複数のトピックの時間的依存関係を考慮した時系列混合モデル

Time Series Mixture Model Considering Dependence to Multiple Topics

佐々木 謙太郎

Kentaro Sasaki

名古屋大学工学研究科

Graduate School of Engineering Nagoya University

sasaki@cplx.cse.nagoya-u.ac.jp

吉川 大弘

Tomohiro Yoshikawa

(同 上)

yoshikawa@cse.nagoya-u.ac.jp

古橋 武

Takeshi Furuhashi

(同 上)

furuhashi@cse.nagoya-u.ac.jp

keywords: topic model, trend analysis, time series document

Summary

This paper proposes a mixture model that considers dependence to multiple topics. In time series documents such as news, blog articles, and SNS user posts, topics evolve with depending on one another, and they can die out, be born, merge, or split at any time. The conventional models cannot model the evolution of all of the above aspects because they assume that each topic depends on only one previous topic. In this paper, we propose a new mixture model which assumes that a topic depends on previous multiple topics. This paper shows that the proposed model can capture the topic evolution of death, birth, merger, and split and can model time series documents more adequately than the conventional models.

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るため, 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ことを目的とした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 [Ahmed 10, Blei 06, 岩田 09, Ahmed 08, Si 13, Wei 07]. ここでトピックモデルとは, bag-of-words を仮定した, 文書の生成過程をモデル化した確率的言語モデルである. 代表的なトピックモデルのひとつとして, Latent Dirichlet Allocation (LDA) がある [Blei 03]. LDA では, 各文書がそれぞれ固有のトピック比率を持ち, その比率に従ってトピックが生成され, さらにそれらのトピックに固有の単語分布に従って各単語が生成されることをモデル化している. また時系列トピックモデルとは, 時間発展を考慮したトピックモデルであり, トピックの比率やトピックの内容に相当する単語分布が, 時間に依存すると仮定したものである. 時系列トピックモデルを用いることにより, 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可能

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際, 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. また逆に, 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についての記事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ある. 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トピック k にのみ依存すると仮定している [Ahmed 10, Blei 06, 岩田 09, Ahmed 08]. この仮定の上では, 各トピックは独立に発展していくことになり, 実際のトピックの結合や分離といった発展を捉えることができない.

本論文では, トピックの発生と消滅を捉えることが可能である Temporal Dirichlet Process Mixture (TDPM) を拡張した, Multi-dependent Temporal Dirichlet Process Mixture (MdTDPM), およびその学習方法を提案し, 上

述の問題に対処する．提案モデルは，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と仮定することで，実際のトピックの発展をより適切に捉えることができる．また，提案モデルを用いることで，時系列文書において発生/消滅/結合/分離を含めたトピックの追跡が可能となる．実際のニュース記事を用いた実験により，提案手法が既存の手法と比べて，より適切に時系列文書のモデル化が可能であることを示す．

2. 従 来 モ デ ル

2.1 Dirichlet Process Mixture Model

本節では，代表的なノンパラメトリックベイズモデルであり，提案モデルのベースでもある Dirichlet Process Mixture Model (DPM) について説明する．DPM は，Dirichlet Process (DP) を事前分布に導入した無限混合モデルであり，データの複雑さに応じて自動的にトピック数を推定することが可能である．ここで，DP は確率分布に対する分布であり [Ferguson 73]，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， $G \sim DP(\gamma, G_0)$ 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，主に Stick-Breaking Process [Sethuraman 94] と Chinese Restaurant Process (CRP) [Aldous 85] の二つがあるが，後述する学習方法であるギブスサンプリング [Neal 00] との相性のよさから，本稿では CRP を用いる．CRP において，1 から $i-1$ 番目のデータ $x_{1:i-1}$ が属するトピック $z_{1:i-1}$ が決まっているとき， i 番目のデータ x_i が属するトピック z_i は以下の確率に従う．

$$P(z_i = k | z_{1:i-1}, \gamma) = \frac{m_k}{i + \gamma - 1} \quad (1)$$

$$P(z_i = k_{new} | z_{1:i-1}, \gamma) = \frac{\gamma}{i + \gamma - 1} \quad (2)$$

