Online Topic Model for Twitter Considering Dynamics of User Interests and Topic Trends

Kentaro Sasaki, Tomohiro Yoshikawa, Takeshi Furuhashi
Graduate School of Engineering Nagoya University
sasaki@cmplx.cse.nagoya-u.ac.jp
yoshikawa, furuhashi@cse.nagoya-u.ac.jp

Abstract

Latent Dirichlet allocation (LDA) is a topic model that has been applied to various fields, including user profiling and event summarization on Twitter. When LDA is applied to tweet collections, it generally treats all aggregated tweets of a user as a single document. Twitter-LDA, which assumes a single tweet consists of a single topic, has been proposed and has shown that it is superior in topic semantic coherence. However, Twitter-LDA is not capable of online inference. In this study, we extend Twitter-LDA in the following two ways. First, we model the generation process of tweets more accurately by estimating the ratio between topic words and general words for each user. Second, we enable it to estimate the dynamics of user interests and topic trends online based on the topic tracking model (TTM), which models consumer purchase behaviors.

1 Introduction

Microblogs such as Twitter, have prevailed rapidly in our society recently. Twitter users post a message using 140 characters, which is called a tweet. The characters limit allows users to post tweets easily about not only personal interest or real life but also public events such as traffic accidents or earthquakes. There have been many studies on how to extract and utilize such information on tweets (Diao et al., 2012; Pennacchiotti and Popescu, 2011; Sakaki et al., 2010; Weng et al., 2010).

Topic models, such as latent Dirichlet allocation (LDA) (Blei et al., 2003) are widely used to identify latent topic structure in large collections of documents. Recently, some studies have applied LDA to Twitter for user classification (Pennacchiotti and Popescu, 2011), detection of influential users (Weng et al., 2010), and so on. LDA is a generative document model, which assumes that each document is represented as a probability distribution over some topics, and that each word has a latent topic. When we apply LDA to tweets, each tweet is treated as a single document. This direct application does not work well because a tweet is very short compared with traditional media such as newspapers. To deal with the shortness of a tweet, some studies aggregated all the tweets of a user as a single document (Hong and Davison, 2010; Pennacchiotti and Popescu, 2011; Weng et al., 2010). On the other hand, Zhao et al. (2011) proposed "Twitter-LDA," which is a model that considers the shortness of a tweet. Twitter-LDA assumes that a single tweet consists of a single topic, and that tweets consist of topic and background words. Zhao et al. (2011) show that it works well at the point of semantic coherence of topics compared with LDA. However, as with the case of LDA, Twitter-LDA cannot consider a sequence of tweets because it assumes that samples are exchangeable. In Twitter, user interests and topic trends are dynamically changing. In addition, when new data comes along, a new model must be generated again with all the data in Twitter-LDA because it does not assume online inference. Therefore, it cannot efficiently analyze the large number of tweets generated everyday. To overcome these difficulties, a model that considers the time sequence and has the capability of online inference is required.

In this study, we first propose an improved model based on Twitter-LDA, which assumes that the ratio between topic and background words differs for each user. This study evaluates the proposed method based on perplexity and shows the efficacy of the new assumption in the improved model. Second, we propose a new topic model called "Twitter-TTM" by extending the improved

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1977–1985,
October 25-29, 2014, Doha, Qatar.

model based on the topic tracking model (TTM) (Iwata et al., 2009), which models the purchase behavior of consumers and is capable of online inference. Finally, we demonstrate that Twitter-TTM can effectively capture the dynamics of user interests and topic trends in Twitter.

2 Improvement of Twitter-LDA

2.1 Improved-Model

Figure 1(a) shows the graphical representation of Twitter-LDA based on the following assumptions. There are K topics in Twitter and each topic is represented by a topic word distribution. Each user has his/her topic interests ϕ_u represented by a distribution over K topics. Topic k is assigned to each tweet of user u depending on the topic interests ϕ_u. Each word in the tweet assigned by topic k is generated from a background word distribution θ_B or a topic word distribution θ_k. Whether the word is a background word or a topic word is determined by a latent value y. When $y = 0$, the word is generated from the background word distribution θ_B, and from the topic word distribution θ_k when $y = 1$. The latent value y is chosen according to a distribution π. In other words, the ratio between background and topic words is determined by π.

In Twitter-LDA, π is common for all users, meaning that the rate between background and topic words is the same for each user. However, this assumption could be incorrect, and the rate could differ for each user. Thus, we develop an improved model based on Twitter-LDA, which assumes that π is different for each user, as shown in Figure 1(b). In the improved model, the rate between background and topic words for user u is determined by a user-specific distribution π_u. The improved model is expected to infer the generative process of tweets more efficiently.

