Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/080137

MB-ToT: An Effective Model for Topic Mining in Microblogs

Shaopeng Liu[1], *Jian Yin*[1,*], *Jia Ouyang*[1], *Yun Huang*[1] *and Piyuan Lin*[2]

[1] School of Information Science and Technology, Sun Yat-Sen University, 510006 Guangzhou, China
[2] College of Software Technology, South China Agricultural University, 510642 Guangzhou, China

Received: 20 Jun. 2013, Revised: 27 Oct. 2013, Accepted: 28 Oct. 2013
Published online: 1 Jan. 2014

Abstract: Topic mining on microblogging sites with sheer scale of instance messages and social network information, such as Twitter, is a hard and challenging problem. Although many text mining techniques and generative probabilistic models have been developed for static plain-text corpus, they are inclined to achieve unsatisfactory results in microblogs without considering that microblogs are temporally sequential and concerned with social network information. In this paper, we propose a novel topic model, MicroBlog-Topics over Time (MB-ToT), which aims for comprehensive topic analysis in microblogs. Firstly, we assume each topic is a mixture distribution influenced by both word co-occurrences and timestamps of microblogs. This allows MB-ToT to capture the changes of each topic over time. Subsequently, we apply users' intrinsic interests, social contact relations and #hashtags to improve the topic mining result. Finally, we present a Gibbs sampling implementation for the inference of MB-ToT. We evaluate MB-ToT and compare it with the state-of-the-art methods on a real dataset. In our experiments, MB-ToT outperforms the state-of-the-art methods by a large margin in terms of both perplexity and KL-divergence. We also show that the quality of the generated latent topics of MB-ToT is promising.

Keywords: topic mining, microblog, generative model, social network, temporal analysis

1 Introduction

Microblogging services have received tremendous popularity and become a new form of information infrastructure in recent years. Twitter, in particular, allows users to conveniently publish short messages (i.e. tweets) for a variety of purposes, such as personal status updating and information news sharing. It is reported that the number of tweets posted daily had exceeded 340 million in 2012[1]. How to exploit useful information from such a high-volume data stream has become a hard and challenging problem. In order to help users to get the big picture, the approach for topic mining is urgently in demand [1,2,3].

In this paper, we study the problem of topic mining in microblogs with sheer scale and social network information. Many text mining techniques and generative probabilistic models have been developed for static plain-text corpus. For instance, Latent Dirichlet Allocation (LDA) [4] is proposed to analyze static collections of document. TwitterMonitor [5] is a system that performs trend detection over the plain text of Twitter stream. Each emerging topic is represented by a group of bursty keywords that suddenly appear in a time interval at an unusually high speed. A principal disadvantage of these models is ignoring both temporal metadata and social network information to some extent.

Microblogs are temporally sequential and concerned with specific social network information in fact. Firstly, each tweet is commonly associated with a timestamp representing its creating time. Thus, topics in microblogs are internally dynamic instead of static. It is difficult to distinguish different topics only by their word co-occurrences because they may consist of similar words. A reasonable assumption is that topics are mixture distributions influenced by both word co-occurrences and timestamps of tweets. Topics over Time (ToT) [6] and Topics Over Nonparametric Time (TONPT) [7] are such effective topic models that analyze topics with timestamps by beta distributions and a Dirichlet process mixture of normals, respectively. Secondly, we should resort to the social network information (i.e., users'

[1] http://en.wikipedia.org/wiki/Twitter

* Corresponding author e-mail: issjyin@mail.sysu.edu.cn

intrinsic interests, social contact relations and #hashtags) to mine better topics. The rationale behind is that the social network information can reveal the underlying motivation to create and share content, and provides an elegant way to reorganize topically diverse tweets into a single "document" with compact topic distribution, making topic mining from microblogs much simpler and treatable. In particular, many tweets discuss about users' personal encounters and interests rather than global events, thus users' intrinsic interests play an important role to model topic distributions. Moreover, there exist tweets such as conversation and retweet messages consisting of structured information on social network. The topic distributions for such tweets are usually relevant to their mention users besides plain contents. MicroBlog-Latent Dirichlet Allocation (MB-LDA) [8] applies both conversation and retweet relations to generate more precise topics in microblogs. #Hashtags, the human-annotated semantic tags in tweets, are another resource to upgrade the topic mining result. From a tag #job, we can infer that the tweet is probably about topics of job, recruit, employ, etc. A natural solution is to collect tweets consisting of tags to train the topic distributions of tags and then use them to infer the topic distributions of new tweets with same tags. In [9], the authors integrated the #hashtags as weakly supervised information into the topic modeling algorithms.

To make use of the characteristics of tweets and produce comprehensive topic analysis, we propose a novel topic model, MicroBlog-Topics over Time (MB-ToT), which takes both temporal metadata and social network information into consideration. Firstly, we assume each topic is a mixture distribution affected by both word co-occurrences and timestamps of tweets. Note that MB-ToT models absolute continuous temporal information rather than discretizes time or makes Markov assumptions over state transitions in time. This allows MB-ToT to capture various skewed shapes of rising and falling topic prominence. Subsequently, we combine users' intrinsic interests, social contact relations and #hashtags into MB-ToT because they play a supporting role in the topic mining result. Finally, we present a fast and effective Gibbs sampling implementation for the inference of MB-ToT. We conduct extensive experiments on a real dataset to evaluate MB-ToT from three different perspectives: the perplexity of held-out content, the KL-divergence and the quality of the generated latent topics. The promising experiment results show that our model clearly outperforms its competitors.

