Context-aware Topic Modeling for Content Tracking in Social Media

Jinjing Zhang[1], Jing Wang[2], and Li Li[1,*]

[1] School of Computer and Information Science, Southwest University, Chongqing, China
[2] Economy and Technology Developing District, Henan, China
1476509610@qq.com,382876766@qq.com,lily@swu.edu.cn

Abstract. Content in social media is difficult to analyse because of its short and informal feature. Fortunately, some social media data like tweets have rich hashtags information, which can help identify meaningful topic information. More importantly, hashtags can express the context information of a tweet better. To enhance the significant effect of hashtags via topic variables, this paper, we propose a context-aware topic model to detect and track the evolution of content in social media by integrating hashtag and time informationnamed hashtag-supervised Topic over Time(hsToT). In hsToT, a document is generated jointly by the existing words and hashtags(the hashtags are treated as topic indicators of the tweet). Experiments on real data show that hsToT capture hashtags distribution over topics and topic changes over time simultaneously. The model can detect the crucial information and track the meaningful content and topics successfully.

Keywords: topic model, content evolution, topic over time, social media.

1 Introduction

In recent years, some conventional topic models such as LDA[1] and PLSA[2] have been proposed successfully in mining topics for a diverse range of document genres. However, for the data in tweets, they always fail to achieve high quality underlying topics because of its short and informal feature. Likely, there are several types of metadata could help identify the contents of tweets, such as the associated short url, picture, and #hashtag[3][4]. Among these metadata types, hashtags always play crucial roles in content analysis. Hashtags can not only express the context information of a tweet to the fullest, but also act as weakly-supervised information when sampling topics from certain tweets. Meanwhile, hashtags enrich the expressiveness of topics.

Motivated by the above, we propose a context-aware topic model to identify and track the evolution of the contents in social media named hashtag-supervised Topic over Time(hsToT). This model extends the classical LDA[1] by integrating the hashtags and time information. In hsToT, the distribution of hashtags over topics directly affects the topic sampling for a document. A topic is defined as

a set of words that is highly correlated. In addition, in order to capture topic evolution over time for our method, we model each topic with a multinomial distribution over timestamps and uses a beta distribution over a time span covering all the data.

The remainder of paper is organized as following. Section 2 review several representative works. Section 3 presents our approach in detail. We show the data preparation and discuss the experiment results in Section 4. The final section concludes the work.

2 Related Work

Topic model in social media. As a powerful text mining tool, topic models have successfully applied to text analysis. Unfortunately, traditional topic models LDA[1] and PLSI[2] do not work well with the messy form of data in Twitter. To overcome the noise in tweets, [5] merges user's tweets as a document. However, they ignore the content detection from topics. Thus, [6] proposes a probabilistic model that a topic depends on not only the users preference but also the preference of users; TCAM[7][8] focuses on analyzing user behaviors combining usersintrinsic interests and temporal context; Except the user features, mLDA[3] utilizes multiple contexts such as hashtags and time to discover consensus topics. While these works focus far more on user interests rather than content mining. Besides, some works take advantage semi-structured information, such as TWDA[9] and MA-LDA[10]. However, these methods ignore the dynamic nature of contents in social media.

Topic over time. To capture the topics change over time, qualitative evolution and quantitative evolution are two main analysis patterns. Qualitative evolution focus on some aspects of a topic like word distribution, inter-topic correlation, vocabulary, etc. [11] uses state space model to model the time variation. DTM[12] and TTM[13] are two typical models. However, the time must be discretized and the length of time intervals must be determined. Quantitative evolution focuses on the amount of data related to some topic at some timestamp, and models the time variation as an attribute of topics. The pioneering works in the literature are TOT[14], COT[4] and [15] where each topic is associated with a beta distribution over time. In this paper, we prefer quantitative evolution and replace the beta distribution with Dirichlet distribution so that the parameters could be simply estimated by Gibbs sampling.

3 Modeling Content Evolution in Social Media

In this section, firstly we introduce some preliminaries, especially the parameters used in the method would be interpreted. Then we explain the details of our hashtag based topic modeling solutions, namely hsToT.

3.1 Preliminaries

As the most popular topic model, Latent Dirichlet Allocation (LDA) has achieved numerous successful extensions. hsToT is LDA-based model, and also a probabilistic generative model. While different from LDA, hsToT includes two additional variables, namely hashtags and timestamps. We can discover the content

through a cluster of hashtags that frequently occur with a topic, then the content over time can be observed through topic distribution over timestamps.

