

# A Biterm-based Dirichlet Process Topic Model for Short Texts

Yali Pan, Jian Yin<sup>1</sup>, Shaopeng Liu, Jing Li

Department of Computer Science  
Sun Yat-Sen University  
Guangzhou, P. R. China  
Email: panyali@mail2.sysu.edu.cn

**Abstract**—Topic models are prevalent in many fields (e.g. context analysis), which are applied to discovering the latent topics. In document modeling, conventional topic models (e.g. latent Dirichlet allocation and its variants) do well for normal documents. However, the severe data sparsity problem makes the topic modeling in short texts difficult and unreliable. To tackle this problem, an effective approach (biterm topic model) has been proposed recently which learns topics by directly modeling the generation of word co-occurrence patterns at corpus-level rather than at document-level. But it requires human intervention for determining the number of topics. In this paper, we propose a Dirichlet process based on word co-occurrence to make topic mining from short texts more automatically. Meanwhile, we design a Markov chain Monte Carlo sampling scheme for posterior inference in our model which is an extension of the sampling algorithm based on Chinese restaurant process. Finally, we conduct experiments on real data. The results show that our method outperforms the baseline on quality of topic and perplexity and it is more flexible.

**Keywords**- *Dirichlet Process; Clustering; Biterm; Short Texts; Topic Mining;*

## I. INTRODUCTION

Mining the hidden topics on the Web has been a focus over the past decade. With the development of social media, it is crucial that we need to keep up with the hot topics and huge data brings us rich information which has tremendous commercial value as well. However, the lack of rich context in short texts makes the topic modeling more challenging. In general, conventional topic models reveal the latent topics by implicitly capturing the document-level word co-occurrence patterns. Therefore, directly applying these models on short texts will suffer from the severe data sparsity problem [8]. Therefore, in conventional methods, including Bayesian parametric models (e.g. pLSI [2], LDA [1]) and Bayesian nonparametric models (e.g. HDP [5], dHDP [9]), we always struggle to combine these short texts with extra information or to make stronger assumptions before training. Different from these, BTM [11] modeled the whole corpus by extracting the word co-occurrence patterns. It strengthens the accuracy of capturing the topics and could be easy to implement.

Inspired by BTM, we propose a Bayesian nonparametric approach, which called biterm-based Dirichlet process (BDP) to tackle the sparsity problem. We describe a representation of Dirichlet process in terms of a Chinese restaurant process (CRP [4]) and design a Markov chain Monte Carlo (MCMC)

algorithm for posterior inference in BDP. Finally, we conduct experiments on a real-world short text collection. Experimental results show BDP can discover more prominent and coherent topics than baseline methods from two perspectives: the quality of topics and the perplexity.

This paper is organized as follows: Section 2 gives a brief review of related works, including popular topic models, Dirichlet process and its extensions. Section 3 discusses our method and we present the experimental results in section 4. At last, in Section 5, some conclusions are made.

## II. RELATED WORKS

A topic model is a type of generative probabilistic model for discovering the latent semantic structure from a collection of documents. In the last decade, many complicated topic models are proposed as variants and extensions of LDA or pLSI. They are useful on normal texts, but do not work well on short texts. To improve the efficiency of these topic models, some early studies mainly focused on three approaches as follows: (1) aggregate short texts into lengthy pseudo-documents before training [8]; (2) assume that a short document only covers a single topic [10]; (3) exploit external knowledge to enrich the representation of short text [7]. All the above methods may become invalid in a general case since most of them need extra knowledge and the single-topic assumption is often unreasonable in real data.

In [11], Yan proposed alternative solution. The model regards the whole corpus as a mixture of topics where each biterm (word-pair) is drawn from a specific topic independently. However, as a parametric model, BTM has its limitations. It is not efficient to decide the number of topics, especially when the number is dynamic and the dataset is large. For capturing the appropriate number of topics, researchers have proposed many nonparametric models. Among these models, HDP makes a prominent role which is a hierarchy generalization of the Dirichlet process. Dirichlet Process [12] is well suited for the problem of placing prior distribution on mixture components in mixture modeling. Meanwhile, it provides an interesting way to understand group assignments and models for clustering effects. Generally, there are three approaches to construct Dirichlet process: stick-breaking, Pólya urn model and Chinese restaurant process. To be suitable in HDP, CRP has been extended to Chinese restaurant franchise (CRF) which means that there are two level CRPs in HDP, corpus-level and document-level. Different from HDP, we apply Dirichlet process only at the corpus-level and regard the whole corpus as a biterm set.