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

1. MB-ToT, a novel generative probabilistic model for topic mining in Microblogs, is proposed.
2. A fast and effective Gibbs sampling implementation for the inference of MB-ToT is presented.
3. Extensive experiments on a real dataset are conducted. The results demonstrate that MB-ToT outperforms the state-of-the-art methods.

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

2 Related Work

There exists extensive studies for topic mining in the literature, starting with the Topic Detection and Tracking program (TDT) [10] which aims to detect and track topics in news corpus. The previous solutions to this problem are clustering-based techniques [11]. Later on, the generative probabilistic models are introduced into use. LDA [4] assumes that documents can be treated as mixtures of topics, each of which is a probability distribution over words. Based on the hierarchical Bayesian analysis of the original texts, LDA can successfully explore underlying semantic structures of documents.

It is appropriate that LDA deems words to be exchangeable to identify semantic structures within each document. However, the implicit assumption of exchangeable documents is too strong to hold in all cases. Document collections such as scholarly journals, news articles, and tweets reflect evolving contents. In [12], the authors devised a probabilistic model to incorporate both content and time information in a unified framework. Dynamic topic model (DTM) [13] uses Gaussian time series to capture the evolution of topics in a sequentially organized corpus of documents. Continuous Time Bayesian Networks (CTBNs) [14] is composed of a Bayesian network and a continuous transition model. The structured stochastic processes that evolve over continuous time are described by a Markov assumption to avoid various granularity problems due to the time discretization. ToT [6] assumes that each topic is influenced by both word co-occurrences and temporal information. Unlike DTM and CTBNs, ToT parameterizes a continuous distribution over time associated with each topic, making it available to capture the change of topics' occurrence and correlations over time. TONPT [7], which is a supervised topic model based on ToT, uses a nonparametric density estimator to model topic-conditional timestamp distributions. Supervised latent Dirichlet allocation (sLDA) [15] is a derivative of LDA, adding a response variable (i.e. the timestamp) connected to each document.

With the rapid development of Microblogging services, how to dig out latent topics and internal semantic structures from tweets has become a hot research field [1,5,9]. Different from other data collections, tweets contain plenty of social network information and temporal metadata. Topic mining in microblogs can utilize such specific information to produce competitive results. In [16], a novel approach to detect in real-time emerging topics on Twitter was

presented. Firstly, the authors selected emerging terms by a novel aging theory and then studied the social relationships in the user network by Page Rank algorithm. Finally, a keyword-based topic graph was built for emerging topic detection. Labeled LDA [17], a scalable implementation of a partially supervised learning model, was introduced to characterize users and tweets. In [18], a mixture latent topic model framework was designed to model user posting behavior in Twitter. The model mainly combines three factors (i.e., breaking news, posts from social friends and user's intrinsic interests) and yields encouraging results. TimeUserLDA [19] considers both temporal information of microblog posts and users' personal interests to find bursty topics from the tweet streams. MB-LDA [8] utilizes both contact relation and document relation to help topic mining in microblogs. Therefore, it can find not only the topics of microblogs, but also the topics focused by contacts. TM-LDA [20] was presented to efficiently model the topics and topic transitions that naturally arise in a tweet stream. The key step of TM-LDA is to learn the transition parameters among topics by minimizing the prediction error on the topic distribution in sequent microblogs. #Hashtags were integrated into the topic modeling algorithms to obtain better coherent topics [9]. However, none of the topic models mentioned above takes both temporal metadata and social network information into consideration adequately.

3 MB-ToT Model

In this section, we introduce a novel generative probabilistic model MB-ToT for topic mining in microblogs. Firstly, we discuss about the impact of both temporal metadata and social network information on the topic mining result. Secondly, we describe the framework of MB-ToT. Finally, we infer MB-ToT by a Gibbs sampling implementation. The notations used in MB-ToT are summarized in Table 1.

3.1 Influence of Temporal Metadata and Social Network Information

The temporal metadata of tweets is a key factor in the topic analysis. Assuming there exist two topics in microblogs, one is about WWDC 2011 and the other is about WWDC 2012. Both topics discuss about the "Apple Worldwide Developers Conference" held annually in California by Apple Inc. Since WWDC is primarily used by Apple to showcase its new softwares and technologies for developers, the word co-occurrences of these two topics are extremely similar, such as "apple", "iphone", "ios", etc. Being unaware of the one-year separation between WWDC 2011 and WWDC 2012, the topic models are inclined to confound them. One solution to distinguish such topics is using the temporal information during the topic modeling process.

Tweets are so different from plain-text corpus that they contain plenty of social network information, including users' intrinsic interests, social contact relations and #hashtags. The social network information also influences the topic mining result significantly. In particular, many tweets are about users' personal encounters and interests rather than global events, thus the topic distribution of such tweet is inferred from its author's intrinsic interests. But for tweets associated with identical social contact relation, their topic distributions are depended on the contact. Note that, the social contact relation is defined as the latent semantic relationship between tweets (starting with symbol "@" or "RT") and contacts. Assuming two tweets: "@mashable the onion launches a news iphone app ... wo any news lol" and "RT @mashable How much does the app charge?", if we recognize the social contact relation, we can build a relationship between these two seemingly unrelated tweets and infer that "app" in the latter tweet refers to "news iphone app". Now #hashtags occur frequently in tweets and become another useful resource to improve the topic mining result. Although #hashtags are not originally part of tweets, they evolve as a convenience tool for the organization of tweets. Assuming a tweet: "#autism gene study", if we add "#autism" into the topic analysis, we can learn that the "gene study" is apparently related to the topic distribution of "autism".

