Adaptation of Topic Model to New Domains Using Recursive Bayes

Ying-Lang Chang
Department of Computer Science and Information Engineering
National Cheng Kung University, Taiwan
ylchang@chien.csie.ncku.edu.tw

Jen-Tzung Chien
Department of Computer Science and Information Engineering
National Cheng Kung University, Taiwan
jtchien@mail.ncku.edu.tw

Abstract

In real-world applications, the topic model using latent Dirichlet allocation (LDA) should be adaptive to meet the unknown domain knowledge. Compensating such domain mismatch assures a robust document representation at different time stamps. This study presents a *recursive Bayes* algorithm to implement an adaptive topic model (ATM) in which LDA is evolved to new domains from adaptation data epoch by epoch without waiting for long batch data. By properly characterizing LDA parameters using the conjugate priors, the reproducible prior/posterior distributions are derived for efficient implementation of ATM. In the experiments, ATM continuously captured the topic evolution and consistently improved the document modeling and categorization from new domain data.

1 Introduction

Latent Dirichlet allocation (LDA) [2] is popular as a mixture topic model for document representation. LDA characterizes the hierarchy of words, topics and documents and generalizes the probabilistic latent semantic analysis based topic model [6] by merging the *Dirichlet priors* for the mixture topics. LDA is deteriorated when the topics and domains are changed in new documents. A possible solution to this circumstance is to retrain LDA using new documents. Such batch training is time consuming and is not generalized well in ill-posed conditions. A flexible LDA should be adaptive at different time stamps instead of performing batch training. This study presents a sequential model to continuously adapt LDA to meet the changing domains from newly-collected documents. Previously, a dynamic topic model (DTM) [3] was proposed to capture the evolution of topics in a single large document collection. A Gaussian noise model was adopted to chain the topic parameters in a state space model. Also, a continuous-time DTM [8] was developed to improve time resolution by a Brownian motion model, and was adopted for dating documents. These DTMs used the distribution of previous topics to evolve new topics in a single document collection. No Bayesian model adaptation was performed. Considering the limited amount of new documents, which are observed incrementally, the information systems are continuously adapted to new topics and domains. Adapting topic model at different epochs is helpful for enhancing system robustness. The online learning algorithms were exploited for speech recognition [4] and information retrieval [5]. The reproducible Dirichlet densities were formulated to build a recursive Bayes (RB) algorithm. In addition, an online LDA [1] was proposed to incrementally capture the thematic patterns and identify the dynamics of topics in text streams. This paper presents a novel adaptive topic model (ATM) where the incremental adaptation is performed to trace the latent topics over time and simultaneously update the LDA parameters and hyperparameters at different learning epochs. Adopting the conjugate prior using Dirichlet density activates an RB learning, which is employed to fulfill a rapid implementation of ATM.

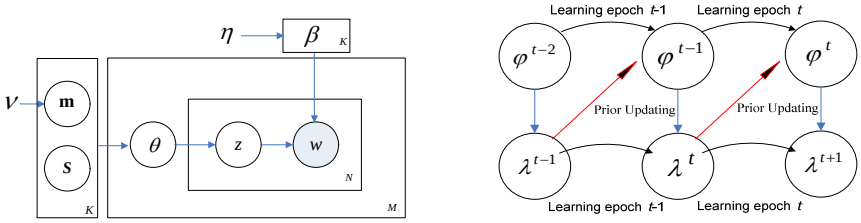
2 Adaptive topic model

2.1 Model construction

An adaptive topic model is presented to meet new domains at different learning epochs. At the *t*-th epoch, the documents $\mathbf{w}^{(t)}$ are collected for model adaptation. Figure 1(a) displays the graphical model of ATM. The hyperparameters $\varphi = \{\nu, \eta\}$ are used to express the variations of model parameters $\lambda = \{\mathbf{m}, s, \beta\}$. In an RB learning procedure [4][5], the parameters $\lambda^t = \{\mathbf{m}^t, s^t, \beta^t\}$ at epoch *t* are recursively estimated by maximizing *a posteriori* distribution, which is equivalent to the product of a likelihood function of current data $\mathbf{w}^{(t)}$ and a prior density given the hyperparameters $\varphi^{t-1} = \{\nu^{t-1}, \eta^{t-1}\}$ updated at epoch $t-1$. The incremental learning is depicted in Figure 1(b). The data generation is described as follows:

1. For each epoch *t* with adaptation data $\mathbf{w}^{(t)}$:
   - Choose the word distributions associated with *K* topics *z* by $\{\beta_k^t \sim \text{Dir}(\eta_k^{t-1})\}$.
   - Choose the mean of the Dirichlet parameters α of topic mixture θ by $\mathbf{m}^t \sim \text{Dir}(\nu^{t-1})$.
2. For each document $\mathbf{w}_d^{(t)}$ and each of *N* words $w_{dn}^{(t)}$:
   The generation process is the same as that of LDA [2].

