情報論的学習理論テクニカルレポート 2009
Technical Report on Information-Based Induction Sciences 2009 (IBIS2009)

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

Online Multiscale Dynamic Topic Models

岩田 具治*† 山田 武士* 櫻井 保志* 上田 修功*
Tomoharu Iwata Takeshi Yamada Yasushi Sakurai Naonori Ueda

Abstract: We propose an online topic model for sequentially analyzing the time evolution of topics in document collections. Topics naturally evolve with multiple timescales. For example, some words may be used consistently over one hundred years, while other words emerge and disappear over periods of a few days. Thus, in the proposed model, current topic-specific distributions over words are assumed to be generated based on the multiscale word distributions of the previous epoch. Considering both the long-timescale dependency as well as the short-timescale dependency yields a more robust model. We derive efficient online inference procedures based on a stochastic EM algorithm, in which the model is sequentially updated using newly obtained data; this means that past data are not required to make the inference. We demonstrate the effectiveness of the proposed method in terms of predictive performance and computational efficiency by examining collections of real documents with timestamps.

Keywords: Topic model, Time series analysis, Stochastic EM algorithm

1 はじめに

近年，トピックモデルを用いた学術論文や新聞記事，ブログ等の文書集合の時間発展を解析する技術が注目されている [1, 4, 6, 10, 13, 17, 18, 19]．トピックモデルは，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る．トピックモデルでは，ある文書に含まれる各単語は，文書固有のトピック比率に従ってあるトピックを選択した後，そのトピックに固有の単語出現確率分に従って生成される．代表例として，Probabilistic Latent Semantic Analysis（PLSA）[8] や Latent Dirichlet Allocation（LDA）[5] があり，時間発展解析だけでなく，情報検索 [5] や協調フィルタリング [9]，可視化 [11] など，様々な分野に適用されている．

本稿では，複数の時間スケールでのトピックの発展を解析するためのトピックモデル，多重スケール時間発展トピックモデル (Multiscale Dynamic Topic Model: MDTM)，を提案する．トピックは複数の時間スケールで発展する．このことを新聞記事データにおける「政治トピック」を例にして考える．「憲法」「国会」「総理大臣」など百年以上の長期間に渡って頻出する単語もあれば，国会議員の名前など数十年の期間に渡って出現する単語，審議中の法案名など数日のみしか出現しない単語もある．このトピックの多重スケール性を考慮するため，提案モデルでは，ある時刻のトピック固有の単語分布は，一時刻前に推定された複数の時間スケールでの単語分布を基に生成されると仮定する．これにより，短期間の依存性とともに長期間の依存性もモデルに組み込むことができるため，情報損失を削減でき，モデルの信頼性を高めることができる．

また，確率的 EM アルゴリズムを用いた，提案モデルの効率的なオンライン学習法を提案する．新たに得られたデータを用いてモデルを逐次的に更新することができるため，過去のデータを記憶する必要がなく，記憶容量・計算量を削減することができる．また，提案法では，各トピック，各時刻，各スケール毎に依存性をデータから学習するため，トピックの時間発展を柔軟に追跡することが可能である．

これまでにいくつかの文書集合におけるトピックの時間発展を解析する手法が提案されている [1, 4, 10, 18]

*NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories
†e-mail: iwata@cslab.kecl.ntt.co.jp

表 1: Notation

Symbol	Description
D_t	number of documents at epoch t
$N_{t,d}$	number of words in the dth document at epoch t
W	number of unique words
$w_{t,d,n}$	nth word in the dth document at epoch t, $w_{t,d,n} \in \{1, \cdots, W\}$
Z	number of topics
$z_{t,d,n}$	topic of the nth word in the dth document at epoch t, $z_{t,d,n} \in \{1, \cdots, Z\}$
S	number of scales
$\boldsymbol{\theta}_{t,d}$	multinomial distribution over topics for the dth document at epoch t, $\boldsymbol{\theta}_{t,d} = \{\theta_{t,d,z}\}_{z=1}^{Z}$, $\theta_{t,d,z} \geq 0$, $\sum_z \theta_{t,d,z} = 1$
$\boldsymbol{\phi}_{t,z}$	multinomial distribution over words for the zth topic at epoch t, $\boldsymbol{\phi}_{t,z} = \{\phi_{t,z,w}\}_{w=1}^{W}$, $\phi_{t,z,w} \geq 0$, $\sum_w \phi_{t,z,w} = 1$
$\hat{\boldsymbol{\omega}}_{t,z}^{(s)}$	multinomial distribution over words for the zth with scale s topic at epoch t, $\hat{\boldsymbol{\omega}}_{t,z}^{(s)} = \{\hat{\omega}_{t,z,w}^{(s)}\}_{w=1}^{W}$, $\hat{\omega}_{t,z,w}^{(s)} \geq 0$, $\sum_w \hat{\omega}_{t,z,w}^{(s)} = 1$

が，これらの手法は多重スケール時間発展を考慮するものではない．多重スケール性を考慮するものとして Multiscale Topic Tomography Model (MTTM)[13] が提案されている．しかし，MTTM はポアソン過程を用いたモデルであり，LDA[5] などのトピックモデルで用いられるディリクレ多項分布とは異なるモデル化であり，また，オンライン学習ができないという問題点がある．トピックモデルではなく特異値分解を用いて複数の時系列データを解析する手法も提案されている [14] が，特異値分解では正規分布ノイズを仮定しているため，文書や購買ログなどの離散データの解析には適さない [8].