3.2 MB-ToT Framework

MB-ToT first views each topic as a mixture distribution influenced by both word co-occurrences and timestamps of microblogs, and then integrates the social network information into the topic modeling process based on three observations: (1) If tweets contain identical #hashtag, they are inclined to share its inherent topic distribution. (2) If tweets are associated with identical social contact relation, they prefer to discuss about the topics concerned by this contact. (3) If tweets are posted by the same user, they often reflect the user's intrinsic interests and are more likely to share the same topic distribution. In other words, drawing the topic distribution depends on the category of the tweet. For reasons of model simplicity and effectiveness, MB-ToT assumes each tweet belong to one category, which is determined by the social network information. Different social network information has different priority in the category determining process: the priority of the #hashtag is the highest, followed by the social contact relation and the user's intrinsic interests. For instance, suppose that a tweet associated with both #hashtag and social contact relation, and then its category only depends on the #hashtag. MB-ToT makes use of the temporal metadata in the same way as ToT does, it can create a topic with a broad time distribution and draw a distinction between

Table 1: Notations Used in MB-ToT

SYMBOL	DESCRIPTION
D,K,V	number of tweets, topics, and unique words, respectively
U,C,H	number of users, contacts, and #hashtags, respectively
x_d	category of tweet d, 0 for tweets with #hashtag, 1 for conversation or retweet tweets, and 2 for otherwise
N_d	number of words in tweet d
n_k^v	number of words v are assigned to topic k
n_u^k, n_c^k, n_h^k	number of words associated with user u, contact c, and #hashtag h for topic k, respectively
$\theta_u, \theta_c, \theta_h$	the multinomial distribution of topics to user u, contact c, and #hashtag h, respectively
$\alpha, \gamma, \eta, \beta$	Dirichlet priors for $\theta_u, \theta_c, \theta_h, and \phi_k$, 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 tweet d
$z_{d,i}$	topic of ith word in tweet 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 tweet d
$t_{d,i}$	the timestamp associated with ith word in tweet d
$\bar{t_k}, s_k^2$	the sample mean and variance of timestamps belonging to topic k, respectively

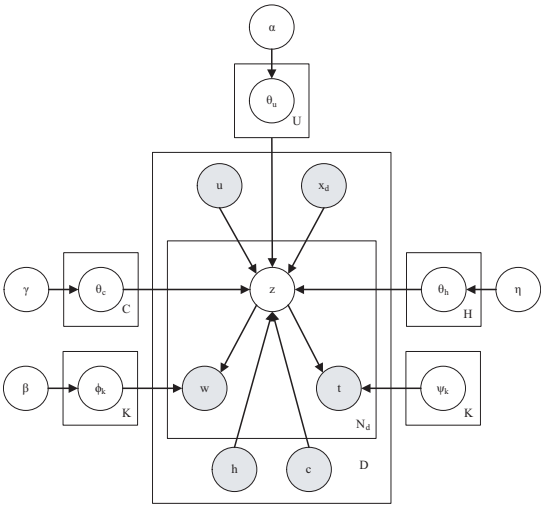

Fig. 1: Bayesian Graphical Framework of MB-ToT

topics due to their changes over time. The beta distribution is an appropriate choice to describe various skewed shapes of rising and falling topic prominence in microblogs. The Bayesian graphical framework of MB-ToT is illustrated in Figure 1.

Let θ_d denote the topic distribution of tweet d, then the detailed generative process of MB-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 microblog $d \in [1,D]$:
  (a)Determine the category x_d;
  (b)Draw a multinomial θ_d;
    i.If $x_d = 0$, and the first #hashtag is h, then draw a multinomial θ_h from a Dirichlet prior η, and assign the value of θ_h to θ_d;
    ii.Else if $x_d = 1$, and the first contact is c, then draw a multinomial θ_c from a Dirichlet prior γ, and assign the value of θ_c to θ_d;
    iii.Else if $x_d = 2$, and the author is u, then draw a multinomial θ_u from a Dirichlet prior α, and 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 tweet d, its posterior distribution of topics θ_d depends on the information from three modalities: the user's intrinsic interests, the social contact relation and the #hashtag.

$$P(\theta_d|\alpha,\gamma,\eta) = \begin{cases} P(\theta_h|\eta), & x_d = 0, \\ P(\theta_c|\gamma), & x_d = 1, \\ P(\theta_u|\alpha), & x_d = 2. \end{cases} \quad (1)$$

The joint probability distribution of tweet d is:

$$P(\mathbf{w},\mathbf{t},\mathbf{z}|\alpha,\gamma,\eta,\beta,\Psi) = \begin{cases} P(\mathbf{w}|\mathbf{z},\beta)P(\mathbf{t}|\mathbf{z},\Psi)P(\mathbf{z}|\eta), & x_d = 0, \\ P(\mathbf{w}|\mathbf{z},\beta)P(\mathbf{t}|\mathbf{z},\Psi)P(\mathbf{z}|\gamma), & x_d = 1, \\ P(\mathbf{w}|\mathbf{z},\beta)P(\mathbf{t}|\mathbf{z},\Psi)P(\mathbf{z}|\alpha), & x_d = 2. \end{cases} \quad (2)$$

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

$$\begin{aligned} \theta_d = \theta_h|\eta &\sim Dirichlet(\eta),\ if\ x_d = 0, \\ \theta_d = \theta_c|\gamma &\sim Dirichlet(\gamma),\ if\ x_d = 1, \\ \theta_d = \theta_u|\alpha &\sim Dirichlet(\alpha),\ if\ x_d = 2, \\ \phi_k|\beta &\sim Dirichlet(\beta), \\ z_{d,i}|\theta_d &\sim Multinomial(\theta_d), \\ w_{d,i}|\phi_{z_{d,i}} &\sim Multinomial(\phi_{z_{d,i}}), \\ t_{d,i}|\psi_{z_{d,i}} &\sim Beta(\psi_{z_{d,i}}). \end{aligned}$$

3.3 MB-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, tweet categories, words, users, contacts, and #hashtags are observed variables, while the topic structure and its changes over time are hidden variables.

