TM-ToT: An Effective Model for Topic Mining from the Tibetan Messages

Chengxu Ye[1,2,⋆], Wushao Wen[1,3], and Ping Yang[4]

[1] School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510006, China
[2] School of Computer Science, Qinghai Normal University, Xining 810008, China
[3] School of Software, Sun Yat-sen University, Guangzhou 510006, China
[4] School of Life and Geography Sciences, Qinghai Normal University, Qinghai Xining 810008, China
ycx@qhnu.edu.cn, wenwsh@mail.sysu.edu.cn, ypycx@163.com

Abstract. The microblog platforms, such as Weibo, now accumulate a large scale of data including the Tibetan messages. Discovering the latent topics from such huge volume of Tibetan data plays a significant role in tracing the dynamics of the Tibetan community, which contributes to uncover the public opinion of this community to the government. Although topic models can find out the latent structure from traditional document corpus, their performance on Tibetan messages is unsatisfactory because the short messages cause the severe data spasity challenge. In this paper, we propose a novel model called TM-ToT, which is derived from ToT (Topic over Time) aiming at mining latent topics effectively from the Tibetan messages. Firstly, we assume each topic is a mixture distribution influenced by both word co-occurrences and messages timestamps. Therefore, TM-ToT can capture the changes of each topic over time. Subsequently, we aggregate all messages published by the same author to form a lengthy pseudo-document to tackle the data sparsity problem. Finally, we present a Gibbs sampling implementation for the inference of TM-ToT. We evaluate TM-ToT on a real dataset. In our experiments, TM-ToT outperforms Twitter-LDA by a large margin in terms of perplexity. Furthermore, the quality of the generated latent topics of TM-ToT is promising.

Keywords: Topic mining, microblog, Tibetan message, TM-ToT.

1 Introduction

Recent years we have witnessed an unprecedented growth of Weibo, which is a popular Chinese microblogging service that enables users to post and exchange short text messages (up to 140 characters). It was launched on 14 August 2009, and has more than 500 million registered users in 2012. Messages can be published through the website interface, SMS, or a wide range of apps for mobile

[⋆] Corresponding author.

C. Zong et al. (Eds.): NLPCC 2014, CCIS 496, pp. 438–448, 2014.

devices. Therefore, Weibo facilitates real-time propagation of information. In particular, about 100 million messages are posted each day on Weibo. This makes it an ideal information network, which can tell people what they care about as it is happening in the society [1].

The whole Weibo data consist of multilingual texts, including the Tibetan messages, which are an essential part of this microblog platform. Such a vast amount of user-generated short messages in the Tibetan language implies a great opportunity for business providers, advertisers, social observers, data mining researchers, as well as governments. In this paper, our goal is to mine the latent topics from the Tibetan data, which plays a significant role in tracing the dynamics of the Tibetan community, contributing to uncover the public opinion of this community to the government.

Topic models can be a wise choice to discover latent topics from the large scale of document collections, such as scholarly journals and news articles. Most existing models are developed from the Latent Dirichlet Allocation (LDA) [2], whose basic idea is that each document is a finite mixture of topics and each topic is described by a distribution over words. The LDA-based models project each document into a low dimensional space where their latent semantic structure can be uncovered easily. Unfortunately, directly using conventional topic models to the Tibetan messages can result in unsatisfactory performance. The major reasons are two-fold. Firstly, the LDA-family models ignore the temporal information in the Tibetan messages. This contradicts the fact that the messages are composed of both plain texts and timestamps. In fact, assuming that each topic is a mixture distribution influenced by both word co-occurrences and timestamps sounds more reasonable in the microblog application. Secondly, the short messages cause severe data spasity challenge, which makes the performance deteriorate significantly because the traditional topic models implicitly capture the document-level word co-occurrence patterns to reveal topics.