ここで， m_k はトピック k に属するデータの数であり， k_{new} は $z_{1:i-1}$ の中で一度も出現しない新しいトピックである．CRP は，各データを客，各トピックをテーブルと見立てて，中華料理店において新しく入店した客は，既に m_k 人の客が座っているテーブルに $m_k / (i + \gamma - 1)$ の確率で着席し，誰も座っていないテーブルに $\gamma / (i + \gamma - 1)$ の確率で着席することに例えられる．

CRP を用いると，DPM における M 個のデータ生成過程は， $z = z_{1:M}$ として以下のように表現することができる．

$$(i) z \sim CRP(\gamma)$$

$$(ii) \text{for } k = 1, \dots, \infty, \phi_k | G_0 \sim G_0$$

$$(iii) \text{for } i = 1, \dots, M, x_i \sim p(x | \phi_{z_i})$$

2.2 Temporal Dirichlet Process Mixture

DPM では，すべてのデータは交換可能であると仮定している．しかし，このような仮定は時間と共に変化するような時系列文書や音楽，ビデオデータに適していない．これに対して Ahmed らは，DPM を拡張して，データを一定の期間ごとに分割し，同一期間内ではデータは交換可能であるが，異なる期間ではデータの交換はできないと仮定した Temporal Dirichlet Process Mixture (TDPM) を提案している [Ahmed 08]．これにより，TDPM はトピックの発生・消滅を扱うことのできるモデルとなっている．

TDPM では，CRP の代わりに Recurrent Chinese Restaurant Process (RCRP) を用いる．RCRP は，CRP を拡張してトピックの人気度の時間依存性を考慮できるようにしたもので，時刻 t における i 番目のデータ $x_{t,i}$ が属するトピック $z_{t,i}$ は以下の確率に従う．

$$P(z_{t,i} = k \in I_t \cup I_{t-1} | z_{t-1}, z_{t,1:i-1}, \gamma) = \frac{m_{t-1,k} + m_{t,k}}{M_{t-1} + i + \gamma - 1} \quad (3)$$

$$P(z_{t,i} = k \notin I_t \cup I_{t-1} | z_{t-1}, z_{t,1:i-1}, \gamma) = \frac{\gamma}{M_{t-1} + i + \gamma - 1} \quad (4)$$

ここで， $m_{t,k}$ は時刻 t において， $i-1$ 番目までデータのうち，トピック k に属するものの数， M_t は時刻 t におけるデータの総数， z_t は時刻 t におけるトピックの系列を表す．ただし， $k \notin I_t$ ならば $m_{t,k} = 0$ ， $k \notin I_{t-1}$ ならば $m_{t-1,k} = 0$ とする．また， I_t は時刻 t において少なくとも一つのデータが属するトピックの集合である．RCRP では，データ $x_{t,i}$ がトピック k に属する確率は，現在の時刻におけるトピック k の人気度 $m_{t,k}$ だけでなく，一時刻前の人気度 $m_{t-1,k}$ も影響する．すなわち，一時刻前で人気だったトピックは，現在の時刻においても人気になりやすい．

RCRP を用いると，TDPM における時刻 t でのデータの生成過程は，以下のように表現することができる．

$$(i) z_t \sim RCRP(\gamma, z_{t-1})$$

$$(ii) \text{for } k \in I_t,$$

$$\phi_{t,k} | \phi_{t-1,k} \sim P(\phi_{t,k} | \phi_{t-1,k}) \text{ if } k \in I_{t-1} \text{ or}$$

$$\phi_{t,k} | G_0 \sim G_0(\phi_{t,k}) \text{ if } k \notin I_{t-1}$$

$$(iii) \text{for } i = 1, \dots, N_t,$$

$$x_{t,i} \sim p(x | \phi_{t,z_i})$$

3. 提 案 手 法

本章では，TDPM を拡張し，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いう仮定を加えた新しいモデル MdTDPM と，その学習方法を提案する．また，提案モデルを用いたトピックの追跡について説明する．

