1I2-4

時系列文書における話題追跡のためのトピックモデルに関する検討

A Study on Topic Model for Topic Tracking in Time Series Documents

佐々木 謙太朗 *1 Sasaki Kentaro　吉川 大弘 *2 Yoshikawa Tomohiro　古橋 武 *2 Furuhashi Takeshi

名古屋大学大学院工学研究科
Graduate School of Engineering Nagoya University

There are some topic models for tracking topic evolution in time series documents such as news or blog articles. In these articles, topic can die, be born, merge, or split at any time. Though several topic models can model some of these evolution of topics, death, birth, merger, and split, none of them can model everything because they assume that the number of topics is fixed at any time or each topic depends on only the previous one. In this paper, we propose the topic model that allows flexible number of topics and considers dependence to multiple topics.

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．これらの文書の内容をすべて把握することは困難であり，いつどのような事が話題になり，それがどのように発展したかを追跡するための研究が数多く報告されている．それらの中でも，時系列トピックモデルに関する研究が近年注目され，また成果を挙げている [Blei 06, Si 13]．時系列トピックモデルは，時間発展を考慮したトピックモデルであり，時間の経過に伴う文書集合中のトピックの発展を追跡することができる手法である．

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．しかし既存のモデルの多くは，ある時刻におけるトピック k が，その前の時刻におけるある特定のトピックにのみ依存すると仮定している [Blei 06, Si 13]．しかしこの仮定では，各トピックが独立に発展していくことになり，実際のトピックの発展を適切に追跡することができないと考えられる．

本稿では，ある時刻におけるトピックが，一時刻前の複数のトピックに依存すると仮定し，かつ各時刻におけるトピックの数が自動で推定される時系列トピックモデルを提案する．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，また実際のニュース記事における話題の追跡が可能であることを示す．

2. 提案手法

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，Dirichlet Process Mixture (DPM)[Antoniak 74] に加えたモデルを提案する．

2.1 Dirichlet Process Mixture

初めに，Dirichlet Process (DP) について説明する．DP は確率分布に対する分布であり，基底分布 G_0 と集中度パラメータ γ によって定義される．離散確率分布 G が DP に従う時，

連絡先: 佐々木謙太朗，名古屋大学工学部工学研究科，名古屋市千種区不老町，052-789-2793，052-789-3166，sasaki@cmplx.cse.nagoya-u.ac.jp

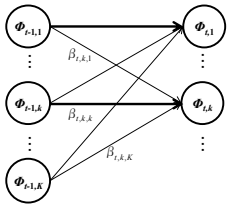

図 1: 提案モデルにおけるトピックの依存関係

$G \sim DP(\gamma, G_0)$ と表記する．集中度パラメータ γ が大きいほど，得られる確率分布 G は基底分布 G_0 に近い離散分布となる．DP の構成法としては，本稿では Chinese Resutaurant Process (CRP) を用いる．CRP を用いると，DPM(DP に基づく混合モデル）におけるデータ生成過程は以下のように表現することができる．

1. $\boldsymbol{z} \sim \mathrm{CRP}(\gamma)$
2. $\phi_k | G \sim G$
3. for $i = 1, ..., n$, $x_i \sim p(x|\phi_{z_i})$

2.2 提案モデル

本稿では，発生や消滅も考慮した話題の追跡を目的として，DPM に時間発展を考慮した仮定を加えたモデルを新たに提案する．まず，DPM を代表的な言語モデルである Dirichlet Mixture (DM) に基づいて拡張することを考える．時刻 t における文書 d を，その文書が含む単語の集合 $\boldsymbol{w}_{t,d} = \{w_{t,d,i}\}_{i=1}^{N_{t,d}}$ によって表す．DM では，各文書にはそれぞれ一つのトピック $z_{t,d}$ が割り当てられ，そのトピックに対応する単語分布 $\phi_{z_{t,d}}$ に従って各単語 $w_{t,d,n}$ が生成される．また，単語分布 $\phi_{z_{t,d}}$ は β をハイパーパラメータとするディリクレ分布に従って生成される．DM における文書の生成過程に基づいて，DPM を言語モデルに拡張すると，文書の生成過程は以下のように表現することができる．

