Vol.2017-CH-115 No.1
2017/8/4

新聞記事に対するトピックモデルの適用とトピックの時系列変化に関する考察

山田 太造[1,2,a)]

概要：本論文では 2010 年から 2015 年の 6 年間に発行された新聞記事を対象にトピックモデル LDA(Latent Dirichlet Allocation) を適用し，検出されたトピックの時系列変化について考察する．LDA を用いた，イベントとそれに関連する記事を自動的に収集・提示する方法，および関連するトピックの提示方法についても示す．また，トピックに属する用語の時系列変化によるトピックの時系列変化を示す．さらに，本手法の地域研究への適用可能性や今後の展開について述べる．

キーワード：新聞記事, トピックモデル, LDA, 時系列変化, 地域研究

A Study on Application of Topic Model to Newspaper Articles and Time Series Change of Topic

Taizo YAMADA[1,2,a)]

Abstract: In the paper, we apply LDA (Latent Dirichlet Allocation) as a topic model to newspaper articles issued in 6 years from 2010 to 2015 and consider the time series change of detected topics. We show a method of automatically collecting and presenting events and articles which are related to topics using LDA. We also show the time series change of topics by time series change of terms belonging to the topics. In addition, we describe the applicability of the method to area studies and future works.

Keywords: newspaper, topic mode, LDA, time series change, area studies

1. はじめに

新聞は政治，経済，事件，国際情勢，文化，スポーツなど幅広いジャンルのニュースについてその動向を伝えるとともに，日本中・世界中に関わる事項・事象だけでなく，特定の地域に関わるそれらについて報じる．各種のニュース・イベントなどは twitter や facebook などの SNS，blog，各種ウェブサイトなどにより，ウェブ上で報じられることが今では普遍的であるが，それでも新聞はニュースを伝えるメディアとしては高く信頼できると考えられる．総務省情報通信政策研究所の調査 [1] によると，テレビ，新聞，インターネット，雑誌のうち，メディアの重要度に関する調査では，全年代を通じてテレビが最も高く，次いでインターネットだったが，メディアの信頼度に関する調査では，新聞が最も高かった．

本研究では，グローバルな情報を持ち合わせながらローカルな情報も報じていく新聞を対象に，そこで報じられている内容から話題を自動的に検出し，その話題の時系列変化を分析していく手法を述べる．話題検出ではトピックモデルの 1 つである LDA（Latent Dirichlet Allocation）を用いた．話題は同一ながら時間とともに出現する用語は変化していく．それについても例示する．新聞全体の中での変化だけでなく，特定の地域に絞った場合での変化についても述べる．さらに，LDA による新聞データの分析手法

[1] 東京大学史料編纂所
Historiographical Institute The University of Tokyo
[2] 東京大学地震火山史料連携研究機構
Collaborative Research Organization for Historical Materials on Earthquakes and Volcanoes The University of Tokyo
[a)] t_yamada@hi.u-tokyo.ac.jp

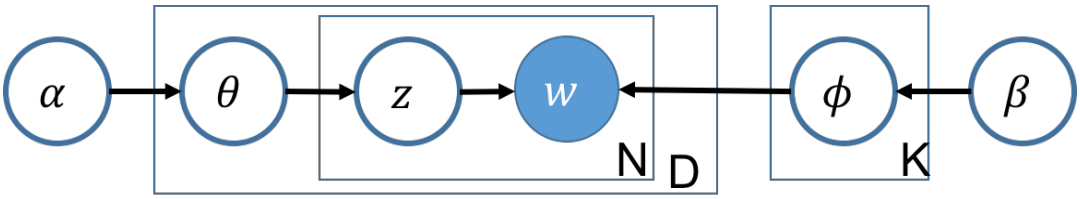


図 1 LDA のグラフィカルモデル
Fig. 1 Graphical model for LDA

の地域研究への適用可能性についても考察する．

2. 新聞データ

本研究では，新聞データとして 2010 年から 2015 年の 6 年間の毎日新聞記事（CD-毎日新聞 2010〜2015 データ集 [*1]）を使用した．このデータは記事を単位として管理されている．記事データは ID 番号，記事見出しキーワード（表記・ヨミ），本文キーワード（表記・ヨミ），掲載面種別コード，写真等の有無，掲載日付・ページ，索引記事番号，記事見出し，朝夕刊区別，記事本文などを項目として持つ．記事の件数は 2010 年分：92,547 件，2011 年分：96,563 件，2012 年分：110,587 件，2013 年分：106,305 件，2014 年分：102,448 件，2015 年分：98,474 件だった．