Considering the characteristics of the Tibetan messages (i.e., the rich temporal information and the severe data sparsity problem), we propose a novel model called TM-ToT, which is derived from ToT (Topic over Time) [3], aiming at mining latent topics effectively from the Tibetan messages. Firstly, we assume each topic is a mixture distribution influenced by both word co-occurrences and messages timestamps. Therefore, TM-ToT can capture the changes of each topic over time. Subsequently, we aggregate all messages published by the same author to form a lengthy pseudo-document to tackle the data sparsity problem. Finally, we present a Gibbs sampling implementation for the inference of TM-ToT. We evaluate TM-ToT on a real dataset. In our experiments, TM-ToT outperforms ToT by a large margin in terms of perplexity. Furthermore, the quality of the generated latent topics of TM-ToT is promising.

To sum up, the major contributions of our work are as follows:

- We study the problem of mining latent topics from the Tibetan messages to help government follow public opinion.
- We propose a novel topic model named TM-ToT and devise a Gibbs sampling for posterior inference.

– Extensive experiments on a real dataset are conducted. The results demonstrate the superiority of the proposed method.

The rest of this paper is organized as follows: the related work is discussed in Section 2; a novel topic model, TM-ToT, is proposed for topic mining from the Tibetan messages in Section 3; experimental results and discussions are presented in Section 4; finally we conclude our work in Section 5.

2 Related Work

Topic models are used to group words in a large-scale corpus into a set of relevant topics and have received increasing attention in the last decade. In this section, we briefly introduce the related work about topic models, from the basic models (i.e., LDA and ToT) to several complicated variants suitable for microblog applications.

LDA [2] is the most popular generative probabilistic model, which recognizes each document as a finite mixture over an underlying set of topics and describes each topic as a distribution of words that tend to co-occur. Through the elaborate probabilistic inference, LDA can successfully explore the hidden semantic structure of the corpus. One limitation of LDA is that it deems documents to be exchangeable during the topic modeling process. This assumption may not be tenable in some cases where the topics' occurrence and correlations change significantly over time, such as scholarly journals. ToT [3] relaxes the document exchangeable assumption by representing the mixture distribution over topics using both word co-occurrences and the text's timestamp. Experiments on real-world data sets demonstrate ToT can discover more salient topics associated with specific events and clearly localized in time.

The microblog services have gained increasing popularity and a large scale of user-generated content have been accumulated. Topic models seem appropriate to mine topics from textual documents with an unprecedented scale because of their principled mathematical foundation and effectiveness in exploratory content analysis. However, the conventional topic models usually fail to achieve satisfactory results when applied to the microblog data because the user-generated short messages cause the severe data spasity challenge. To improve the performance of topic modeling for social media, researchers take the data sparsity into consideration and develop several extensions of LDA. Ramage et al. [4] proposed a scalable implementation of a partially supervised learning model called Labeled LDA to characterize users and messages. Cha et al. [5] incorporated popularity in topic models for social network analysis. The authors argued that a popular user has very important meaning in the microblog dataset and should be carefully handled to refine the probabilistic topic models. Zhao et al. [6] develops a user-based aggregation method, Twitter-LDA, to integrate the tweets published by individual user into a lengthy pseudo-document before training LDA. Zhang et al. [7] introduced a novel probabilistic generative model MicroBlog-Latent Dirichlet Allocation (MB-LDA) for large scale microblog mining. The MB-LDA utilizes both contactor relevance relation and document relevance relation to improve the topic mining result. Tang et al. [8] pointed out that classical topic models will suffer from significant problems of data sparseness when applied to social media. Resorting to other types of information beyond word co-occurrences at the document level can significantly improve the performance of topic modeling. Therefore, the authors proposed a general solution that is able to exploit multiple types of contexts without arbitrary manipulation of the structure of classical topic models. Diao et al. [9] claimed that users on microblogs often talk about their daily lives and personal interests besides talking about global popular events. Therefore, they proposed a novel topic model that considers both temporal information of messages and users' personal interests. The model assumes that each message only consists of a single topic rather than a mixture of topics.

However, none of the topic models mentioned above is specifically tailored to the Tibetan messages, which is an important part of user-generated content in Weibo platform. for social network analysis, such as government follow public opinion. We thus propose a novel topic model named TM-ToT to extract interesting hidden topics from the Tibetan messages, by taking both temporal and context information into consideration.

