User Based Aggregation for Biterm Topic Model

Weizheng Chen, Jinpeng Wang, Yan Zhang , Hongfei Yan and Xiaoming Li

School of Electronic Engineering and Computer Science, Peking University, China
{cwz.pku,wjp.pku,yhf1029}@gmail.com, zhy@cis.pku.edu.cn, lxm@pku.edu.cn

Abstract

Biterm Topic Model (BTM) is designed to model the generative process of the word co-occurrence patterns in short texts such as tweets. However, two aspects of BTM may restrict its performance: 1) user individualities are ignored to obtain the corpus level words co-occurrence patterns; and 2) the strong assumptions that two co-occurring words will be assigned the same topic label could not distinguish background words from topical words. In this paper, we propose Twitter-BTM model to address those issues by considering user level personalization in BTM. Firstly, we use user based biterms aggregation to learn user specific topic distribution. Secondly, each user's preference between background words and topical words is estimated by incorporating a background topic. Experiments on a large-scale real-world Twitter dataset show that Twitter-BTM outperforms several state-of-the-art baselines.

1 Introduction

In recent years, short texts are increasingly prevalent due to the explosive growth of online social media. For example, about 500 million tweets are published per day on Twitter[1], one of the most popular online social networking services. Probabilistic topic models (Blei et al., 2003) are broadly used to uncover the hidden topics of tweets, since the low-dimensional semantic representation is crucial for many applications, such as product recommendation (Zhao et al., 2014), hashtag recommendation (Ma et al., 2014), user interest tracking (Sasaki et al., 2014), sentiment analysis (Si et al., 2013). However, the scarcity of context and the noisy words restrict LDA and its variations in topic modeling over short texts.

Previous works model topic distribution at three different levels for tweets: 1) document, the standard LDA assumes each document is associated with a topic distribution (Godin et al., 2013; Huang, 2012). LDA and its variations suffer from context sparsity in each tweet. 2) user, user based aggregation is utilized to alleviate the sparsity problem in short texts (Weng et al., 2010; Hong and Davison, 2010). In these models, all the tweets of the same user are aggregated together as a pseudo document based on the observation that the tweets written by the same user are more similar. 3) corpus, BTM (Yan et al., 2013) assumes that all the biterms (co-occurring word pairs) are generated by a corpus level topic distribution to benefit from the global rich word co-occurrence patterns.

As far as we know, how to incorporate user factor into BTM has not been studied yet. User based aggregation has proven effective for LDA. But unfortunately, our preliminary experiments indicate that simple user-based aggregation for BTM will generate incoherent topics. To distinguish between commonly used words (e.g., *good*, *people*, etc) and topical words (e.g., *food*, *travel*, etc), a background topic is often incorporated into the topic models. Zhao et al. (2011) use a background topic in Twitter-LDA to distill discriminative words in tweets. Sasaki et al. (2014) reduce the perplexity of Twitter-LDA by estimating the ratio between choosing background words and topical words for each user. They both make a very strong assumption that one tweet only covers one topic. Yan et al. (2015) use a background topic to distinguish between common biterms and bursty biterms, which need external data to evaluate the burstiness of each biterm as prior knowledge. Unlike those above, we incorporate a background

[1] See https://about.Twitter.com/company

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 489–494,
Beijing, China, July 26-31, 2015.

topic to absorb non-discriminative common words in each biterm. And we also estimate the user's preference between common words and topical words. Our new model is named as Twitter-BTM, which combines user based aggregation and the background topic in BTM. Finally, experiments on a Twitter dataset show that Twitter-BTM not only can discover more coherent topics but also can give more accurate topic representation of tweets compared with several state-of-the-art baselines.

We organize the rest of the paper as follows. Section 2 gives a brief review for BTM. Section 3 introduces our Twitter-BTM model and its implementation. Section 4 describes experimental results on a large-scale Twitter dataset. Finally, Section 5 contains a conclusion and future work.