The newest domain knowledge can be traced through an incremental learning procedure. The overfitting problem is avoided and the model mismatch is resolved at different time stamps.



(a) (b)

Figure 1: (a) Graphical modeling and (b) incremental learning using ATM.

2.2 Dirichlet priors

To realize the RB learning, the hyperparameters $\varphi = \{\nu, \eta\}$ of Dirichlet density are adopted to represent the distortion of model parameters $\lambda = \{\mathbf{m}, s, \beta\}$ due to domain mismatch. At learning epoch *t*, the Dirichlet prior is used to model the topic-based word distribution $\beta_k^t \sim \text{Dir}(\eta_k^{t-1})$. However, the LDA parameter α^t serves as Dirichlet parameter of topic mixture θ_d, which cannot be directly characterized by an Dirichlet density. This situation is manipulated by transforming α^t into a mean $\mathbf{m}^t = [m_1^t, \cdots, m_K^t]^T = [\alpha_1^t / s^t, \cdots, \alpha_K^t / s^t]^T$ and a precision $s^t = \sum_k \alpha_k^t$ [7], and so the properties $\alpha_k^t = s^t m_k^t$ and $\sum_k m_k^t = 1$ are held. Estimating α_k^t is equivalent to separately estimate m_k^t, which represents the mean of topic mixtures over *M* documents $(\bar{\theta}_k = \sum_{d=1}^M \theta_{dk} / M)$, and estimate the precision parameter s^t. Therefore, the Dirichlet prior $\mathbf{m}^t \sim \text{Dir}(\nu^{t-1})$ with hyperparameter ν^{t-1} is adopted due to $\sum_k m_k^t = 1$.

2.3 Incremental learning

To perform the incremental learning, ATM parameters $\lambda^t = \{\mathbf{m}^t, s^t, \beta^t\}$ at learning epoch *t* are recursively estimated by maximizing the product of likelihood function $p(\mathbf{w}^{(t)} \mid \lambda^t)$ and prior density $p(\mathbf{m}^t, \beta^t \mid \varphi^{t-1} = \{\nu^{t-1}, \eta^{t-1}\})$. The hyperparameters are updated by $\varphi^{t-1} \to \varphi^t = \{\nu^t, \eta^t\}$ and are adopted when the next adaptation documents $\mathbf{w}^{(t+1)}$ are enrolled at epoch $t+1$. According to the variational inference using VB-EM procedure [2], the lower bound of the logarithm of posterior distribution is optimized by employing the factorized variation distribution $q(\theta, z \mid \gamma, \varphi) = q(\theta \mid \gamma) q(z \mid \phi)$ where γ and ϕ are parameters associated with latent variables θ and *z*, respectively. In the VB-E step, the derived parameter $\tilde{\phi}$ is the same as that in LDA. Differently, the parameter $\gamma = \{\gamma_k\}$ is derived by $\tilde{\gamma}_k \propto s^t + \log m_k^t + \sum_{n=1}^N \phi_{nk}$. In VB-M

step, the lower bound given the variational distribution $q(\theta, z \mid \hat{\gamma}, \hat{\varphi})$ is maximized to obtain

$$\tilde{m}_k^t \propto \sum_{d=1}^{M}\sum_{n=1}^{N} \phi_{dnk} w_{dn}^{(t)} \exp\{(\Psi(\gamma_k) - \Psi(\sum_{j=1}^{K}\gamma_j)\} + (v_k^{t-1} - 1) \tag{1}$$

$$(\tilde{s}^t)^{-1} = (s^t)^{-1} - \Psi(s^t) + \sum_{k=1}^{K} \tilde{m}_k^t \Psi(s^t \tilde{m}_k^t) - \sum_{k=1}^{K} \tilde{m}_k^t \log \bar{\theta}_k \tag{2}$$

