6C-1

時間変化に伴うトピックの発生や消滅も考慮したトピックモデルに関する一検討

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 †

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 [2, 3]．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 [1]．

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 [2, 3]．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，発生/消滅/結合/分離も含めたトピックの時間発展を考慮したトピックモデルを提案する．実験により，提案モデルが互いに依存し合いながら発展していく時系列文書中のトピックを解析可能であることを示す．

2　提案手法

2.1　提案モデル

本稿では，互いに依存しあ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．まず，LDA における文書の生成過程について説明する．

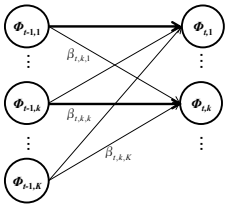

図 1: 提案モデルにおけるトピックの依存関係

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．

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

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) [3] 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，MDTM と同様に確率的 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．

2.2　提案モデルによる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

表 1: 各日における各トピックの単語上位 5 位

	8月1日	8月2日	8月3日
トピック1	バルセロナ，全体，近藤雄二，水泳，世界選手権	バルセロナ，全体，競泳，近藤雄二，水泳	バルセロナ，水泳，世界選手権，競泳，近藤雄二
トピック2	男性，発表，女性，7月，県警	発表，男性，7月，東京，午後	発表，7月，午前，男性，分頃
トピック3	福島，昨年，検査，問題，7月	同容疑者，ロシア，モスクワ，一時，亡命	日本，発表，政府，ワシントン，中国
トピック4	広島，容疑者，7月，大阪，東京	7月，発表，昨年，今年，6月	発表，昨年，7月，東京，今年
トピック5	日本，韓国，米国，政府，7月	日本，政府，中国，7月，提案	日本，政府，午前，7月，岩国

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．

3 実験

3.1 実験データ

実際のニュース記事を対象として，提案モデルを用いたトピックの時間発展の解析を行った．本実験では，ニュースサイト「YOMIURI ONLINE（読売新聞）」における 2013 年 8 月 1 日から 8 月 3 日までの 300 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 word を取り除いた．提案モデルのトピック数は 5 とし，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．

表 1 に，各日の各トピックで推定された単語分布から，それぞれ出現しやすい単語上位 5 つを示す．また図 2 に，学習によって推定されたトピック間の依存度を示す．表 1，図 2 より，トピック 1 はバルセロナで開催された世界水泳選手権に関するものであり，他のトピックとは独立して発展していることがわかる．また，トピック 2 はその日の事件に関するトピック，トピック 4 は地域のニュースに関するトピックだと考えられる．これらの記事は，基本的に毎日一定量書かれているため，この二つのトピックはそれぞれ依存度が高くなっていると考えられる．トピック 3 は 8 月 1 日では地域のトピックであり，翌日はトピック 4 に結合し，代わりに新たなトピックが発生している．この 8 月 2 日におけるトピック 3 は，元米中央情報局職員スノーデン氏のロシア亡命に関するトピックであり，この日大々的に取り上げられていた．さらに翌日の 8 月 3 日では，このトピックの影響力は比較的弱まり (消滅)，政治に関するトピック 5 から分離して生成されていることがわかる．

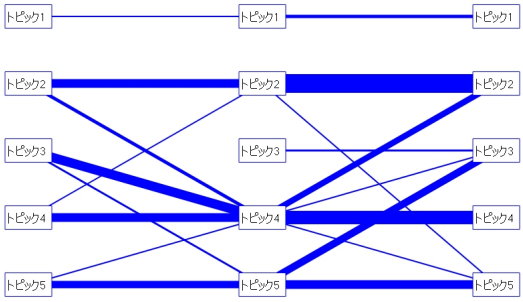

図 2: 提案モデルによって推定された各トピックの時間的な依存関係．エッジが太いほど依存度が高いことを示している．ただし，依存度が 20 以下の場合はエッジを表示していない．

4 おわりに

本稿では，互いに依存しあうトピックの時間発展を解析するためのトピックモデルを提案した．実験により，実際のニュース記事において発生，消滅，結合，分離も含めたトピックの時間発展を追跡可能であることを示した．今後は，他のモデルとの比較を行い，提案モデルの有効性をさらに検証していく予定である．

参考文献

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

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

[3] 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