2 提案法

2.1 モデル

本稿で用いる表記法を表 1 に示す．時刻 t の第 d 文書は，その文書に含まれる単語集合 $\boldsymbol{w}_{t,d} = \{w_{t,d,n}\}_{n=1}^{N_{t,d}}$ によって表現される．ここで時刻 t は離散変数とし，一日や一年など単位時間を任意に設定することができる．

まず，提案モデルの基となる LDA について説明する．LDA では，各文書が固有のトピック比率 $\boldsymbol{\theta}_{t,d}$ を持つとする．単語 $w_{t,d,n}$ は，潜在トピック $z_{t,d,n}$ をトピック比率 $\boldsymbol{\theta}_{t,d}$ に従い選択した後，トピック固有の単語分布 $\boldsymbol{\phi}_{t,z_{t,d,n}}$ に従って生成される．図 1(a) に LDA のグラフィカルモデルを示す．ここで，塗潰し円は観測変数，中抜き円は潜在変数，矢印は依存関係，矩形は繰り返しを表す．トピック比率 $\boldsymbol{\theta}_{t,d}$ と単語分布 $\boldsymbol{\phi}_{t,z}$ はそれぞれ独立同一なディリクレ分布に従うと仮定するため，LDA では時間発展が考慮されていない．

提案モデルでは，多重スケール時間発展を考慮するため，時刻 t におけるトピック z の単語分布 $\boldsymbol{\phi}_{t,z}$ は，時刻 $t-1$ における複数の時間スケールでの単語分布 $\{\hat{\boldsymbol{\omega}}_{t-1,z}^{(s)}\}_{s=1}^{S}$ を基に生成されると仮定する．ここで $\hat{\boldsymbol{\omega}}_{t-1,z}^{(s)}$ は時刻 $t-1$ におけるスケール s でのトピック z の単語分布を表す．具体的には，単語分布 $\boldsymbol{\phi}_{t,z}$ の事前分布として，平均が多重スケール単語分布の重み付き和である以下のディリクレ分布を用いる．

$$\boldsymbol{\phi}_{t,z} \sim \mathrm{Dirichlet}(\sum_{s=0}^{S} \lambda_{t,z,s} \hat{\boldsymbol{\omega}}_{t-1,z}^{(s)}), \tag{1}$$

ここで $\lambda_{t,z,s} > 0$ は重みである．また，ゼロ確率問題を回避するため，スケール $s = 0$ の単語分布として一様分布を仮定する $\hat{\omega}_{t,z,w}^{(s=0)} = W^{-1}$．重みは，新たに得られたデータと過去に学習したモデルを用いて，各時刻，各スケール，各トピック毎に推定する．重みの推定法については 2.2 節で述べる．推定された時刻 $t-1$ の多重スケール単語分布 $\{\hat{\boldsymbol{\omega}}_{t-1,z}^{(s)}\}_{s=1}^{S}$ は時刻 t においてはハイパーパラメータとみなされる．多重スケール単語分布の推定については，2.3 節で説明する．

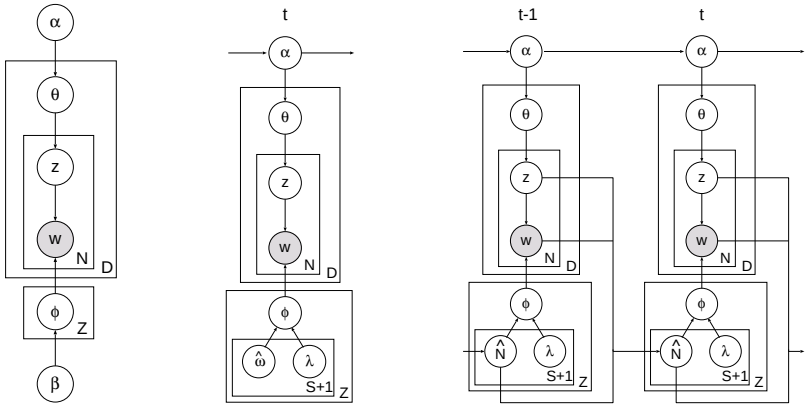
スケールの設定方法はいくつか考えられるが，計算効率のため，また，最適なスケール設定は未知であるため，$\hat{\boldsymbol{\omega}}_{t,z}^{(s)}$ は時刻 $t - 2^{s-1} + 1$ から t までの単語分布を表す，という単純なスケール設定を考える．この設定における多重スケール単語分布イメージを図 2 に示す．単語分布はスケールが長いほど平滑化され，スケールが短いほど尖った形になる．このような様々なスケールの情報を事前分布に利用することにより情報損失を削減でき，モデルをより頑健に学習することができる．