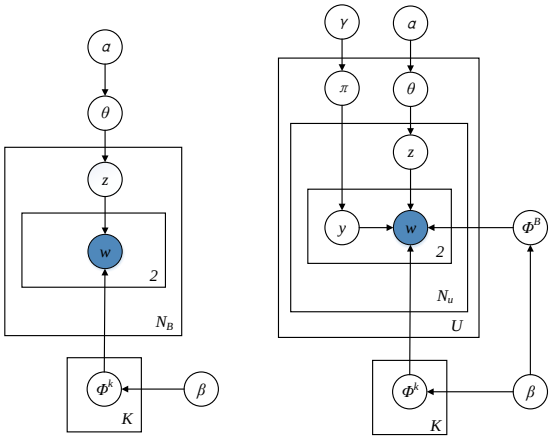
2 BTM

There are two major differences between BTM and LDA (Yan et al., 2013). For one thing, considering a topic is a mixture of highly correlated words, which implies that they often occur together in the same document, BTM models the generative process of the word co-occurrence patterns directly. Thus a document made up of n words will be converted to C_n^2 biterms. For another, LDA and its variants suffer from the severe data sparsity in short documents. BTM uses global co-occurrence patterns to model the topic distribution over corpus level instead of document level.

The graphical representation of BTM (Yan et al., 2013) is shown in Figure 1(a). It assumes that the whole corpus is associated with a distributions θ over K topics drawn from a Dirichlet prior $Dir(\alpha)$. And each topic t is associated with a multinomial distribution ϕ^t over a vocabulary of V unique words drawn from a Dirichlet prior $Dir(\beta)$. The generative process for a corpus which consists of N_B biterms $\mathbb{B} = \{b_1, ..., b_{N_B}\}$, where $b_i = (w_{i_1}, w_{i_2})$, is as follows:

1 For each topic t=1,...,T
  (a) Draw $\phi^t \sim Dir(\beta)$
2 For the whole tweets collection
  (a) Draw $\theta \sim Dir(\alpha)$
3 For each biterm b = 1,...,N_B
  (a) Draw $z_b \sim Multi(\theta)$
  (b) Draw $w_{b,1}, w_{b,2} \sim Multi(\phi^{z_b})$

In the above process, z_b is the topic assignment latent variable of biterm b. To infer the parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, collapsed Gibbs sampling algorithm (Griffiths and Steyvers, 2004) is used for approximate inference.



(a) BTM (b) Twitter-BTM

Figure 1: Graphical representation of (a) BTM, (b) Twitter-BTM

Compared with the strong assumption that a short document only covers a single topic (Diao et al., 2012; Ding et al., 2013), BTM makes a looser assumption that two words will be assigned the same topic label if they have co-occurred. Thus a short document could cover more than one topic, which is more close to the reality. But this assumption causes another issue, those commonly used words and those topical words are treated equally. Obviously it is inappropriate to assign same topic label to those words.

3 Twitter-BTM

In this Section, we introduce our Twitter-BTM model. Figure 1(b) shows the graphical representation of Twitter-BTM. The generative process of Twitter-BTM is as follows:

1 Draw $\phi^B \sim Dir(\beta)$
2 For each topic t=1,...,T
  (a) Draw $\phi^t \sim Dir(\beta)$
3 For each user u=1,...,U
  (a) Draw $\theta^u \sim Dir(\alpha), \pi^u \sim Beta(\gamma)$
  (b) For each biterm b = 1,...,N_u
    (i) Draw $z_{u,b} \sim Multi(\theta^u)$
    (ii) For each word n = 1,2
      (A) Draw $y_{u,b,n} \sim Bern(\pi^u)$
      (B) if $y_{u,b,n} = 0$ Draw $w_{u,b,n} \sim Multi(\phi^B)$
      if $y_{u,b,n} = 1$ Draw $w_{u,b,n} \sim Multi(\phi^{z_{u,b}})$

In the above process, user u's topic interest θ^u is a multinomial distribution over K topics drawn from a Dirichlet prior $Dir(\alpha)$. The background topic B is associated with a multinomial distribution ϕ^B drawn from a Dirichlet prior $Dir(\beta)$. The assumption that each user has a different preference between topical words and background words is shown to be effective in (Sasaki et al., 2014). We adopt this assumption in Twitter-BTM. User u's preference is represented as a Bernoulli distribution with parameter π^u drawn from a beta prior $Beta(\gamma)$. N_u is the number of biterms of user u, $z_{u,b}$ is the topic assignment latent variable of user u's biterm b. For user u and his/her biterm b, n=1 or 2, we use a latent variable $y_{u,b,n}$ to indicate the word type of the word $w_{b,n}$. When $y_{u,b,n} = 1$, $w_{b,n}$ is generated from topic $z_{u,b}$. When $y_{u,b,n} = 0$, $w_{b,n}$ is generated from the background topic B.