2.2 Experiment for Improved Model

We performed an experiment to compare the predictive performances of LDA, TTM, and the improved model shown in Section 2.1. In this experiment, LDA was applied as the method to aggregate all tweets of a user as a single document. The original Twitter data set contains 14,305 users and 292,105 tweets collected on October 18, 2013. We then removed words that occurred less than 20 times and stop words. Retweets[1] were treated as the same as other general tweets because they reflected the user's interests. After the above preprocessing, we obtained the final dataset with 14,139 users, 252,842 tweets, and 7,763 vocabularies. Each model was inferred with collapsed Gibbs sampling (Griffiths and Steyvers, 2004) and the iteration was set at 500. For a fair comparison, the hyper parameters in these models were optimized in each Gibbs sampling iteration by maximizing likelihood using fixed iterations (Minka, 2000).

This study employs perplexity as the evaluation index, which is the standard metric in information retrieval literature. The perplexity of a held-out test set is defined as

$$perplexity = exp\Big(-\frac{1}{N}\sum_{u}\log p(\mathbf{w_u})\Big) \quad (1)$$

where $\mathbf{w_u}$ represents words are contained in the tweets of user u and N is the number of words in the test set. A lower perplexity means higher predictive performance. We set the number of topics K at 50, 100, 150, 200, and 250 and evaluated the perplexity for each model in each K via a 10-fold cross-validation.

The results are shown in Table 1, which shows that the improved model performs better than the other models for any K. Therefore, the new assumption of the improved model, that the rate between background and topic words is different for each user, could be more appropriate. LDA performance worsens with an increase in K because the aggregated tweets of a single user neglect the topic of each tweet.

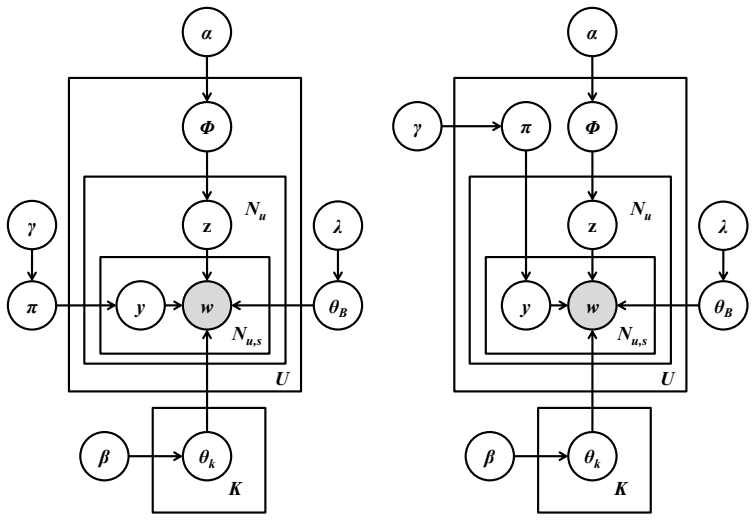
Table 2 shows examples of the tweets of users with high and low rates of background words. The users with a high background words rate tend to use basic words that are often used in any topics, such as "like," "about," and "people," and they tend to tweet about their personal lives. On the other hand, for users with a low background words rate, topical words are often used such as "Arsenal," "Justin," and "Google". They tend to tweet about their interests, including music, sports, and movies.

3 Twitter-TTM

3.1 Model Extension based on Topic Tracking Model

We extend the improved model shown in Section 2.1 considering the time sequence and capabil-

[1] Republishing a tweet written by another Twitter user.



(a) Twitter-LDA (b) Improved-model

Figure 1: Graphical representation of Twitter-LDA and Improved-model

Table 1: Perplexity of each model in 10 runs

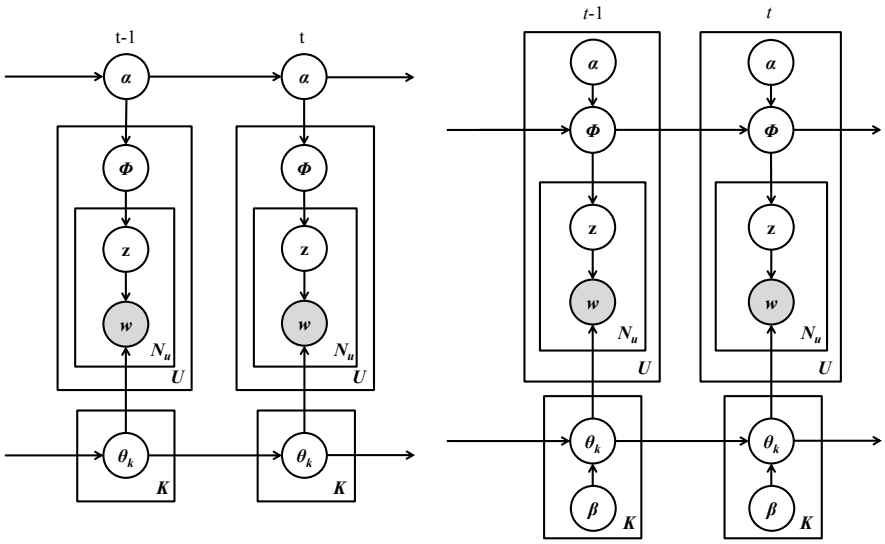
Number of topic K	LDA	Twitter-LDA	Improved-model
50	1586.7 (14.4)	2191.0 (28.4)	**1555.3** (36.7)
100	1612.7 (11.9)	1933.9 (23.6)	**1471.7** (22.3)
150	1635.3 (11.2)	1760.1 (15.7)	**1372.3** (20.0)
200	1655.2 (13.0)	1635.4 (22.1)	**1289.5** (13.3)
250	1672.7 (17.2)	1542.8 (12.5)	**1231.1** (11.9)