LDA ではトピック比率 $\boldsymbol{\theta}_{t,d}$ はディリクレ分布に従う．一方，提案モデルにおいては，トピック比率の時間発展を考慮するため，ディリクレ分布のパラメータ $\boldsymbol{\alpha}_t = \{\alpha_{t,z}\}_{z=1}^{Z}$ を，一時刻前のディリクレ分布のパラメータに依存させる．具体的には，以下のガンマ事前分布を用いる．

$$\alpha_{t,z} \sim \mathrm{Gamma}(\gamma\alpha_{t-1,z}, \gamma), \tag{2}$$

ここで，平均は $\alpha_{t-1,z}$，分散は $\alpha_{t-1,z}/\gamma$ である．この事前分布を用いることにより，新しいデータを観測しなければ，トピック比率の平均は一時刻前の平均と同じになる．パラメータ γ はトピック比率分布の一貫性を調節するものである．

上記の議論をまとめると，提案モデルでは，時刻 $t-1$ における多重スケールパラメータ $\hat{\boldsymbol{\Omega}}_{t-1} = \{\{\hat{\boldsymbol{\omega}}_{t-1,z}^{(s)}\}_{s=0}^{S}\}_{z=1}^{Z}$，ディリクレ分布のパラメータ $\boldsymbol{\alpha}_{t-1} = \{\alpha_{t,z}\}_{z=1}^{Z}$ が与えられたとき，以下の過程によって，時刻 t における文書集合 $\boldsymbol{W}_t = \{\boldsymbol{w}_{t,d}\}_{d=1}^{D_t}$ が生成されると仮定する．



(a) LDA (b) MDTM (c) online MDTM

図 1: Graphical models of (a) latent Dirichlet allocation, (b) the multiscale dynamic topic model, and (c) its online inference version.

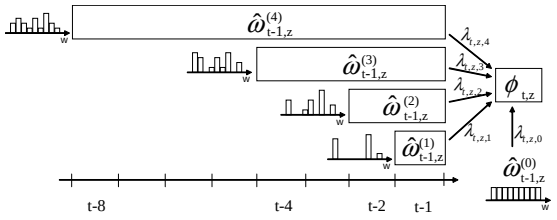


図 2: Illustration of multiscale word distributions at epoch t with $S = 4$. Each histogram shows $\hat{\boldsymbol{\omega}}^{(s)}_{t-1,z}$, which is a multinomial distribution over words with timescale s.

1. For each topic $z = 1, \cdots, Z$:
   (a) Draw word probability $\boldsymbol{\phi}_{t,z} \sim \text{Dirichlet}(\sum_s \lambda_{t,z,s} \hat{\boldsymbol{\omega}}^{(s)}_{t-1,z})$,
   (b) Draw topic proportion prior $\alpha_{t,z} \sim \text{Gamma}(\gamma\alpha_{t-1,z}, \gamma)$,
2. For each document $d = 1, \cdots, D_t$:
   (a) Draw topic proportions $\boldsymbol{\theta}_{t,d} \sim \text{Dirichlet}(\boldsymbol{\alpha}_t)$,
   (b) For each word $n = 1, \cdots, N_{t,d}$:
      i. Draw topic $z_{t,d,n} \sim \text{Multinomial}(\boldsymbol{\theta}_{t,d})$,
      ii. Draw word $w_{t,d,n} \sim \text{Multinomial}(\boldsymbol{\phi}_{t,z_{t,d,n}})$.

図 1(b) に提案モデルのグラフィカルモデルを示す.

2.2 オンライン学習

提案モデルは，確率的 EM アルゴリズム [2] を用いることにより，効率的にオンライン学習することが可能である．時刻 t における文書集合 $\boldsymbol{W}_t$ と一時刻前に推定された多重スケールパラメータ集合 $\hat{\boldsymbol{\Omega}}_{t-1}$ が得られているとする．文書集合および潜在トピック集合の同時分布は下式で表される．

$$P(\boldsymbol{W}_t, \boldsymbol{Z}_t, \boldsymbol{\alpha}_t | \boldsymbol{\alpha}_{t-1}, \gamma, \hat{\boldsymbol{\Omega}}_{t-1}, \boldsymbol{\Lambda}_t) = P(\boldsymbol{Z}_t|\boldsymbol{\alpha}_t) P(\boldsymbol{W}_t|\boldsymbol{Z}_t, \hat{\boldsymbol{\Omega}}_{t-1}, \boldsymbol{\Lambda}_t) P(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, \gamma), \quad (3)$$

ここで $\boldsymbol{Z}_t = \{\{z_{t,d,n}\}_{n=1}^{N_{t,d}}\}_{d=1}^{D_t}$ はトピック集合，$\boldsymbol{\Lambda}_t = \{\{\lambda_{t,z,s}\}_{s=0}^{S}\}_{z=1}^{Z}$ は重み集合を表す．提案モデルでは，単語分布の事前分布として，共役事前分布であるディリクレ分布を用いているため，単語分布パラメータ $\{\boldsymbol{\phi}_{t,z}\}_{z=1}^{Z}$ を積分消去することができる．上式第一項は $P(\boldsymbol{Z}_t|\boldsymbol{\alpha}_t) = \prod_{d=1}^{D_t} \int P(\boldsymbol{z}_{t,d}|\boldsymbol{\theta}_{t,d}) P(\boldsymbol{\theta}_{t,d}|\boldsymbol{\alpha}_t) d\boldsymbol{\theta}_{t,d}$ であり，$\{\boldsymbol{\theta}_{t,d}\}_{d=1}^{D_t}$ を消去することで下式となる．

$$P(\boldsymbol{Z}_t|\boldsymbol{\alpha}_t) = \left(\frac{\Gamma(\sum_z \alpha_{t,z})}{\prod_z \Gamma(\alpha_{t,z})} \right)^D \prod_d \frac{\prod_z \Gamma(N_{t,d,z} + \alpha_{t,z})}{\Gamma(N_{t,d} + \sum_z \alpha_{t,z})}, \quad (4)$$

