The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

IMS-DTM: Incremental Multi-Scale Dynamic Topic Models *

Xilun Chen, K. Selçuk Candan
Arizona State University
Tempe, AZ, USA
xilun.chen@asu.edu and candan@asu.edu

Maria Luisa Sapino
University of Torino
Torino, Italy
mlsapino@di.unito.it

Abstract

Dynamic topic models (DTM) are commonly used for mining latent topics in evolving web corpora. In this paper, we note that a major limitation of the conventional DTM based models is that they assume a predetermined and fixed scale of topics. In reality, however, topics may have varying spans and topics of multiple scales can co-exist in a single web or social media data stream. Therefore, DTMs that assume a fixed epoch length may not be able to effectively capture latent topics and thus negatively affect accuracy. In this paper, we propose a Multi-Scale Dynamic Topic Model (MS-DTM) and a complementary Incremental Multi-Scale Dynamic Topic Model (IMS-DTM) inference method that can be used to capture latent topics and their dynamics simultaneously, at different scales. In this model, topic specific feature distributions are generated based on a multi-scale feature distribution of the previous epochs; moreover, multiple scales of the current epoch are analyzed together through a novel multi-scale incremental Gibbs sampling technique. We show that the proposed model significantly improves efficiency and effectiveness compared to the single scale dynamic DTMs and prior models that consider only multiple scales of the past.

Introduction

Web and social media data evolve over time reflecting events and trending topics in the real world: a topic may last active for a long time or may end abruptly after a short and intense activity. Consequently, understanding the temporal scales of these topics and leveraging this information for inference can potentially help make better decisions and recommendations based on web data.

Recently, probabilistic models for discovering latent patterns in data have drawn significant attention due to the attractiveness of the underlying theory and the practical effectiveness of the probabilistic approaches to data analysis Topic models (TM) are a good example for the successful application of probabilistic techniques for discovery of latent patterns, including in scientific analysis (Chang et al. 2009), image analysis (Wang and Mori 2009), and web social media data (Ahmed et al. 2011) analysis. The basic topic model does not consider time; i.e., it assumes that the data corpus is fixed; therefore, *dynamic topic model*s (DTMs) extend the idea by allowing the data corpus to evolve in *epochs* and by chaining topics from consecutive epochs together to track their evolution in time (Blei and Lafferty 2006). DTM and its variations have been successfully applied in many domains with evolving data, including in the analysis of web and social media data streams (Wang, Blei, and Heckerman 2008). Figure 1 shows an example set of topics extracted from a scientific data stream.

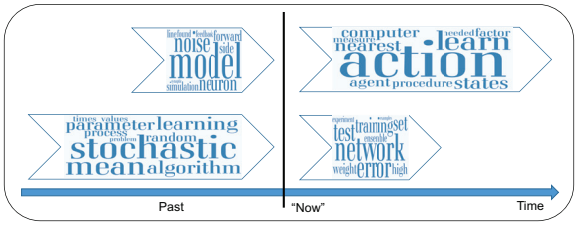


Figure 1: Topics and topic relationships discovered by the proposed multi-scale dynamic topic model at different scales over time (NIPS dataset)

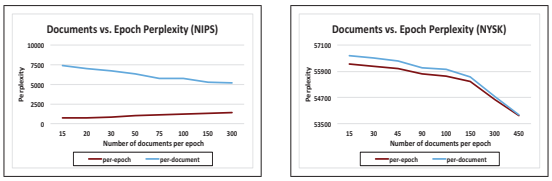


Figure 2: Document- and epoch-level perplexities of different data as a function of the epoch size; different data have different perplexity behaviors; for the NIPS data, the per-epoch and per-document accuracies behave differently

Limitations of the DTM

We note that a major limitation of most existing DTM approaches is that they assume a predetermined and fixed span (or epoch) of topics, whereas an evolving document corpus may contain topics of different temporal scales and, moreover, topics at one scale may impact the prediction of the topics at another scale.

Difficulty of Picking an Epoch Size. One of the major chal-

*This work is partially funded by NSF grants #1610282, #1633381, #1318788, #1339835 and also supported in part by the NSF I/UCRC through the NSF grant #0856090.

lenges is that the relationship between epoch size and accuracy is not trivial to establish. We illustrate this using Figure 2, which shows how document- and epoch-level perplexities vary as a function of the epoch size:

- For the NIPS corpus, representing scientific web collections, epoch-level perplexity increases with the epoch size; i.e, when using larger epochs, it gets more difficult to develop models that describe these epochs using a fixed number of target topics.
- Apple Stock data, representing financial data streams, shows a different behavior: for both epoch- and document-level perplexities, the model gets better as larger time periods are considered This is because the data is too complex to explain accurately focusing on a small time period.

This figure illustrates not only that different data and document streams have different perplexity behaviors, but also that for some data per-epoch and per-document accuracies may behave differently when the epoch size change.

Co-existence and Interdependence of Topics of Different Lengths. It is important to note that, in general, an optimal epoch size may not exist since (a) the relevant time *span may not be fixed and topics of different temporal spans may co-exist in the data stream as an active topic may last for a long time or may abruptly end after a short period*. Moreover, (b) topics of one temporal scale may be predictive of subsequent topics of different temporal scales; in other words, an active topic may be predicted by a mixture of past topics, some of which with long spans and some of which having emerged only recently. Consequently, selecting the appropriate epoch size is not a trivial task:

- if the epoch length is too large, then (since all data objects in the same time epoch are exchangeable) we cannot discover fine grained dynamics (as well as larger patterns that depend on these dynamics) in the data stream;
- if the epoch length is too small, this will not only increase the computational cost during the inference process, but epochs that are too fine grained may not enable us to observe larger/longer patterns in the data stream.

