2B3-NFC-02a-1

東日本大震災時のツイートのトピック系列の可視化と分析

Analysis and Visualization of Topic Series Using Tweets in Great East Japan Earthquake

北田 剛士 *1 Takeshi KITADA
風間 一洋 *1 Kazuhiro KAZAMA
榊 剛史 *2 Takeshi SAKAKI
鳥海 不二夫 *3 Fujio TORIUMI
栗原 聡 *4 Satoshi KURIHARA
篠田 孝祐 *4 Kosuke SHINODA
野田 五十樹 *5 Itsuki NODA
斉藤 和己 *6 Kazumi SAITO

*1和歌山大学 Wakayama University
*2株式会社ホットリンク Hottolink,Inc.
*3東京大学 The University of Tokyo
*4電気通信大学 The University of Electro-Communications
*5産業技術総合研究所 AIST
*6静岡県立大学 University of Shizuoka

Twitter is an important method to observe real-world movements or events in real-time. However, it is very difficult to understand discussion topics and their transition continuously on Twitter. There are some researches to discover topic transition by using a generative topic model such as latent Dirichlet allocation (LDA) or dynamic topic model (DTM). For very large dataset of Twitter, a adequate number of topics is also big and users cannot understand them well. In this paper, we propose a method to extract topic sequences from tweets by using PLDA and cosine similarity and visualize a part of topic sequences after filtering and ranking dynamically. Futhermore, we apply the method to tweet archive in the Great East Japan Earthquake and show that proposed method is suitable for grasping discussion topics.

1. はじめに

現実社会で起こった出来事を観測する手段の中でも，Twitterは様々なユーザの意見をリアルタイムに知ることができる点で特に重要になりつつある．Twitterでは，情報はフォロー関係によって作られるソーシャルネットワーク上の同類性が高い近傍ユーザの間を伝播することから，ユーザのタイムラインにはそのユーザに適した情報が選別されて表示される．しかし，それゆえにTwitter空間全体で現在どのような議論話題（discussion topic）が話し合われているかを的確に把握し続けることは難しく，議論・話題を系統的に整理し，その変化を追跡する技術が望まれている．

例えば，2011年に起きた東日本大震災では，Twitterなどのソーシャルメディアが活用されたことが報告されている[1]．どのようなトピックが話し合われたかについては，高頻度で使用された単語や発生したと思われる事象を表す単語を手掛かりにしても，ある程度は推定することができるが，大規模災害時のような極限状況下においては，想定外の事象が発生している可能性が高い．また，予測可能な地震や原発事故のトピックの場合でも，地震関連なら被害状況や救援活動，原発事故なら事故への対処や原発の是非などの異なるサブトピックへの分岐も起こっているのではないかと予想でき，このようなトピックの分岐・収束が観測できれば，災害対策にも役立てることができると考えられる．

そこで，我々は現実の大規模なTwitterアーカイブにおける議論話題の把握が可能な実用的なツールの実現を目指して，時間区間ごとにLDAを使って抽出したトピックの時系列関係を表すグラフ構造であるトピック系列を可視化し，条件を変えることでさまざまな側面から観察できる手法を提案した[2]．本稿では，さらに実際に東日本大震災前後の約3億6千万ツイートに適用して，得られたトピック系列から東日本大震災時の議論話題の分析を試みる．

連絡先: 風間 一洋 (kazama@ingrid.org)
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. トピック系列の抽出

2.1 トピック系列

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，議論・話題を把握するために，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，文書の単語群の類似性に基づいて分類した文書クラスタをトピックとみなす手法，バーストした単語をトピックとみなす手法があるが，これらの手法で話題の内容を知るためには文書群を要約したり，単語から関連する単語や文書群を抽出しなければならない．本稿では，データに隠された潜在的なトピックを確率的に推定するトピックモデルを用いる．トピックモデルでは，同一の話題に関連している単語集合をトピックとみなすことから，そのままでも内容や変遷を理解しやすい特徴を持つ．

代表的なトピックモデルはBleiらが提案したLDA[3]であるが，そのままでは時系列データを扱えない．例えば，Brei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[4]．ただし，DTMではトピック間の関係は一次元であり，トピックの内容の変化は観測できても，並行しているサブトピックを観測することは難しい．そこで，芹沢らや藤田らと同様に，時間区間ごとにLDAを使ってトピックを抽出し，さらに隣接する時間区間の類似度が高いトピックを結合することで，トピックの時間的変遷を求める[5, 6]．本稿では，このトピックの時系列関係を表すグラフ構造を**トピック系列**と呼ぶ．