We adopt collapsed Gibbs Sampling to estimate the parameters. Because of the limitations of space, we leave out the details about the sampling algorithm. Since we can't get a document's distribution over topics from the parameters estimated by Twitter-BTM directly, we utilize the following formula (Yan et al., 2013) to infer the topic distribution of document d. Given a document d whose author is user u:

$$P(z = t|d) = \sum_{i}^{N_b} P(z = t|b_i)P(b_i|d) \quad (1)$$

Now the problem is converted to how to estimate $P(b_i|d)$ and $P(z = t|b_i)$. $P(b_i|d)$ is estimated by empirical distribution in d:

$$P(b_i|d) = \frac{N_{b_i}}{N_b} \quad (2)$$

where N_{b_i} is the number of biterm b_i occurred in d, N_b is the total number of biterms in d. We can apply Bayes' rule to compute $P(z = t|b_i)$ via following expression:

$$\frac{\theta_t^u \left[\pi^u \phi^B_{w_{i,1}} + (1-\pi^u)\phi^t_{w_{i,1}}\right]\left[\pi^u \phi^B_{w_{i,2}} + (1-\pi^u)\phi^t_{w_{i,2}}\right]}{\sum_k \theta_k^u \left[\pi^u \phi^B_{w_{i,1}} + (1-\pi^u)\phi^k_{w_{i,1}}\right]\left[\pi^u \phi^B_{w_{i,2}} + (1-\pi^u)\phi^k_{w_{i,2}}\right]} \quad (3)$$

4 Experiments

In this Section, we describe our experiments carried on a Twitter dataset collected form 10th Jun, 2009 to 31st Dec, 2009. Stop words and words occur less than 5 times are removed. We also filter tweets which only have one or two words. All letters are converted into lower case. The dataset is divided into two parts. The first part whose statistics is shown in Table 1 is used for training. The second part which consists of 22,496,107 tweets is used as the external dataset in topic coherence evaluation task in Section 4.1.

We compare the performance of Twitter-BTM with five baselines:

- LDA-U, user based aggregation is applied before training LDA.
- Twitter-LDA (Zhao et al., 2011), which makes a strong assumption that a tweet only covers one topic.
- TwitterUB-LDA (Sasaki et al., 2014), an improved version of Twitter-LDA, which models the user level preference between topical words and background words.
- BTM (Yan et al., 2013), the Biterm Topic Model.
- BTM-U, a simplified version of Twitter-BTM without background topic.

For all the above models, we use symmetric Dirichlet priors. The hyperparameters are set as follows: for all the models, we set $\alpha = 50/K$, $\beta = 0.01$; for Twitter-LDA, TwitterUB-LDA and Twitter-BTM, we set $\gamma = 0.5$. We run Gibbs sampling for 400 iterations.

DataSet	Twitter
#tweets	1,201,193
#users	12,006
#vocabulary	71,038
#avgTweetLen	7.04

Table 1: Summary of dataset

Perplexity metric is not used in our experiments since it is not a suitable evaluation metric for BTM (Cheng et al., 2014). The first reason is that BTM and LDA optimize different likelihood. The second reason is that topic models which have better perplexity may infer less semantically topics (Chang et al., 2009).

4.1 Topic Coherence

We use PMI-Score (Newman et al., 2010) to quantitatively evaluate the quality of topic component.

K	50			100		
method	Top5	Top10	Top20	Top5	Top10	Top20
LDA-U	2.83±0.07	1.93±0.06	1.40±0.04	3.11±0.09	1.89±0.09	1.15±0.04
Twitter-LDA	2.58±0.04	1.90±0.03	1.39±0.03	2.97±0.20	1.98±0.09	1.44±0.06
TwitterUB-LDA	2.57±0.05	1.87±0.07	1.45±0.04	3.07±0.11	2.05±0.05	1.45±0.05
BTM	2.88±0.14	2.01±0.09	1.44±0.08	3.25±0.14	2.13±0.06	**1.49±0.06**
BTM-U	2.92±0.10	1.89±0.05	1.33±0.04	3.03±0.07	1.95±0.05	1.34±0.07
Twitter-BTM	**3.04±0.10**	**2.05±0.08**	**1.47±0.05**	**3.27±0.12**	**2.15±0.08**	1.48±0.05

Table 2: PMI-Score of different topic models

Equation (4) defines PMI (Pointwise Mutual Information) for two words w_i and w_j:

$$PMI(w_i, w_j) = \log \frac{P(w_i, w_j) + \epsilon}{P(w_i)P(w_j)} \quad (4)$$

ϵ is an extremely small constant (Stevens et al., 2012), which is equal to 10^{-12} in this paper. The word probabilities and the co-occurrence probabilities are computed on the large-scale external dataset empirically. Here we use the second part Twitter dataset as the external dataset. Then for a topic t and its top T words ranked by topic-word probability ϕ_w^t, the PMI-Score of topic t is defined as follow:

$$PMI - Score(t) = \frac{1}{T(T-1)} \sum_{1 \le i < j \le T} PMI(w_i, w_j) \quad (5)$$