ここで $\Gamma(\cdot)$ はガンマ関数，$N_{t,d,z}$ は時刻 t の第 d 文書でトピック z が割り当てられた単語数，$N_{t,d} = \sum_z N_{t,d,z}$ を表す．同様に，第二項は，$\{\boldsymbol{\phi}_{t,z}\}_{z=1}^{Z}$ を積分消去することにより，下式となる．

$$P(\boldsymbol{W}_t|\boldsymbol{Z}_t, \hat{\boldsymbol{\Omega}}_{t-1}, \boldsymbol{\Lambda}_t) = \prod_z \frac{\Gamma(\sum_s \lambda_{t,z,s})}{\prod_w \Gamma(\sum_s \lambda_{t,z,s} \hat{\omega}^{(s)}_{t-1,z,w})} \times \frac{\prod_w \Gamma(N_{t,z,w} + \sum_s \lambda_{t,z,s} \hat{\omega}^{(s)}_{t-1,z,w})}{\Gamma(N_{t,z} + \sum_s \lambda_{t,z,s})}, \quad (5)$$

ここで $N_{t,z,w}$ は時刻 t において単語 w にトピック z が割り当てられた数，$N_{t,z} = \sum_w N_{t,z,w}$ を表す．第三項

は，(2) を用いることにより，下式となる．

$$P(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1},\gamma)=\prod_z \frac{\gamma^{\gamma\alpha_{t-1,z}}\alpha_{t,z}^{\gamma\alpha_{t-1,z}-1}\exp(-\gamma\alpha_{t,z})}{\Gamma(\gamma\alpha_{t-1,z})}. \tag{6}$$

潜在トピックは Collapsed ギブスサンプリング [7] を用いることにより，効率的に割り当てることができる．表記の簡素化のため $j=(t,d,n)$ とする．トピック z_j を除く全ての変数が与えられたとき，z_j は同時分布 (3) から導かれる下式の確率に従ってサンプリングされる．

$$\begin{aligned}&P(z_j=k|\boldsymbol{W}_t,\boldsymbol{Z}_{t\backslash j},\boldsymbol{\alpha}_t,\hat{\boldsymbol{\Omega}}_{t-1},\boldsymbol{\Lambda}_t)\\ &\propto \frac{N_{t,d,k\backslash j}+\alpha_{t,k}}{N_{t,d\backslash j}+\sum_z\alpha_{t,z}}\frac{N_{t,k,w_j\backslash j}+\sum_s\lambda_{t,s,k}\hat{\omega}^{(s)}_{t-1,k,w_j}}{N_{t,k\backslash j}+\sum_s\lambda_{t,s,k}},\end{aligned} \tag{7}$$

ここで $\backslash j$ はサンプル j を除いたときの回数を表す．

重み $\boldsymbol{\Lambda}_t$ およびハイパーパラメータ $\boldsymbol{\alpha}_t$ は不動点反復法により同時分布 (3) を最大にすることにより推定できる [12]．このとき $\boldsymbol{\Lambda}_t$ の更新式は

$$\lambda_{t,z,s}\leftarrow\lambda_{t,z,s}\frac{\sum_w\hat{\omega}^{(s)}_{t-1,z,w}M_{t,z,w}}{M_{t,z}}, \tag{8}$$

となる．ここで $M_{t,z,w}=\Psi(N_{t,z,w}+\sum_{s'}\lambda_{t,z,s'}\hat{\omega}^{(s')}_{t-1,z,w})-\Psi(\sum_{s'}\lambda_{t,z,s'}\hat{\omega}^{(s')}_{t-1,z,w})$, $M_{t,z}=\Psi(N_{t,z}+\sum_{s'}\lambda_{t,z,s'})-\Psi(\sum_{s'}\lambda_{t,z,s'})$. を表し，$\Psi(\cdot)$ はディガンマ関数 $\Psi(x)=\frac{\partial\log\Gamma(x)}{\partial x}$ である．また $\boldsymbol{\alpha}_t$ の更新式は

$$\alpha_{t,z}\leftarrow\frac{\gamma\alpha_{t-1,z}-1+\alpha_{t,z}\sum_d(\Psi(N_{t,d,z}+\alpha_{t,z})-\Psi(\alpha_{t,z}))}{\gamma+\sum_d\Psi(N_{t,d}+\sum_{z'}\alpha_{t,z'})-\Psi(\sum_{z'}\alpha_{t,z'})} \tag{9}$$