The inference can not be done exactly in MB-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 $P(z_{d,i}|\mathbf{w},\mathbf{t},z_{\neg(d,i)},\alpha,\gamma,\eta,\beta,\Psi)$. Taking advantage of conjugate priors, the joint distribution $P(\mathbf{w},\mathbf{t},\mathbf{z}|\alpha,\gamma,\eta,\beta,\Psi)$ can be resolved into several components:

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

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

$$P(\mathbf{z}|\gamma)=\left(\frac{\Gamma(\sum_{k=1}^{K}\gamma_k)}{\prod_{k=1}^{K}\Gamma(\gamma_k)}\right)^{C}\prod_{c=1}^{C}\frac{\prod_{k=1}^{K}\Gamma(n_c^k+\gamma_k)}{\Gamma(\sum_{k=1}^{K}(n_c^k+\gamma_k))}, \quad (5)$$

$$P(\mathbf{z}|\eta)=\left(\frac{\Gamma(\sum_{k=1}^{K}\eta_k)}{\prod_{k=1}^{K}\Gamma(\eta_k)}\right)^{H}\prod_{h=1}^{H}\frac{\prod_{k=1}^{K}\Gamma(n_h^k+\eta_k)}{\Gamma(\sum_{k=1}^{K}(n_h^k+\eta_k))}, \quad (6)$$

$$P(\mathbf{z}|\alpha)=\left(\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\right)^{U}\prod_{u=1}^{U}\frac{\prod_{k=1}^{K}\Gamma(n_u^k+\alpha_k)}{\Gamma(\sum_{k=1}^{K}(n_u^k+\alpha_k))}. \quad (7)$$

We can conveniently obtain the conditional probability by using the chain rule. Note that different categories of tweets have different conditional probability values. In particular, when $x_d=0$,

$$\begin{aligned}&P(z_{d,i}|\mathbf{w},\mathbf{t},z_{\neg(d,i)},\alpha,\gamma,\eta,\beta,\Psi)\\&=P(z_{d,i}|\mathbf{w},\mathbf{t},z_{\neg(d,i)},\eta,\beta,\Psi)\\&\propto\frac{n_{z_{d,i}}^{w_{d,i}}+\beta_{w_{d,i}}-1}{\sum_{v=1}^{V}(n_{z_{d,i}}^{v}+\beta_v)-1}\times\frac{n_h^{z_{d,i}}+\eta_{z_{d,i}}-1}{\sum_{k=1}^{K}(n_h^k+\eta_k)-1}\times p(t_{d,i}|\psi_{z_{d,i}}),\end{aligned} \quad (8)$$

when $x_d=1$,

$$\begin{aligned}&P(z_{d,i}|\mathbf{w},\mathbf{t},z_{\neg(d,i)},\alpha,\gamma,\eta,\beta,\Psi)\\&=P(z_{d,i}|\mathbf{w},\mathbf{t},z_{\neg(d,i)},\gamma,\beta,\Psi)\\&\propto\frac{n_{z_{d,i}}^{w_{d,i}}+\beta_{w_{d,i}}-1}{\sum_{v=1}^{V}(n_{z_{d,i}}^{v}+\beta_v)-1}\times\frac{n_c^{z_{d,i}}+\gamma_{z_{d,i}}-1}{\sum_{k=1}^{K}(n_c^k+\gamma_k)-1}\times p(t_{d,i}|\psi_{z_{d,i}}),\end{aligned} \quad (9)$$

when $x_d=2$,

$$\begin{aligned}&P(z_{d,i}|\mathbf{w},\mathbf{t},z_{\neg(d,i)},\alpha,\gamma,\eta,\beta,\Psi)\\&=P(z_{d,i}|\mathbf{w},\mathbf{t},z_{\neg(d,i)},\alpha,\beta,\Psi)\\&\propto\frac{n_{z_{d,i}}^{w_{d,i}}+\beta_{w_{d,i}}-1}{\sum_{v=1}^{V}(n_{z_{d,i}}^{v}+\beta_v)-1}\times\frac{n_u^{z_{d,i}}+\alpha_{z_{d,i}}-1}{\sum_{k=1}^{K}(n_u^k+\alpha_k)-1}\times p(t_{d,i}|\psi_{z_{d,i}}).\end{aligned} \quad (10)$$

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

$$\phi_{k,v}=\frac{n_k^v+\beta_v}{\sum_{v=1}^{V}(n_k^v+\beta_v)}, \quad (11)$$

$$\theta_{u,k}=\frac{n_u^k+\alpha_k}{\sum_{k=1}^{K}(n_u^k+\alpha_k)}, \quad (12)$$

$$\theta_{c,k}=\frac{n_c^k+\gamma_k}{\sum_{k=1}^{K}(n_c^k+\gamma_k)}, \quad (13)$$

$$\theta_{h,k}=\frac{n_h^k+\eta_k}{\sum_{k=1}^{K}(n_h^k+\eta_k)}. \quad (14)$$

For the sake of simplicity and speed, Ψ is updated after each Gibbs sample by the method of moments estimates [2], detailed as follows:

$$\hat{\psi}_{k,1}=\bar{t_k}\left(\frac{\bar{t_k}(1-\bar{t_k})}{s_k^2}-1\right), \quad (15)$$

$$\hat{\psi}_{k,2}=(1-\bar{t_k})\left(\frac{\bar{t_k}(1-\bar{t_k})}{s_k^2}-1\right). \quad (16)$$

After finishing the inference process, MB-ToT can detect topics from tweets and assign the most representative words to each topic. Additionally, MB-ToT discovers the changes of each topic over time by a beta distribution with parameters from equation (15) and (16). In summary, MB-ToT offers several perspectives in the topic analysis and is a convenient tool for topic mining in microblogs.