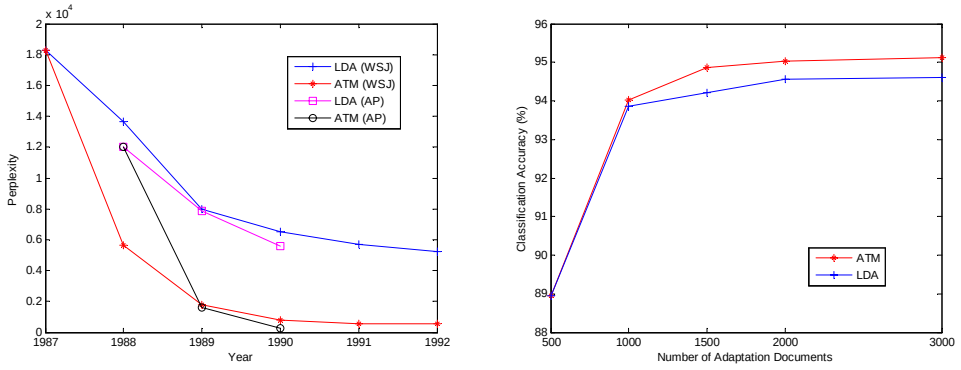
$$\tilde{\beta}_{kw_n}^t \propto \sum_{d=1}^{M}\sum_{n=1}^{N} \phi_{dnk} w_{dn}^{(t)} + (\eta_{kw_n}^{t-1} - 1) \tag{3}$$

where $\Psi(\cdot)$ is a digamma function. Notably, a *closed-form* solution to parameter $\tilde{\alpha}_k^t = \tilde{s}^t \tilde{m}_k^t$ is derived without performing Newton-Raphson algorithm in LDA [2]. Rapid implementation is achieved. New ATM parameters $\tilde{\lambda}^t = \{\tilde{\mathbf{m}}^t, \tilde{s}^t, \tilde{\beta}^t\}$ are then used as the current estimates when running the next VB-EM iteration. The optimal ATM parameters are estimated when VB-EM procedure converges. After performing model adaptation, new hyperparameters of Dirichlet priors are updated by $\varphi^{t-1} \to \varphi^t$ which is derived by arranging the lower bound or equivalently the posterior distribution as a product of Dirichlet densities with the updated hyperparameters

$$\eta_{kw_n}^t = \sum_{d=1}^{M}\sum_{n=1}^{N} \phi_{dnk} w_{dn}^{(t)} + \eta_{kw_n}^{t-1} \tag{4}$$

$$v_k^t = \sum_{d=1}^{M}\sum_{n=1}^{N} \phi_{dnk} w_{dn}^{(t)} \exp\{(\Psi(\gamma_k) - \Psi(\sum_{j=1}^{K}\gamma_j)\} + v_k^{t-1}. \tag{5}$$

Adopting the property of conjugate prior establishes a reproducible prior/posterior pair to fulfill the incremental learning of ATM parameters $\lambda^1 \to \lambda^2 \to \cdots$ and their hyperparameters $\varphi^0 \to \varphi^1 \to \varphi^2 \to \cdots$ which is illustrated in Figure 1(b). Starting from the initial hyperparameters φ^0, ATM *learns without stop* from incrementally observed documents $\{\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \cdots, \mathbf{w}^{(t)}\}$. At each learning epoch, only a small amount of documents $\mathbf{w}^{(t)}$ is required. Such a scenario is practical in real-world information systems.



(a) (b)

Figure 2: Comparison of LDA and ATM in terms of (a) perplexities using WSJ and AP datasets and (b) classification accuracies using Reuters-21578.

3 Experiments

In the experiments, ATM was implemented for adaptive document modeling on two TREC collections WSJ 1987-1992 and AP 1988-1990 [5], which consisted of year-labeled documents. We randomly selected about 3000 and 5000 documents each year from the WSJ and AP collections, respectively. Among these documents, 90% was used for training or adaptation and 10% was used as test data. The stop words were omitted and the stemming was applied. The vocabulary sizes of WSJ and AP were 33835 and 35158, respectively. The number of latent topics was specified by K=20. The incremental learning was performed year by year, i.e. conducting six epochs on WSJ and three epochs on AP. Typically, ATM is viewed as an incremental version of LDA [2], which was originally presented for batch training. Figure 2(a) displays the perplexities of LDA and ATM at different years using WSJ and AP datasets. LDA was trained by using all previous adaptation documents in a single learning epoch. LDA and