3. 用語抽出とトピックモデル

新聞データを特徴づけるため記事本文に対して形態素解析を行い，その結果をもとに記事データに対する Bag-Of-Words を作成した．新聞データにはその項目として本文キーワードや記事未第四キーワードがあるが，出現頻度が把握しづらく，また，本文中には出現しないキーワードが付与されている．そのため，記事本文に対する用語抽出を行う．

本研究では，形態素解析器として mecab[*2]，形態素解析器用辞書として IPADic[*3] を用いた．名詞を用語抽出の対象とした．ただし，代名詞，接尾，副詞可能，形容動詞語幹，ナイ形容詞語幹，接続詞的，非自立は対象外にした．連続する名詞，抽出対象の名詞の直後の接尾，[a-zA-Z]+の連続をチャンクした．出力を Bag-of-Words で表現することにより，抽出した用語とその出現頻度を表現することができる．

トピックの検出ではトピックモデルの 1 つである LDA（Latent Dirichlet Allocation）[2] を用いた．LDA は，統計的に共起しやすい用語の集合がいくつか存在しており，これを潜在トピックとして扱う．以降，潜在トピックを単にトピックと呼ぶ．一つの文書に複数のトピックが存在することを仮定しており，そのトピックの分布をモデル化していく．**図 1** はここで用いた LDA のグラフィカルモデル表現を示す．ここで，青色の円は観測変数，白色の円は未知変数を示し，矩形は繰り返しを，矩形の右下の数字はこの矩形で表す繰り返しの回数を示す．w は先に述べた用語抽出の結果，つまり用語を示す．ここでは唯一観測される変数である．z はトピック，θ はトピック分布，ϕ は用語分布を示す．また α および β は θ および ϕ のパラメータ，つまりハイパーパラメータを示す．文書数を D，文書 d の用語数を N_d としたとき，θ_d および ϕ_k は

$$\theta_d \sim Dir(\alpha) \qquad (d = 1, \ldots, M),$$
$$\phi_k \sim Dir(\beta) \qquad (k = 1, \ldots, K). \tag{1}$$

により生成されると仮定する．ここで $Dir(\cdot)$ はディリクレ分布を示す．トピック $z_{d,i}$ は下記のように生成されることにする．

$$z_{d,i} \sim Multi(\theta_d) \qquad (i = 1, \ldots, N_d) \tag{2}$$

ここで $Multi(\cdot)$ は多項分布を示す．さらに用語 $w_{d,i}$ は下記による生成を仮定する．

$$w_{d,i} \sim Multi(\phi_{z_{d,i}}) \qquad (i = 1, \ldots, N_d) \tag{3}$$

LDA のモデル推定では崩壊型ギブスサンプラを用いた解法が知られており [3]，本研究ではこれを用いてトピックを算出する．

4. トピックモデルの適用と考察

LDA におけるトピック数を 200，崩壊型ギブスサンプリングを 2,000 回繰り返すことでモデルの推定，およびトピックの検出を行った．抽出した用語の異なり数は 2,683,289，出現頻度は 286,288,248 だった．

図 2 は，各トピックに割り当てられた用語の出現頻度の月単位での変化と主なイベントやニュースを示す．この結果から，オリンピック，サッカーワールドカップ，東日本大震災，衆議院・参議院選挙のような大きなイベントやインパクトのあるニュースの生起とそのイベントに関連するトピックの出現頻度は関係があると予想できる．例えば，トピック 3 は東日本大震災，津波，被災地，震災，被災者などが割り当てられており，2011 年 3 月および 4 月に非常に

[*1] http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html
[*2] http://taku910.github.io/mecab/
[*3] https://github.com/neologd/mecab-ipadic-neologd