図1 言語モデルに拡張した TDPM のグラフィカルモデル

3.1 TDPM の言語モデルへの拡張

初めに、時系列文書をモデル化するために、TDPM を言語モデルに拡張することを考える。時刻 t における i 番目の文書 $d_{t,i}$ (単語数 $N_{t,i}$) を、その文書が含む単語の集合 $w_{t,i} = \{w_{t,i,j}\}_{j=1}^{N_{t,i}}$ によって表す。各文書にはそれぞれ一つのトピック $z_{t,i}$ が割り当てられ、そのトピックに対応する単語分布 $\phi_{t,z_{t,i}}$ に従って各単語 $w_{t,i,j}$ が生成されるとする。岩田らの手法 [岩田 10] の考え方に基き、単語分布 $\phi_{t,z_{t,i}}$ は、 $z_{t,i} = k \in I_{t-1}$ のとき、精度 (分散の逆数) が $\beta_{t,k}$ 、平均が一時刻前の単語分布の推定値 $\hat{\phi}_{t-1,k}$ である以下のディリクレ分布に従って生成されるとする。

$$P(\phi_{t,k \in I_{t-1}} | \hat{\phi}_{t-1,k}, \beta_{t,k}) \propto \prod_v \phi_{t,k,v}^{\beta_{t,k} \hat{\phi}_{t-1,k,v} - 1} \quad (5)$$

ここで、 $\phi_{t,k,v}$ は、 $\phi_{t,k}$ における単語 v の生成確率である。 $\beta_{t,k}$ は、一時刻前のトピックへの依存度と考えることができ、これが大きいほど $\phi_{t-1,k}$ に近い分布が生成されやすくなる。 $z_{t,i} = k \notin I_{t-1}$ のときは、時間に依存せずに発生したトピックであると考え、ハイパーパラメータを β とするディリクレ分布に従って生成されるとする。

$$\phi_{t,k} \sim \text{Dirichlet}(\beta) \quad (6)$$

以上の仮定に基づいて、TDPM を言語モデルに拡張したときのグラフィカルモデルを図 1 に示す。また、時刻 t における文書の生成過程は以下のように表現することができる。

- (i) $z_t \sim \text{RCRP}(\gamma, z_{t-1})$
- (ii) for each topic $k \in I_t$,
 $\phi_{t,k} \sim \text{Dirichlet}(\beta_{t,k} \hat{\phi}_{t-1,k})$ if $k \in I_{t-1}$ or
 $\phi_{t,k} \sim \text{Dirichlet}(\beta)$ if $k \notin I_{t-1}$
- (iii) for each document $i = 1, \dots, M_t$,
for each word $j = 1, \dots, N_{t,i}$,
 $w_{t,i,j} \sim \text{Multinomial}(\phi_{z_{t,i}})$

3.2 MdTDPM

MdTDPM では、複数のトピック間の依存関係を考慮するために、 $k \in I_{t-1}$ のとき、上述の単語分布 $\phi_{t,k}$ 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。

$$\phi_{t,k} \sim \text{Dirichlet}\left(\sum_{k'} \beta_{t,k,k'} \hat{\phi}_{t-1,k'}\right) \quad (7)$$

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、これが大きいほど依存度が高いことを示している。

3.3 MdTDPM の学習

MdTDPM の学習には、確率的 EM アルゴリズムを用いる [岩田 10, 岩田 09, Wallach 06]。トピックの依存度 $\beta_{t,k,k'}$ および単語分布の推定値 $\hat{\phi}_{t,k'}$ 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。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 [Neal 00] と、尤度最大化によるトピックの依存度の推定 [Minka 00] 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 i 番目の文書 $d_{t,i}$ に割り当てたトピック $z_{t,i}$ のサンプリングは、以下の式に従って行う。

$$P(z_{t,i} = k | z_{t \setminus i}, z_{t-1}, \mathbf{W}_t, \hat{\Phi}_{t-1}, \beta_t) \propto P(z_{t,i} = k | z_{t \setminus i}, z_{t-1}) P(w_{t,i} | z_{t,i} = k, \mathbf{W}_{t \setminus i}, \hat{\Phi}_{t-1}, \beta_t) \quad (8)$$

