時系列トピックモデルにおける複数のトピックの時間的依存関係の考慮

佐々木　謙太朗　　吉川　大弘　　古橋　武
名古屋大学大学院工学研究科
sasaki@cmplx.cse.nagoya-u.ac.jp

1　はじめに

近年，Web の発展と共に，ニュース記事やブログ記事，SNS におけるユーザの投稿など，時系列的な文書が大量に生成されるようになった．これに伴い，これら時系列文書中のトピックの時間発展の解析を目的として，これまで様々な時系列トピックモデルが提案されている [1, 3, 4]．トピックモデルとは，bag-of-words 表現された文書の生成過程を確率的にモデル化したものであり，代表的なものに Latent Dirichlet Allocation (LDA) がある [2].

時系列文書におけるトピックは，互いに依存し合いながら時間と共に発展していく．例えば，ニュース記事などにおいて書き手が政治に関する事柄を書く時，それまでの政治的動向だけでなく，経済や社会の動向も考慮する場合が考えられる．また，法律の改正といった政治的動向があった場合，それが経済や社会にどのような影響を与えるかといったことが書かれることもある．このように，時間の経過と共に，あるトピックに別のトピックが結合したり，分離して複数のトピックへと発展したりすることがある．また，次第に話題にされなくなり消滅するトピックもあれば，地震のような突発的な出来事に関するトピックが同時多発的に発生したりすることも考えられる．既存のモデルの多くは，ある時刻におけるトピック k は，その前の時刻におけるトピック k にのみ依存すると仮定している [1, 3, 4]．しかしこの仮定では，各トピックは独立に発展していくことになり，実際のトピックの発生や結合といった発展を十分に捉えることができない．

本稿では，複数のトピックが互いに依存し合いながら，時間と共に発展していくことを仮定した時系列トピックモデルを提案する．さらに本稿では，提案モデルを用いた，時系列文書中におけるトピックの発生/消滅/結合/分離を解析する手法について述べる．実験により，提案モデルが既存のモデルよりも適切にトピックの発展をモデル化できることを示す.

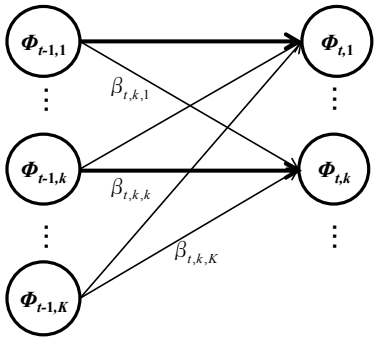

図 1: 提案モデルにおけるトピックの依存関係

2　関連研究

これまでに，時系列文書中のトピックの発展を解析する手法がいくつか提案されている．Dynamic Topic Models (DTM) は代表的な時系列トピックモデルであり，各トピックは独立に発展していくと仮定している [3]．また Multiscale Dynamic Topic Model (MDTM) は，トピックの多重スケール性を考慮したモデルである [4]．これらはいずれも複数のトピックの依存関係を考慮しておらず，トピックの発生/消滅/結合/分離を捉えることはできない．これに対して，トピックの発生，消滅を考慮した *infinite* Dynamic Topic Models (*i*DTM) が提案されている [1]．ただし，*i*DTM もまた各トピックが独立に発展していくと仮定しているため，トピックの結合，分離を捉えることができない．これに対して，提案モデルは複数のトピックの依存関係を考慮しているため，発生/消滅/結合/分離も含めたトピックの発展を捉えることが可能である．

3 提案手法

3.1 提案モデル

本稿では，互いに依存し合うトピックの時間発展を考慮した仮定を LDA に加えたモデルを提案する．初めに，LDA における文書の生成過程について説明する．LDA では，時刻 t における文書 d は，その文書が含む単語の集合 $\boldsymbol{w}_{t,d} = \{w_{t,d,n}\}_{n=1}^{N_{t,d}}$ によって表される．文書は固有のトピック比率 $\theta_{t,d}$ を持ち，この比率に従って文書中の各単語に潜在トピック $z_{t,d,n}$ が割り当てられる．続いて各単語 $w_{t,d,n}$ が，対応するトピックに固有の単語分布 $\phi_{t,z_{t,d,n}}$ に従って生成される．

提案モデルでは，LDA を時間発展を考慮したモデルに拡張するために，単語分布 $\phi_{t,k}$ が，一時刻前のトピックの単語分布 $\{\phi_{t-1,k}\}_{k=1}^{K}$ の重み付き和をハイパーパラメータとする，以下のディリクレ分布から生成されると仮定する．

$$\phi_{t,k} \sim Dirichlet(\sum_{k'} \beta_{t,k,k'} \hat{\phi}_{t-1,k'}) \tag{1}$$