4 Experiments

In this section, MB-ToT is evaluated empirically over a large crawl of Twitter data from three different perspectives: the perplexity of held-out content, the KL-divergence and the quality of the generated latent topics.

4.1 Dataset

To validate MB-ToT, we use a Twitter dataset with 3,845,622 messages collected from September 2009 to January 2010 [21]. Tweets are usually short and informal, which makes the quality of tweets varies a lot from each other, therefore specific data preprocessing techniques are required to filter low-quality tweets. We firstly prepare a stop word list in advance to eliminate stop words in original messages, since stop words appear with high

[2] http://www.itl.nist.gov/div898/handbook/eda/section3/eda366h.htm

Table 2: Description of Dataset

# of tweets	128,183
# of unique words	41,051
# of users	1,639
# of contacts	7,503
# of #hashtags	2,489
# of tweets starting with "@" or "RT"	10,048
# of tweets containing one #hashtag at least	21,736
# of tokens in tweets	1,425,437
average length of each tweet	11
minimal timestamp (seconds)	1251734400.0
maximal timestamp (seconds)	1260280443.0

frequency but do harm to topic mining results. Secondly, we apply the Snowball stemming algorithm[3] to execute the word stem process for reducing inflected (or sometimes derived) words to their stem, base or root form. For example, "stemmer", "stemming", and "stemming" are based on the root word "stem". In this way, words in tweets with the same stem are recognized as synonyms to improve the topic mining result. Thirdly, we filter out words with less than 20 occurrences in our dataset and only keep the tweets with more than 8 terms. Finally, we build a medium dataset containing 128,183 tweets for experiment evaluations. The detail information of our dataset is shown in Table 2.

LDA [4], ToT [6], and MB-LDA [8] are chosen for comparison. For the sake of fairness, the parameters α, and β in all models are set to 0.1 and 0.02, respectively. Both η and γ are set to 0.1 in MB-ToT. The simulations are carried out on an Intel Core Dual PC with 2.93 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 tweets, we use the formula as below:

$$Perplexity(D) = \exp\left(-\frac{\sum_d \sum_i^{N_d} \log p(w_{d,i})}{\sum_d N_d}\right). \quad (17)$$

Figure 2 shows the perplexities for our MB-ToT and other baselines with different number of topics until they reach the convergence after enough iterations. We observe that MB-ToT achieves the best perplexity, followed by MB-LDA. This means that integrating the social network information into the topic modeling process leads to better performance. MB-ToT outperforms MB-LDA because it utilizes not only social contact relations but also #hashtags, which indeed have a positive effect on the topic mining result. The performance of ToT is extremely comparable to LDA. Note that the perplexities of all models do not change significantly when the number of iterations is greater than 200.

[3] http://snowball.tartarus.org/

Table 3: Mean of KL Divergence between Topics (K=50)

	LDA	MB-LDA	ToT	MB-ToT
Mean of KL	3.725	3.708	3.750	**3.806**

4.3 Discriminability of Topics

The discriminative quality of extracted topics is another measurement to evaluate the performance of topic models. We use the pairwise KL-divergence between topics to measure the discriminative quality:

$$KL(\phi_1, \phi_2) = \sum_{w_{d,i}} p(w_{d,i}|\phi_1) \log \frac{p(w_{d,i}|\phi_1)}{p(w_{d,i}|\phi_2)}. \quad (18)$$

The more discriminative two topics are, the large their KL-divergence is. Two topics are identical when the KL-divergence is equal to 0. Table 3 displays the mean of KL-divergence results when the number of topics is 50. MB-ToT achieves the largest value, followed by ToT. This implies that the temporal metadata is the key factor to extract discriminative topics. Both MB-LDA and LDA prefer to discover topics with more duplicated topical words leading to the smaller means of KL-divergence.

4.4 Effectiveness of Latent Topics

The main purpose of topic models for microblogs 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 [18]. We design an experiment to compare quality of latent topics generated by MB-LDA and our model. The reason why we choose MB-LDA for comparison is that MB-LDA is a topic model customized for the Twitter data, while LDA and ToT are proposed for well written documents (e.g., news articles). The number of topics is set to 50 and ten latent topics which are the same salient topics for both models are extracted. Each topic is represented with the top 50 words. Table 4 gives only top 10 topical words for each topic due to the limit of space. We ask four labelers to label which one can better describe a topic. Figure 3 shows the final label result. On average, 65% of topics generated by MB-ToT are labeled better, while MB-LDA receives 35% of the votes. The rationale behind this result is that MB-LDA completely ignores the affect of both temporal metadata and #hashtags which are essential factors on modeling topics in the view of MB-ToT. As a result, MB-ToT discovers more comprehensive topics than MB-LDA.