ここで、 K_{t-1} を時刻 $t-1$ におけるトピックの数として、 $\hat{\Phi}_{t-1} = \{\hat{\phi}_{t-1,k}\}_{k=1}^{K_{t-1}}$, $\beta_t = \{\beta_{t,k}\}_{k=1}^{K_{t-1}}$, $\mathbf{W}_t = \{w_{t,i}\}_{i=1}^{M_t}$ である。また、 $\setminus i$ は i 番目の文書を除くことを表している。依存度 β_t は、不動点反復法を用いて完全尤度 $P(\mathbf{W}_t, z_t | \dots)$ を最大化することによって推定する。不動点反復法における $\beta_{t,k,k'}$ の更新式は、

$$\beta_{t,k,k'}^{\text{new}} = \beta_{t,k,k'} \frac{\sum_v \hat{\phi}_{t-1,k',v} A_{t,k,v}}{\Psi(n_{t,k} + \sum_{k'} \beta_{t,k,k'} \hat{\phi}_{t-1,k',v}) - \Psi(\sum_{k'} \beta_{t,k,k'})} \quad (9)$$

となる。ここで、 $A_{t,k,v} = \Psi(n_{t,k,v} + \sum_{k'} \beta_{t,k,k'} \hat{\phi}_{t-1,k',v}) - \Psi(\sum_{k'} \beta_{t,k,k'} \hat{\phi}_{t-1,k',v})$ であり、 $\Psi(x)$ はディガンマ関数 $\Psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$ である。また、 $n_{t,k,v}$ は時刻 t のトピック k が単語 v に割り当てられた回数であり、 $n_{t,k} = \sum_v n_{t,k,v}$ である。

3.4 トピック追跡

本研究では、ある時刻 t においてトピック k が発生することは、単語分布 $\phi_{t,k}$ が時間に依存しない事前分布から生成されたときであると定義する。同様に消滅については、対象のトピックに対して次の時刻 $t+1$ のどのトピックも依存しない、あるいは依存度が閾値以下の場合に、結合については、対象のトピックが一時刻前 $t-1$ に

おける複数のトピックと依存度が閾値以上の場合に、分離については、対象のトピックに対して次の時刻 $t+1$ の複数のトピックの依存度が閾値以上の場合にそれぞれ起きたと定義する。MdDPM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し、閾値を設けることによって、解析者は以上の定義に従ったトピックの発生、消滅、結合、分離を追跡することができる。

4. 関連研究

提案モデルや TDPM と同様に、DPM を時系列に拡張したモデルとして、Evolutionary Net-Cluster Generative Model (Evo-NetClus)[Sun 10] と continuous Dirichlet Process Mixture (cDPM)[Si 13] がある。Evo-NetClus はネットワークの時間発展を解析するためのモデルであり、時系列文書には直接適用できない。一方 cDPM は、Evo-NetClus をベースにして Twitter におけるトピックの時間発展を追跡できるようにしたモデルであり、Twitter 以外のニュース記事等の時系列文書にも適用可能である。cDPM では、時刻 t におけるトピック k の単語分布 $\phi_{t,k}$ は、一時刻前のある一つの単語分布 $\phi_{t-1,k'}$ に依存するか、あるいは時間に依存せずに β をハイパーパラメータとするディリクレ分布から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。すなわち、 $\phi_{t,k}$ は以下のディリクレ分布に従って生成される。

$$\phi_{t,k} \sim \text{Dirichlet}(V\beta\hat{\phi}_{t-1,k'}) \quad (10)$$

図2に、cDPM、TDPM、MdDPMそれぞれにおけるトピックの発展のイメージを示す。cDPMは、トピックの発生/消滅/分離を捉えることができるものの、依存度は一定である。一方TDPMは、トピックの分離を捉えることはできないが、依存度は学習によって自動的に推定される。これらに対して提案モデルであるMdDPMは、一時刻前の複数のトピックへの依存を考慮しているため、トピックの発生/消滅/分離に加えて結合も考慮することができる。また、それぞれのトピックへの依存度はTDPM同様、学習によって自動的に推定される。

表1 データ全体におけるパープレキシティの平均値（括弧内は標準偏差）

データセット	DPM	cDPM	TDPM	MdDPM
読売新聞	2636(30.5)	2194(33.0)	2297.8(51.6)	2001.6(39.0)
毎日新聞	2318.8(29.6)	1805.7(22.7)	1937.4(42.4)	1608.4(17.5)