The model's PMI-Score is defined as the mean of all the topics' PMI-Score. Table 2 shows the average results over 10 runs of different models. When $K = 50$, Twitter-BTM outperforms all other models significantly. When $K = 100$, The PMI-Score of BTM and Twitter-BTM are very close. BTM-U is worse than BTM, the reason may be that each user's biterm sets provide extremely limited words co-occurring information.

Table 3 shows top 10 words of topic "food" learned by BTM, BTM-U and Twitter-BTM when $K = 50$. We use italic fonts to indicate background words labeled by human judgement. Compared with BTM and BTM-U, Twitter-BTM can rank those background words at lower level. It demonstrates that representative words learned by Twitter-BTM are more coherent and meaningful.

BTM	BTM-U	Twitter-BTM
food	food	vegan
eat	vegan	food
chicken	eat	eat
good	***good***	chicken
vegan	chicken	chocolate
lol	#vegan	cheese
cheese	cream	cream
chocolate	cheese	#vegan
love	chocolate	ice
dinner	ice	dinner

Table 3: Top 10 words of topic food

4.2 Document Representation

Topic models are powerful dimension reduction methods for texts. Given a tweet d, we can infer its probability distribution over K topics with equation (1). Thus d can be represented as a topic probability vector:

$$d = [P(z = 1|d), ..., P(z = K|d)] \quad (6)$$

We use document classification task (Cheng et al., 2014) and document clustering task (Duan et al., 2012) to measure the quality of the documents' topic proportions. Tweets in Twitter have no explicit label information. But some tweets are labeled by one or more hashtags (a type of label whose form is "#keyword") manually by its author to indicate the topic the tweets involve. We follow previous works (Cheng et al., 2014; Wang et al., 2014) and use hashtags as the tweets' labels. Table 4 lists 38 frequent (at least appears in 100 tweets) hashtags relating to certain topic or event manually selected in our dataset.

We choose those tweets which contain only one of these hashtags appear in Table 4 from our original data in the following experiments. When we infer a tweet's topic distribution, the hashtag is ignored. Because it doesn't make sense to use the label information to construct the feature vector directly.

We classify these selected tweets by Random Forest classifier (Breiman, 2001) implemented in

aaliyah afghanistan beatcancer birding blogtalkradio digguser dmv dontyouhate fact giladshalit gno gov green haiku healthcare honduras india iranelection jazz jesus krp lgbt mindsetshift nfl nn oink rhoa slaughterhouse socialmedia tech travel trueblood vegan vegas voss weeklyfitnesschallenge wordpress yyj

Table 4: Hashtags selected for evaluation

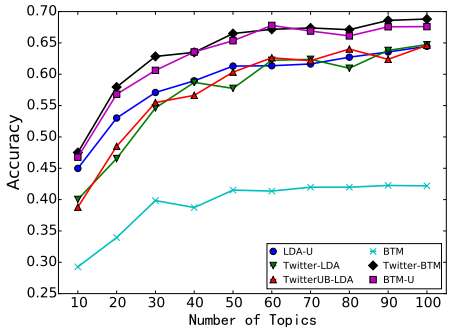


Figure 2: Performance of classification

sklearn [2] python module with 10-fold cross validation. Using accuracy as the evaluation metric, we report the classification performance of different topic models in Figure 2. With the increase of the topic number K, all the models' accuracies are tending to increase. BTM is worse than all other models, which confirms the effectiveness of user based aggregation. Twitter-BTM and BTM-U always outperform LDA-U, Twitter-LDA and TwitterUB-LDA. Twitter-BTM's accuracy is a little higher than BTM-U, which demonstrates that the background topic is helpful to capture more accurate topic representation of documents.

We adopt k-means algorithm implemented in sklearn python module as our clustering method. The number of cluster is set to 38. Considering we have the knowledge of ground truth class assignments of each tweet, and Adjusted Rand Index (ARI) and Normalized Mutual Information are used as cluster validation indices in our experiments. As shown in Figure 3 and Figure 4, The higher ARI and NMI value indicate that Twitter-BTM outperform other models. And BTM performs worse than all other models.