Table 2: Example of tweets of users with high and low rate of background words

High rate of background words	Low rate of background words
I hope today goes quickly	Team Arsenal v will Ozil be
I want to work in a cake	Making Justin smile and laugh as he is working on music
All need your support please	Google nexus briefly appears in Google play store

ity of online inference based on TTM (Iwata et al., 2009). TTM is a probabilistic consumer purchase behavior model based on LDA for tracking the interests of each user and the trends in each topic. Other topic models considering the dynamics of topics include the dynamic topic model (DTM) (Blei and Lafferty, 2006) and topic over time (ToT) (Wang and McCallum, 2006). DTM is a model for analyzing the time evolution of topics in time-ordered document collections. It does not track the interests of each user as shown in Figure 2(a) because it assumes that a user (document) has only one time stamp. ToT requires all the data over time for inference, thus, it is not appropriate for application to continuously generated data such as Twitter. We consider a model must be capable of online inference and track the dynamics of user interests and topic trends for modeling tweets. Since TTM has these abilities, we adapt it to the improved model described in Section 2.

Figure 2(b) shows the graphical representation of TTM. TTM assumes that the mean of user interests at the current time is the same as that at the previous time, unless new data is observed. Formally, the current interest $\phi_{t,u}$ are drawn from the following Dirichlet distribution in which the mean is the previous interest $\hat{\phi}_{t-1,u}$ and the precision is



(a) DTM (b) TTM

Figure 2: Graphical representation of DTM and TTM

$\alpha_{t,u}$

$$p(\phi_{t,u}|\hat{\phi}_{t-1,u},\alpha_{t,u}) \propto \prod_k \phi_{t,u,k}^{\alpha_{t,u}\hat{\phi}_{t-1,u,k}-1} \quad (2)$$

where $\phi_{t,u,k}$ represents the probability that user u is interested in topic k at time t. t is a discrete variable and can be arbitrarily set as the unit time interval, e.g., at one day or one week. The precision $\alpha_{t,u}$ represents the interest persistence of how consistently user u maintains his/her interests at time t compared with the previous time $t-1$. $\alpha_{t,u}$ is estimated for each time period and each user because interest persistence depends on both time and users. As mentioned above, the current topic trend $\theta_{t,k}$ is drawn from the following Dirichlet distribution with the previous trend $\hat{\theta}_{t-1,k}$

$$p(\theta_{t,k}|\hat{\theta}_{t-1,k},\beta_{t,k}) \propto \prod_v \theta_{t,k,v}^{\beta_{t,k}\hat{\theta}_{t-1,k,v}-1} \quad (3)$$

where $\theta_{t,k,v}$ represents the probability that word v is chosen in topic k at time t.

Here our proposed Twitter-TTM adapts the above TTM assumptions to the improved model. That is, we extend the improved model whereby user interest $\phi_{t,u}$ and topic trend $\theta_{t,k}$ depend on previous states. Time dependency is not considered on θ_B and π_u because they can be regarded as being independent of time.

Figures 3 and 4 show the generative process and a graphical representation of Twitter-TTM, respectively. Twitter-TTM can capture the dynamics of user interests and topic trends in Twitter considering the features of tweets online. Moreover, Twitter-TTM can be extended to capture long-term dependences, as described in Iwata et al. (2009).