Because of the above, traditional DTMs that assume a fixed epoch length may not be able to capture emergence of latent topics well.

Contributions: Incremental Multi-Scale Dynamic Topic Model

In this paper, we propose a novel Multi-Scale Dynamic Topic Model (MS-DTM), which allows us to mine the latent topics in a dynamic corpus based on the evidences collected from multiple time scales.

Improving Accuracy through Multi-Scale Inference. Using MS-DTM, one can infer latent topics and their dynamics in multiple scales of time. More specifically, the current epoch's feature distribution prior is based on a weighted average of the past distributions at different scales. Since the impacts of the different time scales of the past are not known *a priori*, these weights are learned by analyzing the inter-dependencies among current topics and past data objects. Moreover, in order to discover dependencies of impact among short and long topics, different scales of current time epochs are also inferred.

Efficient Multi-Scale Learning. An important related challenge is to prevent the multi-scale analysis from significantly increasing the DTM analysis cost. DTM based inference usually involves some form of *expectation maximization* (EM) approach) to discover latent patterns. Due to the inherent cost of EM, optimization techniques, such as Gibbs sampling, are often used to efficiently estimate joint feature distributions. In this paper, we complement these with a novel Incremental Multi-Scale Dynamic Topic Model (IMS-DTM) discovery technique, which supports incremental Gibbs sampling at multiple scales, avoiding the need to independently Gibbs sample for different time scales.

Related Work

Model Learning There are two major approaches to model learning: mixture models (McLachlan and Peel 2000) and latent factor models (Agarwal and Chen 2009; Chen and Candan 2014b; 2014a), and they are widely used in many applications such as image analysis (Wang, Blei, and Heckerman 2008; Wang et al. 2016) and high dimensional data processing (Li et al. 2016; Huang, Candan, and Sapino 2016). The main difference between mixture model and factor model is that, factor model assumes that every observation is of a degree of membership to a cluster, instead of assigning clusters explicitly in the mixture model. Typical factor models include singular value decomposition, non-negative matrix factorization, and tensor decomposition. These techniques are usually used for dimensionality reduction and clustering. One advantage is that the number of clusters is adjustable according to the data, and when the data comes in a streaming fashion, the number of clusters can be increased or decreased.

Dynamic Topic Models and its Extensions Dynamic Topic Models (Blei and Lafferty 2006) extend the basic topic modeling technique into a dynamic/incremental setting, where the topic distribution and word distribution priors are evolving. The main difference between a static topic model and a dynamic topic model is that the static topic model assumes that all the documents are exchangeable for the same set of topics; in contrast, in dynamic topic models, this assumption does not hold because the documents are coming in a streaming manner and the order of documents reflects the evolution of the topics. While the original DTM (Blei and Lafferty 2006) is unsupervised, recent extensions, such as (Blei and McAuliffe 2007), proposed supervised versions by adding a response variable associated with each document, and the documents and response are jointly modeled.

Lots of recent work, extends DTMs in different ways. For example, (Nallapati et al. 2008) considers multiple data sources contributing to the dynamic topic model. In (Wang, Blei, and Heckerman 2008), authors avoid discretization of time and they treat time continuously. (Ahmed et al. 2011) proposes a streaming and distributed unsupervised inference method based on topic modeling of user profiles to support recommendation generation. (Bhadury et al. 2016) scales up

Table 1: Notations used in the paper

Symbol	Description
l_{min}	Length of the smallest time epoch
S	Number of scales
α^t	Dirichlet prior for the topics at time epoch t
K	Number of latent topics to be inferred
W	Vocabulary, the unique word set
N	Total number of words in the corpus
D^t	Documents at time epoch t
$D^{t,s}$	Documents at time epoch t with time scale s
N_i^t	Number of words in the i^{th} document at time epoch t
$w_{i,j}^t$	j^{th} word in the i^{th} document at time epoch t
$k_{i,j}^t$	Topic of j^{th} word in the i^{th} document at time epoch t
θ_i^t	Multinomial distribution over topics for the i^{th} document at time epoch t
ϕ_k^t	Multinomial distribution over words for the k^{th} topics at time epoch t
$\psi_{k,m}^t$	Multinomial distribution over words for the k^{th} document at time epoch t in the m^{th} scale in the past
μ_m^t	Weighting factor for the m^{th} scale in the past at time epoch t

the inference process in the DTMs by a fast and parallelizable algorithm. In a work most related to ours, (Iwata et al. 2010) considers multiple time scales of the past and shows that this can improve the predicting power of the DTM. However, (Iwata et al. 2010) does not consider the fact that also the current epoch (i.e. "now") can be considered at multiple time scales and this would not only enable the user to study topics at multiple time scales simultaneously, but could also be used to improve the overall efficiency through incremental processing.