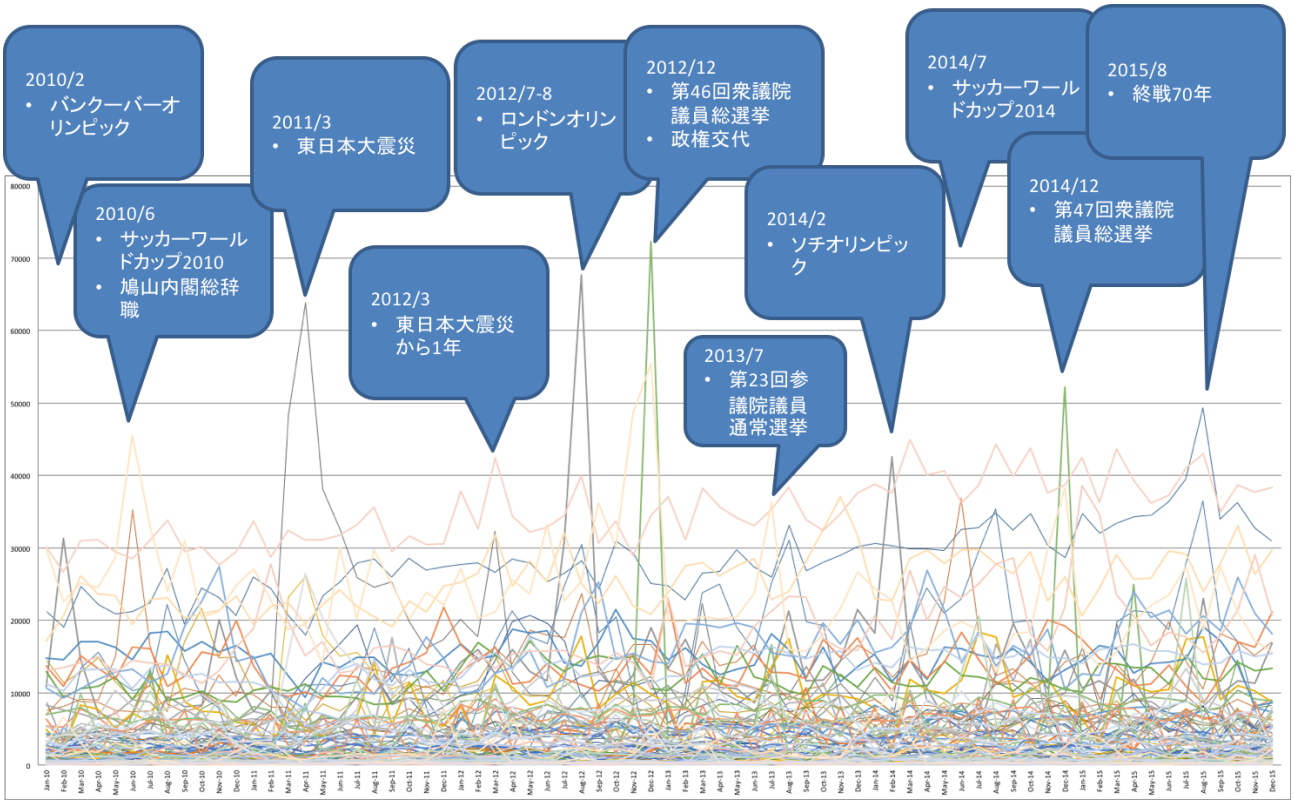
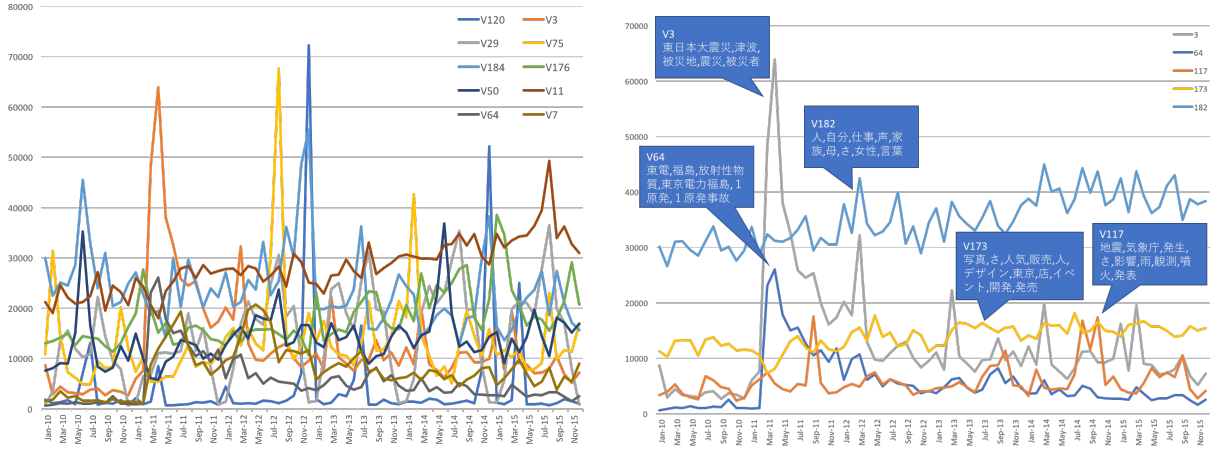


