Steering Time-Dependent Estimation of Posteriors with Hyperparameter Indexing in Bayesian Topic Models

Tomonari Masada[1], Atsuhiro Takasu[2], Yuichiro Shibata[1], and Kiyoshi Oguri[1]

[1] Nagasaki University, 1-14 Bunkyo-machi, Nagasaki-shi, Nagasaki, Japan
{masada,shibata,oguri}@nagasaki-u.ac.jp
[2] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
takasu@nii.ac.jp

Abstract. This paper provides a new approach to topical trend analysis. Our aim is to improve the generalization power of latent Dirichlet allocation (LDA) by using document timestamps. Many previous works model topical trends by making latent topic distributions time-dependent. We propose a straightforward approach by preparing a different word multinomial distribution for each time point. Since this approach increases the number of parameters, overfitting becomes a critical issue. Our contribution to this issue is two-fold. First, we propose an effective way of defining Dirichlet priors over the word multinomials. Second, we propose a special scheduling of variational Bayesian (VB) inference. Comprehensive experiments with six datasets prove that our approach can improve LDA and also Topics over Time, a well-known variant of LDA, in terms of test data perplexity in the framework of VB inference.

Keywords: Bayesian methods, topic models, trend analysis, variational inference, parallelization

1 Introduction

This paper provides a simple and efficient approach to Bayesian analysis of topic time-dependency for large-scale document sets. Our aim is to improve the generalization power of latent Dirichlet allocation (LDA) [3] in terms of test data perplexity by proposing a topic model where we use document timestamps as a key factor for improvement. Our approach is based on the intuition that a careful analysis of word frequency differences among time points will help topic extraction by LDA-like Bayesian models. We propose a simple time-dependent variant of LDA and devise a special scheduling of variational Bayesian (VB) inference [3]. In our model, a different word multinomial distribution is prepared for each time point. When we have T time points and K topics, $T \times K$ word multinomials are prepared in total. As this leads to a large number of parameters, overfitting becomes critical. Our contribution to this issue is two-fold:

1. We propose a non-trivial way of defining Dirichlet priors over $T \times K$ word multinomials. When we define a single common Dirichlet prior over all these

word multinomials, overfitting is too strongly suppressed. Thus we propose an effective way of defining Dirichlet priors over these word multinomials.
2. We propose a special scheduling of VB inference. As an initialization, VB for LDA described in [3] is conducted for a certain number of iterations, and our model is initialized with the estimated parameters. We further introduce another twist as a finalization whose details will be exposed later.

We conduct comprehensive experiments on six datasets, four of which are in English, one in Japanese, and the rest one in Korean. The largest dataset contains 368,000 documents, a set of one-year news articles, and 32,800,000 unique document-word pairs. Since our model has a simple construction, it does not sacrifice the efficiency in computation cost for mathematical sophistication and thus can easily handle large datasets. Our approach requires at most 1.5 times as much computation time as LDA in the framework of VB inference parallelized for multi-core CPU. With this efficiency, our approach can improve the generalization power of LDA and further that of Topics over Time (TOT) [16], a well-known time-dependent variant of LDA, by up to 15 percent.

The rest of the paper is organized as follows. Section 2 discusses previous works related to topical trend analysis. Section 3 provides the details of our proposal. Section 4 presents the settings and the results of our evaluation experiments. Section 5 concludes the paper with discussions and future work.

2 Previous Works

Recent Web analysis has focused on processing *timestamped* documents, because we can observe an immense increase in the number of realtime posts sent to a variety of SNS sites, e.g. Twitter, Facebook, etc. Bayesian approach, one of the mainstreams in text mining, also seeks a method for analyzing time-dependency of topics latent in large-scale document sets to capture salient topical trends.

We briefly introduce LDA [3], a standard document model in Bayesian text mining. With LDA, we can take each document as a conglomerate of multiple semantic contents. LDA characterizes each document by a probability distribution defined over a fixed number K of latent topics. Precisely, LDA attaches a multinomial distribution $\texttt{Multi}(\theta_j)$ defined over the topics $\{c_1, \ldots, c_K\}$ to each of the given documents $\{d_1, \ldots, d_J\}$, where θ_j denotes the parameters $(\theta_{j1}, \ldots, \theta_{jK})$ of the topic multinomial for document d_j. We can regard θ_{jk} as the probability that any word token in document d_j expresses topic c_k, not the other topics. Further, LDA characterizes each topic by the multinomial $\texttt{Multi}(\phi_k)$ defined over the fixed word set $\{v_1, \ldots, v_W\}$, where ϕ_k denotes the parameters $(\phi_{k1}, \ldots, \phi_{kW})$ of the word multinomial for topic c_k. We can regard ϕ_{kw} as the probability that topic c_k is expressed by any token of word v_w, not of the other words.

