Online Multiscale Dynamic Topic Models

Tomoharu Iwata Takeshi Yamada Yasushi Sakurai Naonori Ueda
NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan
{iwata,yamada,yasushi,ueda}@cslab.kecl.ntt.co.jp

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.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; I.2.6 [**Artificial Intelligence**]: Learning; I.5.1 [**Pattern Recognition**]: Model—*Statistical*

General Terms

Algorithms

Keywords

Topic model, Time-series analysis, Online learning

1. INTRODUCTION

Great interest is being shown in developing topic models that can analyze and summarize the dynamics of document collections, such as scientific papers, news articles, and blogs [1, 5, 7, 11, 14, 20, 21, 22]. A topic model is a hierarchical probabilistic model, in which a document is modeled as a mixture of topics, and a topic is modeled as a probability distribution over words. Topic models are successfully used in a wide variety of applications including information retrieval [6], collaborative filtering [10], and visualization [12] as well as the analysis of dynamics.

In this paper, we propose a topic model that permits the sequential analysis of the dynamics of topics with multiple timescales, we call it the *Multiscale Dynamic Topic Model* (MDTM), and its efficient online inference procedures. Topics naturally evolve with multiple timescales. Let us consider the topic 'politics' in a news article collection as an example. There are some words that appear frequently over many years, such as 'constitution', 'congress', and 'president'. On the other hand, some words, such as the names of members in Congress, may appear frequently over periods of tens of years, and other words, such as the names of bills under discussion, may appear for only a few days. Thus, in MDTM, current topic-specific distributions over words are assumed to be generated based on the estimates of multiple timescale word distributions at the previous epoch. Using these multiscale priors improves the predictive performance of the model because the information loss is reduced by considering the long-timescale dependency as well as short-timescale dependency.

The online inference and parameter estimation processes can be achieved efficiently based on a stochastic EM algorithm, in which the model is sequentially updated using newly obtained data; past data does not need to be stored and processed to make new inferences. Some topics may exhibit strong long-timescale dependence, and others may exhibit strong short-timescale dependence. Furthermore, the dependence may differ over time. Therefore, we infer these dependencies for each timescale, for each topic, and for each epoch. By inferring the dependencies from the observed data, MDTM can flexibly adapt to topic dynamics. A disadvantage of online inference is that it can be more unstable than batch inference. With MDTM, the stability can be improved by smoothing using multiple estimates with different timescales.

The remainder of this paper is organized as follows. In Section 2, we formulate a topic model for multiscale dynamics, and describe its online inference procedures. In Section 3, we briefly review related work. In Section 4, we demonstrate the effectiveness of the proposed method by analyzing the dynamics of real document collections. Finally, we present concluding remarks and a discussion of future work in Section 5.

KDD'10, July 25–28, 2010, Washington, DC, USA.

Table 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, $\boldsymbol{\phi}_{t,z} = \{\phi_{t,z,w}\}_{w=1}^{W}$, $\phi_{t,z,w} \geq 0$, $\sum_w \phi_{t,z,w} = 1$
$\boldsymbol{\xi}_{t,z}^{(s)}$	multinomial distribution over words for the zth topic with scale s at epoch t, $\boldsymbol{\xi}_{t,z}^{(s)} = \{\xi_{t,z,w}^{(s)}\}_{w=1}^{W}$, $\xi_{t,z,w}^{(s)} \geq 0$, $\sum_w \xi_{t,z,w}^{(s)} = 1$

2. PROPOSED METHOD

2.1 Preliminaries

In the proposed model, documents are assumed to be generated sequentially at each epoch. Suppose we have a set of D_t documents at the current epoch, t, and each document is represented by $\boldsymbol{w}_{t,d} = \{w_{t,d,n}\}_{n=1}^{N_{t,d}}$, i.e. the set of words in the document. Our notation is summarized in Table 1. We assume that epoch t is a discrete variable, and we can set the time period for an epoch arbitrarily at, for example, one day or one year.

Before introducing the proposed model, we review latent Dirichlet allocation (LDA) [6, 8], which forms the basis of the proposed model. In LDA, each document has topic proportions $\boldsymbol{\theta}_{t,d}$. For each of the $N_{t,d}$ words in the document, topic $z_{t,d,n}$ is chosen from the topic proportions, and then word $w_{t,d,n}$ is generated from a topic-specific multinomial distribution over words $\boldsymbol{\phi}_{z_{t,d,n}}$. Topic proportions $\boldsymbol{\theta}_{t,d}$ and word distributions $\boldsymbol{\phi}_z$ are assumed to be generated according to symmetric Dirichlet distributions. Figure 1 (a) shows a graphical model representation of LDA, where shaded and unshaded nodes indicate observed and latent variables, respectively.

2.2 Model

We consider a set of multiple timescale distributions over words for each topic to incorporate multiple timescale properties. In order to account for the influence of the past at different timescales to the current epoch, we assume that current topic-specific word distributions $\boldsymbol{\phi}_{t,z}$ are generated according to the multiscale word distributions at the previous epoch $\{\boldsymbol{\xi}_{t-1,z}^{(s)}\}_{s=1}^{S}$. Here, $\boldsymbol{\xi}_{t-1,z}^{(s)} = \{\xi_{t-1,z,w}^{(s)}\}_{w=1}^{W}$ represents a distribution over words of topic z with scale s at epoch $t-1$. In particular, we use the following asymmetric Dirichlet distribution for the prior of current word distribution $\boldsymbol{\phi}_{t,z}$, in which the Dirichlet parameter is defined so that its mean becomes proportional to the weighted sum of multiscale word distributions at the previous epoch,

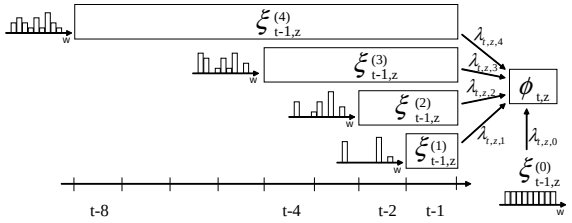


Figure 2: Illustration of multiscale word distributions at epoch t with $S = 4$. Each histogram shows $\boldsymbol{\xi}_{t-1,z}^{(s)}$, which is a multinomial distribution over words with timescale s.