<sup>1</sup> Corresponding author. Email: issjyin@mail.sysu.edu.cn

### III. BITERM-BASED DIRICHLET PROCESS TOPIC MODEL

In [11], a biterm denotes an unordered word-pair co-occurring in a short document. When a biterm is assigned to a topic, the words in it are more likely belong to a same topic than a single word. For example, the words “obama” and “white” co-occur with each other frequently, it is possible that they belong to the same topic “politics”. Inspired by this idea, we explore a new model which we refer to as biterm-based Dirichlet process (BDP). It follows the process as a Dirichlet process mixture model but regards the observations as biterms. Figure 1 illustrates the graphical representation of BTM and BDP.

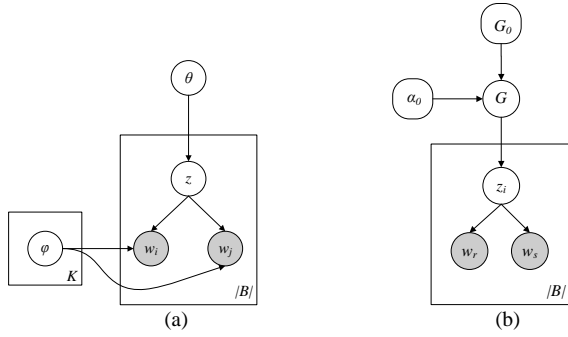


Figure 1: (a) BTM. (b) BDP

Firstly, we need extract two distinct words from all the documents to compose the training data and then directly model the word co-occurrence patterns based on it. In BDP, the observations  $x_i = (w_r, w_s)$  arise as follows:

$$\begin{aligned} z_i | G &\sim G \\ w_r, w_s | z_i &\sim \text{Mult}(z_i) \end{aligned} \quad (1)$$

Where  $w_r$  and  $w_s$  is drawn from multinomial distribution given a topic  $z_i$ .  $\alpha_0$  is the concentration parameter,  $G \square DP(\alpha_0, G_0)$ ,  $G_0 \square \text{Dirichlet}(\eta)$ .

With MCMC sampling, we can easily obtain the global topic distribution  $\theta$  and the topic-word distribution  $\phi$  except the topic-document distribution. However, as [11] assume, we could compute the topic-document distribution as follows:

$$P(z | d) = \sum_x P(z | x) P(x | d) \quad (2)$$

Where  $P(z | x)$  can be easily calculated according to Bayes' formula and  $P(x | d)$  is nearly a uniform distribution over all the biterms in the document. Their expressions are:

$$P(z | x) = \frac{P(z)P(w_r | z)P(w_s | z)}{\sum_z P(z)P(w_r | z)P(w_s | z)} \quad (3)$$

$$P(x | d) = \frac{n_d(x)}{\sum_x n_d(x)} \quad (4)$$

In (3) and (4),  $x = (w_r, w_s)$ ,  $P(z) = \theta_z$  and  $P(w_i | z) = \phi_{w_i | z}$ .  $n_d(x)$  is the frequency of the biterm  $x$  in the document  $d$ .

#### A. Chinese Restaurant Process based on Biterms

Chinese restaurant process (CRP) construction is used to produce samples from the prior distribution over the  $z_i$  in (1). Consider a Chinese restaurant with an unbounded number of tables. A couple enters into the restaurant, then they will choose a table. They sit the old table to share the dish with probability proportional to  $n_k$ , and choose a new table with probability proportional to  $\alpha_0$ . Formally, we define  $z_1, \dots, z_K$  to be the distinct values taken on by  $x_1, \dots, x_{i-1}$  and let  $n_k$  be the number of values  $x_i$  that are equal to  $z_k$  for  $1 \leq i' < i$ . We have:

$$x_i | x_1, x_2, \dots, x_{i-1}, \alpha_0, G_0 \sim \sum_{k=1}^K \frac{n_k}{i-1+\alpha_0} \delta_{z_k} + \frac{\alpha_0}{i-1+\alpha_0} G_0 \quad (5)$$

Where  $x_i = (w_r, w_s)$ .