A remarkable feature of LDA is that all topic multinomial parameters $\{\theta_1, \ldots, \theta_J\}$ are drawn from a single common Dirichlet prior distribution $\texttt{Di}(\alpha)$, where α denotes the set of the K hyperparamters $(\alpha_1, \ldots, \alpha_K)$. Further, all word multinomial parameters $\{\phi_1, \ldots, \phi_K\}$ are drawn from another single common Dirichlet prior $\texttt{Di}(\beta)$, where β denotes the W hyperparamters $(\beta_1, \ldots, \beta_W)$. In

Table 1. Three options for defining word Dirichlet priors.

Option 1	Prepare a different prior $\texttt{Di}(\beta_k)$ for each of the K disjoint groups of word multinomials, where each group corresponds to a different topic c_k and contains T word multinomials $\texttt{Multi}(\phi_{1k}), \ldots, \texttt{Multi}(\phi_{Tk})$. This option gives $K \times W$ word Dirichlet hyperparamters in total.
Option 2	Prepare a different prior $\texttt{Di}(\beta_t)$ for each of the T disjoint groups of word multinomials, where each group corresponds to a different time point s_t and contains K multinomials $\texttt{Multi}(\phi_{t1}), \ldots, \texttt{Multi}(\phi_{tK})$. This option gives $T \times W$ word Dirichlet hyperparamters in total.
Option 3	Prepare a different prior $\texttt{Di}(\beta_{tk})$ for each of the $T \times K$ word multinomials separately. This option gives $T \times K \times W$ word Dirichlet hyperparamters in total.

the following, we simply call $\texttt{Di}(\alpha)$ topic Dirichlet prior and call $\texttt{Di}(\beta)$ word Dirichlet prior. With these two priors, LDA can reduce the diversity among the topic multinomials and also the diversity among the word multinomials, and thus can achieve a generalization power better than PLSI [7].

When making LDA time-dependent, we can consider the following two assumptions [12]: (i) Topic distributions vary along time; (ii) Word distributions vary along time. Many previous works [2, 10, 11, 15] adopt the former assumption. While Srebro et al. [12] give discussions supporting the former, other works adopt the latter [8] or combine both [9]. In this paper, we adopt the latter assumption and prepare a different word multinomial distribution for each time point. In [11], a highly sophisticated modeling is devised, and the former assumption is combined with the assumption of infinite topics [13]. While our approach may be combined with nonparametric approach, we do not pursue this direction in this paper. The topic model in [8] is similar to ours, because this model has a different word multinomial for each time point. However, this work further considers multiscale effects of the past word frequencies at each time point and realizes a flexible time-dependent posterior estimation. Consequently, the model becomes quite complicated and requires an approximated inference, whose implementation becomes intricate when we attempt to achieve a tolerable computation cost. In contrast, our approach seeks an efficient balance between generalization power and computation cost.

3 Method

3.1 Model Construction

In this paper, we model the time-dependency of topics by using a different word multinomial for each of the given time points $\{s_1, \ldots, s_T\}$. Therefore, our model has $T \times K$ word multinomials $\texttt{Multi}(\phi_{tk})$, $t = 1, \ldots, T$, $k = 1, \ldots, K$, where ϕ_{tk} denotes the multinomial parameters $(\phi_{tk1}, \ldots, \phi_{tkW})$. That is, we have $T \times K \times W$ word multinomial parameters in total, as in [8]. However, we take an approach different from [8] in defining Dirichlet priors over the word multinomials.

As is discussed in Section 2, LDA has K word multinomials each corresponding to a different topic and defines a single Dirichlet prior over these K multinomials. Therefore, we first defined a single Dirichlet prior over all $T \times K$ multinomials in our model. However, a preliminary experiment showed that overfitting was too strongly suppressed and that a poor generalization power was obtained.

