ガウス過程を用いた Dynamic Topic Model による時系列分析

緒方学人　　伊東拓真　　阿部秀尚
文教大学情報学部
{b7p31078, b7p31013,hidenao} @ bunkyo.ac.jp

1 はじめに

近年，SNS，Web の普及により，ニュース記事，Twitter のツイート，ブログ記事などの投稿が大量に行われている．これらに伴い，大量の文書集合から，内容を把握することは難しい問題であった．そのため，文章集合の内容を把握できる話題の抽出，時系列的に話題がどのような変化するかを目的とされた **Latent Dirichlet Allocation(LDA)[1]**をはじめとする数々のトピックモデルが多く研究されてきた．LDA とは，各トピックの単語分布と各文書に割り当てられたトピック比率が生成され，これらに従って，サイコロのようにトピックを無作為で選択され，また同様に単語を無作為に選択されるようなモデルである．また時系列トピックモデル とは，トピックの単語分布や文書に割り当てられるトピック比率が時間に依存するモデルである．このようなモデルとしたさまざまな時系列トピックモデルが提案されている．代表的なものとして，話題の隆盛に着目した **Dynamic Topic Model(DTM)[2]**などが挙げられる．

DTM のモデリング要件として，「対象とする現象にどのような性質があるのか」，「データからどのような情報を抽出したいのか」，「時間的に近接する文書集合は類似したトピックを共有する」などが挙げられ，この要件からトピックの単語分布が時系列での変化するモデルが仮定している．**[2]**より，時系列での単語の変化が示されている．

本論文では，ニュース記事を対象にあるトピックが別のトピックに影響している事例があると考えられる．例えば，2020 年アメリカ大統領選挙といった政治的な事例に対して，次の時刻では，株価への影響が実際に起きた経済的な事例があり，また，ある時刻ではスポーツといった話題と政治的な話題が時刻間での関係ないことがあるなど，複数のトピック間に類似性がある場合やない場合がある．既存のトピックモデルではトピックの単語分布のみ時系列変化に依存しているため，複数の類似したトピックと類似していなトピックの時系列変化を捉えることができない．これらに対して，DTM と文書間とトピック間の類似性を測ることができる **Gaussian Process Topic Model(GPTM)[3]** を組み合わせた Gaussian Process-Dynamic Topic Model(GP-DTM)および推論方法を提案する．GP-DTM により，上述の問題への対処を示す．

既存のトピックモデルの各トピックの単語分布は独立しているため，上述の通り複数のトピックは何かしらの時系列的な依存関係があることを仮定し，提案モデルによる推論を行い，関係性を示す．提案モデルによる推論を行うことで，大量の文書集合が生成している複雑な情報社会に対して，時系列的に類似している話題を抽出し，ユーザの理解を目的とする．

2 関連研究

本論文で関連して，本論文と似たような問題設定を行い，Temporal Dirichlet Process Mixture の拡張版，Multi-dependent Temporal Dirichlet Process Mixture(MdTDPM)による複数のトピック間が依存し合いながら時間発展を仮定するモデルが提案されている**[4]**．実験データとして，「YOMIURI ONLINE(読売新聞)」にて 2014 年 4 月 25 日から 2014 年 6 月 16 日までの 4373 件,「毎日新聞のニュース・情報サイト」における 2014 年 7 月 15 日から 8 月 14 日までの 2880 件の 2 つのニュース記事を用いて，複数のトピックへの依存の有用性を示している．

[5]では，ニュース記事からトピック抽出し，時系列で表示し，対象とするトピックと他のトピックの類似性を算出しており，時系列変化への相関性などを考察している．

また，Google Trends のキーワード入力により，人気度の動向や関連トピック，関連キーワード，地域によっての興味度の視覚化などが挙げられる．

3 モデルの概要

3.1 潜在的ディリクレ配分法(LDA)

LDA は，トピック数 k を固定したとき，単語の種類数Vとおくとする．単語のトピック分布は$\phi_k = (\phi_{k,1}, \dots, \phi_{k,V})$と呼ばれる変数に対応し，文書数 M,文書$d$の総単語総数を N とおき．LDA では各文書$d$は複数のトピックから構成され，その構成比率は$\theta_d = (\theta_{d,1}, \dots, \theta_{d,K})$と呼ばれる変数に対応されている．$\phi_k, \theta_d$は確率ベクトルであるため，離散分布であるディリクレ分布によって生成を仮定し，以下のように示される．

$$\begin{cases} \theta_d \sim Dirichlt(\alpha) & (d = 1, \dots, M) \\ \phi_k \sim Dirichlt(\beta) & (k = 1, \dots, K) \end{cases} \tag{1}$$