Formally, we define a set of tweets as $\mathbf{W} = \{\mathbf{d}\}_{d=1}^{M}$. Each tweet is regarded as a document d and has a timestamp t. Suppose that document d is related to a word sequence $\mathbf{w}_d = \{w_1, w_2, ..., w_i, ..., w_N\}$ and a hashtag sequence $\mathbf{h}_d = \{h_1, h_2, ..., h_i, ..., h_L\}$, where N and H is the number of words and hashtags in document d. The rest notations used in this paper are listed in Table 1. For a corpus, T, N, and H are integer constants, while N and H are varying with different document. T is set manually.

Table 1. Notation in hsToT

$M, K,$	number of document and topics respectively
N, H, T	number of words, hashtags and timestamps in a document respectively
z, w, h, t, d	topic, word, hashtag, timestamps and document respectively
θ	multinomial distribution over topics for a hashtag
φ	multinomial distribution over words for a topic
ψ	multinomial distribution over timestamps for a topic
α, β, μ	Dirichlet prior parameters for θ, φ and ψ respectively

3.2 hashtag-supervised Topic over Time

In this subsection, we describe hashtag-supervised Topic over Time (hsToT) to directly uncover the latent relationship among topic, hashtag, and time. In hsTor, hashtags act as the weakly-supervised information in topics sampling. Fig.1 shows the graphical models of the hsToT. In Fig 1, each topic is typically

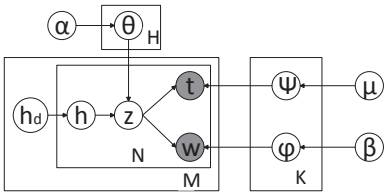

Fig. 1. Graphical model representation of hsToT.

represented by a distribution over words as φ with β as the Direchlet prior. hsToT also includes two distributions over hashtags and timestamps with respect to topics. hsToT do not directly sample the distribution over topics for a document d. Instead, it sample a hashtag's distribution over topics from the $K \times H$ matrix as the topics distribution of the document. Furthermore, the time feature is first discretized and each tweet is annotated with a discrete timestamp label (e.g. day, month, year). Naturally, time modality is captured by the variable t, and consequently topic evolution over time is obtained using multinomial distribution. In particular, each hashtag is characterized by a distribution over topics as θ with α as the Direchlet prior. We allocate a topic assignment z_i and a hashtag assignment h_i for each word w_i in the document d. In hsToT, each word is associated with a "hashtag-topic" assignment pair and a "topic-timestamp" pair. The generative process for hsToT is given as follows, shown in Fig.1. We

use variables z_i, h_i and t_i to represent a certain topic, hashtag, and timestamp associated with the word w_i respectively.

1. For each hashtag $h = 1 : H$, draw the mixture of topics $\theta_h \sim Dir(\alpha)$
2. For each topic $z = 1 : K$, draw the mixture of words $\varphi_z \sim Dir(\beta)$
3. For each topic $z = 1 : K$, draw the mixture of timestamps $\psi_z \sim Dir(\mu)$
4. For each document $d = 1 : M$, draw its words length N, and give its hashtag set $\mathbf{h}_d$
   (a) For each word w_i, $i = 1 : N_d$
      i. Draw a hashtag $h_i \sim Uniform(\mathbf{h}_d)$
      ii. Draw a topic $z_i \sim Mult(\theta_{h_i})$
      iii. Draw a word $w_i \sim Mult(\varphi_{z_i})$
      iv. Draw a timestamp $t_i \sim Mult(\psi_{z_i})$

In hsToT, there are three posterior distributions: hashtag-topic distribution θ, topic-word distribution φ and topic-timestamp distribution ψ. We assume that "topic-word" distribution and "hashtag-topic" distribution are conditionally independent. To efficiently estimate posterior distribution, we employ Gibbs sampling[16].Thus, the joint probability of words, topics, hashtags and timestamps is Eq.1.

$$p(\mathbf{w}, \mathbf{h}, \mathbf{t}, \mathbf{z}|\alpha, \beta, \mu, \mathbf{h_d}) = \mathbf{p}(\mathbf{w}|\mathbf{z}, \beta)\cdot \mathbf{p}(\mathbf{h}|\mathbf{h_d})\cdot \mathbf{p}(\mathbf{t}|\mathbf{z}, \mu)\cdot \mathbf{p}(\mathbf{z}|\mathbf{h_d}, \alpha) \tag{1}$$

In order to infer the hidden variables, we compute the posterior distribution of the hidden variables. The likelihood of a document d is Eq.2 .