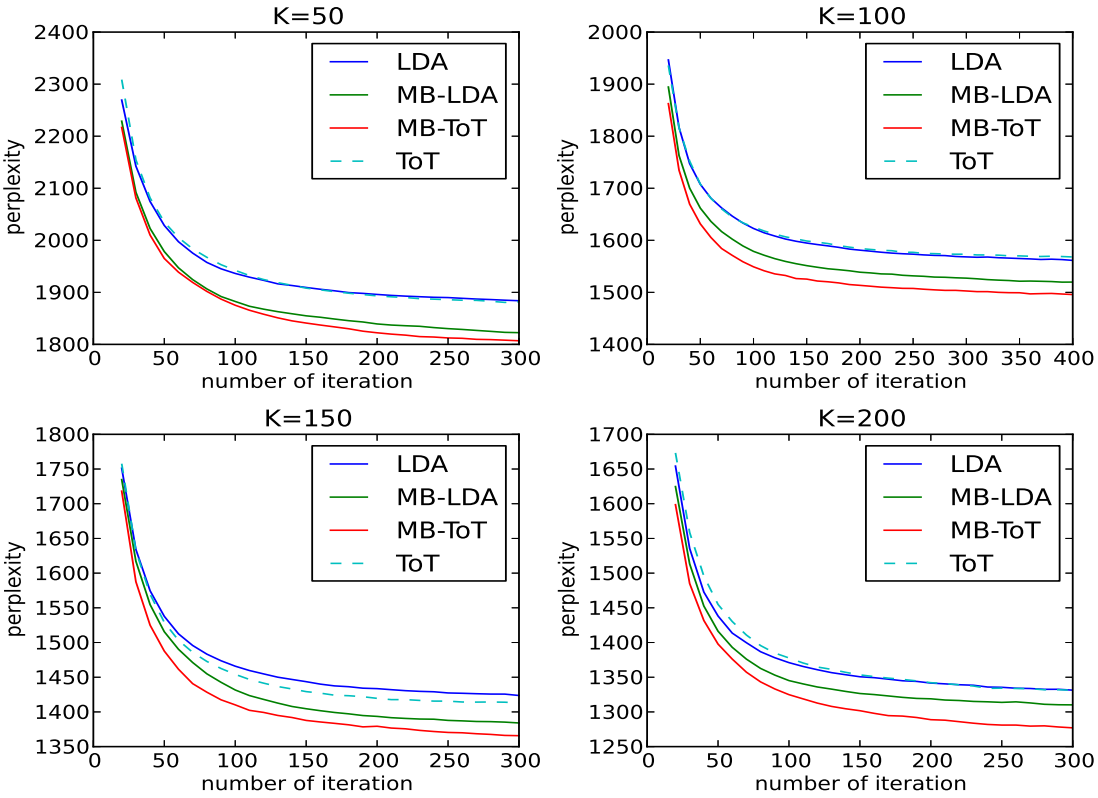

Fig. 2: Perplexities of Different Methods

Table 4: Top 10 Words for Latent Topics (K=50)

Topic 0		**Topic 1**		**Topic 2**		**Topic 3**		**Topic 4**	
MB-ToT	MB-LDA	MB-ToT	MB-LDA	MB-ToT	MB-LDA	MB-ToT	MB-LDA	MB-ToT	MB-LDA
music	music	love	love	job	job	google	google	dog	dog
song	song	god	god	manager	manager	iphone	iphone	pet	product
album	video	thing	#quote	sale	sale	twitter	app	image	skin
video	album	think	thought	service	service	app	apple	cat	care
itune	free	#quote	thing	engineer	engineer	facebook	video	anima	natural
artist	download	help	book	#tech	detail	web	microsoft	photo	cat
hip	itune	brain	live	#job	atlanta	video	mobile	entertainment	pet
hop	mp3	live	heart	detail	system	apple	web	#arizona	beauty
lil	hop	dream	read	atlanta	nurse	microsoft	week	getty	blog
record	hip	quote	quote	system	director	user	user	#coupon	personal
Topic 5		**Topic 6**		**Topic 7**		**Topic 8**		**Topic 9**	
MB-ToT	MB-LDA	MB-ToT	MB-LDA	MB-ToT	MB-LDA	MB-ToT	MB-LDA	MB-ToT	MB-LDA
health	obama	live	today	movie	video	football	football	food	food
bill	health	tonight	live	film	film	week	game	wine	wine
care	bill	show	tonight	twilight	movie	nfl	nfl	recipe	recipe
senate	senate	today	check	movi	fan	fantasy	week	holiday	water
obama	care	com	show	moon	music	pick	sport	thanksgiv	free
house	president	check	tomorrow	star	award	sport	team	coffee	thanksgiv
vote	say	tomorrow	watch	episode	star	rank	win	free	turkey
reform	house	tune	morn	brother	win	game	golf	turkey	bar
healthcare	vote	watch	night	video	concert	say	play	cook	drink
insurance	reform	ticket	tune	digital	movi	top	coach	restaurant	coffee

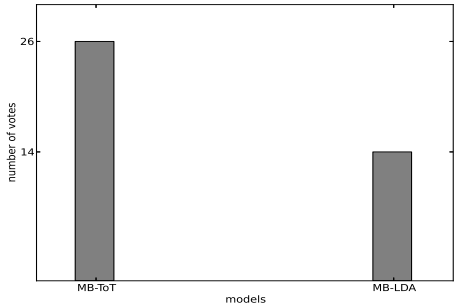

Fig. 3: Labeled Result

4.5 Dynamic Topics Analysis

The ability of modeling the changes of topics over time is very important for topic mining in microblogs. MB-ToT combines the temporal metadata to capture the dynamic topics. Figure 4 illustrates all beta distributions of each topic over time when the number of topics is 50. An immediate and obvious effect of this is to understand more precisely when and how long the topical 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.