以下で今回用いたトピック系列抽出法の概略について述べる[2]．

2.2 議論話題に関するテキストの抽出

Twitter のツイートは多種多様である．ユーザ間の議論だけでなく，個人的な発言や単なる挨拶も存在するし，別のユーザに対する返信であるリプライやフォロワーへの情報転送であるリツイートなども含まる．すなわち，すべてのツイートが議論話題に関係しているわけではなく，LDA のようなトピックを確率的に推定する手法では議論話題と関係のない情報が含まれると抽出性能が極端に低下しやすい．例えば,「おはよう」のような挨拶の交換や「地震だ！」,「バルス」などの現実世界のイベントに対する叫び声などは，文章としての意味は持たない．

そこで，議論話題に関するテキストの抽出を試みる．既存研究においては，Zhao らは，3 単語以下又は 8 回未満しかツイートしていないユーザのツイート，及び 10 回未満の出現頻度又は収集した全ユーザの 70%以上が使用した単語を除外した [7]．また，ツイートの文章が短いために，ユーザの全ツイートをまとめて 1 個の文書として扱ったり [8]，ユーザが固有のトピック比率を持つと仮定した Twitter-LDA 法を用いたりされてきた [7].

本稿では，最初に URL やハッシュタグ，ユーザ名，HTML のタグなどの議論とは関係がない部分を削除する．次に，リツイートのような情報転送は議論ではないと考えて，公式リツイート及び非公式リツイートの元発言部分を取り除く．さらに，議論として成り立つためには，ある程度長い文章でなければならないと考えて，抽出された単語の種類数が閾値より小さい場合は無視する．ここで単語数ではなく種類数を用いるのは，繰り返し表現が多いからである．なお，今回は閾値を 8 とした．

2.3 BoW 表現への変換

LDA は文書を BoW として扱うので，Mecab[9] で日本語形態素解析して得られる単語（形態素）から BoW を作成する．

ただし，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．まず，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した．ただし，標準の IPA 辞書では複合語はうまく処理できない．名詞を連結して自動的に複合語とする手法もあるが，ある名詞を含むさまざまなバリエーションができてしまい，確率的な手法を用いた場合の結果の質が下がることがあるので，新語や流行語，専門用語も正しく解析でき，妥当な複合語が抽出されるように，はてなキーワード [*1] や原子力百科事典 ATOMICA[*2] の用語を辞書に追加して用いた．また，一般的にツイートは口語で書かれるためにうま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
- 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（固有名詞などは含まない）
- 長音符（ー）や「笑」などの 1 文字の単語

2.4 トピックの抽出

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．

∗1 http://d.hatena.ne.jp/keyword/
∗2 http://www.rist.or.jp/atomica/

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．

2.5 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[10] 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 は次のように計算する．

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

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

$$\hat{\beta}_{k,w} = \frac{n_{k,w} + \beta}{n_{k-} + V\beta} \tag{2}$$

$n_{k,w}$ はトピック k における単語 w の出現回数，n_{k-} はトピック k における単語の出現回数の総和，V は全文書における語彙数である．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．

2.6 トピック系列の抽出

隣接する時間区間のトピックの間の連続性を，トピックに含まれる単語の term-score ベクトルのコサイン類似度に基づいて判定する．ただし，トピックとの関係が薄い一般的な単語や使用頻度が低い単語の影響を除去するために，上位 N 件の単語以外のスコアは 0 にする．時間区間 t のトピック n_i^t のベクトル $\boldsymbol{T_i^t}(1 \leq i \leq K)$，時間区間 $t+1$ のトピック n_j^{t+1} のベクトル $\boldsymbol{T_j^{t+1}}(1 \leq j \leq K)$ のコサイン類似度 $csim(\boldsymbol{T_i^t}, \boldsymbol{T_j^{t+1}})$ は次のように計算する．

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

さらに，閾値 S 以上の場合にトピックが連続していると判定し，エッジ $e_{i,j}^{t,t+1}$ を作成する．

最後に，時間区間 t_s から t_e の間のエッジ集合から，トピック系列を以下のように抽出する．

1. $t = t_s$ とする．
2. 各エッジ $e_{i,j}^{t,t+1}$ の始点トピック n_i^t を終点トピックとして含むトピック系列が存在すれば，その $[t, t+1]$ にエッジ $e_{i,j}^{t,t+1}$ を追加する．なければ，新しいトピック系列を生成して追加する．
3. 各エッジ $e_{i,j}^{t,t+1}$ の終点トピック n_j^{t+1} を含むトピック系列が複数あればマージする．
4. $t = t+1$ とし，$t = t_e$ なら終了，そうでなければ 2. に戻る．

2.7 トピック系列の可視化

本稿では，Twitter のツイートのような巨大なデータセットを扱うために，抽出されるトピック数も大きい．例えば，Twitter のトピック抽出の研究では，藤野らは 50，渡邊らは 70，Zhao らは 110 を用いた [6, 11, 7]．さらに，トピックの分岐や統合なども扱うことを考えると，トピック系列が画面に収まりきらないことは容易に想像できる．