Problem Statement

Table 1 presents key notations used in this paper. Dynamic Topic Modeling (DTM) considers a corpus stream that contains documents generated sequentially over a fixed vocabulary set and assumes that the timeline is split into fixed size epochs. At epoch t, there are D^t documents and a document d_i^t in this epoch is represented as a set of words, i.e. $d_i^t = \{w_1^t, w_2^t, ..., w_{|d_i^t|}^t\}$, where $|d_i^t|$ is the vocabulary size of document d_i^t. DTM infers, a set, L^t, of K latent topics and associate a latent topic, $k_{i,j}^t$, to each word/document pair $\langle i, j\rangle$ that occurs at time epoch t. (Blei and Lafferty 2006) addresses this by extending the static latent Dirichlet allocation model along time. Data is divided into slices (epochs) and each slice is modeled with a K component topic model. The collection of topic models are tied by chaining topics and topic proportions across consecutive slices.

DTM at Multiple Scales

As discussed in the Introduction, this basic dynamic topic model has several difficulties, including the fact that the length of the epoch has to be decided ahead of the time. Moreover, basic DTMs do not recognize that multiple time scales may be relevant for the inference task: the list of active topics may be predicted by a mixture of past topics of different lengths and current topics may last for different scales. Therefore, in order to accurately model the latent topics and their dynamics, we need to model multiple scales of past and current topics. We consider a stream of documents generated over a fixed vocabulary:

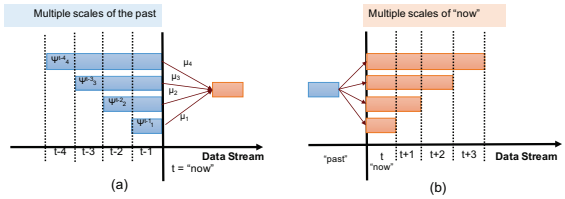


Figure 3: Multi-scale modeling of the (a) past and (b) "now"

- The timeline is split into S many scales of epochs of different lengths. The length of the smallest epoch is l_{min} and the difference of lengths of two consecutive scales s_{h+1} and s_h is also l_{min} (in other words, the length of the scale s is $s \times l_{min}$). All S scales of epoch t start at the same time, but last for different durations.
- At epoch t at scale s, there are $D^{t,s}$ documents and a document $d_i^{t,s}$ in this epoch is represented as a set of words, i.e. $d_i^{t,s} = \{w_1^{t,s}, w_2^{t,s}, ..., w_{|d_i^{t,s}|}^{t,s}\}$, where $|d_i^{t,s}|$ is the vocabulary size of document $d_i^{t,s}$.

The goal of MS-DTM is to infer, for each epoch, t at scale s, a set, $L^{t,s}$, of K latent topics and associate a latent topic, $k_{i,j}^{t,s}$, to each word/document pair $\langle i, j\rangle$ that occurs at time epoch t at scale s.

Incremental Multi-Scale Inference

Due to the non-conjugacy of the Gaussian and multinomial models, inference is often done using approximate techniques, such as variational methods (Blei and Lafferty 2006) or Gibbs sampling (as suggested in (Iwata et al. 2010)), the online inference and parameter estimation can be efficiently achieved by a stochastic EM algorithm, where collapsed Gibbs sampling of latent topics and the maximum likelihood estimation of hyper-parameters are alternately performed). A key difficulty in a multi-scale approach is that the overall work can multiply, rendering the multi-scale approach impractical. Therefore, we need new incremental inference techniques, such as incremental multi-scale Gibbs sampling, to prevent the need to independently Gibbs sample for different time scales.

Multi-Scale Dynamic Topic Model (MS-DTM)

In this section, we introduce our proposed multi-scale DTM (MS-DTM) model that captures evolution of topics of multiple scales. Since multi-scale analysis is performed on past epochs and the current epoch, MS-DTM can be considered in two parts, one dealing with the past and the other dealing with "now".

Multi-Scale Modeling of the Past

In order to model the impacts of the past documents towards the topics in the current epoch, (a) we consider multiple time

scales of word distributions from the previous documents that have been seen and (b) assume that the topic-specific word distribution for the current epoch is a linear combination of the previous word distributions (Iwata et al. 2010). To be more specific, the topic-specific word distribution ϕ_k^t for topic k at the current epoch, t, is computed as a function of the past word distributions as follows:

$$\phi_k^t \sim Dirichlet(f(\psi_{k,1}^{t-1}, \psi_{k,2}^{t-2}, \ldots, \psi_{k,S}^{t-S})), \quad (1)$$

where $f()$ is a function that incorporates the previous word distributions. We enforce that the mean of the Dirichlet parameter for current epoch is proportional to the weighted sum of the word distributions at the previous epochs, i.e.

$$f(\psi_{k,1}^{t-1}, \psi_{k,2}^{t-2}, \ldots, \psi_{k,S}^{t-S}) = \sum_{m=1}^{S} \mu_{k,m}^t \psi_{k,m}^{t-m}. \quad (2)$$

Here, $\mu_{k,m}^t$ are weighting factors that relate the word distributions of the previous epochs to the word distributions of the current epoch and will be learned using the documents in the current time epoch. As visualized in Figure 3(a), these weighting factors enable us to infer the impact of the past scales on the current epoch: if the topics in the current epoch solely depend on the topics in the immediate past, then we would expect that the weighting factor, $\mu_{k,1}^t$, would be large; in contrast, if the current documents are more likely to be influenced by topics of larger temporal scales in the past, then the weighting factors for $m \gg 1$ should be higher. Given this, for each topic $k = 1, 2, ..., K$ at epoch t, we can pick the topic-specific word distribution as $\phi_k^t \sim Dirichlet(\sum_{m=1}^{S} \mu_{k,m}^t \psi_{k,m}^{t-m})$, and, for each document $d_i^t \in D^t$, we can obtain a topic distribution, $\theta_i^t \sim Dirichlet(\alpha^t)$ relying on the Dirichlet prior for the topics at epoch t. Given these, we can then select words in the document by first picking topics, $k \sim Multinomial(\theta_i^t)$, using the topic distribution and then picking corresponding words, $w \sim Multinomial(\phi_k^t)$, using the topic-specific word distribution picked at the beginning.