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

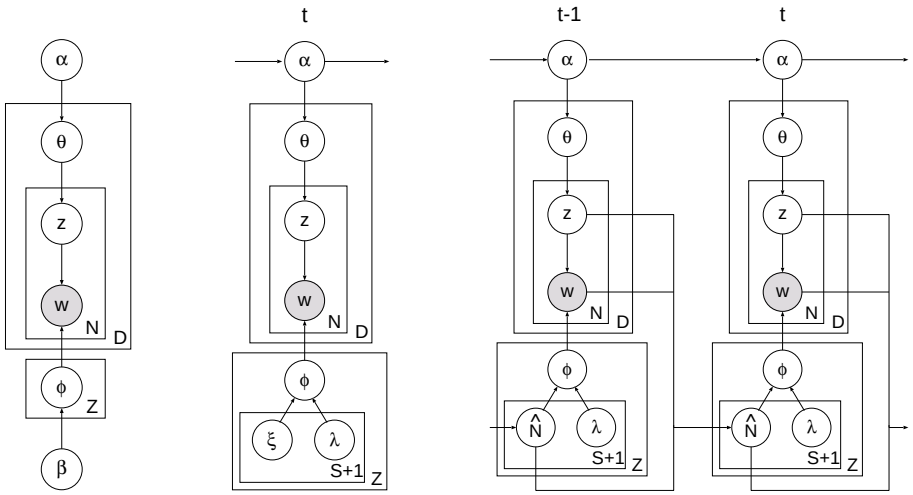
where $\lambda_{t,z,s}$ is a weight for scale s in topic z at epoch t, and $\lambda_{t,z,s} > 0$. By estimating weights $\{\lambda_{t,z,s}\}_{s=0}^{S}$ for each epoch, for each topic, and for each timescale using the current data as described in Section 2.3, MDTM can flexibly respond to the influence on the current distribution of the previous short- and long-timescale distributions. The estimated multiscale word distributions $\{\boldsymbol{\xi}_{t-1,z}^{(s)}\}_{s=1}^{S}$ at the previous epoch are considered as hyperparameters in the current epoch. Their estimation will be explained in Section 2.4.

There are many different ways of setting the scales, but for the simple explanation, we set them so that $\boldsymbol{\xi}_{t,z}^{(s)}$ indicates the word distribution from $t - 2^{s-1} + 1$ to t, where larger s represents longer timescale, and $\boldsymbol{\xi}_{t,z}^{(s=1)}$ is equivalent to the estimate of unit time word distribution $\boldsymbol{\phi}_{t,z}$. We use uniform word distribution $\xi_{t,z,w}^{(s=0)} = W^{-1}$ for scale $s = 0$. This uniform distribution is used to avoid the zero probability problem. Figure 2 illustrates multiscale word distributions with this setting. Word distributions are likely to be smoothed as the timescale becomes long, and be peaked as the timescale becomes short. By using the information presented in these various timescales as the prior for the current distribution with weights, we can infer the current distribution more robustly. In stead of using 2^{s-1} epochs for scale s, we can use any number of epochs. For example, if we know that the given data exhibit periodicity e.g. of one week and one month, we can use the scale of one week for $s = 1$ and one month for $s = 2$. In such case, we can still estimate parameters in the similar way with the algorithm described in Section 2.4. Typically, we do not know the periodicity of the given data in advance, we therefore consider the simple scale setting in the paper.

In LDA, topic proportions $\boldsymbol{\theta}_{t,d}$ are sampled from a Dirichlet distribution. In order to capture the dynamics of topic proportions with MDTM, we assume that the Dirichlet parameters $\boldsymbol{\alpha}_t = \{\alpha_{t,z}\}_{z=1}^{Z}$ depend on the previous parameters. In particular, we use the following Gamma prior for a Dirichlet parameter of topic z at epoch t,

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

where the mean is $\alpha_{t-1,z}$, and the variance is $\alpha_{t-1,z}/\gamma$. By using this prior, the mean is the same as that at the previous epoch unless otherwise indicated by the new data. Parame-



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

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

ter γ controls temporal consistency of the topic proportion prior.

Assuming that we have already calculated the multiscale parameters at epoch $t-1$, $\boldsymbol{\Xi}_{t-1} = \{\{\boldsymbol{\xi}_{t-1,z}^{(s)}\}_{s=0}^{S}\}_{z=1}^{Z}$ and $\boldsymbol{\alpha}_{t-1} = \{\alpha_{t-1,z}\}_{z=1}^{Z}$, MDTM is characterized by the following generative process for the set of documents $\boldsymbol{W}_t = \{\boldsymbol{w}_{t,d}\}_{d=1}^{D_t}$ at epoch t,

1. For each topic $z = 1, \cdots, Z$:
   (a) Draw topic proportion prior $\alpha_{t,z} \sim \text{Gamma}(\gamma\alpha_{t-1,z}, \gamma)$,
   (b) Draw word distribution $\boldsymbol{\phi}_{t,z} \sim \text{Dirichlet}(\sum_s \lambda_{t,z,s}\boldsymbol{\xi}_{t-1,z}^{(s)})$,
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 $w_{t,d,n} \sim \text{Multinomial}(\boldsymbol{\phi}_{t,z_{t,d,n}})$.

Figure 1 (b) shows a graphical model representation of MDTM.

2.3 Online inference

We present an online inference algorithm for MDTM, that sequentially updates the model at each epoch using the newly obtained document set and the multiscale model of the previous epoch. The information in the data up to and including the previous epoch is aggregated into the previous multiscale model. The online inference and parameter estimation can be efficiently achieved by a stochastic EM algorithm [2, 3], in which the collapsed Gibbs sampling of latent topics [8] and the maximum likelihood estimation of hyperparameters are alternately performed [19].