そこで，以下のように，抽出されたトピック系列にフィルタリングとランキングを施した後で，ユーザの意図に合ったトピック系列を優先して表示したり，ユーザが見る視点を対話的に変更することで Twitter のトピック空間を探索できる機能を提供する．

フィルタリングでは，観察したいイベントに合わせた対象期間を設定したり，観察したいトピック系列が一般的・日常的か又は短期的なものかに合わせて継続期間を設定したり，また観察したいトピックをキーワードで指定できる．

ランキングでは，議論の盛り上がりを示唆する継続期間や，議論の多様さを表すエッジ数などをスコアとして用いてランキングし，ランキング上位から画面に収まる限り描画する．

トピック系列のランキング上位から上から下に順に描画するが，時系列変化の理解を容易にするために，トピックに関しては先頭のトピックとピークとなるトピックをそれぞれピンク色と二重線の枠で，またトピックの特徴語としては新出語とフィルタリング時に指定した単語をそれぞれピンク色と太字で強調表示した．なお，表示部は，HTML5[*3] の Canvas 要素と JavaScript を用いて，ネットワーク経由の利用が容易になるように実装した．

3. 東日本大震災時のツイートの分析

3.1 データセット

Twitter API[*4] 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．

*3 http://www.html5.jp
*4 http://apiwiki.twitter.com/

3.2 トピック系列の抽出

現実の巨大なツイートデータを処理するために，LDA を並列分散化した PLDA[12] を用いて，6 並列で実行した．トピック抽出のパラメータは，$K = 50$，$\alpha = 50/K$，$\beta = 0.01$，イテレーション回数は 500 回とした．並列化により，トピック抽出に掛かる時間を大幅に短縮できた．

さらに，トピック系列抽出のパラメータは，類似度計算の対象単語は上位 $N = 10$ 件，類似度の閾値 $S = 0.7$ とした．

3.3 可視化結果の分析

東日本大震災による Twitter への影響を調べる一例として，期間を 3 月 7 日から 23 日，抽出トピック数を $K = 50$，トピック系列の継続期間を 7 日〜13 日と指定した場合の可視化例を，図 1 に示す．

この結果から，単語群が類似したトピックが系列化されていると共に，トピックに関連する単語群も変化し，さらにトピックの分岐や収束が確認できたことから，議論話題の時系列的な内容変化を追跡できることが分かる．なお，トピック数を増やすほど分岐や収束が増えるが，これは同一トピック系列内の異なる事象が並行する別の流れとして観測できるので，より大きいトピック数の場合でも観察が容易になる．

また,「地震」または「原発」というキーワードを指定した場合には，トピック系列数は 7，トピック総数は 39 となり，キーワードに該当するトピック系列だけに絞り込めた代わりに，避難，救助，停電，物資の共有のような派生するようなトピックは観測されなかった．

そこで，トピック系列の継続期間を指定すると，図 1 に示したようにトピック系列数は 14，トピック総数は 145 と増加して，該当期間中に長期に渡って議論されたトピック系列が残るようになり，トピック系列 1 のような仕事やイベントなどの地震による影響についての議論，トピック系列 2 のような支援活動についての議論，トピック系列 6 のような輪番停電に関する議論などが観測できるようになった．

つまり，キーワード指定によるフィルタリングではあるイベントに直接関連するトピック系列を詳細に観察でき，継続期間によるフィルタリングでは関連して派生するような並行するイベントのトピック系列を把握でき，目的に応じたフィルタリング条件の設定が重要であることがわかる．

4. おわりに

本稿では，実ツイートデータから PLDA を用いて妥当な時間で抽出したトピックをコサイン類似度に基づいて時系列的に結合した膨大な数のトピック系列を，フィルタリングとランキングの条件を対話的に変更しながら可視化することでトピック空間の探索を支援する手法を提案した．

さらに，実際に本システムで東日本大震災前後のツイートデータを分析することで，Twitter における議論話題の変遷が観測できること，異なる議論話題が異なるトピック系列またはトピック系列内の別の流れとして表されること，キーワードやトピック系列の継続期間の指定などフィルタリング条件を使い分けることで，ミクロレベルからマクロレベルまでの観察ができることを示した．

今後の一つの課題はフィルタリング機能の高度化である．例えば，図 1 のフィルタリング条件は人間が試行錯誤して決定したが，これでは良い組み合わせ条件を発見するまでに時間が掛かるだけではなく，対象データを熟知していないユーザは適切な条件を発見できない．そこで，例えばあるキーワードのバーストを検出して表示期間や継続期間を決定したり，関連キー