3.2 Model Inference

We use a stochastic expectation-maximization algorithm for Twitter-TTM inference, as described in Wallach (2006) in which Gibbs sampling of latent values and maximum joint likelihood estimation of parameters are alternately iterated. At time t, we estimate user interests $\mathbf{\Phi}_t = \{\hat{\phi}_{t,u}\}_{u=1}^U$, topic trends $\mathbf{\Theta}_t = \{\hat{\theta}_{t,k}\}_{k=1}^K$, background word distribution $\theta_{t,B}$, word usage rate distribution $\pi_{t,u}$, interest persistence parameters $\boldsymbol{\alpha_t} = \{\alpha_{t,u}\}_{u=1}^U$, and trend persistence parameters $\boldsymbol{\beta_t} = \{\beta_{t,k}\}_{k=1}^K$ using the previous time interests $\hat{\mathbf{\Phi}}_{t-1}$ and trends $\hat{\mathbf{\Theta}}_{t-1}$.

We employ collapsed Gibbs sampling to infer the latent variables. Let D_t be a set of tweets and Z_t, Y_t be a set of latent variables z, y at time t. We can integrate the parameters in the joint distribu-

tion as follows:

$$
\begin{aligned}
&p(D_t, Y_t, Z_t | \hat{\boldsymbol{\Phi}}_{t-1}, \hat{\boldsymbol{\Theta}}_{t-1}, \boldsymbol{\alpha_t}, \boldsymbol{\beta_t}, \lambda, \gamma) \\
&= \Big(\frac{\Gamma(2\gamma)}{\Gamma(\gamma)^2}\Big)^U \prod_u \frac{\Gamma(\gamma + n_{t,u,B})\Gamma(\gamma + n_{t,u,K})}{\Gamma(2\gamma + n_{t,u})} \\
&\times \frac{\Gamma(V\lambda)}{\Gamma(\lambda)^V} \frac{\prod_v \Gamma(n_{t,B,v} + \lambda)}{\Gamma(n_{t,B} + V\lambda)} \\
&\times \prod_k \frac{\Gamma(\beta_{t,k})}{\Gamma(n_{t,k} + \beta_{t,k})} \prod_v \frac{\Gamma(n_{t,k,v} + \beta_{t,k}\hat{\theta}_{t-1,k,v})}{\Gamma(\beta_{t,k}\hat{\theta}_{t-1,k,v})} \\
&\times \prod_u \frac{\Gamma(\alpha_{t,u})}{\Gamma(c_{t,u} + \alpha_{t,u})} \prod_k \frac{\Gamma(c_{t,u,k} + \alpha_{t,u}\hat{\phi}_{t-1,u,k})}{\Gamma(\alpha_{t,u}\hat{\phi}_{t-1,u,k})},
\end{aligned} \tag{4}
$$

where $n_{t,u,B}$ and $n_{t,u,K}$ are the number of background and topic words of user u at time t, $n_{t,B,v}$ is the number of times that word v is assigned as a background word at time t, $n_{t,k,v}$ is the number of times that word v is assigned to topic k at time t, $c_{t,u,k}$ is the number of tweets assigned to topic k for user u at time t. In addition, $n_{t,u} = n_{t,u,B} + n_{t,uK}$, $n_{t,B} = \sum_v n_{t,B,v}$, $n_{t,K} = \sum_k n_{t,k} = \sum_k \sum_v n_{t,k,v}$, $n_{t,u} = \sum_k n_{t,u,k}$, and $c_{t,u} = \sum_k c_{t,u,k}$.

Given the assignment of all other latent variables, we derive the following formula calculated from eq.(4) to infer a latent topic,

$$
\begin{aligned}
&p(z_i = k | D_t, Y_t, Z_{t\backslash i}, \hat{\boldsymbol{\Phi}}_{t-1}, \hat{\boldsymbol{\Theta}}_{t-1}, \boldsymbol{\alpha_t}, \boldsymbol{\beta_t}) \\
&\propto \frac{c_{t,u,k\backslash i} + \alpha_{t,u}\hat{\phi}_{t-1,u,k}}{c_{t,u\backslash i} + \alpha_{t,u}} \frac{\Gamma(n_{t,k\backslash i} + \beta_{t,k})}{\Gamma(n_{t,k} + \beta_{t,k})} \\
&\times \prod_v \frac{\Gamma(n_{t,k,v} + \beta_{t,k}\hat{\theta}_{t-1,k,v})}{\Gamma(n_{t,k,v\backslash i} + \beta_{t,k}\hat{\theta}_{t-1,k,v})},
\end{aligned} \tag{5}
$$

where $i = (t, u, s)$, thus z_i represents a topic assigned to the s-th tweet of user u at time t, and $\backslash i$ represents a count excluding the i-th tweet.