表2 $\beta_{t,k,k'}$ ($k = k'$) の初期値とパープレキシティの関係（括弧内は標準偏差）

データセット	$\beta_{t,k,k'} = 10$	$\beta_{t,k,k'} = 100$	$\beta_{t,k,k'} = 1000$	$\beta_{t,k,k'} = 10000$
読売新聞	2103.2(42.0)	2001.6(39.0)	1990.2(50.5)	2049.6(46.4)
毎日新聞	2318.8(29.6)	1608.4(17.5)	1646.4(31.9)	1713.1(34.3)

表3 各時刻で推定されるトピック数の平均値（括弧内は標準偏差）

データセット	DPM	cDPM	TDPM	MdDPM
読売新聞	8.8(1.8)	22.2(4.6)	12.3(2.5)	12.3(1.9)
毎日新聞	11.3(2.5)	31.7(7.9)	18.4(4.0)	15.5(2.3)

5. 実験

実際のニュース記事を対象として、提案モデルに対する評価実験を行った。本実験では、読売新聞と毎日新聞の二つのニュース記事データセットを用いた。読売新聞はニュースサイト「YOMIURI ONLINE (読売新聞)」*1における2014年4月25日から2014年6月16日までの4,373件、毎日新聞はニュースサイト「毎日新聞のニュース・情報サイト」*2における2014年7月15日から8月14日までの2,880件のニュース記事を用いた。前処理として、各データセット中の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。前処理の結果、読売新聞データでは語彙の種類数が6,169、総単語数が114,863、毎日新聞では語彙の種類数が7,167、総単語数が155,760となった。

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おいて一般的に用いられる指標であり、学習によって得られたモデルが、テストデータに含まれる単語 W_{test} をどれだけ予測できるかを表す。

$$\text{perplexity} = \exp\left(-\frac{\log P(W_{test})}{N_{test}}\right) \quad (11)$$

*1 <http://www.yomiuri.co.jp/>

*2 <http://mainichi.jp/>

(a) 読売新聞

(b) 毎日新聞

図3 各時刻のパープレキシティの平均値

(a) 台風

(b) 日本政治・外交

(c) イスラム国

図4 MdTDPMによるトピックの追跡の例。各時刻の各トピックにおいて確率の高い上位10語とその確率を示している。また、リンクは各トピックの依存関係を表しており、リンクが太いほど一方のトピックが一時刻前のもう一方のトピックに強く依存していることを示している。

式(10)において、 N_{test} はテストデータに含まれる単語の総数である。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。比較する従来手法としては、DPM、cDPM、TDPMの三つを用いた。ただしDPMはTDPMと同様に、3.1節に示す手順により言語モデルに拡張したものをを用いた。すべてのモデルに共通するハイパーパラメータ γ は1、 β は0.1とした。TDPMにおける依存度 $\beta_{t,k}$ の初期値は100とし、MdTDPMにおける依存度 $\beta_{t,k,k'}$ の初期値は、 $k=k'$ のとき100、それ以外では β とした。時間の単位は1日とし、各時刻における文書中の単語をランダムに9:1に分割し、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。また、全時刻でのパープレキシティの平均値を算出し、データ全体のパープレキシティとした。これを10回試行したときのデータ全体のパープレキシティの平均値を表1に、各時刻のパープレキシティの平均値を図3に示す。提案モデルの計算時間は、1試行あたり平均で2321.5s(毎日新聞データ)、2722.6s(読

売新聞データ)であった。

表1より、いずれのデータセットにおいても、MdTDPMの性能が従来モデルを上回っ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。DPMの性能が他のすべてのモデルと比較して劣るのは、時間発展を考慮していないためであり。また、cDPMがTDPMよりも性能がよいのは、トピックの分離を考慮しているためであると考えられる。一方で、MdTDPMがcDPM、TDPMの両方と比較して性能が改善しているのは、複数のトピックの依存関係を考慮することで、時間変化に伴うトピックの発展を柔軟に追跡できたためだと考えられる。また、図3より、MdTDPMはTDPMと比較して、ほぼすべての時刻においてパープレキシティが低くなっていることがわかる。MdTDPMとTDPMの違いは、トピックが独立に発展するか、互いに依存し合いながら発展するかであるため、このことから複数のトピックへの依存を考慮することが有用であるといえる。また、MdTDPM

