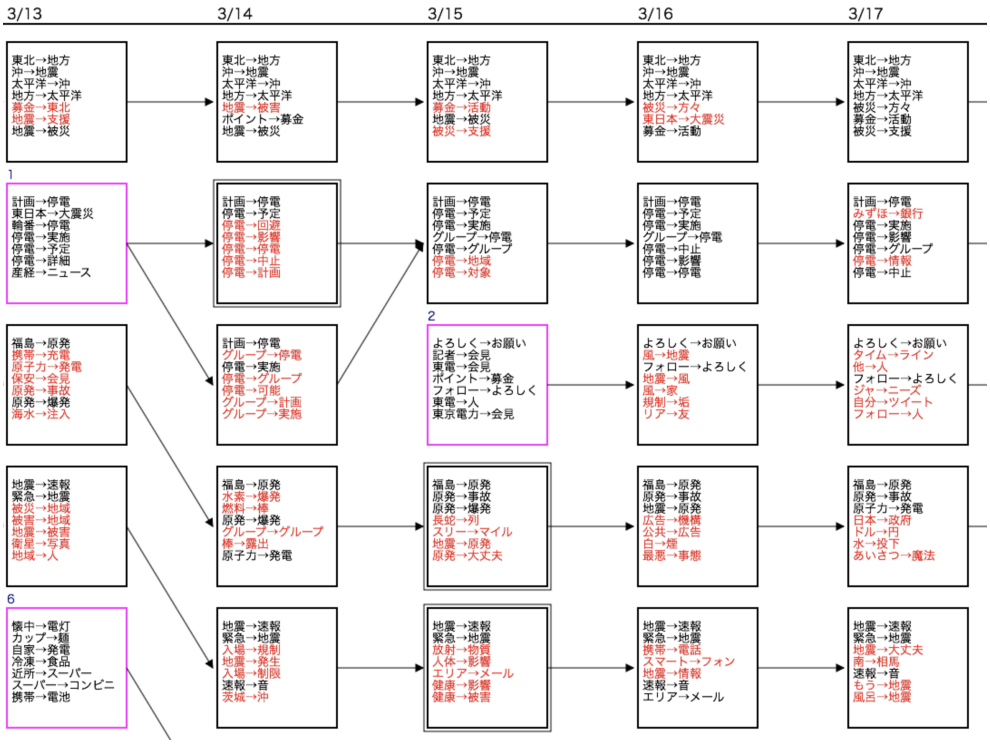

図 1: 本研究のトピック系列の可視化結果 (一部)

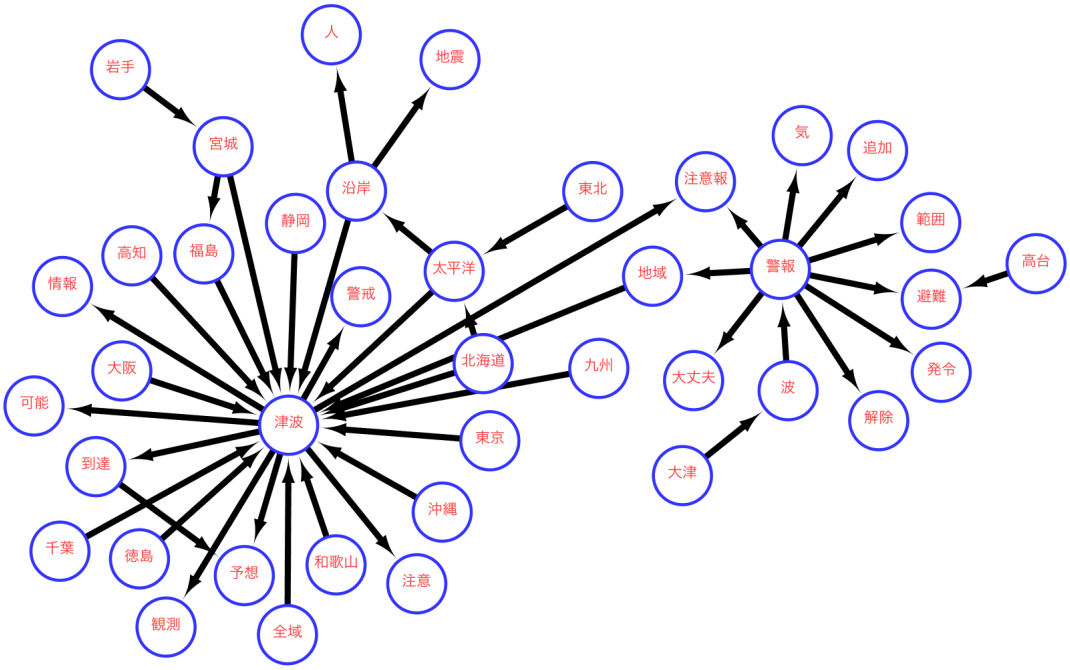

図 2: 単語 2-gram のネットワーク可視化結果 (津波に関するトピック)