#### B. Parameters Inference

In this section, we describe a MCMC sampling scheme for posterior inference given observations  $\mathbf{x} = \{x_1, x_2, \dots, x_{|B|}\}$  where  $x_i$  is a biterm  $(w_r, w_s)$ ,  $|B|$  is the size of biterm set,  $\mathbf{z} = \{z_1, z_2, \dots, z_K\}$ ,  $K$  is the number of the existed topics.  $\mathbf{x}^{-i}$  means all the observation except  $x_i$ .

First of all, we should define the conditional probability of  $x_i$  given all the biterms in topic  $z_k$  except  $x_i$ . According to (1), we have:

$$\begin{aligned} f_k^{-x_i}(x_i) &= p(x_i | \mathbf{x}^{-i}, \mathbf{z}) = \frac{p(\mathbf{x} | \mathbf{z})}{p(\mathbf{x}^{-i} | \mathbf{z})} = \\ &= \frac{\int f(x_i | z_k) \prod_{i' \neq i, i'=z_k} f(x_{i'} | z_k) g(z_k | \eta) d(z_k)}{\int \prod_{i' \neq i, i'=z_k} f(x_{i'} | z_k) g(z_k | \eta) d(z_k)} \end{aligned} \quad (6)$$

Where  $f(x_i | z_k) = f(w_r | z_k) f(w_s | z_k)$ ,  $f(\cdot)$  is the density of  $\text{Mult}(z)$  and  $g(\cdot)$  is the density of  $\text{Dirichlet}(\eta)$ . Taking advantage of conjugate priors, (6) can be simplified as follows:

$$f_k^{-x_i}(x_i = (w_r, w_s)) = \frac{(n_{w_r | z_k} + \eta)(n_{w_s | z_k} + \eta)}{(\sum_w n_{w | z_k} + N\eta)^2} \quad (7)$$

Where  $n_{w | z_k}$  is the number of word  $w$  given topic  $z_k$ , and  $N$  is the number of words in the vocabulary.

1) *Generate  $k_i$* : Considering exchangeability, we treat  $k_i$  as the last variable being sampled.  $k_i$  is the index of topic  $z_k$  which the observation  $x_i$  is assigned to. Combined with the likelihood of generating  $x_i$ , the conditional posterior for  $k_i$  is:

$$\begin{aligned} p(k_i = k | \mathbf{z}^{-i}, \mathbf{x}) &\propto \\ &\begin{cases} n_k f_k^{-x_i}(x_i), & k = [1, \dots, K] \\ \alpha_0 f_{k_{\text{new}}}^{-x_i}(x_i), & k = k_{\text{new}} \end{cases} \end{aligned} \quad (8)$$

Where  $n_k$  is the number of biterns assigned to topic  $z_k$  except  $x_i$ .

2) *Update  $\phi$  and  $\theta$* : Conditioned on  $\mathbf{z}$  and  $\mathbf{x}$ , the posterior distribution for  $z_k$  is dependent only on the data items:

$$p(z_k | \mathbf{z}, \mathbf{x}, \phi^{-k}) \propto g(z_k | \eta) \prod_{i: k_i = k} f(x_i | z_k) \quad (9)$$

After topic assignments for bitern and word occurrences, we can compute the topic-word distribution  $\phi$  and global topic distribution  $\theta$  simply as follows:

$$\phi_{w|z} = \frac{n_{w|z} + \eta}{\sum_w n_{w|z} + N\eta} \quad (10)$$

$$\theta_z = \frac{n_z + \alpha_0}{|B| + K\alpha_0} \quad (11)$$

We show the above process is shown in Algorithm 1.