ここで $\beta_{t,k,k'}$ は，時刻 t におけるトピック k の，一時刻前のトピック k' への依存度を表しており，$\beta_{t,k,k'} > 0$ である．これが大きいほどトピック k' への依存度が高いことを示している．また $\hat{\phi}_{t-1,k'}$ は，時刻 $t-1$ におけるトピック k' の単語分布の推定値である．このディリクレ事前分布は，トピックの時間発展を複数の時間スケールでモデル化する Multiscale Dynamic Topic Model (MDTM) [4] における単語分布 $\phi_{t,k}$ の事前分布と類似するが，MDTM は同一のトピックの時間的依存性を考慮しているのに対して，提案モデルは複数のトピック間の時間的依存性を考慮している点で異なる．トピックの依存度 $\beta_{t,k,k'}$ および単語分布の推定値 $\hat{\phi}_{t-1,k'}$ は，確率的 EM アルゴリズムを用いることで逐次推定することができる．

3.2 学習

提案モデルでは，MDTM と同様に，確率的 EM アルゴリズムを用いることによりオンライン学習することが可能である．確率的 EM アルゴリズムでは，ギブスサンプリングによる潜在トピックの推定と，不動点反復法によるトピックの一時刻前への依存度の推定を交互に繰り返す．ギブスサンプリングにおいて，時刻 t におけるトピック z_i は，位置 i 以外の情報を用いて以下の式で更新される．

$$\begin{aligned} &p(z_i = k | \boldsymbol{w}_t, \boldsymbol{z}_{t\backslash i}, \hat{\boldsymbol{\Phi}}_{t-1}, \alpha, \boldsymbol{\beta_t}) \\ &\propto \frac{n_{t,d,k\backslash i} + \alpha}{n_{t,d\backslash i} + D_t \alpha} \frac{n_{t,k,v_i\backslash i} + \sum_{k'} \beta_{t,k,k'} \hat{\phi}_{t-1,k',v_i}}{n_{t,k\backslash i} + \sum_{k'} \beta_{t,k,k'}} \end{aligned} \tag{2}$$

ここで，D_t は時刻 t における文書の数，$n_{t,d,k}$ は時刻 t において文書 d でトピック k に割り当てられた単語の数，$n_{t,k,v}$ は時刻 t において単語 v にトピック k が割り当てられた数を表し，$\sum_k n_{t,d,k} = n_{t,d}$，$\sum_v n_{t,k,v} = n_{t,k}$ である．また，$\backslash i$ は位置 i を除いた時のカウントを表す．

依存度 $\beta_{t,k,k'}$ は，以下の更新式を用いた不動点反復法により推定される．

$$\beta_{t,k,k'}^{new} = \beta_{t,k,k'} \frac{\sum_v \hat{\phi}_{t-1,k',v} B_{t,k',v}}{\Psi(n_{t,k} + \sum_{k'} \beta_{t,k,k'}) - \Psi(\sum_{k'} \beta_{t,k,k'})} \tag{3}$$

$$\begin{aligned} B_{t,k',v} = \ &\Psi(n_{t,k,v} + \sum_{k'} \beta_{t,k,k'} \hat{\phi}_{t-1,k',v}) \\ &- \Psi(\sum_{k'} \beta_{t,k,k'} \hat{\phi}_{t-1,k',v}) \end{aligned} \tag{4}$$

(2) 式による潜在トピックの推定と，(3) 式による依存度の推定を十分回数繰り返した後，トピック k の単語分布は MAP 推定によって以下の式で推定される．

$$\hat{\phi}_{t,k,v} = \frac{n_{t,k,v} + \sum_{k'} \beta_{t,k,k'} \hat{\phi}_{t-1,k',v}}{n_{t,k} + \sum_{k'} \beta_{t,k,k'}} \tag{5}$$

3.3 トピックの時間発展の解析

提案モデルにより各時刻のトピックの依存度および単語分布を推定することで，時間変化に伴うトピックの発生，消滅，結合，分離を解析することができる．地震など突発的な出来事により新しく発生したトピックは，前の時刻のトピックとの関連が薄いと考えられる．したがって，一時刻前のどのトピックとも依存度が低いトピックは，新たに発生したトピックとみなすことができる．同様に消滅については，あるトピックに対する次の時刻のトピックの依存度がすべて低い場合に，結合については，あるトピックが一時刻前の複数のトピックと依存度が高い場合に，分離については，あるトピックに対して次の時刻の複数のトピックの依存度が高い場合にそれぞれ起きたとみなすことができる．