3 TM-ToT Model

In this section, we introduce a novel generative probabilistic model TM-ToT for topic mining from Tibetan messages. We first describe the framework of TM-ToT, and then design a Gibbs sampling to infer our model. The notations used in TM-ToT are summarized in Table 1.

Table 1. Notations Used in TM-ToT

SYMBOL	DESCRIPTION
D,U,K,V	number of messages, users, topics, and unique words, respectively
N_d	number of words in message d
n_k^v	number of words v are assigned to topic k
n_u^k	number of words associated with user u
θ_u	the multinomial distribution of topics to user u
α, β	Dirichlet priors for θ_u and ϕ_k, respectively
ϕ_k	the multinomial distribution of words to topic k
ψ_k	the beta distribution of timestamps to topic k
$w_{d,i}$	ith word in message d
$z_{d,i}$	topic of ith word in message d
$z_{\neg(d,i)}$	topic assignments for all words except $w_{d,i}$
$w_{\neg(d,i)}$	all words except $w_{d,i}$ in message d
$t_{d,i}$	the timestamp associated with ith word in message d
$\bar{t}_k, s_k^2$	the sample mean and variance of timestamps belonging to topic k, respectively

3.1 TM-ToT Framework

Topic models achieve promising performance when documents present sufficient and meaningful signals of word co-occurrences. However, they fail to perform effectively when applied to user-generated short messages where the word co-occurrences are limited and noisy. Therefore, we should resort to rich context information (e.g., time and authorship) beyond word co-occurrences within plain texts to improve the quality of topic modeling in social media. Motivated by this intuition, the Tibetan messages are grouped into different subsets by their authors (an author refers to the user who posts a message) and each subset can be seemed as a pseudo-document, whose topic distribution inherently reflects the user's intrinsic interests. To utilize the temporal information, we assume each topic as a mixture distribution influenced by both word co-occurrences and timestamps of the Tibetan messages.

According to the above analysis, we devise a novel topic model named TM-ToT to deal with the topic modeling task for the Tibetan messages. Our proposed method firstly assembles messages written by the same user as a pseudo-document to alleviate the data sparity problem. Then, TM-ToT utilizes the temporal information during the topic modeling process in the same way as ToT does, making it possible to create a topic with a broad time distribution and draw a distinction between topics due to their changes over time. In TM-ToT, the beta distribution seems to be an appropriate choice to describe various skewed shapes of rising and falling topic prominence in social media. The Bayesian graphical framework of TM-ToT is illustrated in Figure 1.

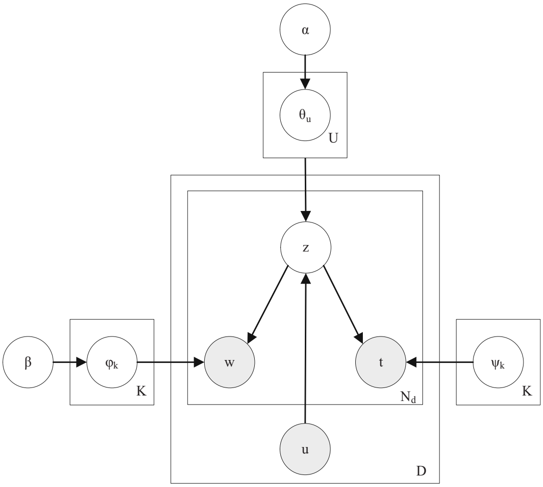


Fig. 1. Bayesian Graphical Framework of TM-ToT

Let θ_d denote the topic distribution of message d, then the detailed generative process of TM-ToT is as follows:

1. For each topic $k \in [1, K]$:
   (a) Draw a multinomial ϕ_k from a Dirichlet prior β;
2. For each message $d \in [1, D]$, published by user u:
   (a) Draw a multinomial θ_u from a Dirichlet prior α;
   (b) Assign the value of θ_u to θ_d;
   (c) For each word $i \in [1, N_d]$:
      i. Draw a topic $z_{d,i}$ from the multinomial θ_d;
      ii. Draw a word $w_{d,i}$ from the multinomial $\phi_{z_{d,i}}$;
      iii. Draw a timestamp $t_{d,i}$ from the beta $\psi_{z_{d,i}}$;