**Algorithm 1:** MCMC sampling algorithm for BDP

**Input:** bitern set  $B$ , hyperparameters  $\alpha_0, \eta$

**Output:** number of topics  $K$ , multinomial parameter  $\phi$  and  $\theta$

**REPEAT**

**FOR**  $x \in B$  **DO**

Draw the topic index  $k$  from (8)

**IF**  $k = k_{\text{new}}$  **THEN**

$K = K + 1$

**END IF**

Update  $n_z, n_{w|z}, n_{w|z}$

**END FOR**

Compute  $\phi$  and  $\theta$  as (10) and (11)

**UNTIL**  $iter = N_{\text{iter}}$  or other conditions

**RETURN**  $K, \phi, \theta$

#### IV. EXPERIMENTS

In this section, we study BDP topic model on real data, Tweets2011 which contained approximately 16 million tweets sampled between January 23rd and February 8th, 2011. After data preprocessing, we left 164,121 tweets which include 16,021 distinct words in vocabulary and 1,483,966 words in the whole corpus. In BDP, we get 5,684,168 biterns. To demonstrate the effectiveness of our approach, we take three topic models as baseline methods: LDA, HDP and BTM. We evaluate the performance on two typical metrics: quality of topics and perplexity. After parameters optimization, in LDA and BTM, we set the number of topics  $K=50$ , hyperparameters  $\alpha = 0.5$ ,  $\beta = 0.01$ . And in HDP, we set the base Dirichlet parameter  $\eta = 0.5$ , hyperparameters  $\gamma$  and  $\alpha_0$  with Gamma priors:  $\gamma \propto \Gamma(1, 0.1)$ ,  $\alpha_0 \propto \Gamma(1, 1)$ . And In BDP, we set the base Dirichlet parameter  $\eta = 0.01$ , hyperparameters  $\alpha_0$  with Gamma priors:  $\alpha_0 \propto \Gamma(1, 1)$ .

##### A. Quality of Topics

We sample 2 topics shared discovered by all the test methods and collect the top 10 words for each topic (shown

in Table I). Then we use the metric *topic coherence* to assess topic quality more objectively:

$$C(t; V^{(t)}) = \sum_{m=2}^M \sum_{l=1}^{m-1} \log((D(v_m^{(t)}, v_l^{(t)}) + 1) / D(v_l^{(t)})) \quad (12)$$

Higher scores mean better performance. To evaluate the overall quality of a topic set, we calculated the average coherence score (demonstrated in Table II) for each method:

$$C_{\text{avg}} = \frac{1}{K} \sum_{t_k} C(t_k; V^{(t_k)}) \quad (13)$$

We can see that our method is better than the baseline methods in terms of average coherence score. In Table I, the topics discovered by our model represent good intelligibility. However, when detecting the detail in Fig. 3, we find that the topics in BDP have a wide range of quality. Though BDP determines the value of  $K$ , there is a problem to ensure high quality for all the topics.

TABLE I. TOP 10 WORDS FOR SELECTED TOPICS

Topic 1			
LDA	HDP	BTM	BDP
egypt jan25 mubarak news police egyptian al cairo tahrir internet	egypt jan25 news mubarak obama people police state egyptian tahrir	egypt jan25 mubarak egyptian cairo tahrir people al police protesters	egypt jan25 mubarak people egyptian cairo tahrir police al protesters
Topic 2			
LDA	HDP	BTM	BDP
game team great fans superbowl steelers packers fan nfl won	video game win super music bowl rt live team fan	super bowl green green packers steelers yellow black superbowl game bay	game green super packers nfl team superbowl pittsburgh win sports

TABLE II. THE AVERAGE OF TOPIC COHERENCE SCORES

	LDA	HDP	BTM	BDP
AVG.	-897.72	-765.29	-758.86	-711.42

##### B. Perplexity

The perplexity of a held-out test set consisting of  $M$  documents is defined to be:

$$\text{perplexity}(D_{\text{test}}) = \exp\left(-\sum_j \sum_i^J \log p(w_{ji})\right) / \sum_j N_j \quad (14)$$

A lower perplexity score always indicates better generalization performance. The results are shown in Fig. 2. It shows that BDP has a lower perplexity, but it is trapped in local optimum like HDP does. Even though it is not the best one, BDP provides a stable performance and works well at

the beginning of iterations. Moreover, the result shows that BDP is able to find more diverse topics than any other baselines.