α, βは K 次元，V 次元ベクトルのパラメータである．また，$z_{d,i}$ は各文書dの n 番目の単語に対する潜在トピックであり，各単語に割り当てられる．そして$w_{d,i}$ は各トピックの単語分布ϕ_kのうちどれから生成されたかを示す変数である．$z_{d,i}$, $w_{d,i}$ は離散値であるため，生成分布として，カテゴリー分布を仮定し，以下のように示される．

$$z_{d,i} \sim Categorical(\theta_d) \tag{2}$$

$$w_{d,i} \sim Categorical\left(\phi_{z_{d,i}}\right) \tag{3}$$

3.2 ガウス過程状態空間モデル

まずガウス過程状態空間モデルを説明する前に，状態空間モデルを説明する．状態空間モデルとは，時系列モデルの一種で観測データと観測データの裏側に存在する潜在状態の変動を推定するモデルであり，DTM や本論文の提案モデルである GP-DTM のモデリングの際に扱うモデルである．式で表すと，下記のように示される．

$$x_t = f(x_{t-1}) + \varepsilon_f \, , y_t = g(x_t) + \varepsilon_g \tag{4}$$

式(4)に着目すると，ここでtは時間を表し，x_{t-1}, x_tには依存関係があると仮定したモデルであり，左の方程式を状態方程式と呼び，状態遷移関数$f(x)$と独立した過程誤差ε_fで成り立つ．また，右の方程式を観測方程式と呼び，観測関数$g(x)$と独立した観測誤差ε_gで成り立っている．過程誤差と観測誤差は正規分布から生成され，$\varepsilon_{\cdot} \sim N(0\,, \sigma_{\cdot}^2 I)$と定式化でき，

$$\begin{cases} x_t \sim N\left(f(x_{t-1})\,, \sigma_f^2 I\right) \\ y_t \sim N\left(g(x_t)\,, \sigma_g^2 I\ \right) \end{cases} \tag{5}$$

と表すことができ，式(3)と同じように考えることはできるが，$x_1 \sim N(f(x_0)\,, \sigma_f^2 I)$と一緒に考える必要がある．式(5)は DTM，GP-DTM のモデル化するのに扱うので，今後式(5)で説明を行う．

次にガウス過程とは，任意の自然数 N に対して，入力$x_n = (x_1, \dots, x_N)$に対して，出力$f(x_n) = (f(x_1), \dots, f(x_N))$が平均$\mu$，分散共分散行列$K = k(x_n, x_{n'})$をもつ多変量正規分布$N(\mu, K)$に従うとする．つまり，

$$f(x) \sim GP(\mu, K) \tag{6}$$

と表すことができる．ここで，分散共分散行列$K = k(x_n, x_{n'})$はカーネル行列と呼び，x 座標とする入力x'が与えられたら，y 座標とするx'の付近にあるデータ点と近い値をとる行列である．

そしてガウス過程状態空間モデルでは，状態空間モデルの状態遷移関数fおよび観測関数gをガウス過程によりモデル化を行い，下記のように示される．

$$f(x) \sim GP\left(\mu\,, K_f\right) \tag{7}$$

$$g(x) \sim GP\left(\mu, K_g\ \right) \tag{8}$$

その他は，通常の状態空間モデルの式(5)と同じように定式化される．ガウス過程状態空間モデルはガウス過程により，非線形な時間発展を考えることができるノンパラメトリックのモデルであり，表現力が高い．また複数の時系列に対して，類似性を測ることが可能である．

3.3 Dynamic Topic Model(DTM)

Dynamic Topic Model(DTM)とは，LDA と状態空間モデルを組み合わせたモデルであり，時刻$t-1$とtの依存関係があるとことを考慮し，時刻tにおけるトピックの単語分布$\phi_{t,k}$の時系列変化をモデリングする手法である．定式化を下記のように示される．

$$\theta_{t,d} \sim Dirichlt(\alpha_t) \tag{9}$$

文書のトピック比率$\theta_{t,d}$はLDAと同様にディリクレ分布より生成され，トピックの単語分布$\phi_{t,k}$は時刻$t-1$とtの依存関係を考慮するため，上述の状態空間モデルにより定式化される．

$$x_{t,k} \sim N\left(x_{t-1,k}, \sigma^2 I\right) \tag{10}$$

式(9)はV次元の実ベクトル$x_{t,k}$を正規分布より生成され，正規分布から生成分布と仮定しているため，要素和が1とは限らず，負の値を取ってしまう．そのため，式(9)にSoft-max関数を利用して変換する．

$$\phi_{t,k,v} = \frac{\exp\left(x_{t,k,v}\right)}{\sum_{v'=1}^{V} \exp\left(x_{t,k,v'}\right)} \tag{11}$$