ATM consistently reduced the perplexities year after year with increasing number of adaptation documents. ATM had significantly lower perplexities than LDA on both WSJ and AP. Conducting incremental learning was beneficial for capturing the dynamics of topics and domains at various time stamps. LDA-based batch training suffered from high computation and could not effectively match the domain information at different epochs. In our evaluation, ATM additionally spends 33% computation time compared to LDA when running at the same computer. Table 1 shows top ten words of two topics extracted by ATM and their evolution at different learning epochs. The frequent words using ATM were changed epoch by epoch while the evolution of frequent words using DTM [3] was analyzed in a single epoch. The RB learning in ATM alleviates the overfitting problem and should be more efficient than the state space learning in DTM. In addition, LDA and ATM were carried out for binary classification of documents. The categories of EARN and Not EARN in Reuters-21578 database were investigated. This dataset contained 8000 documents with 15818 vocabulary words. We sampled 500, 1000, 1500, 2000 and 3000 documents as the adaptation data and performed incremental learning at different time stamps. At each stamp, an additional test set consisting of about 100 documents was used. The experimental setup was kept the same as that in evaluation of perplexity. The classification accuracies using LDA and ATM were compared in Figure 2(b). ATM consistently performed better than LDA at different epochs.

Table 1: Top ten words associated with two topics using ATM at different epochs.

Topic 1					**Topic 2**				
WSJ 87-88	**WSJ 87-89**	**WSJ 87-90**	**WSJ 87-91**	**WSJ 87-92**	**WSJ 87-88**	**WSJ 87-89**	**WSJ 87-90**	**WSJ 87-91**	**WSJ 87-92**
industry	industry	industry	industry	product	said	said	said	said	said
sail	general	product	product	industry	year	new	new	new	new
general	product	general	general	cost	new	increase	state	state	state
month	close	close	close	general	million	state	invest	increase	increase
product	cost	cost	cost	close	sail	loss	loss	invest	invest
close	major	major	major	major	increase	make	rise	investor	investor
share	month	world	world	world	make	operator	operator	term	operator
trade	world	Japan	foreign	foreign	bank	invest	term	loss	term
cost	maker	maker	level	level	base	rise	investor	operator	loss
company	commit	level	job	job	state	interest	bill	long	long

4 Conclusions

We proposed an efficient approach to adaptively represent documents at different time stamps. The ATM model parameters and hyperparameters were recursively estimated by maximizing the posterior distribution given the current data and the previous hyperparameters. The Dirichlet priors were allocated in different levels including *topic-based language model, topic mixtures* and even the *hyperparameters of topic mixtures*. A new VB-EM algorithm was developed to derive the closed-form solutions to variational parameters and model parameters, and so the Newton-Raphson algorithm was not needed. An online learning procedure was established for robust document categorization. The experimental results on TREC and Reuters datasets showed that the performance of ATM was much better than that of LDA at different epochs. In the future, more comparisons between DTM and ATM will be conducted.

References

[1] AlSumait, L., Barbara, D. & Domeniconi, C. (2008) On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking. In *Proceedings of International Conference on Data Mining*, pp. 3-12.

[2] Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research* 3(5): 993-1022.

[3] Blei, D. M. & Lafferty, J. D. (2006) Dynamic topic model. In *Proceedings of the International Conference on Machine Learning* 148, pp. 113-120.

[4] Chien, J.-T. (1999) Online hierarchical transformation of hidden Markov models for speech recognition. *IEEE Transactions on Speech and Audio Processing* 7(6): 656-667.

[5] Chien, J.-T. & Wu, M.-S. (2008) Adaptive Bayesian latent semantic analysis. *IEEE Transactions on Audio, Speech and Language Processing* 16(1): 198-207.

[6] Hofmann, T. (1999) Probabilistic latent semantic indexing. In *Proceedings of ACM SIGIR*, pp. 35-44.

[7] Minka, T. (2000) Estimating a Dirichlet distribution. *Technical Report*, MIT.

[8] Wang, C., Blei, D. & Heckerman, D. (2008) Continuous time dynamic topic models. In *Proceedings of Uncertainty in Artificial Intelligence*, pp. 579-586.