Therefore, in another preliminary experiment, we tested the three options in Table 1. As a result, Option 1 achieved success for many datasets. Option 1 defines a common Dirichlet prior $\texttt{Di}(\beta_k)$ over the T word multinomials $\texttt{Multi}(\phi_{1k}), \ldots, \texttt{Multi}(\phi_{Tk})$ for each k. β_k refers to the hyperparameters $(\beta_{k1}, \ldots, \beta_{kW})$ of $\texttt{Di}(\beta_k)$ prepared for topic c_k. As the T word multinomials $\texttt{Multi}(\phi_{1k}), \ldots, \texttt{Multi}(\phi_{Tk})$ are drawn from the same prior, we have a smoothing only separately for each topic, not over all word multinomials. Therefore, we can achieve a more moderate smoothing than a single common prior. However, Option 3 also gave impressive results for some datasets. When we take Option 3, the hyperparameters of the word Dirichlet priors are endowed with fully fine-grained indices β_{tkw}. Since these indices are as fine-grained as the indices of the word multinomial parameters ϕ_{tkw}, Option 3 gives a smoothing effect more moderate than Option 1 and thus showed poor results due to overfitting for several datasets. However, since Option 3 was effective for some datasets, we combine Option 1 with Option 3 in our inference, as will be described in Section 3.2. The detailed results of the preliminary experiments will be given in Section 4.4.

The topic model proposed in [8] adopts Option 3 and does not consider Option 1 and Option 2. That is, the hyperparameters of the word Dirichlet priors are endowed with fully fine-grained indices β_{tkw}. The model in [8] seems to avoid overfitting with multiscale analysis, which can exploit the interaction among word Dirichlet priors attached to different time points. This may cause a smoothing effect and thus may lead to a good generalization power. In contrast, we avoid such complication in modeling and consider various ways of indexing the hyperparameters of the word Dirichlet priors, as in Table 1, to obtain a special scheduling of VB inference where some of these options are combined.

3.2 Posterior Inference

For posterior inference, we adopt variational Bayesian (VB) inference [3]. One reason of this choice is that parallelization is easier than collapsed Gibbs sampling (CGS) [5] and collapsed variational Bayesian (CVB) inference [14]. Many operations in VB inference are embarrassingly parallel like EM alogorithm [4], and thus our approach can scale up to larger datasets. Another reason is that VB achieves a generalization power comparable with CGS and CVB [1].

Due to space limitation, we only give an outline of the formula derivation for our VB, which is similar to that for LDA [3]. Let $(\iota_{j1}, \ldots, \iota_{jK})$ be the parameters of the variational Dirichlet posterior defined over the topics $\{c_1, \ldots, c_K\}$ and attached to document d_j. Intuitively, ι_{jk} tells how strongly the word tokens in document d_j express topic c_k. Further, let $(\zeta_{tk1}, \ldots, \zeta_{tkW})$ be the parameters of the variational Dirichlet posterior defined over the words $\{v_1, \ldots, v_W\}$ and attached to the pair of time point s_t and topic c_k. Intuitively, ζ_{tkw} tells how

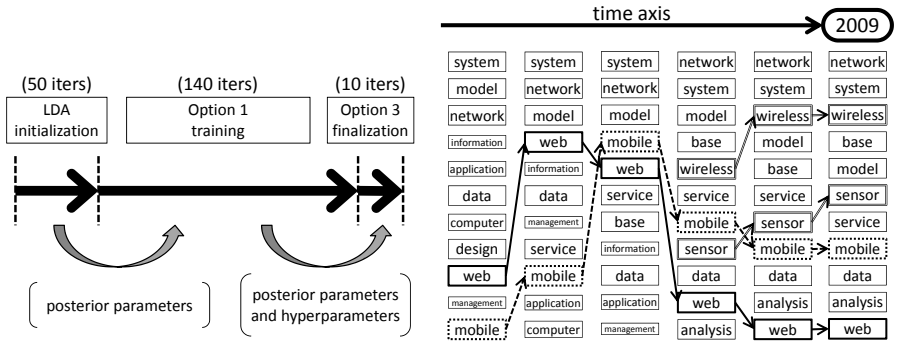

Fig. 1. The proposed inference scheduling for our model (left panel) and an example of the topical trends extracted by our approach (right panel). Each column in the right panel corresponds to a different time point (year) and includes the words sorted in the decreasing order of ζ_{tkw} (cf. Eq. (4)) from top to bottom.

strongly topic c_k is expressed by the tokens of word v_w at time point s_t. With a variational approximation, we obtain the lower bound $\mathcal{L}$ of the log evidence as:

$$\begin{aligned}\mathcal{L} = &\sum_{j,w,k} n_{jw}\pi_{jwk}\Big\{\Psi(\iota_{jk}) - \Psi(\sum_k \iota_{jk}) + \Psi(\zeta_{t_jkw}) - \Psi(\sum_w \zeta_{t_jkw}) - \log \pi_{jwk}\Big\} \\ &- \sum_{j,k}\Big[\Gamma(\alpha_k) - \Gamma(\iota_{jk}) - (\alpha_k - \iota_{jk})\Big\{\Psi(\iota_{jk}) - \Psi(\sum_k \iota_{jk})\Big\}\Big] \\ &- \sum_{t,k,w}\Big[\Gamma(\beta_{kw}) - \Gamma(\zeta_{tkw}) - (\beta_{tw} - \zeta_{tkw})\Big\{\Psi(\zeta_{t_jkw}) - \Psi(\sum_w \zeta_{t_jkw})\Big\}\Big]\ , \quad (1)\end{aligned}$$