図 2 LDA によるトピックの検出
Fig. 2 Topic detection by LDA

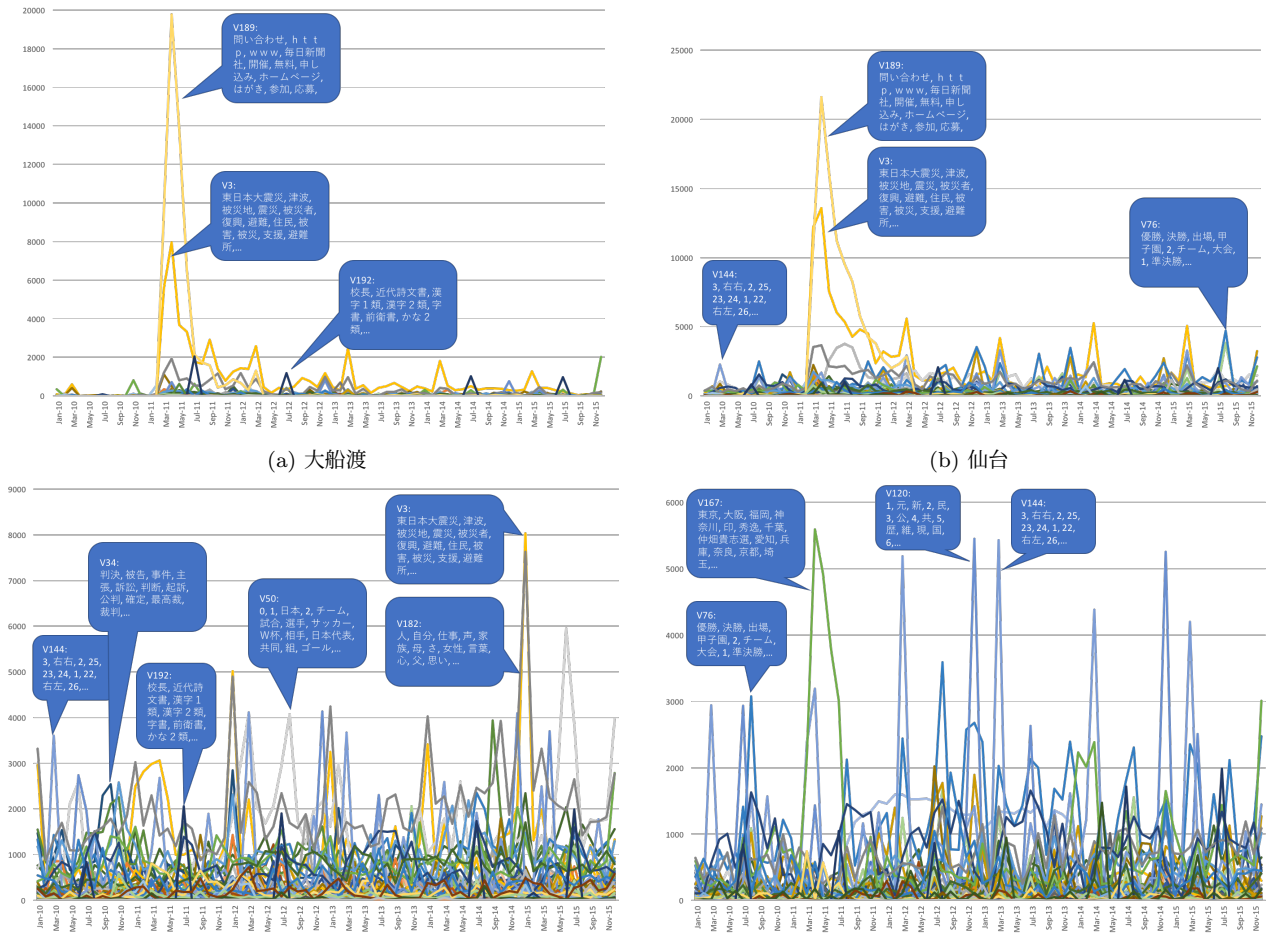


(a) 分散による上位 10 件のトピック　(b) 地震に関するトピック（V3）と類似するトピック

図 3 上位 10 件のトピックと地震に関するトピック
Fig. 3 Top 10 topics by variances and topics concerning earthquakes

高い出現頻度を示していることがわかる．同様に，トピック 120 は国政選挙開票結果，トピック 75 はオリンピックに関係する用語が割り当てられており，そのイベントの生起と高い出現頻度を示すタイミングが合致する．

各トピックにおいて，割り当てられた用語の出現頻度でソートした場合，上位 5 件は，トピック 182（人，自分，仕事，声，家族，母など），トピック 11（日本，人，世界，言葉，時代，戦争，人々など），トピック 166（問題，必要，調査，指摘，国，説明，対応，検討など），トピック 184（首相，民主党，自民党，党，選挙，国民，批判，国会，政府など），トピック 176（ロシア，米国，イラン，シリア，大統領，イスラエル，可能性，死亡など）だった．トピック 182 や 11 は毎日新聞におけるコラムに関係すると考えられ，毎日の新聞に掲載されていることもあり，頻度は高いもののその分散は大きくない．これに対し，**図 3**（a）は分散値でソートしたときの上位 10 件のトピックを示す．順に，トピック 120（国政選挙開票結果; 1，元，新，2，民，3，公，4，共など），トピック 3（地震），トピック 29（プロ野



(a) 大船渡　(b) 仙台

(c) 神戸　(e) 熊本

図 4　地名に関係するトピックの時系列変化

Fig. 4　Time-series changes of topics concerning place names