We assume the set of documents $\boldsymbol{W}_t$ at current epoch t, and estimates of parameters from the previous epoch $\boldsymbol{\alpha}_{t-1}$ and $\boldsymbol{\Xi}_{t-1}$ are given. The joint distribution on the set of documents, the set of topics, and the topic proportion priors given the parameters are defined as follows,

$$\begin{aligned} &P(\boldsymbol{W}_t, \boldsymbol{Z}_t, \boldsymbol{\alpha}_t | \boldsymbol{\alpha}_{t-1}, \gamma, \boldsymbol{\Xi}_{t-1}, \boldsymbol{\Lambda}_t) \\ &\quad = P(\boldsymbol{\alpha}_t | \boldsymbol{\alpha}_{t-1}, \gamma) P(\boldsymbol{Z}_t | \boldsymbol{\alpha}_t) P(\boldsymbol{W}_t | \boldsymbol{Z}_t, \boldsymbol{\Xi}_{t-1}, \boldsymbol{\Lambda}_t), \end{aligned} \quad (3)$$

where $\boldsymbol{Z}_t = \{\{z_{t,d,n}\}_{n=1}^{N_{t,d}}\}_{d=1}^{D_t}$ represents a set of topics, and $\boldsymbol{\Lambda}_t = \{\{\lambda_{t,z,s}\}_{s=0}^{S}\}_{z=1}^{Z}$ represents a set of weights. The first term on the right hand side of (3) is as follows using (2),

$$P(\boldsymbol{\alpha}_t | \boldsymbol{\alpha}_{t-1}, \gamma) = \prod_z \frac{\gamma^{\gamma\alpha_{t-1,z}} \alpha_{t,z}^{\gamma\alpha_{t-1,z}-1} \exp(-\gamma\alpha_{t,z})}{\Gamma(\gamma\alpha_{t-1,z})}, \quad (4)$$

where $\Gamma(\cdot)$ is the gamma function. We can integrate out the multinomial distribution parameters in MDTM, $\{\boldsymbol{\theta}_{t,d}\}_{d=1}^{D_t}$ and $\{\boldsymbol{\phi}_{t,z}\}_{z=1}^{Z}$, by taking advantage of Dirichlet-multinomial conjugacy. The second term is calculated by $P(\boldsymbol{Z}_t | \boldsymbol{\alpha}_t) = \prod_{d=1}^{D_t} \int P(\boldsymbol{z}_{t,d} | \boldsymbol{\theta}_{t,d}) P(\boldsymbol{\theta}_{t,d} | \boldsymbol{\alpha}_t) d\boldsymbol{\theta}_{t,d}$, and we have the following equation by integrating out $\{\boldsymbol{\theta}_{t,d}\}_{d=1}^{D_t}$,

$$P(\boldsymbol{Z}_t | \boldsymbol{\alpha}_t) = \left(\frac{\Gamma(\sum_z \alpha_{t,z})}{\prod_z \Gamma(\alpha_{t,z})} \right)^D \prod_d \frac{\prod_z \Gamma(N_{t,d,z} + \alpha_{t,z})}{\Gamma(N_{t,d} + \sum_z \alpha_{t,z})}, \quad (5)$$

where $N_{t,d,z}$ is the number of words in the dth document assigned to topic z at epoch t, and $N_{t,d} = \sum_z N_{t,d,z}$. Similarly, by integrating out $\{\boldsymbol{\phi}_{t,z}\}_{z=1}^{Z}$, the third term is given as follows,

$$\begin{aligned} P(\boldsymbol{W}_t | \boldsymbol{Z}_t, \boldsymbol{\Xi}_{t-1}, \boldsymbol{\Lambda}_t) = &\prod_z \frac{\Gamma(\sum_s \lambda_{t,z,s})}{\prod_w \Gamma(\sum_s \lambda_{t,z,s} \xi_{t-1,z,w}^{(s)})} \\ &\times \frac{\prod_w \Gamma(N_{t,z,w} + \sum_s \lambda_{t,z,s} \xi_{t-1,z,w}^{(s)})}{\Gamma(N_{t,z} + \sum_s \lambda_{t,z,s})}, \end{aligned} \quad (6)$$

where $N_{t,z,w}$ is the number of times word w was assigned to topic z at epoch t, and $N_{t,z} = \sum_w N_{t,z,w}$.

The inference of the latent topics $\boldsymbol{Z}_t$ can be efficiently computed by using collapsed Gibbs sampling [8]. Let $j = (t, d, n)$ for notational convenience, and z_j be the assignment of a latent topic to the nth word in the dth document

at epoch t. Then, given the current state of all but one variable z_j, a new value for z_j is sampled from the following probability,

$$P(z_j = k|\boldsymbol{W}_t, \boldsymbol{Z}_{t\backslash j}, \boldsymbol{\alpha}_t, \boldsymbol{\Xi}_{t-1}, \boldsymbol{\Lambda}_t) \propto \frac{N_{t,d,k\backslash j} + \alpha_{t,k}}{N_{t,d\backslash j} + \sum_z \alpha_{t,z}} \frac{N_{t,k,w_j\backslash j} + \sum_s \lambda_{t,s,k}\xi^{(s)}_{t-1,k,w_j}}{N_{t,k\backslash j} + \sum_s \lambda_{t,s,k}}, \quad (7)$$

where $\backslash j$ represents the count yielded by excluding the nth word in the dth document.

The parameters $\boldsymbol{\alpha}_t$ and $\boldsymbol{\Lambda}_t$ are estimated by maximizing the joint distribution (3). The fixed-point iteration method described in [13] can be used for maximizing the joint distribution as follows,

$$\alpha_{t,z} \leftarrow \frac{\gamma\alpha_{t-1,z} - 1 + \alpha_{t,z}\sum_d(\Psi(N_{t,d,z} + \alpha_{t,z}) - \Psi(\alpha_{t,z}))}{\gamma + \sum_d(\Psi(N_{t,d} + \sum_{z'}\alpha_{t,z'}) - \Psi(\sum_{z'}\alpha_{t,z'}))}, \quad (8)$$

where $\Psi(\cdot)$ is a digamma function defined by $\Psi(x) = \frac{\partial \log\Gamma(x)}{\partial x}$, and,

$$\lambda_{t,z,s} \leftarrow \lambda_{t,z,s}\frac{\sum_w \xi^{(s)}_{t-1,z,w}A_{t,z,w}}{B_{t,z}}, \quad (9)$$

where

$$A_{t,z,w} = \Psi(N_{t,z,w} + \sum_{s'}\lambda_{t,z,s'}\xi^{(s')}_{t-1,z,w}) - \Psi(\sum_{s'}\lambda_{t,z,s'}\xi^{(s')}_{t-1,z,w}), \quad (10)$$

$$B_{t,z} = \Psi(N_{t,z} + \sum_{s'}\lambda_{t,z,s'}) - \Psi(\sum_{s'}\lambda_{t,z,s'}). \quad (11)$$

By iterating Gibbs sampling with (7) and maximum likelihood estimation with (8) and (9), we can infer latent topics while optimizing the parameters. Since MDTM uses the past distributions as the current prior, the label switching problem [17] is not likely to occur when estimated $\lambda_{t,z,s}$ is high, which implies current topics strongly depend on the previous distributions. Label switching can occur when estimated $\lambda_{t,z,s}$ is low. By allowing low $\lambda_{t,z,s}$, which is estimated from the given data at each epoch and each topic, MDTM can adapt flexibly to changes even if existing topics disappear and new topics appear in midstream.