Multi-Scale Modeling of "Now"

In the previous subsection, we have shown how MS-DTM models the multinomial distribution over words for each topic and relates them, through weighting factors, to the current multinomial distribution over words for these topics. While the above model relates multiple scales of the past with a single scale of now, since "now" can also be considered at multiple scales, we need to extend the model to also account for multiple current scales.

As visualized in Figure 3(b), we achieve this by associating S scales to each topic t. All these S scales start concurrently; however, they span different durations. As stated in Section , if the length of the smallest epoch at scale 1 is l_{min}, then the length of the epoch at scale s is equal to $s \times l_{min}$. Given this, we revise the generative process for the various scales of epoch t as follows:

For each current scale $s = 1, 2, ..., S$ at epoch t

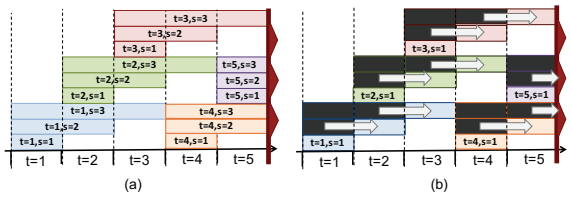


Figure 4: (a). Naive vs. (b). Incremental multi-scale Gibbs sampling

(a) For each topic $k = 1, 2, ..., K$ at epoch t scale s
- $\phi_k^{t,s} \sim Dirichlet(\sum_{m=1}^{S} \mu_{k,m}^t \psi_{k,m}^{t-m})$

(b) For each document $d_i^{t,s} \in D^{t,s}$ at epoch t scale s

i. Draw $\theta_i^{t,s} \sim Dirichlet(\alpha^{t,s})$

ii. For each word in the document $d_i^{t,s}$
- Select a topic $k \sim Multinomial(\theta_i^{t,s})$
- Select a word $w \sim Multinomial(\phi_k^{t,s})$

As we see here, MS-DTM takes into account multiple scales of the past and current epoch, as well as their interdependencies.

IMS-DTM: Incremental Multi-Scale Dynamic Topic Model Inference

In the previous section, we presented the proposed multi-scale DTM (MS-DTM) model, which associates multiple scales to each epoch and analyzes the relationships among the topics in these scales. This is visualized in Figure 4(a). In this example, each epoch is associated with three scales that start simultaneously, but last for different durations. From this example, however, it should be clear that if we naively execute the inference process (e.g., we sample for epochs of all scales) the amount of work will increase significantly. In this section, we discuss how to avoid this potential difficulty within the proposed multi-scale approach.

Overlaps across Scale-Epoch pairs

We can readily notice in Figure 4(a) is that the different scales of different epochs overlap with each other: in fact, in this example where we have 3 scales for each epoch, up to 6 scale-epoch pairs may overlap (see epochs $t = 3, 4, 5$ in the figure).

A second thing that we can easily notice from Figure 4(a) is that the smallest scale of a given epoch, t, is covered by not only the larger scales of the same epoch, but also the larger scales of the epochs that pre-date it. In general, given S scales, the smallest scale of a given epoch, t, is covered by, $S-1$ scales of the same epoch as well as by the scales $j+1$ through S of the time epoch $t-j$, for $1 \leq j \leq S-1$. Therefore, the total amount of overlaps of the smallest scale at epoch t can be computed as $S + \sum_{j=1}^{S-1} \Big(S - (j+1) + 1\Big)$. In the above example, since $S = 3$, this would lead to

$3+(6/2)=6$ overlaps, which can be confirmed by counting the overlaps for epochs $t=3,4,5$.

While this looks like it can cause serious efficiency problems (a given document may be relevant for a quadratic number of overlapping epoch-scale pairs), Figure 4(b) shows that this is not the case.

If the Gibbs sampling performed for the smallest scale at epoch t can be leveraged also for the larger scales, this can help eliminate the need to collect a large number of Gibbs samples. We discuss how to enable this reuse next.

Multi-Scale Collapsed Gibbs Sampling

Collapsed Gibbs sampling is a common technique to infer latent topics. It integrates out the variables controlling multinomial distribution over topics documents, i.e. θ and multinomial distribution over words for topics, i.e. ϕ, while only the latent topic variable k is sampled. In particular, the topic assignment of word v is sampled according to its conditional distribution,

$$P(k_v|k_{\mathcal{N}\setminus v},W_{\mathcal{N}}) \propto \frac{n^{w_v}_{k_v,\mathcal{N}\setminus v}+\beta}{n^{(\cdot)}_{k_v,N\setminus v}+W\beta} \times \frac{n^{d_v}_{k_v,\mathcal{N}\setminus v}+\alpha}{n^{(d_v)}_{\cdot,\mathcal{N}\setminus v}+K\alpha}, \quad (3)$$

where α and β are Dirichlet prior for the topics and words respectively, $\mathcal{N}\setminus v$ is the set minus, $n^{w_v}_{k_v,\mathcal{N}\setminus v}$ is the number of times that word w_v is assigned to topic k_v, and $n^{d_v}_{k_v,\mathcal{N}\setminus v}$ is the number of times a word in document d_v is assigned to topic k_v. Given this, collapsed Gibbs sampling iterates through all words in all documents to approximate the posterior distribution $P(k_{\mathcal{N}}|W_{\mathcal{N}})$.