となる．

各時刻のデータが与えられる毎に，ギブスサンプリング (7) と同時分布最大化 (8)，(9) を繰り返すことにより，提案モデルをオンライン学習することができる．

2.3 多重スケール単語分布推定

多重スケール単語分布 $\omega^{(s)}_{t,z,w}$ は，確率的 EM アルゴリズムにより推定された潜在トピックを用いて推定できる．$\omega^{(s)}_{t,z,w}$ は時刻 $t-2^{s-1}+1$ から t におけるトピック z での単語 w の出現確率であるため，その推定値は下式により得られる．

$$\hat{\omega}^{(s)}_{t,z,w}=\frac{\hat{N}^{(s)}_{t,z,w}}{\sum_w\hat{N}^{(s)}_{t,z,w}}=\frac{\sum_{t'=t-2^{s-1}+1}^{t}\hat{N}_{t',z,w}}{\sum_w\sum_{t'=t-2^{s-1}+1}^{t}\hat{N}_{t',z,w}}, \tag{10}$$

ここで $\hat{N}^{(s)}_{t,z,w}$ は時刻 $t-2^{s-1}+1$ から t までのトピック z において単語 w の期待出現回数であり，$\hat{N}_{t,z,w}$ は時刻 t における期待出現回数である．期待出現回数は $\hat{N}_{t,z,w}=N_{t,z}\hat{\phi}_{t,z,w}$ により計算できる．ここで $\hat{\phi}_{t,z,w}$ は $\phi_{t,z,w}$ の推定値を表し，

$$\hat{\phi}_{t,z,w}=\frac{N_{t,z,w}+\sum_s\lambda_{t,s,z}\hat{\omega}^{(s)}_{t-1,z,w}}{N_{t,z}+\sum_s\lambda_{t,s,z}}, \tag{11}$$

により得られる．式 (10) において，期待値 $\hat{N}_{t,z,w}$ ではなく，実際の値 $N_{t,z,w}$ を用いることも可能だが，$\hat{\omega}^{(s=1)}_{t,z,w}$ と $\phi_{t,z,w}$ の推定値を

$$\hat{\omega}^{(s=1)}_{t,z,w}=\frac{\hat{N}_{t,z,w}}{\sum_w\hat{N}_{t,z,w}}=\hat{\phi}_{t,z,w}, \tag{12}$$

のように一致させるため，期待値を用いた．

期待値 $\hat{N}^{(s)}_{t,z,w}$ は下式のように，一時刻前の値 $\hat{N}^{(s)}_{t-1,z,w}$ から逐次的に計算できるため，2^{s-1} 回の加算は必要はなく，二回の加算のみで推定可能である．

$$\hat{N}^{(s)}_{t,z,w}\leftarrow\hat{N}^{(s)}_{t-1,z,w}+\hat{N}_{t,z,w}-\hat{N}_{t-2^{s-1},z,w}. \tag{13}$$

しかしながら，$\hat{N}^{(s)}_{t,z,w}$ の更新には時刻 $t-2^{S-1}$ から $t-1$ までの $\hat{N}_{t,z,w}$ が必要であるため，記憶容量はスケール数に対して指数的に増加し，多重スケール単語分布の更新に計 $O(2^{S-1}ZW)$ の記憶容量が必要となる．そこで，長期スケールの更新頻度を減らし，近似的に期待値を求めることにより，記憶容量を削減するアルゴリズム（表 3）を提案する．$\hat{N}^{(s)}_{t,z,w}$ を毎時刻更新するのではなく，2^{s-1} 時刻毎に更新することにより，一時刻前の $\hat{N}^{(s)}_{t-1,z,w}$ の近似値と新たに得られたデータから，$\hat{N}^{(s)}_{t,z,w}$ の近似値を計算することができる．つまり，各時刻の期待値 $\hat{N}_{t,z,w}$ の保持が不要で，記憶容量を $O(SZW)$ に抑えることができ，記憶容量はスケール数に対して線形に増加するため，長期スケールの単語分布もモデルに組み込むことが可能である．図 4 に，期待値 $\hat{N}^{(s)}_{t,z,w}$ の提案近似更新図を示す．各矩形は $\hat{N}_{t',z,w}$ を表し，矩形内の数字は t' を表す．各時刻の各行は $\hat{N}^{(s)}_{t,z,w}$ を表し，塗潰し矩形は値が更新されたことを表す．長期スケールの単語分布の変化は短期スケールの単語分布の変化よりも遅いと考えられるため，長期スケールの更新頻度を減らすことは妥当であると考えられる．図 1(c) にオンライン推定する場合の提案モデルののグラフィカルモデルを示す．

単語分布のディリクレ事前分布のパラメータとして，(1) にあるように，多重スケール単語分布の重み付き和を用いた．このパラメータは，下式の様に，各時刻の単語分布の重み付き和として表現することができる．

$$\begin{aligned}\sum_{s=1}^{S}\lambda_{t,z,s}\hat{\omega}^{(s)}_{t-1,z,w} &= \sum_{s=1}^{S}\lambda_{t,z,s}\frac{\sum_{t'=t-2^{s-1}}^{t-1}\hat{N}_{t',z,w}}{\sum_w\sum_{t'=t-2^{s-1}}^{t-1}\hat{N}_{t',z,w}}\\ &= \sum_{t'=t-2^{S-1}}^{t-1}\lambda'_{t,z,t'}\hat{\phi}_{t',z,w},\end{aligned} \tag{14}$$