As shown in the above process, for each message d, its posterior distribution of topics θ_d depends on the authorship information.

$$P(\theta_d|\alpha) = P(\theta_u|\alpha). \tag{1}$$

The joint probability distribution of message d is:

$$P(\mathbf{w}, \mathbf{t}, \mathbf{z}|\alpha, \beta, \Psi) = P(\mathbf{w}|\mathbf{z}, \beta)P(\mathbf{t}|\mathbf{z}, \Psi)P(\mathbf{z}|\alpha). \tag{2}$$

To sum up, the formal description of the generative process in TM-ToT is:

$$\begin{aligned}
\theta_d = \theta_u|\alpha &\sim Dirichlet(\alpha) \\
\phi_k|\beta &\sim Dirichlet(\beta) \\
z_{d,i}|\theta_d &\sim Multinomial(\theta_d) \\
w_{d,i}|\phi_{z_{d,i}} &\sim Multinomial(\phi_{z_{d,i}}) \\
t_{d,i}|\psi_{z_{d,i}} &\sim Beta(\psi_{z_{d,i}})
\end{aligned}$$

3.2 TM-ToT Inference

The key issue for generative probabilistic models is to infer the hidden variables by computing their posterior distribution given the observed variables. As show in Figure 1, the temporal metadata, words and users are observed variables, while the topic structure and its changes over time are hidden variables.

The inference can not be done exactly in TM-ToT. We employ Gibbs sampling to perform approximate inference due to its speediness and effectiveness. In the Gibbs sampling procedure, we need to calculate the conditional distribution $P(z_{d,i}|\mathbf{w}, \mathbf{t}, z_{\neg(d,i)}, \alpha, \beta, \Psi)$. Taking advantage of conjugate priors, the joint distribution $P(\mathbf{w}, \mathbf{t}, \mathbf{z}|\alpha, \beta, \Psi)$ can be resolved into several components:

$$P(\mathbf{w}|\mathbf{z}, \beta) = \Big(\frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)}\Big)^K \prod_{k=1}^{K} \frac{\prod_{v=1}^{V} \Gamma(n_k^v + \beta_v)}{\Gamma(\sum_{v=1}^{V}(n_k^v + \beta_v))} \tag{3}$$

$$P(\mathbf{t}|\mathbf{z}, \Psi) = \prod_{d=1}^{D} \prod_{i=1}^{N_d} P(t_{d,i}|\psi_{z_{d,i}}) \tag{4}$$

$$P(\mathbf{z}|\alpha) = \Big(\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)}\Big)^U \prod_{u=1}^{U} \frac{\prod_{k=1}^{K} \Gamma(n_u^k + \alpha_k)}{\Gamma(\sum_{k=1}^{K}(n_u^k + \alpha_k))} \tag{5}$$

We can conveniently obtain the conditional probability by using the chain rule.

$$\begin{aligned} & P(z_{d,i}|\mathbf{w},\mathbf{t},z_{\neg(d,i)},\alpha,\beta,\Psi) \\ = & P(z_{d,i}|\mathbf{w},\mathbf{t},z_{\neg(d,i)},\alpha,\beta,\Psi) \\ \propto & \frac{n_{z_{d,i}}^{w_{d,i}}+\beta_{w_{d,i}}-1}{\sum_{v=1}^{V}(n_{z_{d,i}}^{v}+\beta_v)-1} \times (n_{z_{d,i}}^{v}+\alpha_{z_{d,i}}-1) \times p(t_{d,i}|\psi_{z_{d,i}}) \end{aligned} \tag{6}$$

We sample the posterior distribution using Gibbs sampling until it reaches a convergence for all messages. Then, we obtain the multinomial parameters as follows:

$$\phi_{k,v} = \frac{n_k^v + \beta_v}{\sum_{v=1}^{V}(n_k^v + \beta_v)} \tag{7}$$

$$\theta_{u,k} = \frac{n_u^k + \alpha_k}{\sum_{k=1}^{K}(n_u^k + \alpha_k)} \tag{8}$$