Since the assignments of the words to the topics may change with new data. Therefore, collapsed Gibbs sampling cannot be directly used when data evolves. (Canini, Shi, and Griffiths 2009) expands collapsed Gibbs sampling to the cases where the set of documents evolves over time, by relying on a decayed MCMC approach: In particular, it keeps a *rejuvenate list*, which contains the topic assignments of some previously seen words. When new documents arrive, it re-samples the topic variables for the words in the rejuvenate list and new samples may alter the word-topic assignments.

We adopt (Canini, Shi, and Griffiths 2009) to develop an incremental, smulti-scale Gibbs sampler. In particular, for each epoch, t, we apply collapsed Gibbs sampling on the documents, $D^{t,1}$, in smallest scale, $s=1$. Then, second scale is incrementally sampled from scale $s=1$, while each later scale can be incrementally sampled from its previous scale. In other words, to obtain the collapsed Gibbs sampling for the S scales of epoch t, we apply collapsed Gibbs sampling on documents in $D^{t,1}$, for each scale $s=2$ to S, we compute $\Delta = diff(D^{t,s}, D^{t,s-1})$, and apply incremental Gibbs sampling on documents in Δ.

Multi-Scale Online Inference

In this section, we discuss how we implement efficient online inference in IMS-DTM. In particular, to achieve efficient online inference using the proposed multi-scale DTM (MS-DTM) model, we build on a stochastic EM based method and incorporate temporal scales. The latent topics are inferred by using incremental multi-scale Gibbs sampling and Dirichlet hyperparameters are determined by maximum likelihood estimation. Below we describe this process.

Formulating the Joint Probability Our first step is to formulate the joint probability of documents and topics

$$P(D^{t,s},Z^{t,s}|\alpha^{t,s},\mu^{t,s},E^{t,s}) = P(Z^{t,s}|\alpha^{t,s}) \times P(D^{t,s}|Z^{t,s},\mu^{t,s},E^{t,s}), \quad (4)$$

where $D^{t,s}$ is the set of documents at time t and scale s, $Z^{t,s}$ is a set of topics, and $E^{t,s}$ is the multiscale matrix containing multinomial distribution over topics, i.e. $E^{t,s}=[\psi^{t,s}_{k,m}]$, and $\mu^{t,s}$ is a vector of weighting factors corresponding to $E^{t,s}$.

Integrating out the Multinomials Given this, we can take advantage of the Dirichlet-multinomial conjugacy and integrate out the multinomial distribution parameter, $\theta_{i,t}$ to rewrite the first term on the right hand side as follows:

$$P(Z^{t,s}|\alpha^{t,s}) = \prod_d \left(\frac{\Gamma(\sum_{z=1}^K \alpha_z^{t,s})}{\prod_{z=1}^K \Gamma(\alpha_z^{t,s})} \times \frac{\prod_z \Gamma(N^{t,s}_{d,z}+\alpha_z^{t,s})}{\Gamma(N_d^{t,s}+\sum_z \alpha_z^{t,s})} \right), \quad (5)$$

Here $N^{t,s}_{d,z}$ is the number of times a specific word is assigned to topic z from document d at time epoch t and scale s, and $N_d^{t,s}$ indicates the total number of times that a word has been assigned to each topic, i.e. $N_d^{t,s}=\sum_z N^{t,s}_{d,z}$.

In fact, the second term on the right hand side can also be rewritten by integrating out the multinomial distribution parameter, $\phi_{i,t}$:

$$P(D^{t,s}|Z^{t,s},\mu^{t,s},E^{t,s}) = \prod_z \Big(\frac{\Gamma(\sum_{m=1}^S \mu^{t,s}_{z,m})}{\prod_w \Gamma(\sum_{m=1}^S \mu^{t,s}_{z,m}\psi^{t,s}_{z,m,w})} \times \frac{\prod_w \Gamma(N^{t,s}_{z,w}+\sum_{m=1}^S \mu^{t,s}_{z,m}\psi^{t,s}_{z,m,w})}{\Gamma(N_z^{t,s}+\sum_{m=1}^S \mu^{t,s}_{z,m})} \Big), \quad (6)$$

where $N^{t,s}_{z,w}$ is the number of times a specific word w appears in topic z at time epoch t and epoch scale s and N_z^t indicates the total number of words appeared in topic z; i.e. $N_z^{t,s}=\sum_w N^{t,s}_{z,w}$.

Applying Multi-Scale Incremental Collapsed Gibbs Sampling Next, we use collapsed Gibbs sampling to sequentially sample each topic variable, depending on the current state of all other variables to see that the new probability for the topic assignment, $P(z_x^{t,s}=j|D^{t,s},Z^{t,s}_{\setminus x},E^{t,s},\mu^{t,s})$, is proportional to

$$\left(\frac{N^{t,s}_{j,w_j\setminus x}+\sum_{m=1}^S \mu^t_{j,m}\psi^{t,s}_{j,m,w_j}}{N^{t,s}_{k\setminus x}+\sum_{m=1}^S \mu^{t,s}_{j,m}} \right) \times \left(\frac{N^{t,s}_{d,j\setminus x}+\alpha_j^{t,s}}{N^{t,s}_{d\setminus x}+\sum_j \alpha_j^{t,s}} \right).$$

Here, index symbol x denotes the quadruple (t,s,d,n), which corresponds to the n^{th} word from document d at time epoch t at scale s, and $\setminus x$ means excluding the count of n^{th} word from document d at time epoch t and epoch scale s. The first ratio shows the probability of word w_j under topic j, using weighted multi-scale distribution from the past, and the second ratio shows the probability of topic j in document d at time t and scale s.