$$\begin{aligned} p(\mathbf{w}_d|\theta, \varphi, \psi, \mathbf{h}_d) &= \prod_{i=1}^{N} p(w_i|\theta, \varphi, \psi, \mathbf{h}_d) \\ &= \prod_{i=1}^{N_d} \sum_{j=1}^{L_d} \sum_{k=1}^{K} \sum_{s=1}^{T} p(w_i, z_i = k, h_i = j, t_i = s|\theta, \varphi, \psi, \mathbf{h}_d) \\ &= \prod_{i=1}^{N_d} \sum_{j=1}^{L_d} \sum_{k=1}^{K} \sum_{s=1}^{T} p(w_i|z_i = k,\varphi)p(z_i = k|h_i = j, \theta)p(t_i = s|z_i = k, \psi)p_{jh_i} \\ &= \prod_{i=1}^{N_d} \sum_{j=1}^{L_d} \sum_{k=1}^{K} \sum_{s=1}^{T} \varphi_{w_i,k}\theta_{k,j}\psi_{s,k}p_{jh_i} \end{aligned} \tag{2}$$

Where p_{jh_i} represents the probability of $h_i = j$ when sampling a hashtag h_i from $\mathbf{h}_d$. The generating probability of the corpus is:

$$p(\mathbf{W}|\theta, \varphi, \psi, \mathbf{h}) = \prod_{d=1}^{M} p(\mathbf{w}_d|\theta, \varphi, \psi, \mathbf{h}_d) \tag{3}$$

The estimation method of posterior distributions in hsToT is Eq.4.

$$p(z_i = k, h_i = j|\mathbf{w}, \mathbf{h}^{-i}, \mathbf{t}, \mathbf{z}^{-i}) \propto \frac{n^k_{w_i,-i}+\beta}{n^k_{w',-i}+V\beta} \times \frac{n^j_{k,-i}+\alpha}{n^j_{k',-i}+K\alpha} \times \frac{n^k_{t_i,-i}+\mu}{n^k_{t',-i}+T\mu} \tag{4}$$

Where $-i$ means assignments except for current word in the current document d. In Eq.4, $n^j_{k,-i}$ is the number of words assigned to topic k in a document d, $n^j_{k',-i}$ is total number of words in a document d; $n^j_{w_i,-i}$ is the number of words assigned to topic k and hashtags j, $n^j_{k',-i}$ is total number of words assigned to topic k; $n^j_{t,-i}$ is the number of time words assigned to topic k and hashtags j,

$n^j_{t',-i}$ is total number of words assigned to topic k and hashtags j. Finally, when the sampling process reaches the convergence, we can get the results of φ, θ, and ψ by:

$$\varphi_{w,k} = \frac{n^k_w+\beta}{n^k_{w'}+V\beta} \quad \theta_{k,j} = \frac{n^j_k+\alpha}{n^j_{k'}+K\alpha} \quad \psi_{s,k} = \frac{n^k_t+\mu}{n^k_{t'}+T\mu} \tag{5}$$

From the generative process of hsToT, time modality is involved in topic discovery. However, this may impact the homogeneity of topics because time modality is assumed having the same "weight" in word modality. In practice, it is not. To address this issue, we adopt the same strategy as in TOT[14] where a balancing hyperparameter is introduced in order to balance word and time contribution in topic discovery. Naturally, we set the hyperparameter as the inverse of the number of words n_d.

4 Experiments

Experiments are conducted based on evaluation on results of topic detection and topic evolution over time. We take three topic models TOT[14], COT[4] and hgToT, which is our another model as the baselines. (hgToT is generally similar to COT in the usage of hashtags, but whose time modeling methold changes the beta distribution into the multinomial distribution.) TOT is extended from LDA by adding a Beta distribution over timestamps for topics. COT is extended from TOT by adding a multinomial distribution over hashtags for topics.

4.1 Data Preparation

The experiments are conducted on a twitter data set, named "TREC2011"[3]. The original data contains nearly 16 millions tweets posted from January 23rd to February 8th in 2011. Each tweet includes a user id and a timestamp. The process of the raw data is similar to the steps in [17]. The properties of the dataset are given in Table 2.To guarantee the convergence of Gibbs sampling, all results were obtained after 1000 iterations. Timestamps is divided by days. The hyperparameters in generative models (hsToT, hgToT, TOT) are set as $50/K$ for α and 0.04, 0.04, 0.01 for β, γ and μ respectively[11].

Table 2. Dataset properties

♯ tweets	♯ Unique words	♯ hashtags	♯ Average words	♯ Average hashtags
304,480	12,160	98,649	5.11	1.42

4.2 Evaluation on Topic Detection

To assess the effectiveness of topic models, a typical metric like perplexity on a held-out test set[17] has been widely used. The perplexity represents the performance of document modeling by comprehensively estimate the results of $p(z|d)$ and $p(w|z)$. Another automatic evaluation metrics *coherent score*[18] is proposed to measure the quality of topics from the perspective of topic visualization and semantic coherent. In this paper, we choose perplexity and coherent score as evaluation criteria.