where π_{jwk}, satisfying $\sum_k \pi_{jwk} = 1$, refers to the approximated posterior probability that a token of word v_w in document d_j expresses topic c_k. n_{jw} denotes the number of the tokens of word v_w in document d_j. Further, Γ (resp. Ψ) denotes the gamma (resp. digamma) function, and $t_j \in \{1, \ldots, T\}$ is the index of the timestamp of document d_j.

From the partial derivatives of $\mathcal{L}$, we obtain the following update formulas:

$$\pi_{jwk} \propto \exp\Big\{\Psi(\iota_{jk}) - \Psi(\sum_k \iota_{jk}) + \Psi(\zeta_{t_jkw}) - \Psi(\sum_w \zeta_{t_jkw})\Big\}\ , \quad (2)$$

$$\iota_{jk} = \alpha_k + \sum_w n_{jw}\pi_{jwk}\ , \quad (3)$$

$$\zeta_{tkw} = \beta_{kw} + \sum_{\{j:t_j=t\}} n_{jw}\pi_{jwk}\ , \quad (4)$$

$$\alpha_k = \Psi^{-1}\Big(\sum_k \alpha_k + \sum_j \{\Psi(\iota_{jk}) - \Psi(\sum_k \iota_{jk})\}/J\Big)\ , \quad (5)$$

$$\beta_{kw} = \Psi^{-1}\Big(\sum_w \beta_{kw} + \sum_t \{\Psi(\zeta_{tkw}) - \Psi(\sum_w \zeta_{tkw})\}/T\Big)\ , \quad (6)$$

where Ψ^{-1} is the inverse of the digamma function.

While we used the formulas above in a preliminary experiment, any start from a random initialization was likely to find a poor local optimum. Therefore, we propose a special *initialization*. We first conduct VB for LDA with a random initialization by ignoring all timestamps. After a fixed number of iterations (50 iterations in our experiments) of this VB inference, we initialize the parameters of our model with the parameters estimated by this VB for LDA and start the VB for our model by using the update formulas shown above.

In the VB for LDA as an initialization, we use the following update formulas:

$$\pi_{jwk} \propto \exp\left\{\Psi(\iota_{jk}) - \Psi(\sum_k \iota_{jk}) + \Psi(\eta_{kw}) - \Psi(\sum_w \eta_{kw})\right\} , \tag{7}$$

$$\iota_{jk} = \alpha_k + \sum_w n_{jw}\pi_{jwk} , \tag{8}$$

$$\eta_{kw} = \beta_w + \sum_j n_{jw}\pi_{jwk} , \tag{9}$$

where η_{kw} is the approximated posterior parameter telling how strongly topic c_k is expressed by the tokens of word v_w. These three formulas correspond to Eqs. (2), (3), and (4), respectively. In our special scheduling, we first conduct the VB inference for LDA by using Eqs. (7), (8), and (9) and then initialize the parameters π_{jwk} and ι_{jk} of our model with the estimations obtained by this VB for LDA. At the same time, we initialize the posterior parameters ζ_{tkw}, $k = 1, \ldots, K$, $w = 1, \ldots, W$ of our model as $\zeta_{tkw} = \eta_{kw}$ for each t.

As is discussed in Section 3.1, Option 3 in Table 1 sometimes gave impressive results in the preliminary experiment. Therefore, we use Option 3 as a *finalization* of our VB inference. After conducting VB for LDA as an initialization, we train our model with Option 1 for a large enough number of iterations (140 iterations in our experiments). Then, we use the hyperparameters β_{kw} to initialize the hyperparameters β_{tkw} in Option 3. Precisely, we set $\beta_{tkw} = \beta_{kw}$ for each t. After this, we conduct a small number of iterations of VB inference update (10 iterations in our experiments) with Option 3 as a finalization. The update formula for β_{tkw} in Option 3 can be written as follows:

$$\beta_{tkw} = \Psi^{-1}\Big(\sum_w \beta_{tkw} + \Psi(\zeta_{tkw}) - \Psi(\sum_w \zeta_{tkw})\Big) . \tag{10}$$

With respect to the effectiveness of Option 3, Section 4.4 includes detailed discussions based on the experimental results.