2.4 Efficient estimation of multiscale word distributions

By using the topic assignments obtained after iterating the stochastic EM algorithm, we can estimate multiscale word distributions. Since $\xi^{(s)}_{t,z,w}$ represents the probability of word w in topic z from $t - 2^{s-1} + 1$ to t, the estimation is as follows,

$$\xi^{(s)}_{t,z,w} = \frac{\hat{N}^{(s)}_{t,z,w}}{\sum_w \hat{N}^{(s)}_{t,z,w}} = \frac{\sum_{t'=t-2^{s-1}+1}^{t}\hat{N}_{t',z,w}}{\sum_w\sum_{t'=t-2^{s-1}+1}^{t}\hat{N}_{t',z,w}}, \quad (12)$$

where $\hat{N}^{(s)}_{t,z,w}$ is the expected number of times word w was assigned to topic z from $t - 2^s + 1$ to t, and $\hat{N}_{t,z,w}$ is the expected number of times at t. The expected number is calculated by $\hat{N}_{t,z,w} = N_{t,z}\hat{\phi}_{t,z,w}$, where $\hat{\phi}_{t,z,w}$ is a point estimate of the probability of word w in topic z at epoch t. Although we integrate out $\phi_{t,z,w}$, we can recover its point estimate as follows,

$$\hat{\phi}_{t,z,w} = \frac{N_{t,z,w} + \sum_s \lambda_{t,z,s}\xi^{(s)}_{t-1,z,w}}{N_{t,z} + \sum_s \lambda_{t,z,s}}. \quad (13)$$

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**

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

While it is simpler to use the actual number of times, $N_{t,z,w}$, instead of the expected number of times, $\hat{N}_{t,z,w}$, in (12), we use the latter in order to constrain the estimate of $\xi^{(s=1)}_{t,z,w}$ to be the estimate of $\phi_{t,z,w}$ as follows,

$$\xi^{(s=1)}_{t,z,w} = \frac{\hat{N}_{t,z,w}}{\sum_w \hat{N}_{t,z,w}} = \hat{\phi}_{t,z,w}. \quad (14)$$

Note that the value $\hat{N}^{(s)}_{t,z,w}$ can be updated sequentially from the previous value $\hat{N}^{(s)}_{t-1,z,w}$ as follows,

$$\hat{N}^{(s)}_{t,z,w} \leftarrow \hat{N}^{(s)}_{t-1,z,w} + \hat{N}_{t,z,w} - \hat{N}_{t-2^{s-1},z,w}. \quad (15)$$

Therefore, $\hat{N}^{(s)}_{t,z,w}$ can be updated through just two additions instead of 2^{s-1} additions.

However, to update $\hat{N}^{(s)}_{t,z,w}$, we still need to store values $\hat{N}_{t,z,w}$ from $t-2^{S-1}$ to $t-1$, which means that $O(2^{S-1}ZW)$ memory is required in total for updating multiscale word distributions. Since the memory requirement increases exponentially with the number of scales, this requirement prevents us from modeling long-timescale dynamics. Thus, we consider approximating the update by decreasing the update frequency for long-timescale distributions as in Algorithm 3; this reduces the memory requirement to $O(SZW)$, which is linear against the number of scales. Figure 4 illustrates approximate updating $\hat{N}^{(s)}_{t,z,w}$ with $S = 3$ from $t = 4$ to $t = 8$. Each rectangle represents $\hat{N}_{t',z,w}$, where the number represents t'. Each row at each epoch represents $\hat{N}^{(s)}_{t,z,w}$, and shaded rectangles represent that the values that differ from the previous values. $\hat{N}^{(s)}_{t,z,w}$ is updated at every 2^{s-1}nd epoch. Since the dynamics of a word distribution for a long-timescale is considered to be slower than that for a short-timescale, this approximation, decreasing the update frequency for long-timescale distributions, is reasonable. Updating $\hat{N}^{(s)}_{t,z,w}$ with this approximation requires us to store only the previous $\hat{N}^{(s-1)}_{t,z,w}$ values, and so the memory requirement is $O(SZW)$. Figure 1 (c) shows a graphical model representation of online inference in MDTM.

For the Dirichlet prior parameter of the word distribution, we use the weighted sum of the multiscale word distributions as in (1). The parameter can be rewritten as the weighted sum of the word distributions for each epoch as follows,

$$\sum_{s=1}^{S}\lambda_{t,z,s}\xi^{(s)}_{t-1,z,w} = \sum_{t'=t-2^{S-1}}^{t-1}\lambda'_{t,z,t'}\hat{\phi}_{t',z,w}, \quad (16)$$

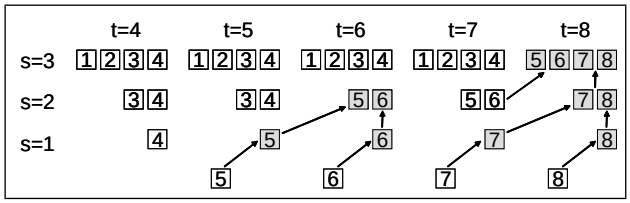


Figure 4: Illustration of approximate updating $\hat{N}^{(s)}_{t,z,w}$ from $t = 4$ to $t = 8$ with $S = 3$.

where

$$\lambda'_{t,z,t'} = \sum_{s=\lceil \log_2(t-t'+1)+1 \rceil}^{S} \frac{\lambda_{t,z,s} \sum_w \hat{N}_{t',z,w}}{\sum_w \sum_{t''=t-2^{s-1}}^{t-1} \hat{N}_{t'',z,w}}, \quad (17)$$

is its weight. See Appendix for the derivation. Therefore, the multiscale dynamic topic model can be seen as an approximation of a model that depends on the word distributions for each of the previous epochs. By considering multiscale word distributions, the number of weight parameters $\mathbf{\Lambda}_t$ can be decreased from $O(2^{S-1}Z)$ to $O(SZ)$, and this leads to more robust inference. Furthermore, the use of multiscaling also decreases the memory requirement from $O(2^{S-1}ZW)$ to $O(SZW)$ as described above.