球; 先発，勝，敗，直球，チーム，球，試合，一回，巨人など），トピック 75（オリンピック），トピック 184（政治動向）だった．トピック 3 は出現頻度順でも 11 位であり分散値で 2 位だった．また，トピック 184 は出現頻度順で 4 位，分散値で 5 位だった．トピックモデル適用により自動的に検出したトピックにおいて，出現頻度および分散値の高いトピックはこの間の日本を象徴するイベント・ニュースだったと言える可能性が高い．

図 3（b）は地震に関係するトピック 3 と類似するトピックを示す．ここで，各トピックに割り当てられた用語の出現頻度（実際には tf-idf により重み付けを行う）をそのトピックの特徴ベクトルとし，次式用いて類似度を算出した．

$$sim(u,v) = cos(u,v) = \frac{\sum_i u_i v_i}{\sqrt{\sum_i u^2} \cdot \sqrt{\sum_i v^2}} \qquad (4)$$

トピック 3 と最も類似したのはトピック 117 であり，地震，気象庁，発生，影響，雨，観測，噴火，発表など地震・噴火等の気象庁発表に関係する．次はトピック 64 であり，東電，福島，放射性物質，東京電力福島，原発，原発事故が割り当てられており，福島第一原子力発電所事故に関係する．式（4）による結果ではあるが，直感的な結果と合致する．またトピック 3 の時系列変化とも相関があるかもしれない波形を示した．しかしながら，それ以外のトピック（182，173）は時系列変化に相関があるとはいえず，また，直感的にも関係するトピックとは思えない．トピックの類似性については，時系列変化の相関性をもとに再考する必要があると考えている．例えば，原発，稼働，政府，東電，関電，電力，福島，電力会社，必要などが割り当てられているトピック 7 は，その出現頻度が示す波形はトピック 64 やトピック 3 に近い．しかしながら式（4）による類似度ではトピック 182 などよりも上位ではなかった．トピック 7 とトピック 3 の関係をうまく表現できる方法があれば，より深化したトピックの類似性を導くことが可能であろう．