1: $\hat{N}^{(1)}_{t,z,w} \leftarrow \hat{N}_{t,z,w}$
2: **for** $s = 2, \cdots, S$ **do**
3: **if** $t \bmod 2^{s-1} = 0$ **then**
4: $\hat{N}^{(s)}_{t,z,w} \leftarrow \hat{N}^{(s-1)}_{t,z,w} + \hat{N}^{(s-1)}_{t-1,z,w}$
5: **else**
6: $\hat{N}^{(s)}_{t,z,w} \leftarrow \hat{N}^{(s)}_{t-1,z,w}$
7: **end if**
8: **end for**

図 3: Algorithm for the approximate update of $\hat{N}^{(s)}_{t,z,w}$.

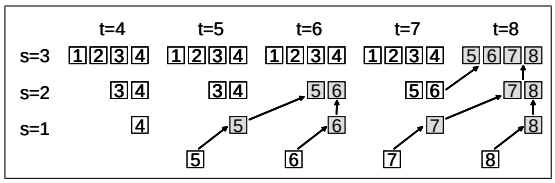


図 4: Illustration of approximate updating $\hat{N}^{(s)}_{t,z,w}$ from $t = 4$ to $t = 8$ with $S = 3$.

ここで重みは

$$\lambda'_{t,z,t'} = \sum_{s=\lceil \log_2(t-t'+1)+1 \rceil}^{S} \frac{\lambda_{t,z,s} \sum_w \hat{N}_{t',z,w}}{\sum_w \sum_{t''=t-2^{s-1}}^{t-1} \hat{N}_{t'',z,w}} \tag{15}$$

である．従って，提案モデルは過去の各時刻の単語分布に依存するモデルとみなすこともできる．提案モデルでは，多重スケールモデルとして考えることにより，重みパラメータ $\mathbf{\Lambda}_t$ の数を $O(2^{S-1}Z)$ から $O(SZ)$ に削減できるため，頑健なモデル推定が可能となる．さらに，近似を用いることにより，上述のように記憶容量も $O(2^{S-1}ZW)$ から $O(SZW)$ に削減できる．

3 実験

提案法を評価するため，NIPS，PNAS，Digg，Addresses の 4 つの実文書データセットを用いて実験を行った．

NIPS データは国際会議 NIPS (Neural Information Processing Systems) の 1987 年から 1999 年までの論文データであり，文書数 1,740，語彙数 14,036 であった．PNAS データは論文誌 Proceedings of the National Academy of Sciences の 1915 年から 2005 年までのタイトルデータであり，文書数 79,477，語彙数 20,534 であった．Digg データはソーシャルニュースサイト Digg (http://digg.com) に投稿された 2009 年 1 月 29 日から 2 月 20 日までのブログ記事データであり，文書数 108,356，語彙数 23,494 であった．Addresses データは 1790 年から 2002 年までのアメリカ大統領の一般教書演説のデータであり，三段落を集めたものを一文書とみなした [18]．このとき文書数 6,413，語彙数 6,759 であった．全データセットにおいてストップワードを省き，単位時間を NIPS，PNAS，Addresses では 1 年，Digg では 1 日とした．

提案法 MDTM を DTM，LDAall，LDAone，LDAonline の 4 手法と比較した．DTM はオンライン学習可能な時間発展トピックモデルであり，提案モデル MDTM でスケール数を $S = 1$ としたときのモデルに対応し，多重時間スケールは考慮していない．LDAall，LDAone，LDAonline は時間発展を考慮しない LDA モデルである．LDAall では過去の全データを学習に用いる．LDAone では一時刻前のデータのみ学習に用いる．LDAonline は LDA をオンライン学習したものである [3]．トピック数は全モデル $Z = 50$ とした．提案モデルにおけるスケール数は、最大スケールの分布がデータ全期間を含むように設定した $S = \lceil \log_2 T + 1 \rceil$．ここで T は時刻数である．また，$\gamma = 1$ とし，ディリクレ分布のパラメータは，過学習を防ぐため，$\alpha_{t,z} \geq 10^{-2}$ の制約のもとで最適化した．各手法の予測誤差をパープレキシティによって評価した．低いパープレキシティは低い予測誤差を表す．全体の 10%の文書をテスト文書とし，テスト文書に含まれる半数の単語を予測した．10 セットの学習・テストデータをランダムに作成し，その平均パープレキシティで評価した．