において、 $k = k'$ のときの $\beta_{t,k,k'}$ の初期値を変化させた時のパーレキシティの平均値を表 2 に示す．表 2 より、いずれのデータセットにおいても、 $\beta_{t,k,k'}$ が 100 ~ 1000 の間で表 1 と同様の結果となることがわかる．cDPM で設定される依存度 $V\beta(6, 169 \text{ or } 7, 167 \times 0.1)$ もこの間の値となることから、 $\beta_{t,k,k'} = V\beta(k = k')$ と設定することもある有効であると考えられる．

表 3 に、各モデルにおいて各時刻で推定されるトピック数の平均値を示す．表 3 より、推定されるトピック数は、DPM が最も少なく、TDPM と MdTDPM は同程度、cDPM が最も多いことがわかる．DPM が最も少ないのは、一時刻前への依存を考慮しないために、これまでのトピックの流れを無視した推定を行うためであると考えられる．一方 cDPM は、トピックの内容の時間依存性は考慮しているものの、TDPM、MdTDPM における RCRP のような、トピックの人気度の時間依存性を考慮する機構がないために、トピックの分離が起きやすく、トピック数が増大したと考えられる．

図 4 に、MdTDPM によるトピック追跡の例を示す．用いたデータは毎日新聞データであり、上述のパラメータで MdTDPM を適用した．図 4 において、台風、日本の政治・外交、イスラム国のトピックと依存度が 100 以上で繋がるトピックをそれぞれ (a), (b), (c) に示しており、結合・分離の閾値も 100 としている．ただし図 4 において、リンクは依存度が 35 以上のものを表示している．図 4(a) は、平成 26 年台風第 11 号^{*3}に関するトピックである．トピック 1 は台風の予報に関するトピックであり、8 月 9 日には、台風からの雨雲による大雨などといった、台風の影響に関するトピック 2 が出現している．また今回設定した閾値では、8 月 10 日において、トピック 1 が分離し、トピック 2 に結合している．翌日には台風が過ぎ去ったが、トピック 1、トピック 2 は依然として残っており、さらにトピック 2 からは、台風の影響により栃木県で発生した竜巻の話題を含むトピック 10 が分離している．トピック 2 が常にトピック 1 に依存しているのは、一時刻前の台風の予報が実際に起きていたためだと考えられる．同様のトピックは TDPM、cDPM でも推定されたが、予報と影響のトピックはそれぞれ独立に発展しており、cDPM においても台風の影響に関するトピック内でしか分離は捉えられていなかった．図 4(b) は、日本の政治・外交に関するトピックである．8 月 8 日から岸田外務大臣が ASEAN 関連外相会議に出席したことから、それに関するトピック 11 がトピック 9 から分離し、以降 8 月 12 日まで一貫して存在している．一方国内の政治に関するトピック 9 は、依存度が 100 を下回ったことで、8 月 10 日以降は表示されていない．図 4(c) は、イスラム国への空爆に関するトピックであり、トピック 7 は他のトピックからほとんど影響を受けておらず、同一トピック間での依存性が高い独立したトピックとして

推定された．図 4(c) については TDPM でも同様に推定されるが、cDPM ではトピックの依存度が固定されているために、このように独立性の高いトピックを見つけることは困難である．このように、提案モデルにより、分離や結合、類似したトピック間の依存関係なども含めたトピック追跡が可能である．また、一時刻前のすべてのトピックへの依存を考慮しているものの、図 4(c) のように独立したトピックも同様に推定可能である．

6. お わ り に

本論文では、あ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た時系列混合モデルを提案し、その学習方法を示した．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．台風の話題を例に、提案モデルが発生/消滅/結合/分離に加えて、他のトピックへの影響まで考慮したトピック追跡が可能であることを示した．

提案モデルによるトピック追跡の際には、3.4 節で述べたように、解析者が適切な閾値を設定する必要がある．この時どのような閾値が適当であるかの検討は、今後の課題である．また、本論文で示した実験でのデータサイズであれば大きな問題とはならないが、提案モデルでは複数のトピックへの依存度を推定しているために、従来モデルよりも計算時間がかかってしまうという問題点がある．これに対しては、学習の途中で依存度の低いトピックへのリンクを断ち切る方法などを検討していく予定である．