[3] http://trec.nist.gov/data/microblog2011.html

Perplexity Perplexity indicates the uncertainty in predicting a single word. The lower the perplexity score is, the higher the performance will be. We compute this metric according to the method[1]. To equilibrate the different usage of hashtags in these methods, the computation of $p(w_d)$ is different. For TOT, the computation method of $p(w_d)$ is same as [1]. While for hsToT, hgToT and COT, $p(w_d) = \sum_{N_d} p(w_i) + \sum_{L_d} p(h_i)$. In this step, we hold out 10% of data for test and train these methods on the remaining 90% data. Fig.2 states the

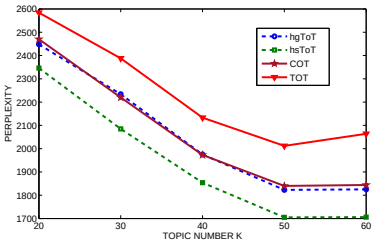

Fig. 2. perplexity results with different topic number K.

perplexity results with topic number $k = 20, 30, 40, 50, 60$. From the Fig.2, hsToT obviously outperforms other methods. This indicates that we can indeed improve the document modeling performance by taking advantage of hashtags especially regarding hashtags as weak-supervised information. In addition, For hsToT, the perplexity value reduces gradually with the increase of topic number, and then tends to stable when $k \geq 50$. While TOT is running into over-fitting. This means that our method is more stable. Based on this observation, the number of topic K is fixed at 50 in the remaining experiments.

Coherent Score To intuitively investigate the quality of topics, we analyze

Table 3. a sample of semantic representation of topic "EGYPT"

hgToT		hsToT		TOT
#egypt,	egypt	#egypt,	egypt	obama
#election,	obama	#mubarak,	people	mubarak
#25-Jan,	#egypt	#election,	obama	#mubarak
#tcot,	mubarak	#turbulence,	mubarak	turbulence
#sotu,	egyptian	#news,	egyptian	egypt

the topics from visualization perspective. For each topic, we take top 5 words or hashtags ordered by $p(w|z)$ or $p(h|z)$) as their semantic representation. By observing the 50 topics, there are two major kind of topics. One is "common topics", which are related to users's daily lives. The other is "time-sensitive topics" which may be some emergencies or hot news events.Overall, compared with TOT and hsToT can discover more meaningful hashtags and words highly related to a topic. Besides, hsToT can detect the content from a topic which TOT can not achieve. Table 3 lists an example of a common topic "EGYPT" learnt by all hsToT and TOT methods.

For TOT, both words and hashtags occur in the "topic-word" distribution. While in hgToT and hsToT, words and hashtags only occur in "topic-word"

and "topic-hashtag" distribution respectively. In addition, topic "EGYPT" also shows that hsToT outperforms TOT in discovering meaningful hashtag i.e., the semantics of topics for a common topic. For example, TOT only discovers one hashtag "#mubarak", whereas hsToT discovered more meaningful hashtags.

In order to quantitatively evaluate the topic quality of all test methods, we further utilize the automated metric, namely *coherent score*. The coherence score is that words belonging to a single concept will tend to co-occur within the same documents[19]. A larger coherence score means the topics are more coherent. Given a topic z and its top n words $V^{(z)} = (v_1^{(z)}, ..., v_n^{(z)})$ ordered by $p(w|z)$, the coherent score can be defined as:

$$C(z; V^{(z)}) = \sum_{t=2}^{n} \sum_{l=1}^{t} \log \frac{D(v_t^{(z)}, v_l^{(z)}) + 1}{D(v_l^{(z)})} \tag{6}$$

where $D(v)$ is the document frequency of word v, $D(v, v')$ is the number of document in which words v and v' co-occurred. The final results are $\frac{1}{K} \sum_k C(z_k; V^{(z_k)})$. For TOT, the average coherent score can be directly captured by this way. While for other three methods, a topic is jointly associated with a multinomial distribution over words and a multinomial distribution over hashtags. Therefore, we also need consider the top n hashtags when computing the coherent score of a given topic. The final average coherent score in these three methods is $\frac{1}{2K} \sum_k (C(z_k; V^{(z_k)}) + C(z_k; H^{(z_k)}))$.