Updating Weighting Factors Given the above, we find the weighting factors for the multi-scale parameters by directly maximizing the joint distribution in Equation 4, using the fix-point iteration method. More specifically,

by taking gradient of the log-likelihood of Equation 6 and setting it to 0, we obtain the following update rule for $\mu_{z,m}^{t,s}$:

$$\mu_{z,m}^{t,s} \leftarrow \frac{\mu_{z,m}^{t,s} \sum_w \psi_{z,m,w}^{t,s} H}{Q}, \tag{7}$$

where $\Psi()$ is the digamma function and we have $H = \Psi\left(N_{z,w}^{t,s} + \sum_m \mu_{z,m}^{t,s}\psi_{z,m,w}^{t,s}\right) - \Psi\left(\sum_m \mu_{z,m}^{t,s}\psi_{z,m,w}^{t,s}\right)$ and $Q = \Psi\left(N_z^{t,s} + \sum_m \mu_{z,m}^{t,s}\right) - \Psi\left(\sum_m \mu_{z,m}^{t,s}\right)$.

Learning the Hyper-parameter, α Finally, to complete the inference, we learn the hyper-parameter α^t from the new data using maximum-likelihood estimation. Again, by taking the gradient of the log-likelihood of Equation 5 and setting the gradient to 0, we obtain the update rule for $\alpha_z^{t,s}$:

$$\alpha_z^{t,s} \leftarrow \frac{\alpha_z^{t,s} \sum_d \left(\Psi\left(N_{d,z}^{t,s} + \alpha_z^{t,s}\right) - \Psi\left(\alpha_z^{t,s}\right)\right)}{\sum_d \left(\Psi\left(N_d^{t,s} + \sum_z \alpha_z^{t,s}\right) - \Psi\left(\sum_z \alpha_z^{t,s}\right)\right)}. \tag{8}$$

Summary Algorithm 1 presents the pseudocode of the proposed iterative multi-scale inference process (IMS-DTM), which leverages the multi-scale update rules introduced above. Through iterative multi-scale Gibbs sampling, IMS-DTM is able to update both the weighting factors and the Dirichlet priors for the topics for all scales of all epochs with minimal overhead.

Algorithm 1 IMS-DTM Algorithm

Input:

Streaming corpus D; Number of topics K; Number of time epochs $\mathcal{T}$; Number of multi scale epochs S; Number of Gibbs sampling iterations $iter$; Hyperparameter update frequency $iterUpdate$;

Output:

The incremental multi-scale dynamic topic model $\mathcal{M}$;

for t = 1 **to** $\mathcal{T}$ **do**
  Initialize Dirichlet prior for the topics at time epoch t as $\alpha^t = \frac{\bar{D}^t \times 0.05}{K}$, where $\bar{D}^t$ is the average document length at time epoch t
  Initialize count variables, Multinomial distribution θ, ϕ and topic assignments.
  Initialize multi-scale parameter μ as $\mu = \frac{1}{S}$
  Initialize Dirichlet prior for the words at time epoch t as $\beta_z^t = \mu_z^t \times E_z^t$
  for i =1 **to** $iter$ **do**
    Update $\theta^{t,1}$, $\phi^{t,1}$, $Z^{t,1}$ using Collapsed GibbsSampling($D^{t,1}$, $\alpha^{t,1}$, $\beta^{t,1}$)
    for s = 2 **to** S **do**
      Update $\theta^{t,s}$, $\phi^{t,s}$, $Z^{t,s}$ using Incremental GibbsSampling($D^{t,s}$, $\alpha^{t,s}$, $\beta^{t,s}$)
      if (i mod $iterUpdate$) == 0 **then**
        Update $\alpha^{t,s}$ using Equation 8
        Update $\mu^{t,s}$ using Equation 7
        Set new word Dirichlet prior as $\beta = \mu \times E$
      end if
    end for
  end for
end for

Experimental Evaluation

In this section, we present experimental evaluations of the efficiency and effectiveness of the proposed IMS-DTM algorithm. All experiments were conducted in Matlab 2015b using an Intel Core i5-2400 machine with 8GB memory. In these experiments, we set the number, S, of scales to 4. The default parameters for Dirichlet prior for the topics for time epoch t is set to $\frac{\bar{D}^t \times 0.05}{K}$, where $\bar{D}^t$ is the average document length at time epoch t.

We compare the proposed IMS-DTM (we refer to this as *multi-past multi-current DTM* (**MPMC**)) with multi-scale dynamic topic model proposed in (Iwata et al. 2010). Since it considers only multiple scales of the past, we refer to this approach as *multi-past single-current DTM* (**MPSC**). We also consider a baseline dynamic topic model with *single-past*[1] *and single-current scales* (**SPSC**) proposed in (Blei and Lafferty 2006) and a *single-past multi-current* (**SPMC**) approach, implemented based on (Canini, Shi, and Griffiths 2009). In the rest of this section, we refer to our approach as *multi-past multi-current* approach.