3. RELATED WORK

A number of methods for analyzing the evolution of topics in document collections have been proposed, such as the dynamic topic model [5], topic over time [21], online latent Dirichlet allocation [1], and topic tracking model [11]. However, none of the above methods take account of multiscale dynamics. For example, the dynamic topic model (DTM) [5] depends only on the previous epoch distribution. On the other hand, MDTM depends on multiple distributions with different timescales. Therefore, with MDTM, we can model the multiple timescale dependency, and so infer the current model more robustly. Moreover, while DTM uses a Gaussian distribution to account for the dynamics, the proposed model uses conjugate priors. Therefore, inference in MDTM is relatively simple compared to that in DTM.

The multiscale topic tomography model (MTTM) [14] can analyze the evolution of topics at various resolutions of timescales by assuming non-homogeneous Poisson processes. In contrast, MDTM models the topic evolution within the Dirichlet-multinomial framework as the same with most topic models including latent Dirichlet allocation [6]. Another advantage of MDTM over MTTM is that it can make inferences in an online fashion. Therefore, MDTM can greatly reduce the computational cost as well as the memory requirements because past data need not be stored. Online inference is essential for modeling the dynamics of document collections, in which large numbers of documents continue to accumulate at any given moment, such as news articles and blogs, because it is necessary to adapt to the new data immediately for topic tracking, and it is impractical to prepare sufficient memory capacity to store all past data. Online inference algorithms for topic models have been proposed [1, 4, 7, 11].

Singular value decomposition (SVD) is used for analyzing multiscale patterns in streaming data [15] as well as topic models. However, since SVD assumes Gaussian noise, it is inappropriate for discrete data such as document collections [9].

4. EXPERIMENTS

4.1 Setting

We evaluated the multiscale dynamic topic model with online inference (MDTM) using four real document collections with timestamps: NIPS, PNAS, Digg, and Addresses.

The NIPS data consists of papers from the NIPS (Neural Information Processing Systems) conference from 1987 to 1999. There were 1,740 documents, and the vocabulary size was 14,036. The unit epoch was set to one year, so there were 13 epochs. The PNAS data consists of the titles of papers that appeared in the Proceedings of the National Academy of Sciences from 1915 to 2005. There were 79,477 documents, and the vocabulary size was 20,534. The unit epoch was set at one year, so there were 91 epochs. The Digg data consists of blog posts that appeared in the social news website Digg (http://digg.com) from January 29th to February 20th in 2009. There were 108,356 documents, and the vocabulary size was 23,494. The unit epoch was set at one day, so there were 23 epochs. The Addresses data consists of the State of the Union addresses from 1790 to 2002. We increased the number of documents by splitting each transcript into 3-paragraph "documents" as done in [21]. We omitted words that occurred in fewer than 10 documents. There were 6,413 documents, and the vocabulary size was 6,759. The unit epoch was set at one year, and excluding the years for which data was missing there were 205 epochs. We omitted stop-words from all data sets.

We compared MDTM to DTM, LDAall, LDAone, and LDAonline. DTM is a dynamic topic model with online inference that does not take multiscale distributions into consideration; it corresponds to MDTM with $S = 1$. Note that DTM used here models dynamics with Dirichlet priors while the original DTM with Gaussian priors. LDAall, LDAone, and LDAonline are based on LDA, and so do not model the dynamics. LDAall is an LDA that uses all past data for inference. LDAone is an LDA that uses just the current data for inference. LDAonline is an online learning extension of LDA, in which the parameters are estimated using those of the previous epoch and the new data [4]. For a fair comparison, the hyperparameters in these LDAs were optimized using stochastic EM as described by Wallach [19]. We set the number of latent topics at $Z = 50$ for all models. In MDTM, we used $\gamma = 1$, and we estimated the Dirichlet prior for topic proportions subject to $\alpha_{t,z} \geq 10^{-2}$ in order to avoid overfitting. We set the number of scales so that one of the multiscale distributions covered the entire period, or $S = \lceil \log_2 T + 1 \rceil$, where T is the number of epochs. We did not compare with the multiscale topic tomography model (MTTM) because the perplexity of MTTM was worse than that of LDA in [14] and MDTM has a clear advantage over MTTM in that MDTM can make inferences in an online fashion.

We evaluated the predictive performance of each model using the perplexity of held-out words,

$$\text{Perplexity} = \exp\left(-\frac{\sum_d \sum_{n=1}^{N^{\text{test}}_{t,d}} \log P(w^{\text{test}}_{t,d,n} | t, d, \mathcal{D}_t)}{\sum_d N^{\text{test}}_{t,d}}\right), \quad (18)$$

where $N_{t,d}^{\text{test}}$ is the number of held-out words in the dth document at epoch t, $w_{t,d,n}^{\text{test}}$ is the nth held-out words in the document, and $\mathcal{D}_t$ represents training samples until epoch t. A lower perplexity represents higher predictive performance. We used half of the words in 10% of the documents as held-out words for each epoch, and used the other words as training samples. We created ten sets of training and test data by random sampling, and evaluated the average perplexity over the ten data sets.

4.2 Results

The average perplexities over the epochs are shown in Table 2, and the perplexities for each epoch are shown in Figure 5. For all data sets, MDTM achieved the lowest perplexity, which implies that MDTM can appropriately model the dynamics of various types of data sets through its use of multiscale properties. DTM had higher perplexity than MDTM because it could not model the long-timescale dependencies. The reason for the high perplexities of LDAall and LDAonline is that they do not consider the dynamics. The perplexity achieved by LDAone is high because it uses only current data and ignores the past information.