1. $\boldsymbol{z} \sim \mathrm{CRP}(\gamma)$
2. $\phi_{t,k}$ ~Dirichlet(β)
3. for $d = 1, ..., D$, for $i = 1, ..., N_{t,d}$,
   - $w_{t,d,i}$ ~Multinomial($\phi_{z_{t,d}}$)

このモデルでは，トピックの数はデータに応じて自動的に推定される．

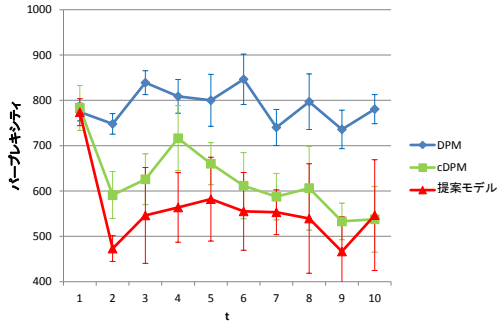

図 2: 各時刻におけるパープレキシティの平均と標準偏差

提案モデルでは，複数のトピック間の依存関係を考慮するために，DPM における単語分布 $\phi_{t,k}$ 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．

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

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．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．

提案モデルと同様に DPM を時系列に拡張したモデルとして，continuous Dirichlet Process Mixuture (cDPM)[Si 13] がある．cDPM では，時刻 t におけるトピックは，一時刻前のある一つのトピックに依存するか，あるいは時間に依存せずに生成されることを仮定している．一時刻前のトピックに依存する場合，その依存度は語彙数を V として $V\beta$ で与えられる．提案モデルは，一時刻前の複数のトピックへの依存を考慮しており，かつそれぞれのトピックへの依存度は学習によって自動的に推定されるという点で異なる．

3. 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．

3.1 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．

$$perplexity = exp\Big(-\frac{1}{N}\sum_{d} \log p(\mathbf{w}_d)\Big) \quad (2)$$

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．

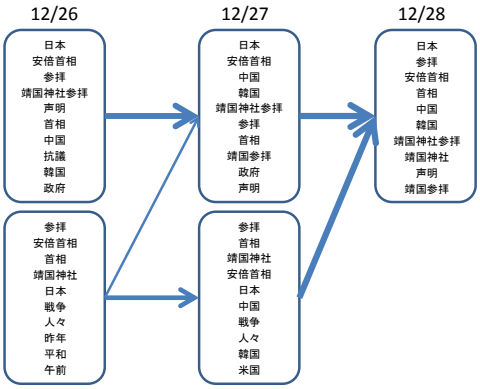

図 3: 靖国神社参拝に関するトピックとその時間発展

比較する従来モデルとしては，DPM と cDPM を用いる．各モデルのパラメータは，[Si 13] を参考に $\gamma = 1$，$\beta = 0.5$ 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と標準偏差を示す．図 2 より，ほとんどの時刻で提案モデルの性能が従来モデルと比べて高いことがわかる．このことから提案モデルにより，複数のトピックへの依存を考慮することで，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．

3.2 トピックの時間発展の解析

図 3 に，提案モデルによって推定された靖国神社参拝に関する話題とその発展を示す．図において，矢印は依存度が 20 以上ある場合に示しており，また太いほど依存度が大きいことを表している．上側のトピックは靖国神社参拝に対する国外の反応，下側は安部首相の行動や考えに関するトピックであると考えられる．12 月 28 日にはこれら二つのトピックが結合し，話題が収束していく様子が捉えられている．実際，28 日には靖国神社参拝に関する記事は少なくなっていた．

4. おわりに

本稿では，時系列文書中の話題を追跡するために，複数のトピックへの依存を考慮した時系列トピックモデルを提案した．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，またニュース記事中の話題の追跡が可能であることを示した．

参考文献

[Antoniak 74] Antoniak.: Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems, The Annals of Statistics, Vol.2, NO.6, 1974

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

[Si 13] Si, J et al.: Exploiting Topic Based Twitter Sentiment for Stock Prediction, ACL'13, 2013.