Evaluation Criteria

To evaluate accuracy of the models, we use the perplexity measure (Blei, Ng, and Jordan 2003). We define *epoch-level perplexity* to assess how well a probability model predicts the epoch. More formally, given a dynamic topic model $\mathbf{M} = \{w, k\}$, the epoch-level perplexity of an epoch $D^t = \{w_i\}$ can be computed as $PLX_{epoch}(t) = P(D^t|\mathbf{M}) = exp\left(-\frac{\sum_{i=1}^{|D^t|} \log p(w_i|\mathbf{M})}{\sum_{i=1}^{|D^t|} N_i}\right)$, where w_i is a word token in the i^{th} document in the epoch and N_i is the total number of words in that document. Similarly, we define *document-level perplexity* of a given epoch as follows: given a dynamic topic model $\mathbf{M} = \{w, k\}$, the document-level perplexity of an epoch $D^t = \{w_i\}$ is $PLX_{doc}(t) = avg_{d \in D^t}\left(P(d|\mathbf{M})\right) = avg_{d \in D^t}\left(exp\left(-\frac{\log p(w_i|\mathbf{M})}{N_i}\right)\right)$. A lower perplexity indicates a better model. Since the goal is often to characterize the epochs, we use *epoch-level perplexity* as the default accuracy measure.

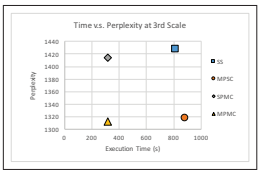
Datasets

In this section, we consider various publicly available datasets, including text streams and numerical time series data.

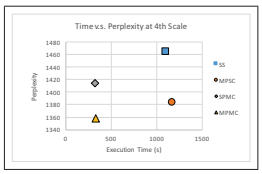
NIPS Data. We obtained the NIPS data, representing web-based scientific data streams, from UCI Machine Learning Repository Bag of Words Data Set. In this text collection, there are 1500 documents, the size of vocabulary set is 12419 and there are approximately 1.9 million words in the corpus. For this data, the default epoch length is 100 documents. The number, K, of latent topics is set to be 50, which is in line with (Iwata et al. 2010) for easy comparison.

NYSK Data. NYSK (New York v. Strauss-Kahn) data set also comes from UCI Machine Learning Repository, representing web-based news data streams, is a collection of English news articles about the case relating to allegations of sexual assault against the former IMF director Dominique

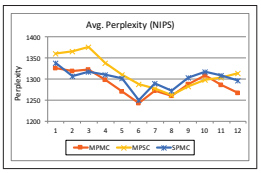
[1]By default, when we refer to single past scale, we consider the smallest of the considered scales.



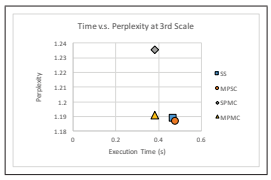
(a) Time v.s. perplexity at 3rd scale (NIPS dataset)



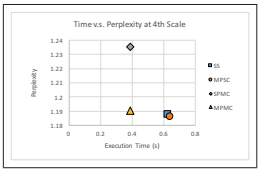
(b) Time v.s. perplexity at 4th scale (NIPS dataset)



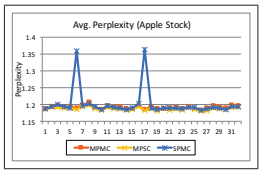
(c) Avg. perplexity over time (NIPS dataset)



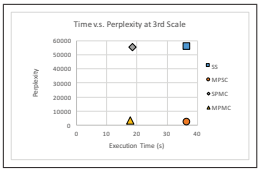
(d) Time v.s. perplexity at 3rd scale (Apple Stock dataset)



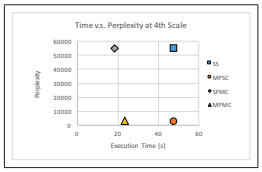
(e) Time v.s. perplexity at 4th scale (Apple Stock dataset)



Avg. perplexity over time (Apple Stock dataset)



(g) Time v.s. perplexity at 3rd scale (NYSK dataset)



(h) Time v.s. perplexity at 4th scale (NYSK dataset)

Figure 5: (a-h) Results for the NIPS dataset, Apple Stock dataset and NYSK dataset

Strauss-Kahn in May, 2011. There are ∼10420 documents in the corpus. For this data, the default epoch length is 45 documents. Since the dataset is more focused (and crawled using just three specific hashtags), we expect the number of latent topics to be lower than NIPS data, and hence we set K to 20.

Apple Stock Data. Apple stock data, representing web-based financial data streams, is in the form of numerical time series from 1981 to 2015 from Quandl website. We use SAX (Lin et al. 2007) to discretize the numerical time series into sets of "documents" such that each "document" has one month worth of closing price data (i.e., 22 5-character SAX words, each corresponding to a moving average of 3 days). The data is split into year-length epochs. For this data set, which tracks stock market price movement, we set the target number, K, of topics to 5 – i.e., we are interested in a few major patterns in the data.

Results

NIPS Dataset Figure 5 (a,b,c) summarizes the results for the NIPS data set[2]. As we see in Figure 5(a,b), the multi-past multi-current scheme (MPMC) provides the lowest execution time, with best accuracy among all four approaches; whereas single scale (SS) results in the worst time/accuracy trade-off. The figure also shows that multi-past single-current (MPSC) provides better accuracy than single-past multi-current (SPMC) and the proposed multi-past multi-current scheme (MPMC) outperforms the MPSC accuracy. The figure also shows that, while gains in accuracy provided by MPSC comes with the penalty of significantly higher execution times than the other approaches, the accuracy gain of MPMC does not come with any execution time penalty. In fact, due to the incremental nature, the amount of time MPMC uses to compute the model for each scale is roughly equal to the time the single scale approach (SS) needs to compute the model for the smallest scale, even though MPMC is able to provide the best overall accuracy. In Figure 5(c), we plot the average perplexity as function of time: the figure shows that at different epochs, MPSC or SPMC might be more advantageous than each other, while the proposed MPMC scheme performs almost always better than the best of MPSC or SPMC.

