Detecting Social Topic by Hashtag-Weighted Topic Model over Time

Jie Qiu[1, a] Li Li[2, b]

[1]School of Computer and Information Science, Southwest University, Chongqing, 400715, China

[2] School of Computer and Information Science, Southwest University, Chongqing, 400715, China

[a]email: cherishyi@163.com [b] lily@swu.edu.cn

Keywords: Hashtag-weighted; Topic model; Twitter

Abstract. Nowadays, more and more social media platforms support hashtags to facilitate information classification. Like Twitter hashtags, a user-initiated hashtag can suggest emotion/mood, convey so much extra information in addition to the actual tweet. Hashtags have been widely used in topic analysis because of its informative effect, but all hashtags are created equally. In the paper, we propose a Hashtag-Weighted Topic Model over Time (HWOT) which assigns hashtags to deal with topic evolving over time with different hashtag weight. To leverage hashtags across topics in a specific time period, the topic of hashtag is represented as a multinomial distribution and the topic over time as a Beta distribution. Our model can uncover the latent relationships among topics, hashtags and time. The weight of the hashtag is learned via a novel context aware weakly supervised approach. Experiments on Twitter dataset show that our model can achieve better performance in terms of model perplexity. It further reveals the change of the topics over time.

Introduction

With the increasing number of social media platforms, people are able to express their thoughts and share their ideas. Especially in Twitter platforms, people post their opinions towards various topics and generate their own hashtags (start with the $\#$) to emphasize their opinions. Mining Twitter dataset is a good way to detect hot news and emergencies. Twitter dataset with both text data and metadata (i.e. hashtag and time information, which can be viewed as features of the corresponding documents) are called semi-structured dataset. Traditional topic models, such as PLSA and LDA always fail to model such semi-structured dataset [4]. To model such dataset needs more consideration on different attributes and relationship between different attributes of semi-structure dataset. Meanwhile, Twitter users are free to generate their own hashtags without considering the language and format of hashtags. It is common that the same hashtag may imply different latent topics at different time. The time feature of Twitter is valuable to detect the topic distribution in a specific timespan.

To cope with these issues, our model is established to achieve high quality when detecting social topics and discovering the relationships among hashtags, topics and time. In this paper, we propose a content aware topic model over time sequence, named Hashtag-Weighted Topic Model over Time (HWOT) to detect the evolution of social topic, which combines the user-generated hashtag and time information in tweets. It is a novel method to alleviate the informal problem in mining Twitter dataset and obtain the hashtag weight of each hashtag in tweet.

Related work

Topic evolution is one of the significant tasks in the field of social media. In order to get satisfactory results, many topic related models were proposed based on the foundation of LDA [1], PLSA [3]. They work well with traditional tasks in plain text mining. However these models fail to model semi-structured dataset

The Author-Topic model (ATM) [6] is a special method to model text via tags by regard tags as authors. Topic over time (TOT) [10] is an LDA-based model to detect the evolution of topic by counting the number of documents related to topic at each timestamp. Here, we refer to the method mentioned in TOT to model time distribution. They are strong baselines of our paper. Labeled LDA [5] manually defines that each topic is restricted to be associated with unique one label. Based on the labeled LDA, Li et al. [4] proposed TWDA to represent the tags with weight to evaluate the importance of the tags by the method of hashtag weight in our paper is based on the context aware method the prior of tags mentioned in TWDA

Our model is similar to theirs, but we both consider the hashtag and time into LDA and a novel process to test hashtag weight in tweets. In addition, we add hashtag weight layer to together hashtag external knowledge base and the corpus semantic analysis to balance different the importance of several hashtags in one document.

Hashtag-weighted Topic Model over Time

HWOT is a probabilistic generative model, describing a process of generating a tweet collection with hashtag weight and time information. The model is showed in Fig.1.