For the sake of simplicity and speed, Ψ is updated after each Gibbs sample by the method of moments estimates:

$$\hat{\psi}_{k,1} = \bar{t_k}\left(\frac{\bar{t}_k(1-\bar{t}_k)}{s_k^2} - 1\right) \tag{9}$$

$$\hat{\psi}_{k,2} = (1-\bar{t_k})\left(\frac{\bar{t}_k(1-\bar{t}_k)}{s_k^2} - 1\right) \tag{10}$$

After finishing the inference process, TM-ToT can detect topics from messages and assign the most representative words to each topic. Additionally, TM-ToT can detect the changes of each topic over time by a beta distribution with parameters from Equation 9 and 10. In summary, TM-ToT is a convenient tool for topic mining from the Tibetan messages.

4 Experiments

In this section, TM-ToT is evaluated empirically over a crawl of Tibetan messages from four different perspectives: the perplexity of held-out content, the quality of the generated latent topics and the dynamic topics analysis.

4.1 Dataset

To validate TM-ToT, we use a Weibo dataset with 70,768 messages from January 2013 to October 2013. Messages are usually short and inaccurate uses of language, which makes the quality of messages varies a lot from each other, therefore specific data preprocessing techniques are required to filter low-quality messages. We firstly use a novel Tibetan word segmentation to divide the original message text into meaningful units. Secondly, we prepare a Tibetan stop word list in advance to remove Tibetan stop words in original messages, since these frequent words do not have much meaning. Thirdly, we filter out words with less

Table 2. Description of Dataset

# of messages	30,260
# of unique words	8,987
# of users	182
# of tokens in messages	411,694
average length of each message	13
minimal timestamp (seconds)	1356969600.0
maximal timestamp (seconds)	1381248000.0

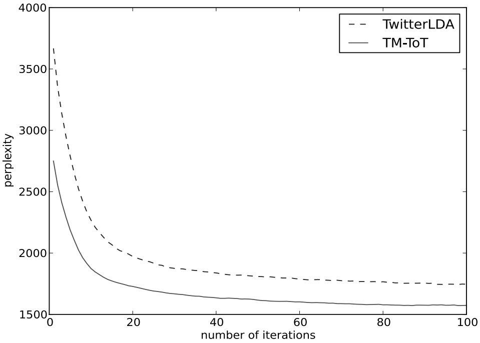


Fig. 2. Perplexities of Different Methods

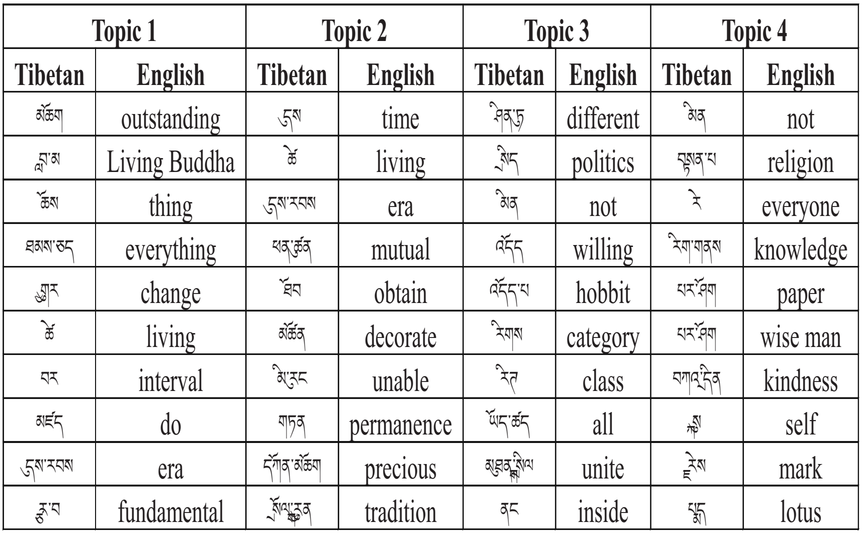
than 10 occurrences in our dataset and only keep the messages with more than 8 terms. Finally, we build a medium dataset containing 30,260 messages collected from 182 selected users for experiment evaluations. The detail information of our dataset is shown in Table 2. The simulations are carried out on an Intel Dual Core PC with 2.67 GHz CPU and 2 GB RAM.