Then, when $z_i = k$ is given, we derive the following formula to infer a latent variable y_j,

$$
\begin{aligned}
&p(y_j = 0 | D_t, Y_{t\backslash j}, Z_t, \lambda, \gamma) \\
&\propto \frac{n_{t,B,v\backslash j} + \lambda}{n_{t,B\backslash j} + V\lambda} \frac{n_{t,u,B\backslash j} + \gamma}{n_{t,u\backslash j} + 2\gamma},
\end{aligned} \tag{6}
$$

$$
\begin{aligned}
&p(y_j = 1 | D_t, Y_{t\backslash j}, Z_t, \hat{\boldsymbol{\Theta}}_{t-1}, \boldsymbol{\beta_t}, \gamma) \\
&\propto \frac{n_{t,k,v\backslash j} + \beta_{t,k}\hat{\theta}_{t-1,k,v}}{n_{t,k\backslash j} + \beta_{t,k}} \frac{n_{t,u,K\backslash j} + \gamma}{n_{t,u\backslash j} + 2\gamma},
\end{aligned} \tag{7}
$$

where $j = (t, u, s, n)$, thus y_j represents a latent variable assigned to the n-th word in the s-th tweet of user u at time t, and $\backslash j$ represents a count excluding the j-th word.

The persistence parameters $\boldsymbol{\alpha_t}$ and $\boldsymbol{\beta_t}$ are estimated by maximizing the joint likelihood eq.(4), using a fixed point iteration (Minka, 2000). The update formulas are as follows:

$$
\alpha_{t,u}^{new} = \alpha_{t,u} \frac{\sum_k \hat{\phi}_{t-1,u,k} A_{t,u,k}}{\Psi(c_{t,u} + \alpha_{t,u}) - \Psi(\alpha_{t,u})}, \tag{8}
$$

where $A_{t,u,k} = \Psi(c_{t,u,k} + \alpha_{t,u}\hat{\phi}_{t-1,u,k}) - \Psi(\alpha_{t,u}\hat{\phi}_{t-1,u,k})$, and

$$
\beta_{t,k}^{new} = \beta_{t,k} \frac{\sum_v \hat{\theta}_{t-1,k,v} B_{t,k,v}}{\Psi(n_{t,k} + \beta_{t,k}) - \Psi(\beta_{t,k})}, \tag{9}
$$

where $B_{t,k,v} = \Psi(n_{t,k,v} + \beta_{t,k}\hat{\theta}_{t-1,k,v}) - \Psi(\beta_{t,k}\hat{\theta}_{t-1,k,v})$. We can estimate latent variables Z_t, Y_t, and parameters $\boldsymbol{\alpha_t}$ and $\boldsymbol{\beta_t}$ by iterating Gibbs sampling with eq.(5), eq.(6), and eq.(7) and maximum joint likelihood with eq.(8) and eq.(9). After the iterations, the means of $\phi_{t,u,k}$ and $\theta_{t,k,v}$ are obtained as follows.

$$
\hat{\phi}_{t,u,k} = \frac{c_{t,u,k} + \alpha_{t,u}\hat{\phi}_{t-1,u,k}}{c_{t,u} + \alpha_{t,u}}, \tag{10}
$$

$$
\hat{\theta}_{t,k,v} = \frac{n_{t,k,v} + \beta_{t,k}\hat{\theta}_{t-1,k,v}}{n_{t,k} + \beta_{t,k}}. \tag{11}
$$

These estimates are used as the hyper parameters of the prior distributions at the next time period $t + 1$.

4 Related Work

Recently, topic models for Twitter have been proposed. Diao et al. (2012) proposed a topic model that considers both the temporal information of tweets and user's personal interests. They applied their model to find bursty topics from Twitter. Yan et al. (2013) proposed a biterm topic model (BTM), which assumes that a word-pair is independently drawn from a specific topic. They demonstrated that BTM can effectively capture the topics within short texts such as tweets compared with LDA. Chua and Asur (2013) proposed two topic models considering time order and tweet intervals to extract the tweets summarizing a given event. The models mentioned above do not consider the dynamics of user interests, nor

1. Draw $\theta_{t,B}$ ∼Dirichlet(λ)
2. For each topic $k = 1, ..., K$,
   (a) draw $\theta_{t,k}$ ∼Dirichlet($\beta_{t,k}\hat{\theta}_{t-1,k}$)
3. For each user $u = 1, ..., U$,
   (a) draw $\phi_{t,u}$ ∼Dirichlet($\alpha_{t,u}\hat{\phi}_{t-1,u}$)
   (b) draw $\pi_{t,u}$ ∼Beta(γ)
   (c) for each tweet $s = 1, ..., N_u$
      i. draw $z_{t,u,s}$ ∼Multinomial($\phi_{t,u}$)
      ii. for each word $n = 1, ..., N_{u,s}$
         A. draw $y_{t,u,s,n}$ ∼Bernoulli($\pi_{t,u}$)
         B. draw $w_{t,u,s,n}$ ∼ Multinomial($\theta_{t,B}$) if $y_{t,u,s,n} = 0$ or Multinomial($\theta_{t,z_{t,u,s}}$) if $y_{t,u,s,n} = 1$