5 Conclusion

In this paper, we investigate the problem of topic modeling over short texts with user factor. User individualities are sacrificed to obtain the corpus level words co-occurrence patterns in BTM. However, unlike LDA, simple user based aggregation will reduce the topic coherence for BTM. To address this problem, we propose Twitter-BTM which loosens the inappropriate assumption that two co-occurring words must have same topic label made in BTM by leveraging user based aggregation and incorporating a background topic in BTM. The experimental results show that Twitter-BTM substantially outperforms BTM.

In the future, we plan to study the influence of other factors such as temporal information to BTM and its variants.

[2] See http://scikit-learn.org/stable/

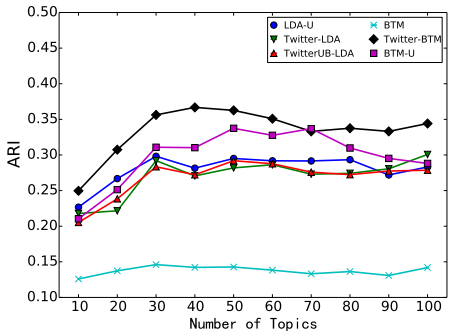


Figure 3: Performance of clustering (ARI)

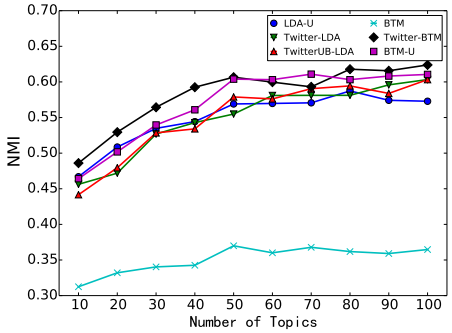


Figure 4: Performance of clustering (NMI)

Acknowledgments

This work is supported by 973 Program with Grant No.2014CB340405, NSFC with Grant No.61272340. Yan Zhang is supported by NSFC with Grant No.61370054. We thank the three anonymous reviewers for their comments and constructive criticism.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296.

Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. 2014. Btm: Topic modeling over short texts. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*.

Qiming Diao, Jing Jiang, Feida Zhu, and Ee-Peng Lim. 2012. Finding bursty topics from microblogs. In *ACL (1)*, pages 536–544. The Association for Computer Linguistics.

Zhuoye Ding, Xipeng Qiu, Qi Zhang, and Xuanjing Huang. 2013. Learning topical translation model for microblog hashtag suggestion. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*.

Dongsheng Duan, Yuhua Li, Ruixuan Li, Rui Zhang, and Aiming Wen. 2012. Ranktopic: Ranking based topic modeling. In *ICDM*, pages 211–220.

Fréderic Godin, Viktor Slavkovikj, Wesley De Neve, Benjamin Schrauwen, and Rik Van de Walle. 2013. Using topic models for twitter hashtag recommendation. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 593–596. International World Wide Web Conferences Steering Committee.

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

Liangjie Hong and Brian D. Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88, New York, NY, USA. ACM.

Zhuoye Ding Qi Zhang XuanJing Huang. 2012. Automatic hashtag recommendation for microblogs using topic-specific translation model. In *24th International Conference on Computational Linguistics*, page 265. Citeseer.

Zongyang Ma, Aixin Sun, Quan Yuan, and Gao Cong. 2014. Tagging your tweets: A probabilistic modeling of hashtag annotation in twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 999–1008. ACM.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics.

Kentaro Sasaki, Tomohiro Yoshikawa, and Takeshi Furuhashi. 2014. Online topic model for twitter considering dynamics of user interests and topic trends. In *Proceedings of the 2014 Conference on Empircal Methods in Natural Language Processing*, pages 1977–1985.

Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. 2013. Exploiting topic based twitter sentiment for stock prediction. In *ACL (2)*, pages 24–29.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961. Association for Computational Linguistics.

Yuan Wang, Jie Liu, Jishi Qu, Yalou Huang, Jimeng Chen, and Xia Feng. 2014. Hashtag graph based topic model for tweet mining. In *2014 IEEE International Conference on Data Mining, ICDM 2014, Shenzhen, China, December 14-17, 2014*, pages 1025–1030.

Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM*, pages 261–270. ACM.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456. International World Wide Web Conferences Steering Committee.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2015. A probabilistic model for bursty topic discovery in microblogs.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer.

Xin Wayne Zhao, Yanwei Guo, Yulan He, Han Jiang, Yuexin Wu, and Xiaoming Li. 2014. We know what you want to buy: a demographic-based system for product recommendation on microblogs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1935–1944. ACM.