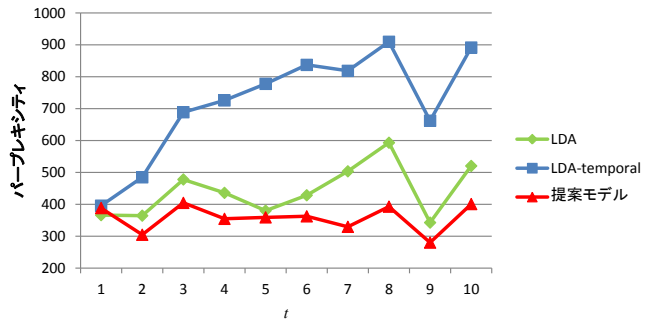

図 2: 各時刻におけるパープレキシティの平均

4 実験

実際のニュース記事を対象として，提案手法の評価実験を行った．本実験では，ニュースサイト「YOMIURI ONLINE（読売新聞）」における 2013 年 12 月 26 日から 2014 年 1 月 4 日までの 669 件のニュース記事を用いた．前処理として，これらニュース記事を形態素解析して名詞だけを抽出し，さらに出現回数が 5 回未満の単語と stop words を取り除いた．

4.1 パープレキシティを用いた評価

パープレキシティを用いて，提案モデルの性能を従来モデルと比較評価した．パープレキシティは，言語モデルの評価によく用いられる指標であり，学習によって得られたモデルが，テストデータ $D_{(test)}$ をどれだけ予測出来るかを表す．

$$perplexity = exp\Big(-\frac{1}{N}\sum_{d}\log p(\mathbf{w}_d)\Big) \qquad (6)$$

ここで，N はテストデータ中の全単語数であり，$\mathbf{w}_d$ は文書 d に含まれる全単語である．パープレキシティが低いほど，モデルの予測性能が高いことを示している．

比較する従来モデルとしては，LDA と LDA-temporal を用いる．LDA-temporal は，提案モデルにおける単語分布 $\phi_{t,k}$ の事前分布を $Dir(\beta_{t,k}\hat{\phi}_{t-1,k})$ に置き換えたモデルである．ここで，$\beta_{t,k}$ は時刻 t におけるトピック k の，一時刻前におけるトピック k への依存度を表しており，LDA-temporal は DTM や MDTM と同様，各トピックが独立に発展していくことを仮定している．各モデルのトピック数は 20 とし，ハイパーパラメータは初期値を 0.1 として確率的 EM アルゴリズムにより推定した．ただし，提案モデルの各時刻におけるトピック k の依存度 $\beta_{t,k,k'}$ の初期値は，$k = k'$ のときは 100，そうでない場合は 0.1 とした．一日を時間の単位とし，各時刻における文書の 90% を学習に用い，残り 10% をテストデータとしてパープレキシティの算出に用いた．これを 10 試行繰り返し，パープレキシティの平均値で評価を行った．

図 2 に，各時刻における各モデルのパープレキシティの平均値を示す．図 2 より，時刻 $t = 2$ 以降で提案モデルの性能が従来モデルに比べて改善していることがわかる．このことから，提案モデルがニュース記事中のトピックの時間発展をより適切にモデル化できているといえる．LDA の性能が提案モデルに劣るのは，時間発展を考慮していないためであると考えられる．また LDA-temporal は，時間発展を考慮しているにも関わらず，LDA よりも性能が悪くなっている．これは，各トピックが独立に発展していくという仮定が，実際のニュース記事におけるトピックの時間発展を十分に考慮できていないためだと考えられる．ニュース記事においては，話題の移り変わりが激しく，ある特定の話題がいつまでも継続して発展することは少ない．したがって，LDA-temporal ではこれらの話題の変化を十分に捉えることができない．これに対して提案モデルは，話題の発生や消滅といった変化を柔軟に捉えることができる．

4.2 トピックの時間発展の解析

提案モデルを用いたトピックの時間発展の解析について評価を行った．提案モデルのトピック数やハイパーパラメータの初期値は 4.1 節と同じものを用いた．

図 3 に，提案手法によって得られたニュース記事の解析例を示す．トピック 6，9 は，12 月 26 日から翌日の 27 日において，靖国神社参拝に関する話題を表すトピックであり，この日，安部首相が靖国神社に参拝したことを報じる記事が多数配信されたことに起因して生成されている．28 日においては，これらの記事の数が減ったことにより，トピック 9 はトピック 6 と結合し，別の話題を表すトピックに変化した（結合）．またトピック 18 は，12 月 30 日において，前日 29 日のどのトピックへの依存度も低くなっており，農薬混入問題に関する話題を表すトピックへと変化している（発生）．これは，12 月 29 日にマルハニチロホールディングスの子会社アクリーズが記者会見を開き，農薬「マラチオン」の検出された商品を自主回収すると発表したことを受け，翌 30 日に大きく報道されたことに起因している．