Figure 3: Generative process of tweets in Twitter-TTM

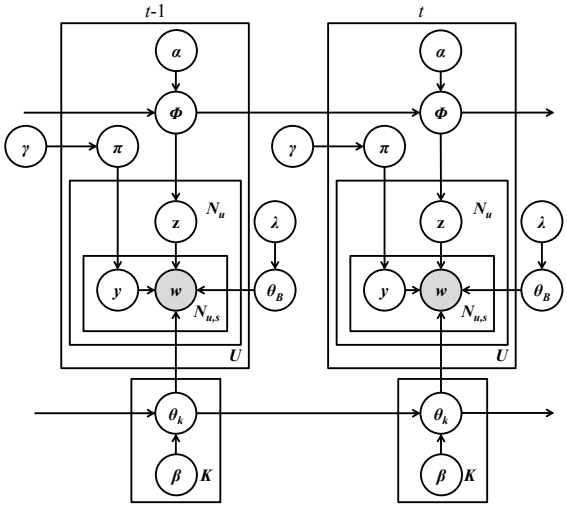


Figure 4: Graphical model of Twitter-TTM

do they have the capability of online inference; thus, they cannot efficiently model the large number of tweets generated everyday, whereas Twitter-TTM can capture the dynamics of user interests and topic trends and has the capability of online inference.

Some online topic models have also been proposed. TM-LDA was proposed by Wang et al. (2012), which can efficiently model online the topics and topic transitions that naturally arise in a tweet stream. Their model learns the transition parameters among topics by minimizing the prediction error on topic distribution in subsequent tweets. However, the TM-LDA does not consider dynamic word distributions. In other words, their model can not capture the dynamics of topic trends. Lau et al. (2012) proposed a topic model implementing a dynamic vocabulary based on online LDA (OLDA) (AlSumait et al., 2008) and applied it to track emerging events on Twitter. An online variational Bayes algorithm for LDA is also proposed (Hoffman et al., 2010). However, these methods are based on LDA and do not consider the shortness of a tweet. Twitter-TTM tackles the shortness of a tweet by assuming that a single tweet consists of a single topic. This assumption is based on the following observation: a tweet is much shorter than a normal document, so a single tweet rarely contains multiple topics but rather a single one.

5 Experiment

5.1 Setting

We evaluated the effectiveness of the proposed Twitter-TTM using an actual Twitter data set. The original Twitter data set contains 15,962 users and 4,146,672 tweets collected from October 18 to 31, 2013. We then removed words that occurred less than 30 times and stop words. After this preprocessing, we obtained the final data set with 15,944 users, 3,679,481 tweets, and 30,096 vocabularies.

We compared the predictive performance of Twitter-TTM with LDA, TTM, Twitter-LDA, Twitter-LDA+TTM, and the improved model based on the perplexity for the next time tweets. Twitter-LDA+TTM is a combination of Twitter-LDA and TTM. It is equivalent to Twitter-TTM, except that the rate between background and topic words is different for each user. We set the number of topics K at 100, the iteration of each model at 500, and the unit time interval at one day. The hyper parameters in these models were optimized in each Gibbs sampling iteration by maximizing likelihood using fixed iterations (Minka, 2000). The inferences of LDA, Twitter-LDA, and the improved model were made for current time tweets.

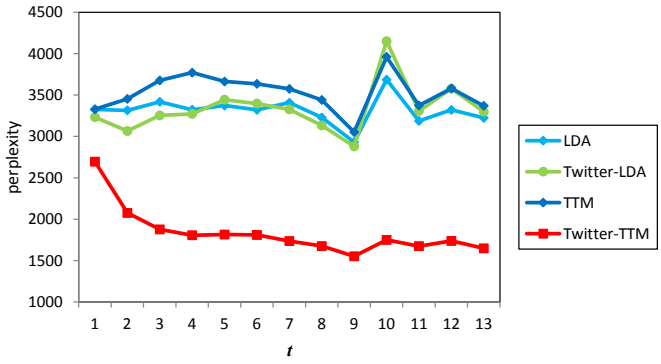
5.2 Result

Figure 5 shows the perplexity of each model for each time, where $t = 1$ in the horizontal axis represents October 18, $t = 2$ represents October 19, ..., and $t = 13$ represents October 31. The perplexity at time t represents the predictive performance of each model inferred by previous time tweets to