◇ 参 考 文 献 ◇

- [Ahmed 08] Ahmed, A. and Xing, E. P.: Dynamic Non-Parametric Mixture Models and the Recurrent Chinese Restaurant Process: with Applications to Evolutionary Clustering, in *8th SIAM International Conference on Data Mining* (2008)
- [Ahmed 10] Ahmed, A. and Xing, E. P.: Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream, in *26th Conference on Uncertainty in Artificial Intelligence* (2010)
- [Aldous 85] Aldous, D.: Exchangeability and Related Topics, in *École d'été de probabilités de Saint-Flour, XIII*, pp. 1–198 Berlin, Springer (1985)
- [Blei 03] Blei, D. M., Ng, A. Y., and Jordan, M. I.: Latent dirichlet allocation, *the Journal of machine Learning research*, Vol. 3, pp. 993–1022 (2003)
- [Blei 06] Blei, D. M. and Lafferty, J. D.: Dynamic topic models, in *23rd international conference on Machine learning*, pp. 113–120 ACM (2006)
- [Ferguson 73] Ferguson, T. S.: A Bayesian analysis of some nonparametric problems, *The annals of statistics*, Vol. 1, No. 2, pp. 209–230 (1973)
- [岩田 09] 岩田 具治, 山田 武士, 櫻井 保志, 上田 修功: 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 (2009)
- [岩田 10] 岩田 具治, 渡部 晋治, 山田 武士, 上田 修功: 購買行動解析のためのトピック追跡モデル, 電子情報通信学会論文誌, Vol. 93, No. 6, pp. 978–987 (2010)

*3 [http://ja.wikipedia.org/wiki/平成 26 年台風第 11 号](http://ja.wikipedia.org/wiki/平成_26_年台風第_11_号)

- [Minka 00] Minka, T.: Estimating a Dirichlet distribution, Technical report, MIT (2000)
- [Neal 00] Neal, R. M.: Markov chain sampling methods for Dirichlet process mixture models, *Journal of computational and graphical statistics*, Vol. 9, No. 2, pp. 249–265 (2000)
- [Sethuraman 94] Sethuraman, J.: A constructive definition of Dirichlet priors, *Statistica Sinica*, Vol. 4, pp. 639–650 (1994)
- [Si 13] Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., and Deng, X.: Exploiting Topic based Twitter Sentiment for Stock Prediction, in *51th Annual meeting of the Association for Computational Linguistics*, pp. 24–29 (2013)
- [Sun 10] Sun, Y., Tang, J., Han, J., Gupta, M., and Zhao, B.: Community evolution detection in dynamic heterogeneous information networks, in *8th Workshop on Mining and Learning with Graphs*, pp. 137–146 ACM (2010)
- [Wallach 06] Wallach, H. M.: Topic modeling: beyond bag-of-words, in *23rd international conference on Machine learning*, pp. 977–984 ACM (2006)
- [Wei 07] Wei, X., Sun, J., and Wang, X.: Dynamic Mixture Models for Multiple Time-Series, in *20th International Joint Conference on Artificial Intelligence*, pp. 2909–2914 (2007)

〔担当委員：福井 健一〕

2014 年 10 月 2 日 受理

著 者 紹 介

佐々木 謙太郎(学生会員)

2013 年 3 月名古屋大学工学部電気電子・情報工学科卒業。同年 4 月，同大学大学院工学研究科博士課程前期課程計算理工学専攻に入学。現在に至る。主として自然言語処理に関する研究に従事。

吉川 大弘(正会員)

1997 年名古屋大学大学院博士課程修了。同年カリフォルニア大学バークレー校ソフトコンピューティング研究所客員研究員。1998 年三重大学工学部助手。2005 年名古屋大学大学院工学研究科 COE 特任准教授。2006 年 10 月同研究科准教授。現在に至る。主としてソフトコンピューティングとその応用に関する研究に従事。博士(工学)。IEEE，情報処理学会，電子情報通信学会，日本知能情報ファジィ学会，進化計算学会各会員。

古橋 武(正会員)

昭和 60 年名古屋大学大学院工学研究科博士後期課程電気系専攻修了。工学博士。平成 16 年名古屋大学大学院工学研究科計算理工学専攻教授。現在に至る。ソフトコンピューティング，感性工学に関する研究に従事。平成 8 年日本ファジィ学会論文賞受賞。IEEE，日本知能情報ファジィ学会，電気学会等の会員。