4.6 Analysis of Social Network Information

Our MB-ToT finds not only the topics of tweets, but also the topics focused by #hashtags, contacts and users, respectively. Table 5 gives several examples of tweets associated with the topic focused by a specific #hashtag, contact and user, respectively. Once the topic distributions of #hashtags, contacts and users have been extracted by MB-ToT, we can carry on some personalization analysis of the tweets. For instance, we can mine topics in which the user (or contact) is interested, and recommend related tweets to the user (or contact). Another example is that we can obtain similar #hashtags according to their topic distributions. In this way, discovering the latent connection between seemingly different #hashtags becomes possible.

5 Conclusions

In this paper, we present and evaluate a time-aware topic model MB-ToT mixed with social network information, for effectively modeling and analyzing the topics that naturally arise in microblogs. Each topic is treated as a mixture distribution influenced by both word co-occurrences and timestamps of microblogs. In light of this, MB-ToT is able to capture the changes in the occurrence of topics. Additionally, users' intrinsic interests, social contact relations and #hashtags are used to strengthen the topic analysis ability of MB-ToT. Finally, the inference of MB-ToT is completed by a Gibbs sampling. Extensive experiments on a real dataset demonstrate that MB-ToT outperforms its competitors.

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. We will also further study the impact of #hashtags for latent topic structures. 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.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 61033010, No. 61272065), Natural Science Foundation of Guangdong Province (No. S2011020001182), Research Foundation of Science and Technology Plan Project in Guangdong Province (No. 2009B090300450, No. 2010A040303004, No. 2011B040200007).
The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] A. Angel, N. Koudas, N. Sarkas, and D. Srivastava, Dense subgraph maintenance under streaming edge weight updates for real-time story identification, Proceedings of the 38th International Conference on Very Large Data Bases, VLDB Endowment, 574-585 (2012).
[2] C. Budak, D. Agrawal, and A. E. Abbadi, Structural trend analysis for online social networks, Proceedings of the 37th International Conference on Very Large Data Bases, VLDB Endowment, 646-656 (2011).
[3] T. Takahashi, R. Tomioka, and K. Yamanishi, Discovering emerging topics in social streams via link anomaly detection, Proceedings of the 11th International Conference on Data Mining, Los Alamitos: IEEE Computer Society, 1230-1235 (2011).
[4] D. Blei, A. Ng, and M. Jordan, Latent dirichlet allocation, The Journal of Machine Learning Research, **3**, 993-1022 (2003).
[5] M. Mathioudakis and N. Koudas, Twittermonitor: Trend detection over the twitter stream, Proceedings of the ACM SIGMOD international conference on Management of data, New York: ACM Press, 1155-1158 (2010).

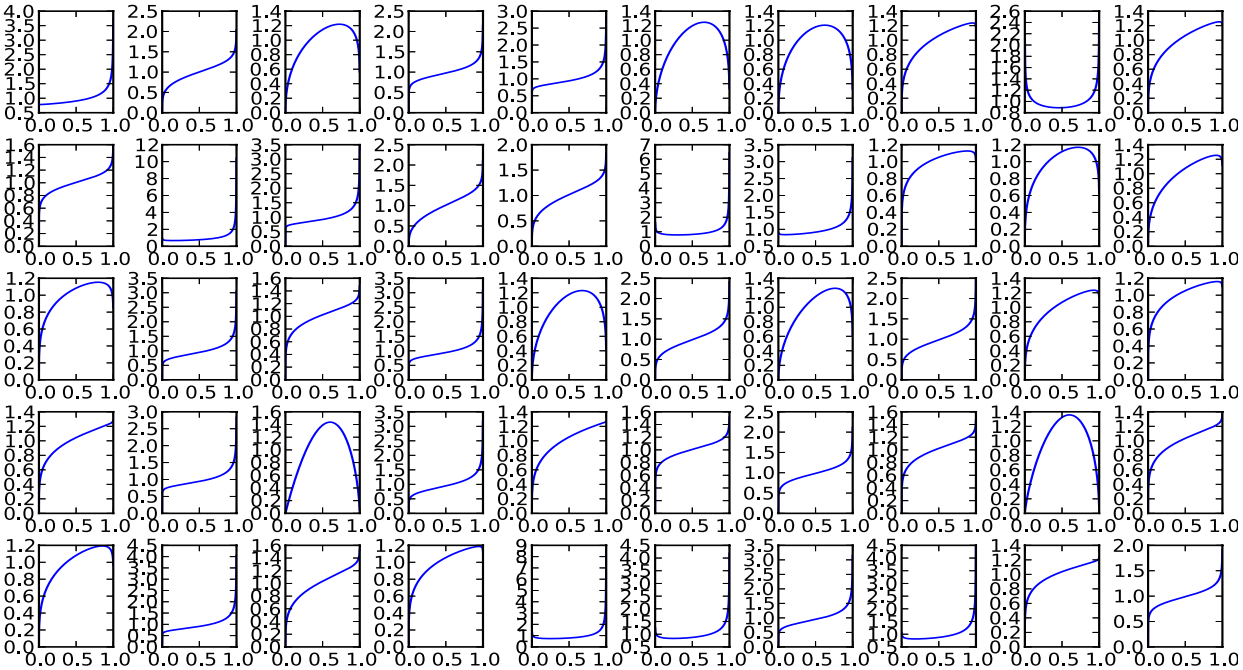

Fig. 4: Beta Distributions of Topics over Time (K=50)