図 4 は指定したキーワードが出現する記事のみを対象としたトピックの時系列変化を示す．この結果ではキーワードとして地名を指定した．具体的には，（a）は “大船渡”，（b）は “ 仙台”，（c）は “神戸”，（d）は “熊本” をキーワードとして指定したときの結果を示す．これにより，指定した地名が出現するトピックの時系列変化が把握できる．（a）および（b）は東日本大震災の発生直後に，トピック 3 が高い頻度を示した．またトピック 189 も同じように出現して

いることがわかった．このトピックは，問い合わせ，ｈｔｔｐ，www，毎日新聞社，開催，無料，申し込み，ホームページ，はがき，参加，応募などが割り当てられており，ボランティアの募集に関係すると思われる．(c)，(d) は (a) や (b) とは大きく異る結果を示した．(c) は野球の結果をしめすトピック 144 やサッカー日本代表に関係するトピック 50 などが頻出した．また地震に関するトピック 3 も高頻度だったが (a) や (b) とは異なり，毎年 1 月に高い頻度を示し，2015 年 1 月ではさらに高頻度だった．これは同じ地震に関するトピックではあるが，阪神淡路大震災に関係すると考えられる．2016 年 4 月の地震により熊本も震災被害が大きいが，対象期間外である．熊本地震が発生する前の熊本に関係するトピックとしては，トピック 167 (『万能川柳』)，トピック 76 (高校野球全国大会関係)，トピック 120，トピック 144 などがあった．この結果になったのは，全国版の新聞記事を用いていることが大きいと思われる．地方版の記事を用いる，もしくは補間するなどにより，結果は大きく変わる可能性が高い．

5. 考察

トピック検出およびその時系列変化により話題・関心事の変化の把握が容易になった．またキーワードでのフィルタリングにより特定の記事のみでのそれらの変化についても把握が可能である．記事内の地名を抽出し，それに対して緯度・経度を付与することができれば，より地域でのニュースの変化がわかりやすくなり，また地域間での，またはグローバルとの比較が可能になると思われる．これが可能になれば洗練された地域研究の素材として扱うことが可能になると考えている．

LDA は k-means 等と同様に教師なし学習 (Unsupervised Learning) の手法の 1 つである．本実験で示すように，分類指標のないままデータを分類することが可能であり，データに潜在している本質的な構造・モデルを検出・推定するために利用することができる．他方，機械学習の手法として，ニューラルネットワークや SVM (Support Vector Machine) などの教師あり学習 (Supervised learning) がある．教師あり学習では分類指標が存在する状態で入力データを分類していく手法である．教師なし学習によりデータのモデル化を進め，それに応じて新たなデータを入力すれば分類可能になると考えられる．これは新聞などのニュースにも適用していくことが可能であろうと考えている．これの実現に向けて取り組む予定である．

本研究では新聞データのみを用いたが，他の新聞や SNS などの他のリソースを組み合わせていくことも検討すべきだと考えている．一紙にて世の中のすべての情報を俯瞰することは難しいためである．また，SNS などの個人による発信と新聞等の比較により，個人による発信の重要性が把握可能になれば，地域研究などにおいても重要な研究リソースとして位置づけられるかもしれない．

6. おわりに

本研究では新聞記事を対象に LDA によるトピック検出の手法を示すとともに，2010 年から 2015 年までのトピックの時系列変化を分析するための可視化について述べた．今後は地域の情報をより洗練し進化したモデルを推定していく予定である．

謝辞 本研究の成果の一部は，JSPS 科研費 26730167，26240049，15H01722，16H01897，および「日 ASEAN 協働による超学際生存基盤研究の推進」事業（京都大学東南アジア地域研究研究所）の助成を受けたものによる．CD-毎日新聞 2010〜2015 データ集を使用した．

参考文献

[1] 総務省情報通信政策研究所: 平成 27 年情報通信メディアの利用時間と情報行動に関する調査報告書，入手先 ⟨http://www.soumu.go.jp/iicp/chousakenkyu/seika/houkoku-since2011.html⟩ (参照 2017-06-23).
[2] D.M.Blei, A.Y.Ng, and M.I.Jordan: Latent Dirichlet Allocation, Journal of Machine Learning Research, vol.3, pp.993-1022(2003).
[3] T.L.Griffiths and M.Steyvers: Finding scientific topics, Proc. of the National Academy of Sciences of the United States of America, vol.101, pp.5228-5235(2004).