The average perplexities over epochs with different numbers of topics are shown in Figure 6. Under the same number of topics, MDTM achieved the lowest perplexities in all of the cases except when $Z = 150$ and 200 in the NIPS data. Even if the number of topics of the other models increases, the perplexities of the other models did not become better than that of our model with fewer topics in PNAS, Digg, and Addresses data. This result indicates that the larger number of parameters of our model is not a major reason for the lower perplexity.

The average perplexities over epochs with different numbers of scales in MDTM are shown in Figure 7. Note that $s = 0$ uses the uniform distribution only, while $s = 1$ uses the uniform distribution and the previous epoch's distribution. The perplexities decreased as the number of scales increased. This result indicates the importance of considering multiscale distributions.

Figure 8 shows the average computational time per epoch when using a computer with a Xeon5355 2.66GHz CPU. The computational time for MDTM is roughly linear against the number of scales. Even though MDTM considers multiple timescale distributions, its computational time is much smaller than that of LDAall which considers a single timescale distribution. This is because that MDTM uses only current samples for inference, in contrast, LDAall uses all samples for inference.

Figure 9 shows the estimated $\lambda_{t,z,s}$ with different numbers of scales s in MDTM. The sum of the values for each epoch and for each topic are normalized to one. The parameters decrease as the timescale lengthens. This result implies that recent distributions are more informative as regards estimating current distributions, which is intuitively reasonable.

Figure 10 shows two topic examples of the multiscale topic evolution in NIPS data analyzed by MDTM. Note that we omit words appeared in the longer timescales from the table. In the longest timescale, basic words for the research field are appropriately extracted, such as 'speech', 'recognition', and 'speaker' in the speech recognition topic, 'control', 'action', 'policy', and 'reinforcement' in the reinforcement learning topic. In the shorter timescale, we can see the evolution of trends in the research. For example, in the speech recognition research, phoneme classification is a popular task until 1995, and probabilistic approaches such as hidden Markov models (HMM) from 1996 are frequently used.

5. CONCLUSION

In this paper, we have proposed a topic model with multiscale dynamics and efficient online inference procedures. We have confirmed experimentally that the proposed method can appropriately model the dynamics in document data by considering multiscale properties, and that it is computationally efficient.

In future work, we could determine the unit time interval and the length of scale automatically from the given data. We assumed that the number of topics was known and fixed over time. We can automatically infer the number of topics by extending the model to a nonparametric Bayesian model such as the Dirichlet process mixture model [16, 18]. Since the proposed method is applicable to various kinds of discrete data with timestamps, such as web access log, blog, and e-mail, we will evaluate the proposed method further by applying it to other data sets.

6. REFERENCES

[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. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *UAI '09*, pages 27–34, 2009.

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

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

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

[7] 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.

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

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

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

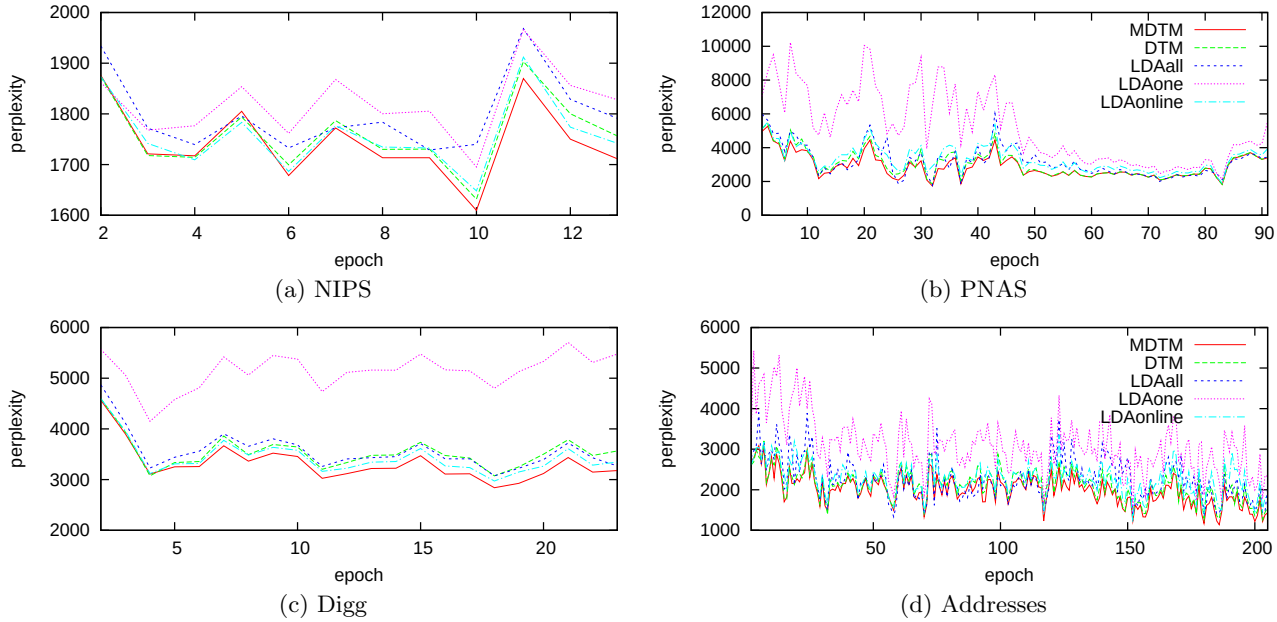
[11] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda. Topic tracking model for analyzing consumer purchase behavior. In *IJCAI '09*, pages 1427–1432, 2009.

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

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

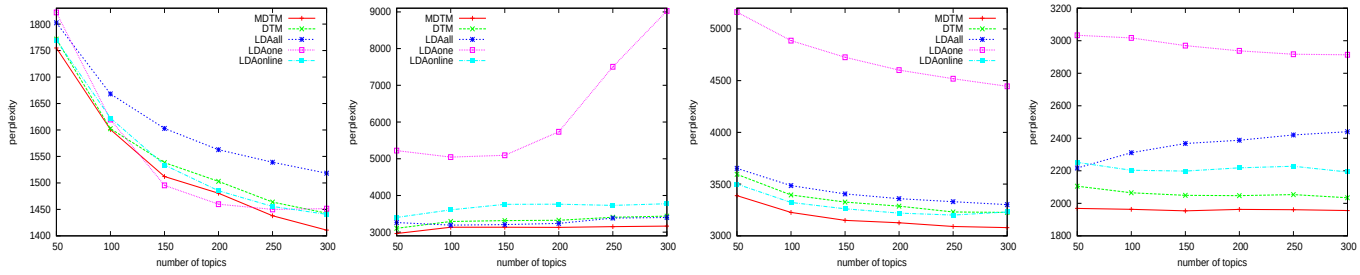
Table 2: Average perplexities over epochs. The value in the parenthesis represents the standard deviation over data sets.