Apple Stock Dataset Figure 5 (d, e, f) summarizes the results for the Apple stock data set. Once again, the multi-past multi-current scheme (MPMC) provides the best time/accuracy performance among all four approaches and the multi-scale approaches improve accuracy relative to single scale execution (SS): While, unlike the other two data sets, the MPSC provides a slightly better perplexity than MPMC; the proposed incremental multi-past multi-current scale approach (MPMC) provides the fastest execution time (Figure 5(d,e)), without incurring errors that the SPMC scheme introduces (Figure 5(f)) – i.e., even in this data set where (as we see in Figure 1) the model tends to get better as larger time periods are considered, once again MPMC provides the best of the both worlds in terms of efficiency and effectiveness.

NYSK Dataset As we see in Figure 5 (g, h), the results for the NYSK data resembles the results for the NIPS data: the multi-past multi-current scheme (MPMC) provides the best time/accuracy performance among all four approaches and the multi-scale approaches improve accuracy relative to single scale execution (SS): the multi-scale incremental nature of MPMC ensures that the amount of time MPMC uses to compute the model for each scale is roughly equal to the time the single scale approach (SS) needs to compute the model for the smallest scale, even though MPMC provides as high accuracy as SPMC at the execution time cost point of the SS.

Conclusion

Data on the web reflect the evolution of the events and topics in the real world. A major limitation of most existing

[2]Due to limitations of space, we only report results for the 3rd and 4th scales; others also produce similar results.

dynamic topic modeling approaches is that they assume a predetermined and fixed span (or epoch) of topics, whereas an evolving document corpus may contain topics of different temporal scales and, moreover, topics at one scale may impact the prediction of the topics at another scale. In this paper, we proposed a novel multi-scale dynamic topic model (MS-DTM), which considers both "past" and "now" in multiple scales. We further developed a multi-scale incremental Gibbs sampling mechanism for incremental multi-scale dynamic topic model (IMS-DTM) inference. Our experiments show that the proposed IMS-DTM provides accuracy and efficiency gains for data streams with evolving topics of varying lengths.

References

Agarwal, D., and Chen, B. 2009. Regression-based latent factor models. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009*, 19–28.

Ahmed, A.; Low, Y.; Aly, M.; Josifovski, V.; and Smola, A. J. 2011. Scalable distributed inference of dynamic user interests for behavioral targeting. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, 114–122.

Bhadury, A.; Chen, J.; Zhu, J.; and Liu, S. 2016. Scaling up dynamic topic models. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, 381–390.

Blei, D. M., and Lafferty, J. D. 2006. Dynamic topic models. In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, 113–120.

Blei, D. M., and McAuliffe, J. D. 2007. Supervised topic models. In *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, 121–128.

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

Canini, K. R.; Shi, L.; and Griffiths, T. L. 2009. Online inference of topics with latent dirichlet allocation. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, Clearwater Beach, Florida, USA, April 16-18, 2009*, 65–72.

Chang, J.; Boyd-Graber, J. L.; Gerrish, S.; Wang, C.; and Blei, D. M. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada.*, 288–296.

Chen, X., and Candan, K. S. 2014a. GI-NMF: group incremental non-negative matrix factorization on data streams. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*, 1119–1128.

Chen, X., and Candan, K. S. 2014b. LWI-SVD: low-rank, windowed, incremental singular value decompositions on time-evolving data sets. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, 987–996.

Huang, S.; Candan, K. S.; and Sapino, M. L. 2016. BICP: block-incremental CP decomposition with update sensitive refinement. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, 1221–1230.

Iwata, T.; Yamada, T.; Sakurai, Y.; and Ueda, N. 2010. Online multiscale dynamic topic models. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, 663–672.

Li, X.; Huang, S.; Candan, K. S.; and Sapino, M. L. 2016. 2pcp: Two-phase CP decomposition for billion-scale dense tensors. In *32nd IEEE International Conference on Data Engineering, ICDE 2016, Helsinki, Finland, May 16-20, 2016*, 835–846.

Lin, J.; Keogh, E. J.; Wei, L.; and Lonardi, S. 2007. Experiencing SAX: a novel symbolic representation of time series. *Data Min. Knowl. Discov.* 15(2):107–144.

McLachlan, G. J., and Peel, D. 2000. Mixtures of factor analyzers. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, 599–606.

Nallapati, R.; Ahmed, A.; Xing, E. P.; and Cohen, W. W. 2008. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, 542–550.

Wang, Y., and Mori, G. 2009. Human action recognition by semilatent topic models. *IEEE Trans. Pattern Anal. Mach. Intell.* 31(10):1762–1774.

Wang, Y.; Wang, S.; Tang, J.; Liu, H.; and Li, B. 2016. PPP: joint pointwise and pairwise image label prediction. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 6005–6013.

Wang, C.; Blei, D. M.; and Heckerman, D. 2008. Continuous time dynamic topic models. In *UAI 2008, Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence, Helsinki, Finland, July 9-12, 2008*, 579–586.