Our three-stage VB inference scheduling is summarized in the left panel of Figure 1. While the number of iterations at each stage is determined based on the preliminary experiments, the important point is that we give a *gradually increasing degree of freedom* to word posterior estimation as VB inference proceeds. In this manner, we *S*teer *T*ime-dependent *E*stimation of *P*osteriors with *HY*perparameter indexing. We call our approach *STEPHY* by concatenating the italicized uppercase letters in the previous sentence.

Table 2. Specifications of six datasets used in our experiments.

	J	W	T	P		J	W	T	P
NIPS	1,740	11,998	13	919,916	TDT	96,256	51,849	123	11,460,231
DBLP	1,235,988	273,173	20	7,814,175	NSF	128,181	25,325	13	10,388,976
DONGA	24,093	71,621	53	7,949,288	YOMI	367,910	84,060	52	32,762,456

4 Experiments

4.1 Datasets

We prepared the six datasets in Table 2 for our experiments. J, W, T, and P in Table 2 are the numbers of documents, different words, different time points, and different document-word pairs, respectively. NIPS is the dataset often used in the experiments for machine learning. We used a well-cleaned version appearing in [6][3]. We regarded each publication year as a different time point. This dataset is far smaller than the other five datasets. DBLP dataset is a part of the XML data available at the DBLP Web site[4]. We regarded the paper title as a document and the publication year as a timestamp. Since we used the papers dated from 1990 to 2009, T is equal to 20. DONGA is a set of Korean news articles issued in 2008 and were downloaded from the politics section of the Donga Ilbo Web site[5]. Each article was processed by KLT morphological analyzer[6] to segment each sentence into word tokens. We regarded each week as a single time point. Consequently, we have 53 different time points. TDT is the dataset prepared for the 2002-2003 Evaluation in Topic Detection and Tracking, Phase 4[7]. We regarded the date of each document as a timestamp. This dataset has the largest number of different time points among the six datasets. NSF is the dataset available at UCI machine learning repository[8]. Each document has an ID (e.g. "a9000006"). We regard its first two digits (e.g. "90") as a timestamp. The resulting timestamps range from 90 (i.e., 1990) to 02 (i.e., 2002). YOMI is a set of the Japanese news articles of Yomiuri newspaper published in 2005[9]. The documents were processed by MeCab[10] morphological analyzer to extract words. As in case of DONGA, we regarded each week as a timestamp. For all datasets, we removed the words of low and high frequency by following a common practice of text mining.

4.2 Settings

For comparison, we also adopt VB inference for both LDA and TOT. The VB for LDA is explained in [3]. For TOT, no preceding works report the update

[3] http://www.cs.huji.ac.il/~amitg/htmm.html
[4] http://dblp.uni-trier.de/xml/dblp.xml.gz
[5] http://news.donga.com/Politics
[6] http://nlp.kookmin.ac.kr/HAM/kor/
[7] http://projects.ldc.upenn.edu/TDT4/
[8] http://archive.ics.uci.edu/ml/
[9] http://www.ndk.co.jp/yomiuri/
[10] http://mecab.sourceforge.net/

formulas of VB. Since the formulas are a slight modification of those of LDA, we omit the details here. We only note that the parameters of per-topic Beta distributions in TOT should be rescaled as in case of CGS [16]. We set the rescaling factor to 0.7 based on preliminary trials. Also for TOT, any start from a random initialization gave a poor generalization power. Therefore, we first train LDA and use the resulting posteriors for initializing TOT as in STEPHY.

The experiments are conducted on a Fedora 12 Linux PC equipped with Intel Core i7 920 CPU at 2.67 GHz and 12 Gbytes of main memory. For all cases in our experiments, this main memory size is enough to store all of the input data and the model parameters. To exploit the full potential of our multi-core CPU, we parallelize the operations in VB with OpenMP library by implementing the inference from scratch. Every execution time reported in this paper is a wall-clock time obtained by running eight threads on the four cores of our CPU.