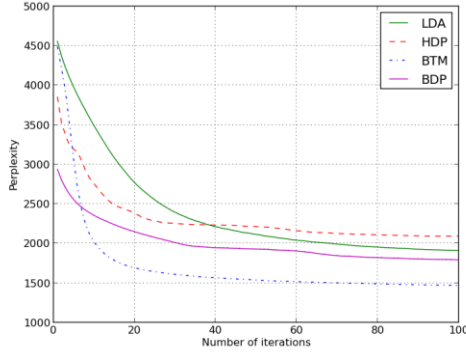


Figure 2: Perplexity of models

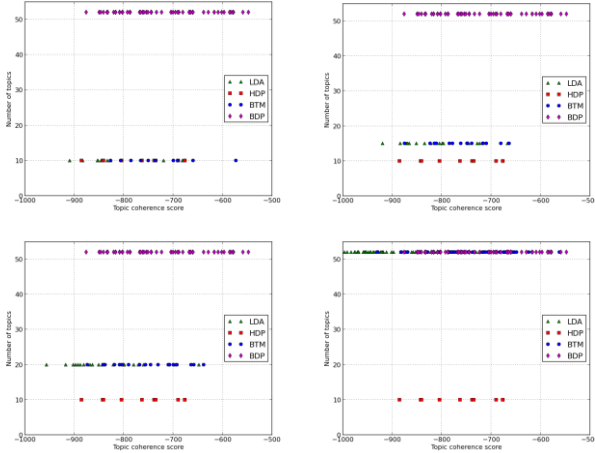


Figure 3: Coherence score for each topic. For LDA and BTM,  $K=10, 15, 20, 52$ ; For HDP,  $K=10$ ; For BDP,  $K=52$ .

## V. CONCLUSION AND FUTURE WORKS

In this paper, we proposed a novel topic model based on Dirichlet process which can capture the topics in short texts well. The idea is inspired by BTM which considers word-occurrence patterns in the whole corpus rather than single word in every document. We conduct experiments on real data to prove the effectiveness of our model. It works well even though the test data is sparse. Besides, the process of its

modeling is more flexible since users need not determine the number of topics.

However, there are still a lot of problems to deal with. For example, the sampling algorithm is labored and costs more memory space in terms of large-scale data. Improving the process of sampling will be significant in our future work. Moreover, it is interesting and efficient to extend our model with external knowledge in real environment.

## ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (61272065), Natural Science Foundation of Guangdong Province (S2011020001182, S2012010009311), Research Foundation of Science and Technology Plan Project in Guangdong Province and Guangzhou City (2011B040200007, 2012A010701013, 11A31090341, 11A53010726, 2011Y5-00004).

## REFERENCES

- [1] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993-1022, 2003.
- [2] T. Hofmann. Probabilistic Latent Semantic Indexing. In *SIGIR*, pages 50-57. ACM, 1999.
- [3] Y. Tech. Dirichlet Process. *Encyclopedia of Machine Learning*, Springer, 2010.
- [4] M. Jordan. Dirichlet Process, Chinese Restaurant Processes, and all that. In *Proceedings of the Tutorial Presentation at the NIPS Conference*, 2005.
- [5] Y. Tech, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet Process. *The Journal of the American Statistical Association*, 101(476):1566-1581, 2006.
- [6] D. Ramage, S. Dumias, and D. Liebling. Characterizing Microblogs with Topic Models. In *International AAAI Conference on Weblogs and Social Media*, 2010.
- [7] X. Phan, L. Nguyen, and S. Horiguchi. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections. In *Proceedings of the 17th international conference on World Wide Web*, 2008.
- [8] L. Hong and B. Davison. Empirical Study of Topic Modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80-88. ACM, 2010.
- [9] L. Ren, D. Dunson, and L. Carin. The Dynamic Hierarchical Dirichlet Process. In *Proceedings of the 25th international conference on Machine Learning*, 2008.
- [10] W. Zhao, J. Jiang, J. Weng, J. He, E. Lim, H. Yan, and X. Li. Comparing Twitter and Traditional Media using Topic Models. *Advances in Information Retrieval*, pages 338-349, 2011.
- [11] X. Yan, J. Guo, Y. Lan, and X. Cheng. A Biterm Topic Model for Short Texts. In *Proceedings of the 22th international conference on World Wide Web*, 2013.
- [12] T. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *Annals of Statistics*, 1(2):209-230, 1973.