Notations in HWOT
D :the number of document
K :the number of topic
V :vocabulary size
H :the number of hashtags
α , β :Direchlet parameters
τ :haperparameter of hashtag assignment
N_d :the length of document d
H_d :the number of hashtags in d
h_d :the collection of hashtag in d
x , z :hashtag and topic assignments
w :the word associated with document
t :the timestamp associated with document
π :the weight of each hashtag in document
λ_1 :the factor 1 associated with π
λ_2 :the factor 2 associated with π

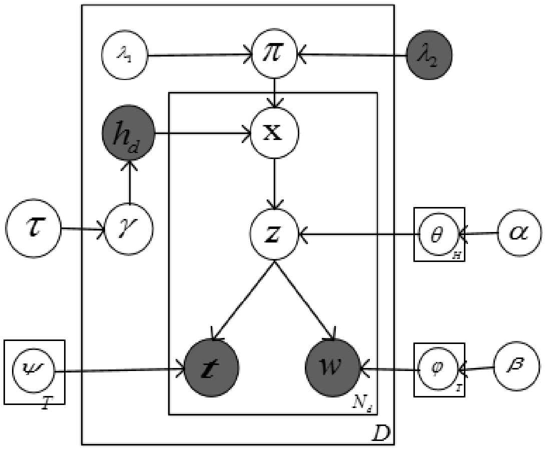

Fig. 1: HWOT: The generative model

In HWOT, we define the whole corpus as $C=\{d_1,d_2,\cdots,d_D\}$, the word sequence of document d is $w_d=\{w_{d_1},w_{d_1},\cdots,w_{d_{N_d}}\}$ and a hashtag sequence of document *d* is $H_d=\{h_{d_1},h_{d_2},,\cdots,h_{d_{Hd}}\}$, in which N_d and H_d is the number of words and hashtags in document d, respectively. Thus we can use a set of triples $D=\{(w_1,h_1,t_1),\cdots,(w_D;\ h_D;\ t_D)\}$ to represent the corpus. In HWOT, all words and hashtags in one document should have the same timestamp *t*. So each latent topic discovered by HWOT has a distribution with the corresponding timestamp to know when this topic emerges.

In particular, we denote a hashtag assignment for each word in the document d as an *N*-dimension vector x_d . We use a Bernoulli variable τ to represent the probability to assign hashtag in the condition of hashtag weight in document *d*. First we assign a hashtag x_{di} randomly from the h_d . Then, we sample the γ ~ *Bernoulli*(τ) to decide whether the current hashtag is the Maximum weight of document *d*. If so, x_{di} equals x_d . If not, the x_{di} is from the probability of multinomial distribution under the probability of γ . The content aware hashtag weight method will be introduced next. After hashtag assignment, we do the topic assignment z_{di} . In HWOT, we use a parameter τ as the factor to model the sampling process of hashtag section. τ is randomness of the probability for the hashtag assignment in current document. Here the τ is the value of 0 to 1. We use such factor to simulate the scene that Twitter users select different hashtag for their tweets.

Here, the content aware means that hashtags in tweet is decided by the whole tweets (without hashtags) and hashtag dictionary (http://www.hashtag.org), which is the twitter wiki for users. For the hashtags in tweets are informal and hard to understand, we first use hashtag dictionary to query the corresponding entity vocabularies of hashtag in tweets. We employ a crawler and regard each unique hashtag as input to obtain corresponding words collection.

In $\omega_{h_{di}} = \{\omega_{1h_{di}}, \omega_{2h_{di}}, \cdots, \omega_{jh_{di}}\}$, *j* is the *j*th entity word of hashtag *i* in document *d*. Taking hashtag #NBA or #nba for example, the results are some related words, such as 'basketball',' team', stars, etc. Next, we use traditional LDA to analysis the whole tweets (without considering weakly supervised information). We set the parameter α = 0:5, β = 0:5 and *k* = 50 the same with the paper [6]. Finally, we match entity words collection and the topic-word and document-topic distribution obtained by the LDA. We orient $\omega_{jh_{di}}$ as target to search the topic-word matrix to get the latent topic, then the topic as target to match the document-topic matrix to get the final probability of the current entity word of hashtag i in document d. Finally, we do the sum of the probability for each entity word of current hashtag. The number of the entity word is decide by the hashtag dictionary. We use the following Eq.1 to represent the process mentioned above.

$$\pi_{h_{di}} = \frac{\sum_{j} p(z_{dn} = t \mid w_{t,n} = \omega_{jh_{di}}, \alpha, \beta)}{c} \tag{1}$$

$\sum_{j} p(z_{dn} = t \mid w_{t,n} = \omega_{jh_{di}}, \alpha, \beta)$ is the weight of each hashtag in document *d*, calculated by the process b. The parameter c is a constant, which is used to limit the value of $\pi_{h_{di}}$ from 0 to 1.