4.3 Evaluation Measure

We evaluate the generalization power of each compared approach by test data perplexity, which tells how well each topic model can generalize to test data. We randomly select 10 percent word tokens from the entire dataset as test word tokens and use them for calculating the perplexity defined as follows:

$$perplexity \equiv \exp\Big\{ -\sum_j \sum_i \log \sum_k \bar{\iota}_{jk} \bar{\zeta}_{t_j k x_{ji}} \Big/ N_{test} \Big\} , \tag{11}$$

where N_{test} is the number of the test word tokens, $t_j \in \{1, \ldots, T\}$ is the index of the timestamp of document d_j, and $x_{ji} \in \{1, \ldots, W\}$ is the index of the word appearing as the ith test word token of document d_j. The summation $\sum_i$ in Eq. (11) is taken only over the test word tokens. Further, $\bar{\iota}_{jk}$ and $\bar{\zeta}_{tkw}$ are the posterior probabilities obtained by normalizing the posterior Dirichlet parameters ι_{jk} and ζ_{tkw}, appearing in Eq. (3) and Eq. (4), as follows:

$$\bar{\iota}_{jk} = \frac{\iota_{jk}}{\sum_k \iota_{jk}} , \quad \bar{\zeta}_{tkw} = \frac{\zeta_{tkw}}{\sum_{w'} \zeta_{tkw'}} . \tag{12}$$

A smaller perplexity corresponds to a better generalization power. The perplexity for LDA and TOT is defined similarly by using η_{kw} in Eq. (9) instead of ζ_{tkw}.

4.4 Preliminary Experiments

Before giving the results of the main experiment comparing STEPHY with LDA and TOT, we overview the results of our preliminary experiments in Table 3 and Table 4 to support the discussions in Section 3.1 and Section 3.2. These tables give the test data perplexity at the 200th iteration, i.e., the final iteration, of the VB inference when we set $K = 50$. Each perplexity is averaged over 20 different execution instances, and the corresponding standard deviation is also presented.

Table 3 shows the effect of our special initialization. The leftmost column includes the tags of the six datasets. When we train our model with Option 1 in

Table 3. Results of the preliminary experiment comparing initialization methods.

	Option 1 (Random init.)	Option 1 (LDA init.)
NIPS	1604.0±9.8	**1394.4±10.2**
DBLP	3299.3±9.4	**3065.9±12.2**
DONGA	3026.2±79.8	**2195.9±23.1**
TDT	3963.7±71.4	**2049.6±15.3**
NSF	3382.5±6.3	**1700.5±13.6**
YOMI	3477.7±76.4	**2768.2±20.1**

Table 4. Results of the preliminary experiment comparing various options for hyperparameter indexing.

	Option 1	Option 2	Option 3	Single common prior
NIPS	1394.4±10.2	1633.9±8.9	1779.0±10.7	**1334.0±11.5**
DBLP	**3065.9±12.2**	3625.3±9.1	3327.5±21.5	3311.8±17.7
DONGA	**2195.9±23.1**	3211.4±13.4	2310.6±19.0	2599.2±19.5
TDT	**2049.6±15.3**	3811.4±12.0	2706.8±15.0	2787.6±41.4
NSF	**1700.5±13.6**	3422.7±4.3	1723.2±12.8	3373.2±4.9
YOMI	**2768.2±20.1**	4847.2±24.6	3130.8±29.8	3041.6±34.0

Table 1 after a random initialization, we obtain the test data perplexity shown in the center column. When we train our model with Option 1 after an initialization using the posterior estimation of LDA, we obtain the perplexity in the rightmost column. Table 3 proves that the initialization with VB for LDA gives a better generalization power than the random initialization for all datasets.

Table 4 shows a comparison between the various ways of indexing the hyperparameters of word Dirichlet priors. We applied the initialization with VB for LDA to each compared case. When we prepare a single common word Dirichlet prior for all $T \times K$ word multinomials, we obtain the perplexity in the rightmost column. When we adopt Option 1, Option 2, and Option 3 in Table 1, we obtain the results in the second, third, and fourth column, respectively.

When we only define a single common word Dirichlet prior, the test data perplexity is poor for many datasets, as is shown in the rightmost column of Table 4. While the perplexity for NIPS dataset is occasionally good, this dataset is far smaller than the other datasets and does not represent the general situation. It seems that a single word Dirichlet prior is enough to cover the topical diversity latent in NIPS dataset. In contrast, the perplexity for NSF dataset is of disastrous level. This may be because overfitting is too strongly suppressed by a single common prior. Option 2 also gives a poor perplexity for many datasets. Option 2 defines a common word Dirichlet prior $\mathtt{Di}(\beta_t)$ over the K word multinomials $\mathtt{Multi}(\phi_{t1}), \ldots, \mathtt{Multi}(\phi_{tK})$ each corresponding to a different topic. This definition is used for each time point s_t separately. Therefore, while Option 2 can differentiate between various time points, the topical diversity is not well captured, because the word posteriors corresponding to different topics share the same Dirichlet prior. In contrast, the perplexity achieved by Option 3 is fairly

Table 5. Test data perplexity and wall-clock computation time after 200 iterations.