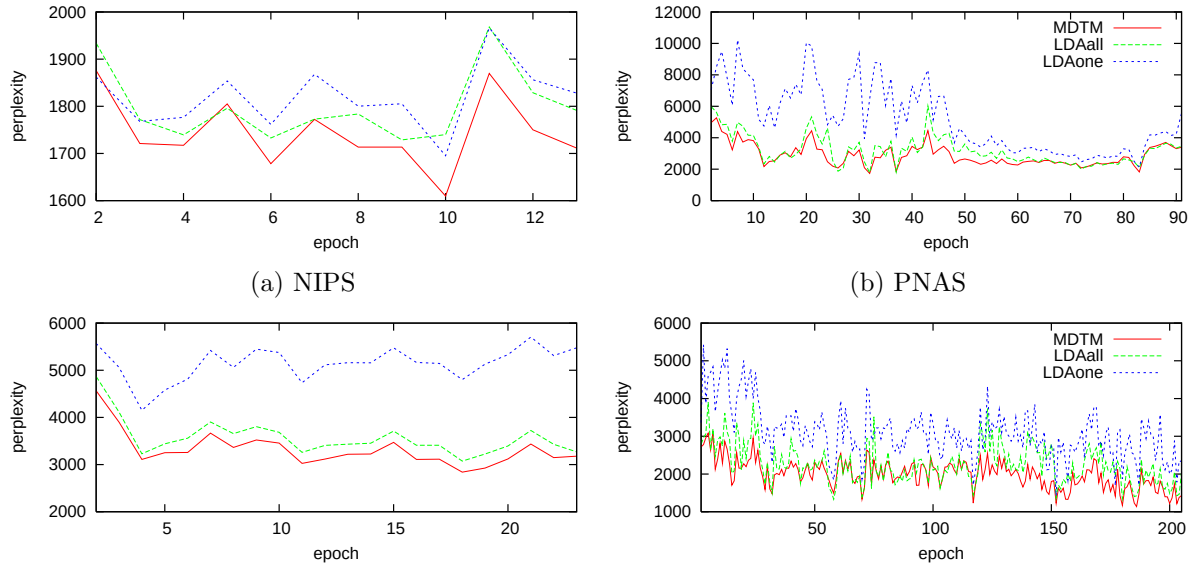
平均パープレキシティを表 2 に，各時刻毎のパープレキシティを図 5 に示す．全データセットにおいて，MDTM が最も低い平均パープレキシティであり，MDTM は多重スケールでの時間発展を考慮することにより，多様な文書データの時間発展を適切にモデル化できることを示す．DTM のパープレキシティが高い理由は，長期間の依存性を考慮していないためであると考えられる．LDAall と LDAonline は時間発展を考慮していないため，高いパープレキシティとなっている．また，LDAone は一時刻のみのデータしか学習に利用せず，過去の情報を無視するため，高いパープレキシティとなっている．

提案法においてスケール数を変化させたときの平均パープレキシティを図 6 に示す．スケール数が増加するに従い，パープレキシティが減少している．この結果は，複数の時間スケールの分布を考慮することの重要性を示している．

図 7 に Xeon5355 2.66GHz CPU の計算機を用いた場合の，一時刻あたりの平均計算時間 (秒) を示す．MDTM の計算時間はスケール数に対して線形に増加している．また，MDTM では多重時間スケールを考慮しているに

表 2: Average perplexities over epochs. The value in the parenthesis represents the standard deviation over data sets.

	MDTM	DTM	LDAall	LDAone	LDAonline
NIPS	**1754.9** (41.3)	1771.6 (37.2)	1802.4 (36.4)	1822.0 (44.0)	1769.8 (41.5)
PNAS	**2964.3** (122.0)	3105.7 (146.8)	3262.9 (159.7)	5221.5 (268.7)	3401.7 (149.1)
Digg	**3388.9** (37.7)	3594.2 (46.4)	3652.6 (27.1)	5162.9 (43.4)	3500.0 (43.6)
Addresses	**1968.8** (56.5)	2105.2 (49.7)	2217.2 (75.3)	3033.5 (70.9)	2251.6 (62.0)



(a) NIPS (b) PNAS

(c) Digg (d) Addresses

図 5: Perplexities for each epoch with MDTM, LDAall, and LDAone.

も関わらず，全データを用いて学習する LDAall に比べて極めて少ない計算量であると言える．

図 8 に推定されたスケール毎の重み $\lambda_{t,z,s}$ を示す．各時刻，各トピックで和が 1 になるように正規化している．スケールが長くなるに従い，重みが減少する傾向がある．この結果は，最近の分布は，現在の分布を推定するためにより重要であるという，直感と合致した結果である．

4 おわりに

本稿では，多重時間スケールでの時間発展を考慮したトピックモデルと，その効率的なオンライン学習法を提案した．また，実験により，提案法は従来法に比べより高い精度で予測できることを示した．今後の課題としては，ディリクレ過程を用いたトピック数の自動推定 [15, 16] や，スケールの自動設定などが考えられる．

参考文献

[1] L. AlSumait, D. Barbara, and C. Domeniconi. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *ICDM '08*, pages 3–12, 2008.

[2] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50(1):5–43, 2003.