	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)



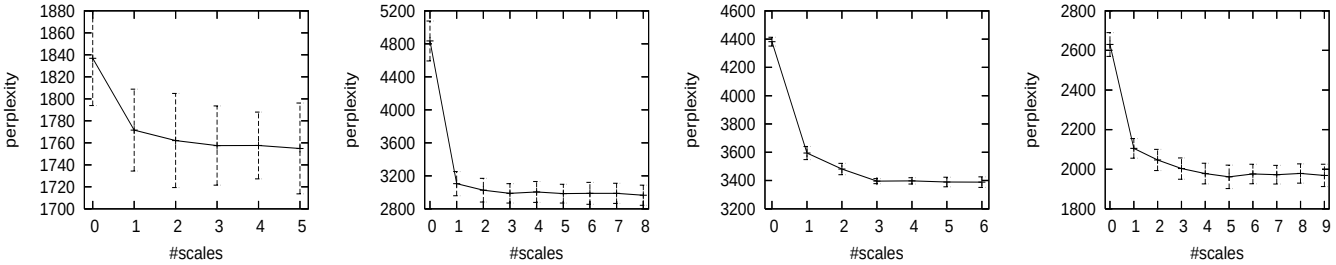
(a) NIPS (b) PNAS (c) Digg (d) Addresses

Figure 5: Perplexities for each epoch.



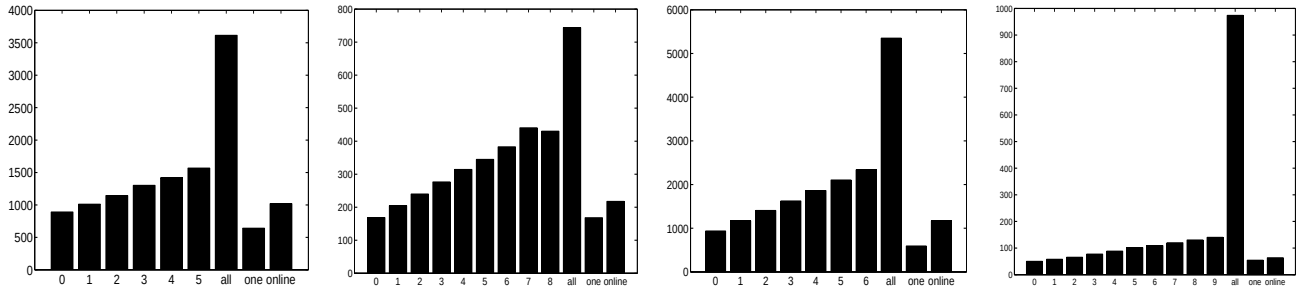
(a) NIPS (b) PNAS (c) Digg (d) Addresses

Figure 6: Average perplexities with different numbers of topics.



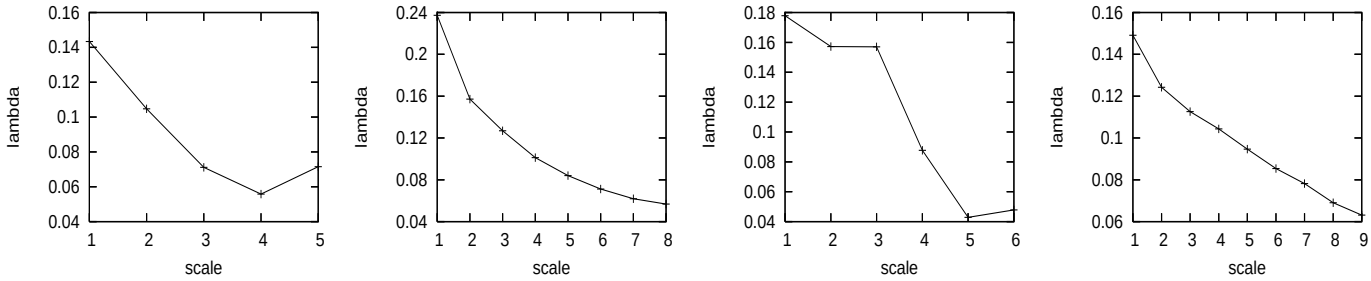
(a) NIPS (b) PNAS (c) Digg (d) Addresses

Figure 7: Average perplexity of MDTM with different numbers of scales.



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

Figure 8: 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

Figure 9: Average normalized weight λ with different scales estimated in MDTM.

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

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

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

[17] M. Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society B*, 62:795–809, 2000.

[18] 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.

[19] H. M. Wallach. Topic modeling: Beyond bag-of-words. In *ICML '06*, pages 997–984, 2006.