	Test data perplexity			Computation time (in sec.)		
	STEPHY	TOT	LDA	STEPHY	TOT	LDA
NIPS	**1407.9±13.8**	1685.2±9.8	1659.9±8.4	489.0 (×1.46)	373.5	334.7
DBLP	**3027.6±17.3**	3439.4±39.6	3446.2±30.3	4737.8 (×1.39)	3700.6	3411.4
DONGA	**2062.2±26.7**	2524.4±25.7	2475.1±24.9	3925.5 (×1.39)	3130.1	2829.2
TDT	**1897.1±31.7**	2005.5±11.6	1988.7±10.7	5354.4 (×1.39)	4292.3	3842.8
NSF	1684.1±18.8	1689.5±12.4	1691.8±14.2	3800.4 (×1.09)	3876.6	3473.8
YOMI	**2671.8±33.3**	2850.2±18.0	2844.2±14.9	13380.5 (×1.18)	12701.5	11390.8

good. Option 3 gives the second best perplexity for DONGA, TDT, and NSF datasets. However, Option 3 requires a large computation time, because Option 3 gives $T \times K \times W$ word Dirichlet hyperparameters in total and thus requires considerable time for the hyperparameter update using Eq. (10). Therefore, we adopt Option 1, giving the best result for all datasets except NIPS, as the main driving force and use Option 3 for finalization. Based on this line of reasoning, we propose an inference scheduling drawn in the left panel of Figure 1.

4.5 Main Experiment

The results of our main experiment are summarized in Table 5. Both test data perplexity and computation time are obtained at the 200th iteration, i.e., at the final iteration, and are averaged over the results of 20 different execution instances. We also include the corresponding standard deviation for test data perplexity. Table 5 shows that STEPHY gives a smaller perplexity than LDA and TOT for all datasets. Especially, for DONGA dataset, the perplexity is reduced by 16.7 percent when compared with LDA. While the margin of improvement is not significantly large only for NSF dataset, we can say that STEPHY can put an improvement into the VB inference framework of LDA-like topic models.

Further, we can compare STEPHY in Table 5 with Option 1 in Table 4. As the left panel of Figure 1 shows, STEPHY conducts 10 iterations with Option 3 as a finalization after the 140 iterations with Option 1. By comparing STEPHY in Table 5 with Option 1 in Table 4, it can be observed that this finalization improves the perplexity for the following five datasets: DBLP ($3027.6 \pm 17.3 < 3065.9 \pm 12.2$), DONGA ($2062.2 \pm 26.7 < 2195.9 \pm 23.1$), TDT ($1897.1 \pm 31.7 < 2049.6 \pm 15.3$), NSF ($1684.1 \pm 18.8 < 1700.5 \pm 13.6$), and YOMI ($2671.8 \pm 33.3 < 2768.2 \pm 20.1$). We can contend that the finalization with Option 3 works.

The comparison experiment also shows that the increase in computation cost brought by our approach is moderate. Table 5 includes the wall-clock computation time each compared method requires for each dataset. The computation time of STEPHY is at most 1.46 times of LDA. With this increase in computation time, we can achieve a significant improvement shown in Table 5. While TOT requires less running time than STEPHY, TOT improves LDA only for DBLP and NSF datasets with a small margin. We can say that STEPHY provides a good balancing between generalization power and computation cost. We

additionally conducted a set of experiments also for the case $K = 100$, i.e., the case where the number of topics is 100. The results, omitted due to space limitation, confirm our conclusion on the efficiency of STEPHY.

The right panel of Figure 1 gives an example of the topical trend extracted by STEPHY from DBLP dataset. Each column corresponds to a different time point and includes the words sorted in the decreasing order of the posterior parameters ζ_{tkw} from top to bottom for one topic arbitrarily selected from the 50 topics. We can interpret the parameter ζ_{tkw} as showing how popularly word v_w is used to express topic c_k at time point s_t. In the right panel of Figure 1, two or three top-ranked words keep their positions over many different time points. However, some explicit topical trends can be observed under these top-ranked words. For example, the word "web" shows a peak around five or six years ago from 2009, and the word "mobile" shows a stable popularity in recent three or four years. Further, the rapid growth of the popularity of the words "wireless" and "sensor" may correspond to the recent rise of the trend related to wireless and sensor networks. Since our approach provides a different word posterior distribution for each time point, this type of trend analysis can be easily conducted only by inspecting the estimated values of ζ_{tkw} along the time axis.

5 Conclusion

In this paper, we propose a simple time-dependent variant of LDA and an effective VB for the proposed model. STEPHY, our total schema for time-dependent topic modeling, improves LDA and TOT in terms of test data perplexity and only increases the computation time of LDA by at most a factor of 1.5.