そして，潜在トピック$z_{t,d,i}$と各トピックの単語分布ϕ_kのうちどれから生成された単語集合$w_{t,d,i}$はLDAと同様に，

$$z_{t,d,i} \sim Categorical\left(\theta_{t,d}\right) \tag{12}$$

$$w_{t,d,i} \sim Categorical\left(\phi_{t,z_{d,i}}\right) \tag{13}$$

3.4 Gaussian Process Dynamic Topic Model

DTMはトピックの単語分布のみ時系列変化に依存しているため，複数の類似したトピックと類似していないトピックの時系列変化を捉えることができない．また時系列変化の類似性を測ることはできないため，複数の文書間，トピック間の類似関係を考慮した時系列トピックモデルを考える．そのため，時系列の類似性を考慮し，表現力が高いガウス過程状態空間モデルとDTMを組み合わせたモデルを提案する．具体的には，LDAの各文書のトピック比率はディリクレ分布によって生成されているため，文書間の相関性や類似性を把握できない．それに対して，トピック比率をガウス過程から生成し，カーネルにより文書間の類似性を把握できるように仮定する．また文書も時系列的に依存関係があることも考慮しながら，状態空間モデルを定式化するが，状態空間モデルは生成分布を正規分布と仮定しているため，本論文では多変量正規分布を仮定し，分散共分散行列により相関性を取り入れることで文書間の相関性を持つことを仮定する．そしてガウス過程からトピック比率を生成していると仮定しているため，3.3節と同様にSoft-max関数を利用して変換する必要がある．

$$g(\eta) \sim GP(\mu, \mathcal{K}_t) \tag{14}$$

$$\eta_{t,d} \sim N_k\left(g\left(\eta_{t-1,d}\right), \Sigma\right) \tag{15}$$

$$\theta_{t,d,k} = \frac{\exp\left(\eta_{t,d,k}\right)}{\sum_{k'=1}^{K} \exp\left(\eta_{t,d,k'}\right)} \tag{16}$$

次にトピックの単語分布もトピック比率と同様に考え，

$$f(x) \sim GP(\mu, \mathcal{K}_t') \tag{17}$$

$$x_{t,k} \sim N_k\left(f\left(x_{t-1,k}\right), \Sigma\right) \tag{18}$$

$$\phi_{t,k,v} = \frac{\exp\left(\eta_{t,k,v}\right)}{\sum_{v'=1}^{V} \exp\left(\eta_{t,k,v'}\right)} \tag{19}$$

と示される．

次に，潜在トピック$z_{t,d,i}$と各トピックの単語分布ϕ_kのうちどれから生成された単語集合$w_{t,d,i}$はDTMと同様に，

$$z_{t,d,i} \sim Categorical\left(\theta_{t,d}\right),\ w_{t,d,i} \sim Categorical\left(\phi_{t,z_{d,i}}\right) \tag{20}$$

4 実験結果と評価

4.1 評価方法

提案モデルの評価として，今回扱うデータであるgooニュースの2020/12/30~2020/1/6の「社会」，「ビジネス」，「政治」，「国際・科学」から各日付，各ジャンルから5件のニュース(合計 160 件)を抽出したBag-of-words 形式のニュースデータである．ニュースデータから8:2に学習用データと評価データに分割し，1~6のトピック数を用いて，LDA，GP-DTMの二つのモデルを6回繰り返し学習し，Perplexityの平均を求め，比較評価を行った．

Perplexityとは，トピックモデルの予測精度を測る評価指標としてよく扱われる．Perplexityは値が低いほど，モデルの予測性能が高いことを示している．下記にPerplexityの式を示す．

$$perplextiy = \exp\left(-\frac{1}{N}\sum_{d}\log p(w_d)\right) \tag{21}$$

モデル	トピック数1	トピック数2	トピック数3	トピック数4	トピック数5	トピック数6
LDA	325.463	340.955	**113.367**	334.89	937.381	748.207
GPDTM	1131.469	**294.775**	1047.251	152003.482	3368.04	100025.145