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

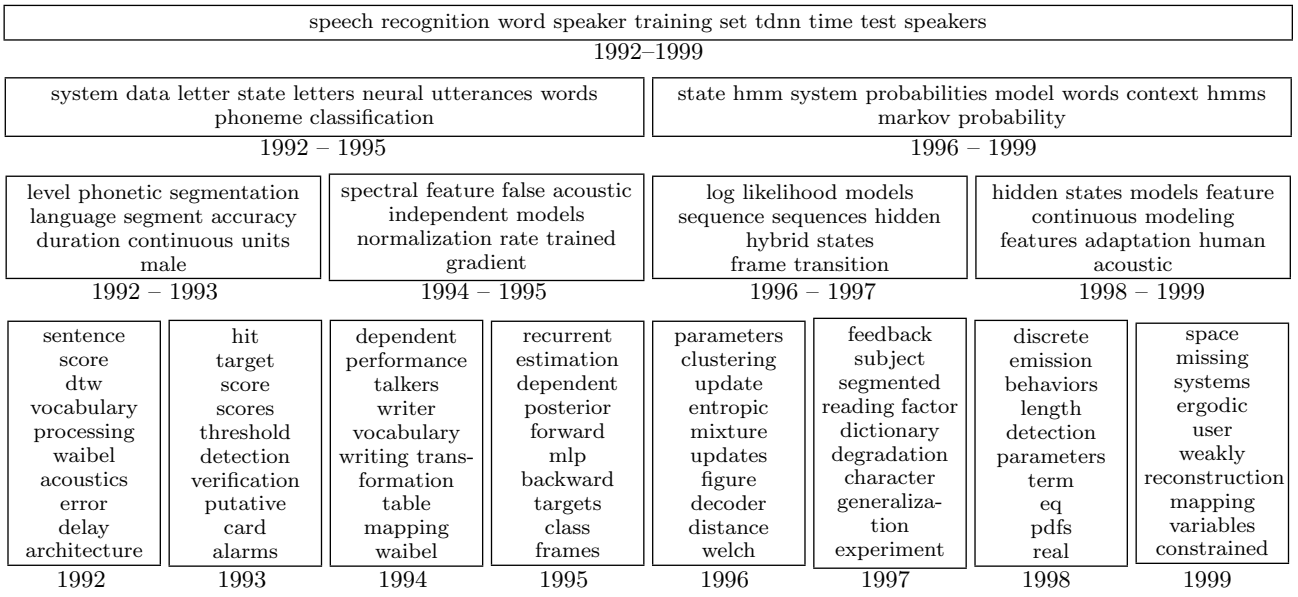
[21] X. Wang and A. McCallum. Topics over time: a non-Markov continuous-time model of topical trends. In *KDD '06*, pages 424–433, 2006.

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

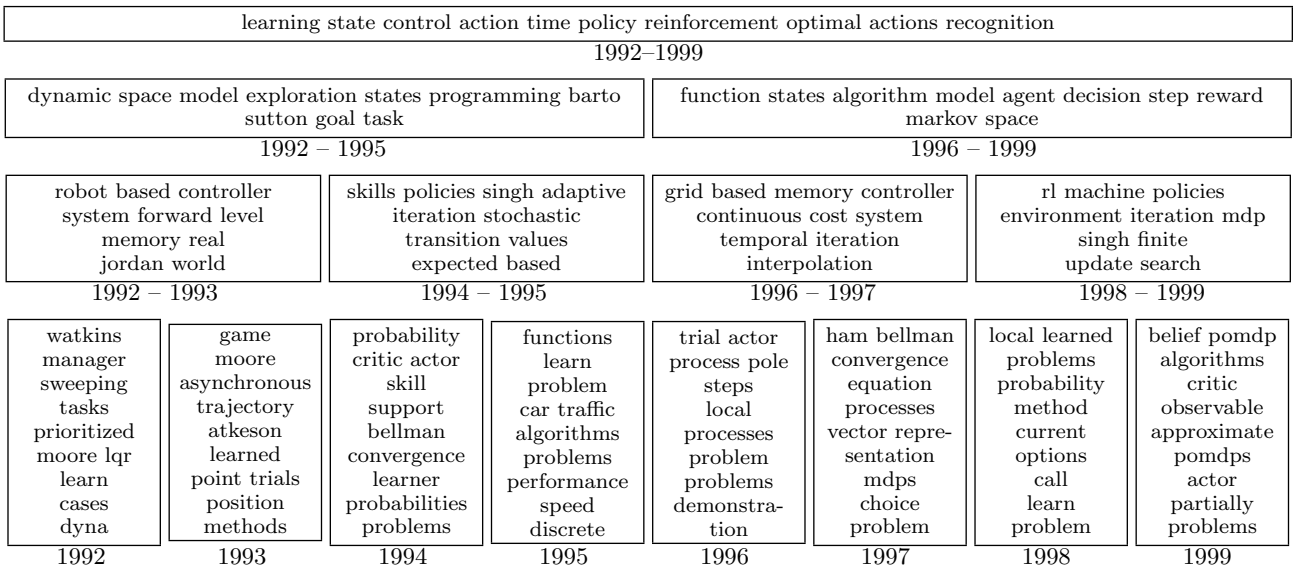
APPENDIX

In this appendix, we give the derivation of (16). Let $\hat{N}^{t-1}_{t-2^{s-1},z} = \sum_w \sum_{t'=t-2^{s-1}}^{t-1} \hat{N}_{t',z,w}$, and $\hat{N}_{t,z} = \sum_w \hat{N}_{t,z,w}$. The Dirichlet prior parameter of the word distribution can be rewritten as the weighted sum of the word distributions for each epoch using (12) as follows,

$$
\begin{aligned}
&\sum_{s=1}^{S} \lambda_{t,z,s} \xi^{(s)}_{t-1,z,w} \\
&= \sum_{s=1}^{S} \lambda_{t,z,s} \frac{\sum_{t'=t-2^{s-1}}^{t-1} \hat{N}_{t',z,w}}{\hat{N}^{t-1}_{t-2^{s-1},z}} \\
&= \sum_{s=1}^{S} \sum_{t'=t-2^{s-1}}^{t-1} \frac{\lambda_{t,z,s}}{\hat{N}^{t-1}_{t-2^{s-1},z}} \hat{N}_{t',z,w} \\
&= \sum_{t'=t-2^{S-1}}^{t-1} \sum_{s=\lceil \log_2(t-t'+1)+1 \rceil}^{S} \frac{\lambda_{t,z,s}}{\hat{N}^{t-1}_{t-2^{s-1},z}} \hat{N}_{t',z,w} \\
&= \sum_{t'=t-2^{S-1}}^{t-1} \sum_{s=\lceil \log_2(t-t'+1)+1 \rceil}^{S} \frac{\lambda_{t,z,s} \hat{N}_{t',z}}{\hat{N}^{t-1}_{t-2^{s-1},z}} \frac{\hat{N}_{t',z,w}}{\hat{N}_{t',z}} \\
&= \sum_{t'=t-2^{S-1}}^{t-1} \lambda'_{t,z,t'} \hat{\phi}_{t',z,w}. \qquad (19)
\end{aligned}
$$



(a) Speech recognition



(b) Reinforcement learning

Figure 10: Two topic examples of the multiscale topic evolution in NIPS data analyzed by MDTM: (a) speech recognition, and (b) reinforcement learning topics. The ten most probable words for each epoch, timescale, and topic are shown.