With respect to the balancing between generalization power and computation cost, we can add the following discussion. STEPHY improves LDA only based on the *intra-epoch* document similarity assumption, i.e., the assumption that the documents having the same timestamp are semantically related to each other. We do not explicitly model any interrelationships of word frequencies over neighboring time points. Therefore, our model does not require an intricate inference. In contrast, many time-dependent topic models are further based on the *inter-epoch* similarity assumption, i.e., the assumption that the documents having different but close timestamps are also semantically related, and intensively exploit the topical dependency over neighboring time points [2, 8–11, 15, 16]. Consequently, the inference requires detailed tricks and becomes less scalable. Our experiments show that, with the intra-epoch similarity, STEPHY achieves an efficient balance between computation cost and generalization power.

STEPHY leaves intact the model construction related to latent topics in LDA. Therefore, one possible future work is to combine our approach with the assumption of infinite topics [13, 11], though we should shift the balance against computational efficiency and check if STEPHY can contribute more than the nonparametric approach that takes advantage of intricate inference.

Another more challenging future work is to apply STEPHY to the probabilistic models where each topic is characterized by a probability distribution

other than multinomial distribution. By steering the time-dependent estimation of posteriors with some hyperparameter indexing strategies like those given in Table 1, we can make a similar proposal for efficiently exploring the parameter space also with respect to those models.

Acknowledgement

This work was supported in part by the Nagasaki University Strategy for Fostering Young Scientists with funding provided by the Special Coordination Funds for Promoting Science and Technology of the Ministry of Education, Culture, Sports, Science and Technology (MEXT).

References

1. Asuncion, A., Welling, M., Smyth, P., Teh, Y.-W.: On smoothing and inference for topic models. In: *Proc. of UAI'09* (2009)
2. Blei, D. M., Lafferty, J. D.: Dynamic topic models. In: *Proceedings of ICML'06*, pp. 113–120 (2006)
3. Blei, D. M., Ng, A. Y., Jordan, M. I.: Latent Dirichlet allocation. *Journal of Machine Learning Research* 3, 993–1022 (2003)
4. Chu, C.-T., Kim, S.-K., Lin, Y.-A., Yu, Y.-Y., Bradski, G., Ng, A. Y., Olukotun, K.: Map-reduce for machine learning on multicore. In: *Proceedings of NIPS'06*, pp. 281–288 (2006)
5. Griffiths, T. L., Steyvers, M.: Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(1), pp. 5228–5235 (2004)
6. Gruber, A., Rosen-Zvi, M., Weiss, Y.: Hidden topic Markov models. In: *Proceedings of AISTATS'07* (2007)
7. Hofmann, T.: Probabilistic latent semantic indexing. In: *Proceedings of SIGIR'99*, pp. 50–57 (1999)
8. Iwata, T., Yamada, T., Sakurai, Y., Ueda, N.: Online multiscale dynamic topic models. In: *Proceedings of KDD'10*, pp. 663–672 (2010)
9. Nallapati, R. M., Ditmore, S., Lafferty, J. D., Ung., K.: Multiscale topic tomography. In: *Proceedings of KDD'07*, pp. 520–529 (2007)
10. Pruteanu-Malinici, I., Ren, L., Paisley, J., Wang, E., Carin, L.: Hierarchical Bayesian modeling of topics in time-stamped documents. *IEEE Trans. Pattern Anal. Mach. Intell.* 32(6), pp. 996–1011 (2010)
11. Ren, L., Dunson, D. B., Carin, L.: The dynamic hierarchical Dirichlet process. In: *Proceedings of ICML'08*, pp. 824–831 (2008)
12. Srebro, N., Roweis, S.: Time-varying topic models using dependent Dirichlet processes. Technical report, Dept. of Computer Science, Univ. of Toronto (2005)
13. Teh, Y.-W., Jordan, M. I., Beal, M. J., Blei, D. M.: Hierarchical Dirichlet processes. *Journal of the American Statistical Association* 101(476), pp. 1566–1581 (2006)
14. Teh, Y.-W., Newman, D., Welling, M.: A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In: *Proceedings of NIPS'06*, pp. 1353–1360 (2006)
15. Wang, C., Blei, D., Heckerman, D.: Continuous time dynamic topic models. In: *Proceedings of UAI'08*, pp. 579–586 (2008)
16. Wang, X.-R., McCallum, A.: Topics over time: A non-Markov continuous-time model of topical trends. In: *Proceedings of KDD'06*, pp. 424–433 (2006)