4.2 Perplexity of Held-out Content

The metric perplexity is a widely used method to measure the performance of a topic model, which indicates the uncertainty in predicting a single word. A lower perplexity indicates better performance. To compute the perplexity of all messages, we use the formula as below:

$$Perplexity(D) = \exp\left(- \frac{\sum_d \sum_i^{N_d} \log p(w_{d,i})}{\sum_d N_d} \right) \tag{11}$$

Twitter-LDA [6] is chosen for comparison. For the sake of fairness, the parameters α, and β in both models are set to 0.1 and 0.01, respectively. Figure 2 shows the perplexities for our TM-ToT and the baseline with 50 latent topics until they reach the convergence after enough iterations. We observe that TM-ToT achieves the best perplexity. This means that integrating the temporal



Topic 1		Topic 2		Topic 3		Topic 4	
Tibetan	English	Tibetan	English	Tibetan	English	Tibetan	English
མཆོག	outstanding	དུས	time	ཤིན་ཏུ	different	མིན	not
བླ་མ	Living Buddha	ཚེ	living	སྲིད	politics	བསྟན་པ	religion
ཆོས	thing	དུས་རབས	era	མིན	not	རེ	everyone
ཐམས་ཅད	everything	ཕན་ཚུན	mutual	འདོད	willing	རིག་གནས	knowledge
འགྱུར	change	ཐོབ	obtain	འདོད་པ	hobbit	པར་ཤོག	paper
ཚེ	living	མཚོན	decorate	རིགས	category	པར་ཤོག	wise man
བར	interval	མི་ནུང	unable	རིཏ	class	བཀའ་དྲིན	kindness
མཛད	do	གཏན	permanence	ཡོད་ཚད	all	[illegible]	self
དུས་རབས	era	དཀོན་མཆོག	precious	མཐུན་སྒྲིལ	unite	རྗེས	mark
རྩ་བ	fundamental	སྲོལ་རྒྱུན	tradition	ནང	inside	པདྨ	lotus

Fig. 3. Top 10 Words for Latent Topics (K=50)

information into the topic modeling process leads to better performance. Note that the perplexities of both models do not change significantly when the number of iterations is greater than 40.

4.3 Effectiveness of Latent Topics

The main purpose of topic models for messages is to find out interesting topics from the overwhelming information. One typical method of judging the effectiveness of topic models is to print words with top weights for the latent topics and judge them by experience [10]. Figure 3 shows the quality of latent topics generated by our model. There are four topics listed out of total 50 topics, each of which is represented with the top 10 words due to the limit of space. We can learn that Topic 1 is about "Living Buddha"; Topic 2 is about "Eternal Time"; Topic 3 is about "Politics"; Topic 4 is about "Religion". The key words of each topic are accurate enough to recognize and these topics are pretty independent with each other.

4.4 Dynamic Topics Analysis

The ability of modeling the changes of topics over time is very important for topic mining in microblogs. TM-ToT combines the temporal information to capture the dynamic topics. Figure 3 illustrates all beta distributions of each topic over time when the number of topics is set to 50. An immediate and obvious effect of this is to understand more precisely when and how long the topical