Parameter estimation

In HWOT, the words, hashtags and timestamps of each document are observed while topics are hidden variables guided by the latent distribution parameters. In order to infer the hidden variables, we need to compute the posterior distribution of the hidden variables. We employ Gibbs Sampling to estimate parameter. The whole corpus's probability is listed as follows:

$$p(w,t,z,x \mid \alpha,\beta,\tau,\lambda,\psi) = \prod_{t=1}^{T} \frac{\Delta(C_Z + \beta)}{\Delta(\beta)} \prod_{t=1}^{T} \frac{\Delta(C_H + \alpha)}{\Delta(\alpha)} \prod_{d=1}^{D} \prod_{i=1}^{Nd} p(t_{d,i} \mid \psi_{Z_{d,i}}) p_{x'x_{d,i}} \tag{2}$$

Here we directly associate time with topics. Each topic k has a latent distribution k with respect to time. For a versatile characteristic of Bate distribution, we need to set timestamp to a range from 0 to 1. The probability of an observed timestamp t can be calculated as:

$$p(t \mid \psi_k) = \frac{\Gamma(\psi_{k,1} + \psi_{k,2})}{\Gamma(\psi_{k,1}) + \Gamma(\psi_{k,2})} (1-t)^{\psi_{k,1}-1} t^{\psi_{k,2}-1} \tag{3}$$

Where t_k and $S^{(t)^2}{}_k$ k are the mean and biased variance of timestamps for the current topic of the word. After interactions, $S^{(t)^2}{}_k$ and $\varphi_{k,i}$ can be evaluated as follows:

$$\varphi_{k1} = t_k \left(\frac{t_k(1-t_k)}{S^{(t)^2}{}_k} - 1\right); \quad \varphi_{k2} = (1-t_k)\left(\frac{t_k(1-t_k)}{S^{(t)^2}{}_k} - 1\right) \tag{4}$$

In Eq.8, *px0xdi* indicates the probability of hashtag assignment *x0* conditioned on the current hashtag assignment of h_d and related to the hashtag-weighted.

$$p_{x'x_{di}} = \begin{cases} p(x_{di} = x_1' \mid x' = Max(\pi_{xdi}), \gamma); \\ p(x_{di} = x_1' \mid x' = Max(\pi_{xdi}), \gamma) \propto \pi_{di} \dfrac{C_{th}^{TH} \neg di + \alpha}{\sum_{k=1}^{T} C_{TH}^{TH} \neg di + T\alpha} \end{cases} \tag{5}$$

Where the x_{di} is the *i*th hashtag in document d. Here, we introduce τ as the parameter to represent the probability that we have to choose the two methods of hashtag assignment in the hashtag collection. If γ = 1, we choose the hashtag with the max weight in the set of the corresponding document. Else we sample the x_d ; *i* from h_d according to Eq.6. This method is susceptible and efficient to the local maximal of the whole corpus.

In Gibbs Sampling procedure, we assume that the topic word distribution and hashtag topic distribution are conditionally independent. We can obtain the sample posterior distribution:

$$p(z_{di}=k, x_{di}=x^{'}, \gamma, t \mid w_{\neg di}, z_{\neg di}, x_{\neg di}, t, \alpha, \beta, \psi) \\ \propto \frac{C_{th}^{TH} \neg di + \alpha}{\sum_{k=1}^{T} C_{th}^{TH} \neg di + T\alpha} \frac{C_{wt}^{WT} \neg di + \beta}{\sum_{k=1}^{T} C_{wt}^{WT} \neg di + W\beta} \frac{\Gamma(\psi_{k,1}+\psi_{k,2})}{\Gamma(\psi_{k,1})+\Gamma(\psi_{k,2})} (1-t)^{\psi_{k,1}-1} t^{\psi_{k,2}-1} p_{x^{'}x_{d,i}} \quad (6)$$

Where C_{wt}^{WT} and C_{th}^{TH} is the matrix of the times about a specific word assigned to a specific topic and the matrix of the times of a specific topic assign to a specific hashtag. After iteratively sampling, the final results of the distribution θ and φ are:

$$\theta_x \propto \frac{C_{th}^{TH} + \alpha}{\sum_{K=1}^{T} C_{th}^{TH} + T\alpha}; \quad \phi_t \propto \frac{C_{wt}^{WT} + \beta}{\sum_{K=1}^{T} C_{wt}^{WT} + W\beta} \quad (7)$$