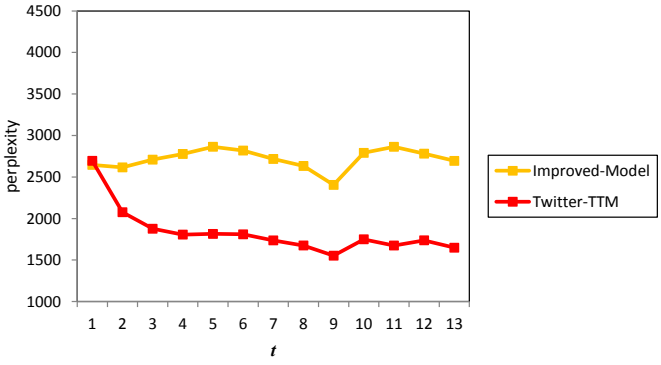
the current time tweets. Note that at $t = 1$, the performance of LDA and TTM, that of Twitter-LDA and Twitter-LDA+TTM, and that of Twitter-TTM and the improved model were found to be equivalent.

As shown in Figure 5(a), the proposed Twitter-TTM shows lower perplexity compared with conventional models, such as LDA, Twitter-LDA, and TTM at any time, which implies that Twitter-TTM can appropriately model the dynamics of user interests and topic trends in Twitter. TTM could not have perplexity lower than LDA although it considers the dynamics. If LDA could not appropriately model the tweets, then the user interests $\hat{\mathbf{\Phi}}_{t-1}$ and topic trends $\hat{\mathbf{\Theta}}_{t-1}$ in the previous time are not estimated well in TTM. Figure 5(b) shows the perplexities of the improved model and Twitter-TTM. From $t = 2$, Twitter-TTM shows lower perplexity than the improved model for each time. The reason for the high perplexity of the improved model is that it does not consider the dynamics. Twitter-TTM also shows lower perplexity than Twitter-LDA+TTM for each time, as shown in Figure 5(c), because Twitter-TTM's assumption that the rate between background and topic words is different for each user is more appropriate, as demonstrated in Section 2.2. These results imply that Twitter-TTM also outperforms other conventional methods, such as DTM, OLDA, and TM-LDA, which do not consider the shortness of a tweet or the dynamics of user interests or topic trends .

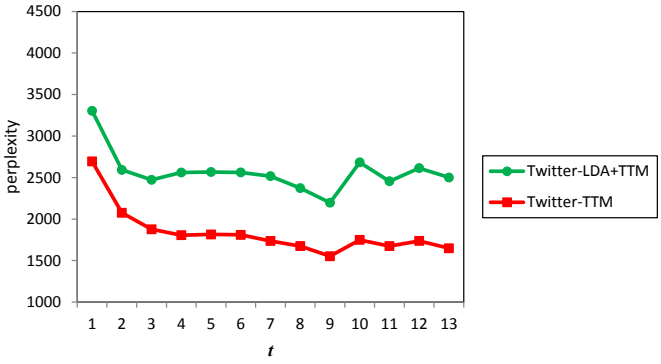
Table 3 shows two topic examples of the topic evolution analyzed by Twitter-TTM, and Figure 6 shows the trend persistence parameters β of each topic at each time. The persistence parameters of the topic "Football" are lower than those of "Birthday" because it is strongly affected by trends in the real world. In fact, the top words in "Football" change more dynamically than those of "Birthday." For example, in the "Football" topic, though 'Arsenal' is usually popular, 'Madrid' becomes more popular on October 24.



(a) Comparison with LDA, Twitter-LDA, and TTM



(b) Comparison with Improved-model



(c) Comparison with Twitter-LDA+TTM

Figure 5: Perplexity for each time

6 Conclusion

We first proposed an improved model based on Twitter-LDA, which estimates the rate between background and topic words for each user. We demonstrated that the improved model could model tweets more efficiently than LDA and Twitter-LDA. Next we proposed a novel probabilistic topic model for Twitter, called Twitter-TTM, which can capture the dynamics of user interests and topic trends and is capable of online inference. We evaluated Twitter-TTM using an actual Twitter data set and demonstrated that it could model more accurately tweets than conventional

methods.

The proposed method currently needs to predetermine the number of topics each time, and it is fixed. In future work, we plan to extend the proposed method to capture the birth and death of topics along the timeline with a variable number of topics, such as the model proposed by Ahmed (Ahmed and Xing, 2010). We also plan to apply the proposed method to content recommendations and trend analysis in Twitter to investigate this method further.

References

Amr Ahmed and Eric P. Xing. 2010. Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)*, 20–29.

Loulwah AlSumait, Daniel Barbará and Carlotta Domeniconi. 2008. On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 3-12.

David M. Blei., and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine learning (ICML)*, 113-120.

David M. Blei, Andrew Y. Ng and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3: 993-1022.

Freddy C. T. Chua and Sitaram Asur. 2013. Automatic summarization of events from social media. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Qiming Diao, Jing Jiang, Feida Zhu and Ee-Peng Lim 2012. Finding bursty topics from microblogs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, 536–544.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. In *Proceedings of the National Academy of Sciences of the United States of America*, 101(1):5228-5235.