The result is shown in Table 4, the number of top words ranges from 5 to 20. From Table 4, hsToT achieves the best performance(with p-value <0.01 by T-test), and hsToT receives the highest coherence score. hgToT and COT outperform TOT by 0.086 on average. It also shows hashtag co-occurrence frequency can help detect more coherent words, and improve the topic coherence greatly. Moreover, compared with the strategy of hashtags in COT, hsToT always receives better results. By observing the different results with the change of top n words(or hashtags), we find that the coherent score of all methods gradually reduces when the number of n increases.

Table 4. Average coherence score on the top n words(hashtags) in topics.

n	5	10	20
hgToT	-62.6	-239.3	-1043.5
hsToT	**-61.5**	**-236.5**	**-1029.2**
COT	-62.7	-239.4	-1042.9
TOT	-64.3	-243.6	-1066.4

5 Conclusions

In this paper, we introduced a new model(hsToT) of tracking topics over time in social media. By considering features such as hashtags, words and timestamps jointly, our model can successfully identify the meaningful topics precisely and track topic changes over time appropriately . Experiments on real dataset illustrate the effectiveness and efficiency of our methods.

Acknowledgments. The work was supported by the Fundamental Research Funds For the Central Universities (No. XDJK2017D059).

References

1. David M Blei, Andrew Y Ng, Michael I Jordan(2003)Latent dirichlet allocation. J Machine Learning Research, 3:993–1022.
2. Thomas Hofmann(1999)Probabilistic latent semantic indexing. In: 22nd ACM SIGIR, pp 50–57.
3. Jian Tang, Ming Zhang, and Qiaozhu Mei(2013)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.
4. Md Hijbul Alam, Woo-Jong Ryu, SangKeun Lee(2014)Context over time: Modeling context evolution in social media. In: Proceedings of the 3rd workshop on data-driven user behavioral modeling and mining from social media, pp 15–18.
5. Liangjie Hong, Brian D Davison(2010)Empirical study of topic modeling in twitter. In: Proceedings of the first workshop on social media analytics, pp 80–88.
6. Younghoon Kim, Kyuseok Shim(2011)Twitobi: A recommendation system for twitter using probabilistic modeling. In: Data mining (ICDM), 2011 IEEE 11th International Conference, pp 340–349.
7. Hongzhi Yin, Bin Cui, Ling Chen, Zhiting Hu, Zi Huang(2014)A temporal context-aware model for user behavior modeling in social media systems. Association for computing machinery. Special interest group on management of data, pp.1543-1554.
8. Hongzhi Yin, Bin Cui, Ling Chen, Zhiting Hu, Xiaofang Zhou(2015)Dynamic user modeling in social media systems. Acm transactions on information systems, 33(3), 10.
9. Shuangyin Li, Guan Huang, Ruiyang Tan, Rong Pan(2013)Tag-weighted dirichlet allocation. In: Data Mining (ICDM), 2013 IEEE 13th International Conference, pp 438–447.
10. Jing Wang, Li Li, Feng Tan, Ying Zhu, Weisi Feng(2015)Detecting hotspot information using multi-attribute based topic model. PloS one, 10(10):e0140539.
11. Mohamed Dermouche, Julien Velcin, Leila Khouas, Sabine Loudcher(2014)A joint model for topic-sentiment evolution over time. In: Data mining (ICDM), 2014 IEEE International Conference, pp 773–778.
12. David M Blei, John D Lafferty(2006)Dynamic topic models. In: Proceedings of the 23rd international conference on Machine learning, pp 113–120.
13. Tomoharu Iwata, Shinji Watanabe, Takeshi Yamada, Naonori Ueda(2009)Topic tracking model for analyzing consumer purchase behavior. In: IJCAI, Citeseer, vol 9, pp 1427–1432.
14. Xuerui Wang, Andrew McCallum(2006)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.
15. Hongzhi Yin, Bin Cui, Hua Lu, Yuxin Huang, Junjie Yao (2013). A unified model for stable and temporal topic detection from social media data. IEEE, International Conference on Data Engineering, Vol.48, pp.661-672.
16. Gregor Heinrich(2005)Parameter estimation for text analysis. Technical report, Technical report.
17. Xueqi Cheng, Xiaohui Yan, Yanyan Lan, Jiafeng Guo. Btm: Topic modeling over short texts. Knowledge and data engineering, IEEE Transactions, 26(12):2928–2941.
18. David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, Andrew McCallum(2011)Optimizing semantic coherence in topic models. In: Proceedings of the conference on empirical methods in natural language processing, pp 262–272.
19. Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng(2013)A biterm topic model for short texts. In: Proceedings of the 22nd international conference on world wide web, pp 1445–1456.