Experiments Analysis

In this section, we detect topic evolution on the Twitter dataset to prove the effectiveness of HWOT. We use the real word dataset called TREC 2011(http://trec.nist.gov/data/microblog2011.html). We select twitter corpus in English, and we filter noisy data by the steps of similar normalization [9]. After data preprocessing, there are 29220 tweets, 10214 distinct words and 84331 hashtags prepared for HWOT. The average length of each document is 6 and each document has more than one hashtag. All of the dataset is used for training our model to get satisfactory results.Then we follow the setting in JST model [2]. The hyperparameters α, β are 0.5 and 0.1. We regard τ as random number range 0 from 1 to improve the performance to sample hashtag of current tweet. Topic number is 50 in the step of case study. We run our model 800 times in Gibbs sampling in Python. For better results, we conduct experiment in improving topic mining performance and showing the evolution of social topic. We compare HGTM with three other models: 1) ATM, the Author Topic Model treats hashtag as author to sample. 2) TOT, the topic over time presents the LDA-based model. 3) HTDA, our model ignores the hashtag weight attribute.

In order to compare the results of HWOT and baselines, we follow the model evolution by using perplexity. As mentioned in [7], the model is generating a density estimation describing the underlying structure of Twitter dataset.

$$perplexity(D_{test}) = \exp\left\{ - \frac{\sum_d \sum_{w \in d} \ln p(w)}{\sum_d |d|} \right\} \quad (9)$$

$$p(w) = \sum_k p(k,w) = \sum_K p(k) p(w \mid k)$$

Here, D_{test} is the collection of the dataset, $|d|$ is the length of document d and $p(w)$ is the probability of each word assigned to topic.

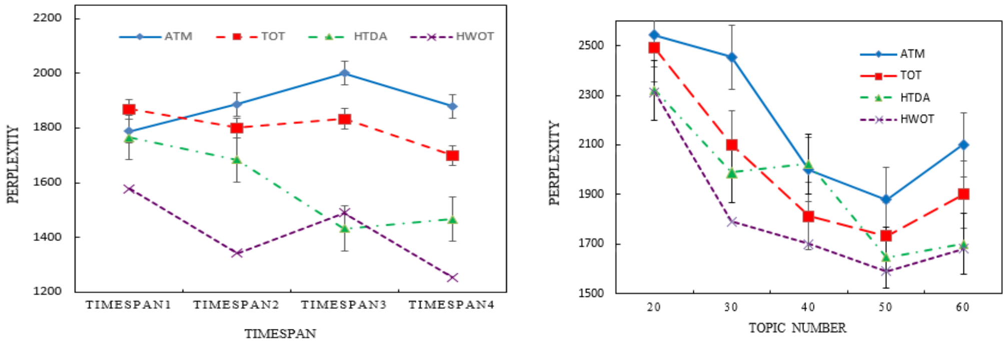

Fig. 2: The perplexity of topics (a) and time (b) in 5 topics

In our model, we evaluate perplexity value of each model in two respective: timespan and topic number. We first conduct models ten times to calculate the average value of perplexity. Fig.2 (a) illustrates that our model performs much better than baselines with the topic numbers of K=20, 30, 40, 50, 60. Besides, the HWOT is more stable than baselines. In Fig.2 (b), we divide the time sequence into four time spans averagely. Obviously, in each timespan, our model shows a lower perplexity than others. At timespan 1 and 2, the perplexity value performs a small difference between our model and baselines. It is because that social events suddenly break out uncertainly.

In order to detect the topic evolution in different layers, we study the quality of topics, hashtags and words discovered by our model. Table 1 to 3 tells 5 topics learned by HWOT. Each topic is denoted by the highest probability word distributed by each word.