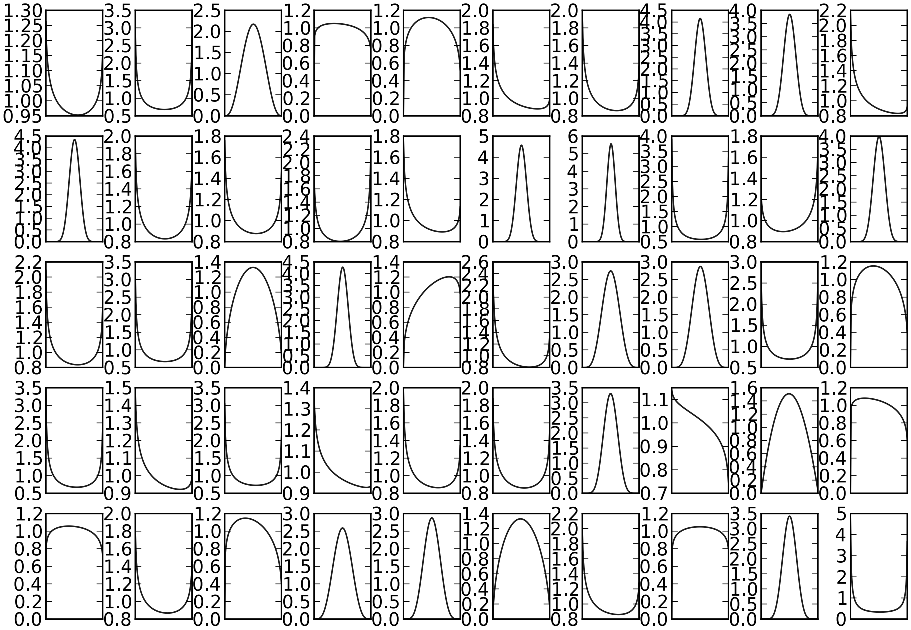


Fig. 4. Beta Distributions of Topics over Time (K=50). Note that, the X-axis is from 0 to 1.

trend was occurring. Although the beta distribution is adopted for representing various skewed schemes of rising and falling topic prominence, there still exist several salient types. For example, the uniform distribution is common in our experiment. Note that, even the similar shape of distributions have different means and variances. Thus, the changes of topics over time can be distinguished easily.

5 Conclusions

In this paper, we present and evaluate a time-aware topic model TM-ToT mixed with user's intrinsic interests, for effectively modeling and analyzing the topics that naturally arise in Tibetan microblogs. TM-ToT is able to capture the changes in the occurrence of topics by assuming that each topic is a mixture distribution influenced by both word co-occurrences and timestamps of microblogs. Moreover, the author relationship information is used to solve the severe data sparsity problem. Finally, the inference of TM-ToT is completed by a Gibbs sampling. Extensive experiments on a real dataset demonstrate that TM-ToT outperforms its competitor.

In the future work, we will focus on investigating more social network information, such as follow/following relations and URLs, to improve the performance of topic models. How to describe the temporal metadata distributions is another interesting direction. Finally, we will devise more elaborate and effective model to merge the social network information.

Acknowledgments. This work is supported by the National Social Science Fund of China (13BXW037)and the Ministry of Education Chunhui Project (Z2011023).

References

1. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: A content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 759–768 (2010)
2. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022 (2003)
3. Wang, X., McCallum, A.: Topics over time: A non-markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 424–433 (2006)
4. Ramage, D., Dumais, S., Liebling, D.: Characterizing microblogs with topic models. In: Proceedings of the International AAAI Conference on Weblogs and Social Media, pp. 130–137 (2010)
5. Cha, Y., Bi, B., Hsieh, C.-C., Cho, J.: Incorporating popularity in topic models for social network analysis. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 223–232 (2013)
6. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. In: Proceedings of the European Conference on Information Retrieval, pp. 338–349 (2011)
7. Zhang, C., Sun, J.: Large scale microblog mining using distributed mb-lda. In: Proceedings of the 21st International Conference Companion on World Wide Web, pp. 1035–1042 (2012)
8. Tang, J., Zhang, M., Mei, Q.Z.: One theme in all views: Modeling consensus topics in multiple contexts. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 5–13 (2013)
9. Diao, Q., Jiang, J., Zhu, F., Lim, E.: Finding bursty topics from microblogs. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp. 536–544 (2012)
10. Xu, Z., Zhang, Y., Wu, Y., Yang, Q.: Modeling user posting behavior on social media. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 545–554 (2012)