Table 5: Examples of Relations between Social Network Information and Topics

Information Modality	Topic	Related Tweets
#job	Topic 2	New #job: Senior Software Engineer Loader Development, C++, Linux, - CyberCoders - Boston, MA?Framingham, MA #jobs #tech
		New #job: SURVEY TAKER-*Part Time*(Receptionist*Administrative Assistant*Retail) -APPLY NOW!-$75*/per ł Online Research (Paid Online...
		New #job: Lead Engineer, Transmission Line Engineer, Engineering Lead - Overhead Transmission Lines, Design - CyberCoders - Boston,...
@OGOchoCinco	Topic 8	@OGOchoCinco Hello Chad, Describe what have been the ”Highlights” of your NFL Career as a Wide Receiver for the Bengals?
		@OGOchoCinco Planning any solid touchdown celebrations for this week’s game against the Steel Curtain?
		RT @OGOchoCinco: People I’m going live on NFL NETWORK in 10 minutes, turn to 212
7429102 (User ID)	Topic 5	Watch Sen. Warner’s town hall on health care reform now - live on http://warner.senate.gov/townhall -staff
		Just finished talking to ABCNews.com’s TopLine about WH mtg on health care and Afghanistan. Check our website for the video soon.
		Heading to Senate floor with 9 freshmen colleagues for our second series of floor speeches on health care - how we can control costs

[6] X. Wang and A. McCallum, Topics over time: a non-markov continuous-time model of topical trends, Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, New York: ACM Press, 424-433 (2006).

[7] D. Walker, K. Seppi, and E. Ringger, Topics over nonparametric time: A supervised topic model using bayesian nonparametric density estimation, Proceedings of the 9th Bayesian Modelling Applications Workshop, 1-10 (2012).

[8] C. Zhang and J. Sun, Large scale microblog mining using distributed MB-LDA, Proceedings of the 21st international conference companion on World Wide Web, New York: ACM Press, 1035-1042 (2012).

[9] X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, and H. Wang, Entity-centric topic-oriented opinion summarization in twitter, Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, New York: ACM Press, 379-387 (2012).

[10] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, Topic detection and tracking pilot study: Final report, Proceedings of Broadcast News Transcription and Understanding Workshop, 194-218 (1998).
[11] J. Allan, Topic detection and tracking: event-based information organization, Kluwer Academic Publishers, Norwell, MA, (2002).
[12] Z. Li, B. Wang, M. Li, and W. Ma, A probabilistic model for retrospective news event detection, Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, New York: ACM Press, 106-113 (2005).
[13] D. Blei and J. Lafferty, Dynamic topic models, Proceedings of the 23rd international conference on Machine learning, New York: ACM Press, 113-120 (2006).
[14] U. Nodelman, C. Shelton, and D. Koller, Continuous time bayesian networks, Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence, San Francisco, CA: Morgan Kaufmann Publishers Inc., 378-387 (2002).
[15] D. M. Blei and J. D. Mcauliffe, Supervised topic models, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, New York: Curran Associates, Inc., 1-8 (2007).
[16] M. Cataldi, L. D. Caro, and C. Schifanella, Emerging topic detection on twitter based on temporal and social terms evaluation, Proceedings of the 10th International Workshop on Multimedia Data Mining, New York: ACM Press, 1-10 (2010).
[17] D. Ramage, S. Dumais, and D. Liebling, Characterizing microblogs with topic models, International AAAI Conference on Weblogs and Social Media, California: The AAAI Press, 130-137 (2010).
[18] Z. Xu, Y. Zhang, Y. Wu, and Q. Yang, Modeling user posting behavior on social media, Proceedings of the 35th annual international ACM SIGIR conference on Research and development in information retrieval, New York: ACM Press, 545-554 (2012).
[19] Q. Diao, J. Jiang, F. Zhu, and E. Lim, Finding bursty topics from microblogs, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Pennsylvania: Association for Computational Linguistics, 536-544 (2012).
[20] Y. Wang, E. Agichtein, and M. Benzi, TM-LDA: efficient online modeling of latent topic transitions in social media, Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, New York: ACM Press, 123-131 (2012).
[21] Z. Cheng, J. Caverlee, and K. Lee, You are where you tweet: a content-based approach to geo-locating twitter users, Proceedings of the 19th ACM international conference on Information and knowledge management, New York: ACM Press, 759-768 (2010).

Shaopeng Liu is currently a Ph.D. candidate at Information Science and Technology School, Sun Yat-Sen University. His major research interests include Data Mining and Machine Learning.

Jian Yin received the B.S., M.S., and Ph.D. degrees from Wuhan University, China, in 1989, 1991, and 1994, respectively, all in computer science. He joined Sun Yat-Sen University in July 1994 and now he is a professor and Ph.D. supervisor of Information Science and Technology School. He has published more than 100 refereed journal and conference papers. His current research interests are in the areas of Data Mining, Artificial Intelligence, and Machine Learning. He is a senior member of China Computer Federation.

Jia Ouyang is currently a Ph.D. candidate at Information Science and Technology School, Sun Yat-Sen University. His major research interests include Data Privacy and Data Mining.

Yun Huang is currently a Ph.D. candidate at Information Science and Technology School, Sun Yat-Sen University. His major research interests include Data Mining and Machine Learning.

Piyuan Lin received his B.S. and M.S degrees, both in computer science, from the University of Electronic Science and Technology of China in 1984 and 1989, respectively. He is currently a senior member of the China Computer Federation, and a professor of College of Informatics, South China Agricultural University. His research interests include bioinformatics, data mining, and network & information security. Professor Lin has published more than 40 research papers.