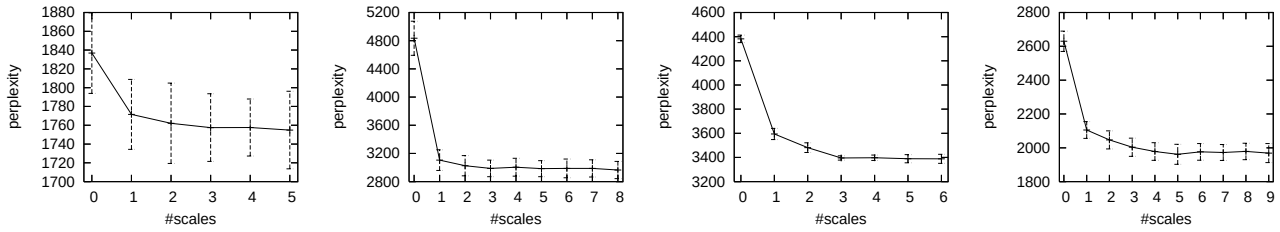
[3] A. Banerjee and S. Basu. Topic models over text streams: A study of batch and online unsupervised learning. In *SDM '07*, 2007.

[4] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML '06*, pages 113–120, 2006.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

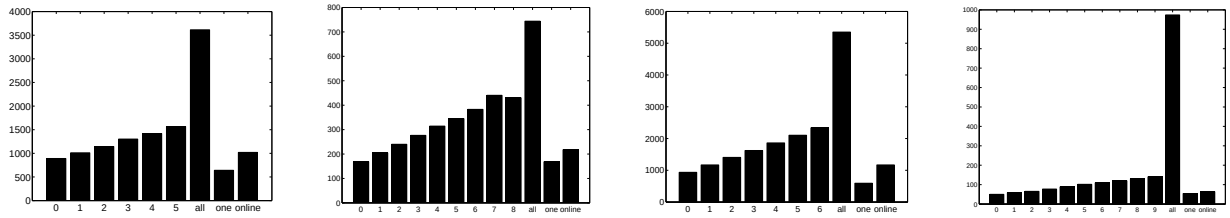
[6] K. R. Canini, L. Shi, and T. L. Griffiths. Online inference of topics with latent Dirichlet allocation. In *AISTATS '09*, volume 5, pages 65–72, 2009.

[7] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 Suppl 1:5228–5235, 2004.



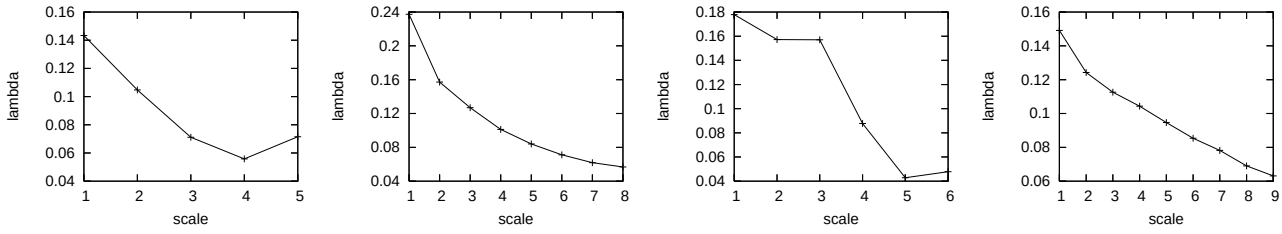
(a) NIPS (b) PNAS (c) Digg (d) Addresses

図 6: Average perplexity of MDTM with different numbers of scales.



(a) NIPS (b) PNAS (c) Digg (d) Addresses

図 7: Average computational time (sec) of MDTM per epoch with different numbers of scales, LDAall, LDAone, and LDAonline.



(a) NIPS (b) PNAS (c) Digg (d) Addresses

図 8: Average normalized weight λ with different scales estimated in MDTM.

[8] T. Hofmann. Probabilistic latent semantic analysis. In *UAI '99*, pages 289–296, 1999.

[9] T. Hofmann. Collaborative filtering via Gaussian probabilistic latent semantic analysis. In *SIGIR '03*, pages 259–266, 2003.

[10] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda. Topic tracking model for analyzing consumer purchase behavior. In *IJCAI '09*, 2009.

[11] T. Iwata, T. Yamada, and N. Ueda. Probabilistic latent semantic visualization: topic model for visualizing documents. In *KDD '08*, pages 363–371, 2008.

[12] T. Minka. Estimating a Dirichlet distribution. Technical report, M.I.T., 2000.

[13] R. Nallapati, W. Cohen, S. Ditmore, J. Lafferty, and K. Ung. Multiscale topic tomography. In *KDD '07*, pages 520–529, 2007.

[14] S. Papadimitriou, J. Sun, and C. Faloutsos. Streaming pattern discovery in multiple time-series. In *VLDB '05*, pages 697–708, 2005.

[15] L. Ren, D. B. Dunson, and L. Carin. The dynamic hierarchical Dirichlet process. In *ICML '08*, pages 824–831, 2008.

[16] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[17] C. Wang, D. M. Blei, and D. Heckerman. Continuous time dynamic topic models. In *UAI '08*, pages 579–586, 2008.

[18] X. Wang and A. McCallum. Topics over time: a non-Markov continuous-time model of topical trends. In *KDD '06*, pages 424–433, 2006.

[19] X. Wei, J. Sun, and X. Wang. Dynamic mixture models for multiple time-series. In *IJCAI '07*, pages 2909–2914, 2007.