さらに，これら以外のトピックに関しても，発生，結合，消滅が頻繁に起きているという解析結果が得られ

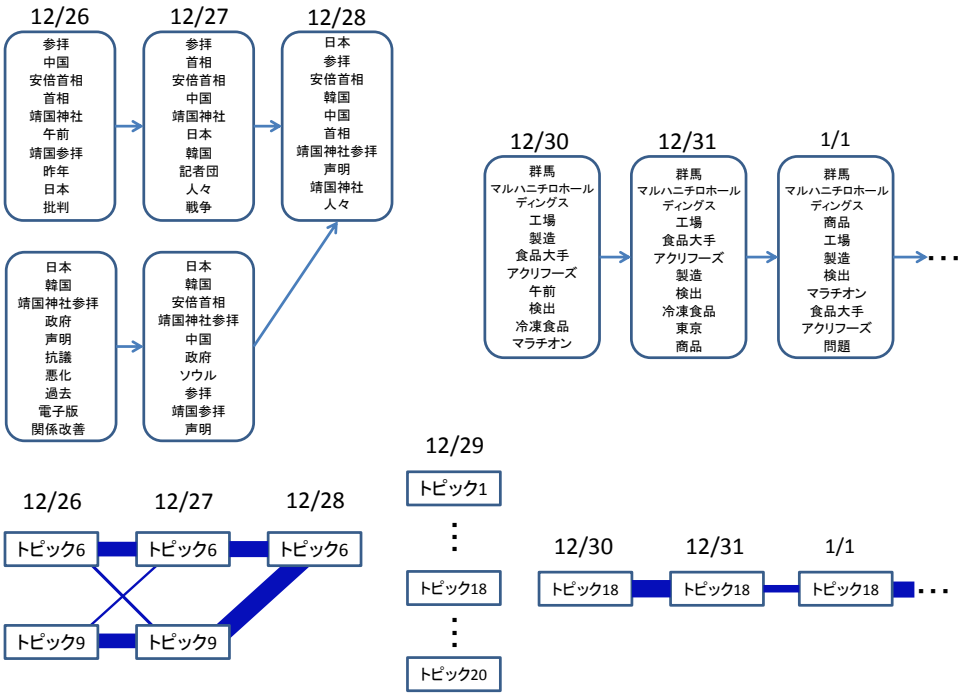

図 3: 提案手法によるトピックの時間発展の解析例．左上は各時刻におけるトピック 6，及びトピック 9，右上はトピック 18 の単語上位 10 語を示している．また，下のグラフは各トピックの依存関係を示しており，エッジが太いほど依存度が高いことを表している．

た．これは，何か大きな出来事があった際に大々的に報じられ，それが徐々に収束していくという，ニュース記事におけるトピックの時間発展を，提案モデルが適切に捉えられていたためだと考えられる．

5　おわりに

本稿では，複数のトピックが互いに依存し合いながら，時間と共に発展していくことを仮定した時系列トピックモデルを提案し，その学習方法を示した．また，提案モデルを用いた，時系列文書におけるトピックの時間発展の解析手法について述べた．実際のニュース記事を用いた実験により，提案モデルが従来のモデルよりも適切にトピックの発展をモデル化できることを示した．さらに，提案手法により，ニュース記事中の話題の発生や，消滅，結合を解析できることを示した．

今後は，Twitter やブログ記事といった，ニュース記事以外の時系列文書に提案手法を適用し，提案手法に対する有効性をさらに検証していく予定である．

6　謝辞

本研究は，文部科学省科学研究費（基盤研究 (C), No.22500088）の補助を得て遂行された．

参考文献

[1] Amr, Ahmed and Eric, P. Xing.: Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream, Proc. of UAI'10, p. 20–29.

[2] Blei, D.M. et al.: Latent dirichlet allocation, Machine Learning Research, Vol. 3, pp. 993-1022, 2003

[3] Blei, D.M. and John D. Lafferty.: Dynamic topic models, Proc. of ICML'06, p. 113-120, 2006

[4] 岩田 具治, et al.: オンライン学習可能な多重スケールでの時間発展を考慮したトピックモデル, 情報論的学習理論テクニカルレポート, 2009