Matthew D. Hoffman, Francis Bach and David M. Blei. 2010. Online learning for latent dirichlet allocation. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 856–864.

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)*, 80–88.

Tomoharu Iwata, Shinji Watanabe, Takeshi Yamada. and Naonori Ueda. 2009. Topic tracking model for analyzing consumer purchase behavior. In *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI)*,1427–1432.

JeyHan Lau, Nigel Collier and Timothy Baldwin. 2012. On-line trend analysis with topic models:# twitter trends detection topic model online. In *Proceedings of the 23th International Conference on Computational Linguistics (COLING)*, 1519–1534.

Thomas P. Minka 2000. Estimating a Dirichlet distribution *Technical report, MIT*.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. A machine learning approach to Twitter user classification. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, 281–288.

Takeshi Sakaki, Makoto Okazaki and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: realtime event detection by social sensors. In *Proceedings of the World Wide Web Conference (WWW)*, 851–860.

Hanna M. Wallach 2006. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 977–984.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-Markov continuous-time model of topical trends. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)*, 424–433.

Yu Wang, Eugene Agichtein and Michele Benz. 2012. TM-LDA: efficient online modeling of the latent topic transitions in social media. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD)*, 123–131.

Jianshu Weng, Ee Peng Lim, Jing Jiang and Qi He. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM)*, 261–270.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan and Xueqi Cheng 2013. A biterm topic model for short texts. In *Proceedings of the World Wide Web Conference (WWW)*, 1445–1456.

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*, 338–349.

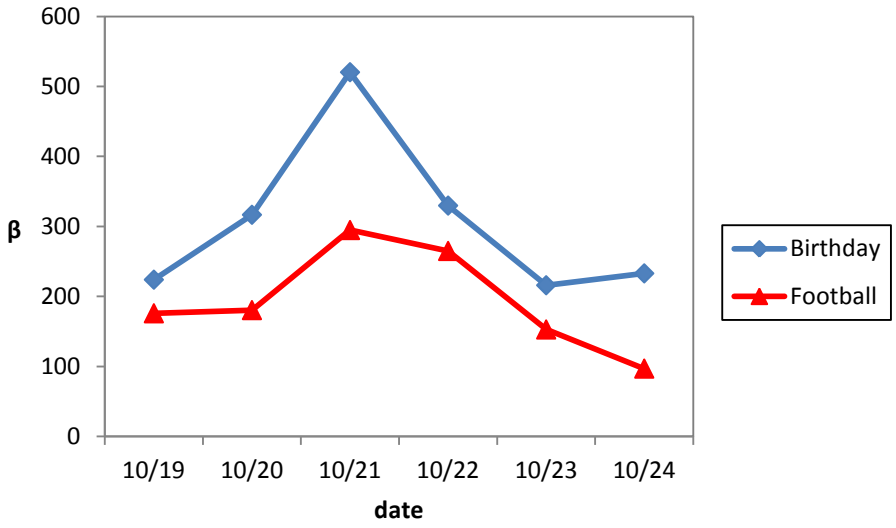


Figure 6: Trend persistence parameters β of each topic at each time estimated by Twitter-TTM

Table 3: Two examples of topic evolution analyzed by Twitter-TTM

Label	Date	Top words
Birthday	10/18	birthday,happy,maria,hope,good,love,thanks,bday,lovely,enjoy
	10/19	happy,birthday,good,hope,thank,enjoy,love,bday,lovely,great
	10/20	birthday,happy,hope,good,love,lovely,great,enjoy,thank,beautiful
	10/21	birthday,happy,hope,good,beautiful,love,lovely,bday,great,thank
	10/22	birthday,happy,hope,good,beautiful,love,bless,thank,today,bday
	10/23	birthday,happy,thank,good,love,hope,beautiful,enjoy,channing,wish
	10/24	birthday,happy,thank,love,hope,good,beautiful,fresh,thanks,jamz
Football	10/18	arsenal,ozil,game,team,cazorla,league,wenger,play,season,good
	10/19	goal,liverpool,gerrard,arsenal,ozil,league,newcastle,suarez,goals,team
	10/20	arsenal,ozil,goal,ramsey,norwich,goals,league,wilshere,mesut,premier
	10/21	arsenal,goal,goals,league,townsend,spurs,player,season,wenger,ozil
	10/22	arsenal,goal,wenger,ozil,league,arsene,goals,birthday,happy,team
	10/23	arsenal,dortmund,ozil,fans,wilshere,borussia,ramsey,lewandowski,giroud,league
	10/24	madrid,goals,ronaldo,cska,real,league,city,moscow,champions,yaya