Table 1: Topic mining performance of topic1, topic 3 and topic 5 of four models. Each topic is shown with top-8 words in the range of word-topic probability.

topic 1-game			topic 3-snow			topic 5-egypt		
HWOT	HTDA	TOT ATM	HWOT	HTDA	ATM TOT	HWOT	HTDA	ATM TOT
game	game	game	snow	snow	snow	egypt	egypt	egypt
super	super	bowl	weather	home	cold	police	police	egypt
bowl	weekend	super	cold	weather	feel	state	president	police
team	bowl	tonight	wind	cold	tomorrow	protest	president	protest
fan	play	blizzard	spring	spring	morning	news	arrest	man
star	tonight	*bear*	morning	morning	*Chicago*	mubarak	news	*news*
show	show	start	*way*	today	blizzard	Obama	*live*	*sound*
play	team	green		*far*	today	protest	right	protest

Table 2: Words-topic distribution of topic 5 in different time span

Timespan1	Timespan2	Timespan3	Timespan4
egypt	egypt	egypt	egypt
people	**mubarak**	**Obama**	mubarak
police	president	**America**	people
state	state	people	police
news	news	occur	news
America	people	news	**turbulence**
Obama	police	state	**crowd**
mubarak	protest	protest	street

Table 1 introduces topic mining performance of topic 1, topic 3 and topic 5 of four models. Each topic is shown with top-10 words in the range of word topic probability. We can see that HWOT can select more related words than baselines (italics words are irrelevant). For the topic 1 "game ", TOT and ATM research some irrelevant words such as "bear" and "hour" without time and hashtag controlling. For the topic 3 "egypt", hashtag weight influences weaker than topic 1 "game" and topic 5 "egypt" since that the topic 3 is oriented to a daily topic "weather", while topic 1 and topic 5 are social events at that time. It means that our model performs better in detecting the suddenly broken social event.

Table 2 shows the change of words distribution in different timespans. These important information reflects how the event developed. Now we take topic 5 as example. The boldface words are the topic evolution related words in the specific timespan. Recalling the events procedure at that time. In the timespan1, Egypt Protest started with people demonstration for the police corruption. So the probability of the words ‘people’ and ‘police’ is high. Then things went worse, the President Mubarak gave up the right of next vote. So here the related words in timespan 2 is ‘president’, ‘mubarak’ and ‘state’. Next, with the development of Egypt Protest, many countries came out to make a statement. So the word ‘state’ and ‘Obama’ occurred in timespan 3. This is the regulation of social events development. Our model mines the evolution of social events by analysis the word distribution to obtain more valuable information.

Conclusion

In this paper, we have proposed a new model called the HWOT to detect the evolution of social topics based on the Twitter dataset. Taking user-initial hashtag as weakly-supervised information to the model is a practical way to track the process of social topic evolution over time, which contributes to semantic analysis, government decision on emergencies and news recommendation. Our model uses a content aware method to illustrate the importance of hashtag in each tweet. The results of evaluation experiments show that hashtag weight and time as weakly-supervised information have impacted on the performance of topic mining and evolution results of social topics. In the future, we will do better in the hashtag weight by adding the retweets and comments information.

Acknowledgment

The work was supported by (XDJK2016E070), (KJ1501405), NSFC (No.61170192) and National High-tech R&D Program of China (No.2013AA013801).

References

[1]. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. the Journal of machine Learning research 3, 993{1022 (2003)

[2]. Dermouche, M., Velcin, J., Khouas, L., Loudcher, S.: A joint model for topic sentiment evolution over time. In: 2014 IEEE International Conference on Data Mining. pp. 773{778. IEEE (2014)

[3]. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 50{57. ACM (1999)

[4]. Li, S., Huang, G., Tan, R., Pan, R.: Tag-weighted dirichlet allocation. In: Data Mining (ICDM), 2013 IEEE 13th International Conference on. pp. 438{447. IEEE (2013)

[5]. Ramage, D., Hall, D., Nallapati, R., Manning, C.D.: Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. pp. 248{256. Association for Computational Linguistics (2009)

[6]. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th conference on Uncertainty in artificial intelligence. pp. 487{494. AUAI Press (2004)

[7]. Wang, J., Li, L., Tan, F., Zhu, Y., Feng, W.: Detecting hotspot information using multi-attribute based topic model. PloS one 10(10), e0140539 (2015)

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

[9]. Yan, X., Guo, J., Lan, Y., Cheng, X.: A biterm topic model for short texts. In: Proceedings of the 22nd international conference on World Wide Web. pp. 1445-1456. International World Wide Web Conferences Steering Committee (2013)