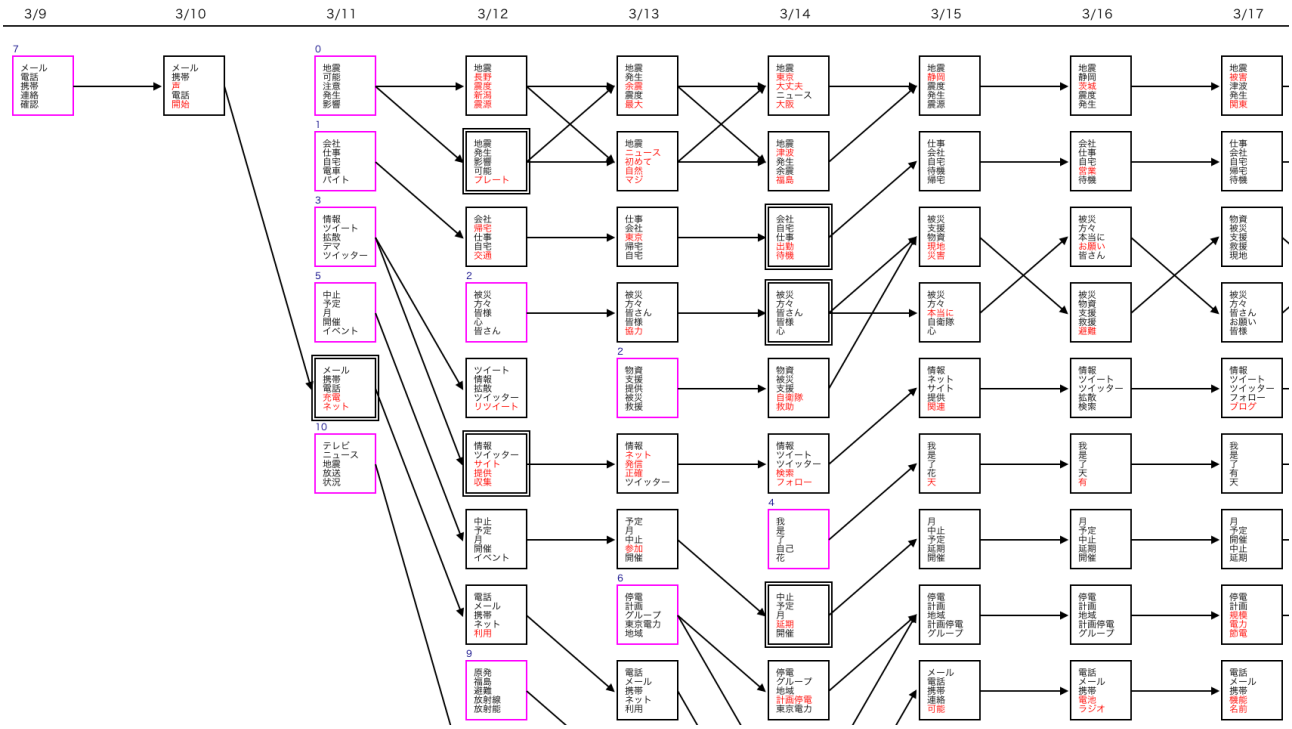

図 1: 東日本大震災直後のトピック系列の可視化の例

ワードの自動的な追加などで，ユーザの情報探索を支援することを検討している．

もう一つの課題は，可視化の改良である．表示されていない下位の単語の確認機能やトピック系列のよりわかりやすい描画法に加えて，単語ではなく具体的な議論内容を知るためのトピック系列中の各トピックに対応したツイートの表示や，議論の盛り上がりの変化を知るためのバースト検出と表示などを予定している．

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 24300064，26330345 の助成を受けた．

参考文献

[1] 総務省. 平成 23 年度情報通信白書. `http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h23/pdf/`, 2011.

[2] 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.

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

[4] David M. Blei and John D. Lafferty. Dynamic topic model. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113–120, 2006.

[5] 芹澤翠, 小林一郎. 文書内のトピック数を考慮したトピック追跡の試み. 第 18 回年次大会発表論文集, pp. 1196–1199. 言語処理学会, 2012.

[6] 藤野巖, 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 2014 C4-2*, 2014.

[7] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jimg He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *Proceeding of the 33rd European conference on Advances in information retrieval (ECIR'11)*, pp. 338–349, 2011.

[8] Liangjie Hong and Brian D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics (SOMA '10)*, pp. 80–88, 2010.

[9] 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.

[10] David M.Blei and John D. Lafferty. *Text Mining: Theory and Applications*, chapter TOPIC MODELS. Taylor and Francis, 2009.

[11] 渡邊恵太, 加藤昇平. ユーザ興味を反映した情報推薦のための潜在的ディリクレ配分法を用いた協調フィルタリング. 信学技報, Vol. 113, No. 429, pp. 15–20, 2014.

[12] 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.