表 1　Perplexity によるモデル比較

式(21)では，N はテストデータ中の総単語数のことを指し，w_dは文書 d に含まれる単語である．

4.2 実験結果

Perplexity の最小値としては，トピック数 2 の提案モデルよりトピック数 3 の LDA が優れていることがわかった．また，時系列の順序にトピックを抽出し，時刻tと時刻$t-1$の互いの類似度を測った．本論文では，2020/12/30~2021/1/6 を対象にコサイン類似度を測った．話題の類似性の平均および時系列のトピックの結果 1~10 位の確率が高いものを図 2 に示す．トピックを時系列順に考察すると，「社会関連」，「社会関連」,「国際関連」,「社会関連」,「政治関連」，「政治関連」，「コロナ関連」，「コロナ関連」のような話題を取り上げられていると解釈できると考える．

したがって，図 2 の時系列トピックの類似関係を考察して，類似関係があるものを図 3 に示す．
上記の図の関係性はコロナ関連で成り立っている．そのことから，社会的な話題から国際的な話題，政治的な話題，コロナ関連の話題の共通として類似しているものはコロナ関連であることが解釈できる．

5 まとめ

本論文では，時系列での各トピックの類似性があることを仮定し，時系列トピックモデルを提案し，そのモデルの推論方法と実際のニュース記事のデータを用いて，実験を行なって，考察した．結果として Perplexity を用いて，モデル選択の評価では，LDA の方が提案モデルより優れていたが，仮定した問題への話題の類似関係を示すことができた．
今後の課題としては，Perplexity の精度を上げ，話題の類似関係を可視化できるようにしたいと考える．Perplexity に対して，提案モデルの近似事後分布の設定やモデル自体の改善，別の推論を検討する必要があると考えられる．

2020/12/30	2020/12/31	2021/1/1	2021/1/2	2021/1/3	2021/1/4	2021/1/5	2021/1/6
最多	それぞれ	繁華街	神奈川	時	日	自粛	自身
十分	確実	英国	大雪	殺害	幹部	こと	風力
拘置所	先	3日	代	野党	菅義偉	指摘	飲食店
制限	警察	ジョージア州	方針	宣言	何	商品	陽性
当局	大学	時代	後	自宅	申請	見通し	ほか
意味	2日	車	そこ	年	円	場合	候補
公表	はず	段階	いずれ	実現	意識	可能性	旅行
時点	最初	目標	幸せ	間	料金	これ	人たち
全国	東京都	判断	午前	核	状況	話	性
通り	病床	2020年	初売り	合意	減少	部	男

表 2　各日付のトピック

2020/12/30	2020/12/31	2021/1/1	2021/1/2	2021/1/3	2021/1/4	2021/1/5	2021/1/6
最多	それぞれ	**繁華街**	**神奈川**	時	日	**自粛**	自身
十分	確実	**英国**	大雪	殺害	幹部	こと	風力
拘置所	先	3日	代	野党	**菅義偉**	**指摘**	**飲食店**
制限	警察	ジョージア州	**方針**	**宣言**	何	商品	陽性
当局	大学	時代	後	自宅	申請	**見通し**	ほか
意味	2日	車	そこ	年	円	**場合**	候補
公表	はず	段階	いずれ	**実現**	意識	可能性	旅行
時点	**最初**	目標	幸せ	間	料金	これ	人たち
全国	**東京都**	判断	午前	**核**	状況	話	性
通り	**病床**	2020年	初売り	合意	減少	部	男

表 3　各日付のトピック類似関係

参考文献

[1]. 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).

[2]. Blei, D. M. and Lafferty, J. D.: Dynamic topic models, Proc. ICML'06, ACM, pp. 113–120 (2006).

[3]. Stefanos Eleftheriadis, Thomas F.W. Nicholson, Marc P. Deisenroth, James Hensman.:Identification of Gaussian Process State Space Models.

[4]. 佐々木 健太郎,吉川　大弘,古橋　武.:複数のトピックの時間依存関係を考慮した時系列トピックモデル(2014).情報処理学会研究報告.Vol.2014-MPS-100 No3

[5].山田大造.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(2017). 情報処理学会研究報告.Vol.2017-CH-115 No1