Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Hashtag-Based Sub-Event Discovery Using Mutually Generative LDA in Twitter

Chen Xing,[†‡] Yuan Wang,[†‡] Jie Liu,[†‡] Yalou Huang,[†‡] Wei-Ying Ma[*]
[†]College of Computer and Control Engineering, Nankai University, Tianjin, China
[‡]College of Software, Nankai University, Tianjin, China
[*]Microsoft Research, Beijing, China
v-chxing@microsoft.com {nkwangyuan,nkjieliu}@gmail.com huangyl@nankai.edu.cn wyma@microsoft.com

Abstract

Sub-event discovery is an effective method for social event analysis in Twitter. It can discover sub-events from large amount of noisy event-related information in Twitter and semantically represent them. The task is challenging because tweets are short, informal and noisy. To solve this problem, we consider leveraging event-related hashtags that contain many locations, dates and concise sub-event related descriptions to enhance sub-event discovery. To this end, we propose a hashtag-based mutually generative Latent Dirichlet Allocation model(MGe-LDA). In MGe-LDA, hashtags and topics of a tweet are mutually generated by each other. The mutually generative process models the relationship between hashtags and topics of tweets, and highlights the role of hashtags as a semantic representation of the corresponding tweets. Experimental results show that MGe-LDA can significantly outperform state-of-the-art methods for sub-event discovery.

Introduction

Social events, including large-scale emergencies(such as the Ebola outbreak in December 2013), political movements(such as Egyptian Revolution of 2011), and global influential sports games(such as Super Bowl), etc., could not only cause hot discussions among people, but also bring about widely-related, sophisticated and long-lasting social effects. In order to handle or to react to these social events, it is necessary for the related crowds to gain as much information as possible about the ongoing situation(Pohl, Bouchachia, and Hellwagner 2012a) and analyze them thoroughly. Twitter, as one of the most popular global social networking sites, attracts people worldwide to record details and share their opinions about social events, hence gradually making itself a very valuable data source for social event analysis. However, global social events usually have huge amount of related tweets, most of which are written in a short, informal and non-declarative style. Moreover, a global social event normally consists of many underlying and subtle constituent parts which lead to different aspects of social effects. Thus these constituent parts of one event require to be noticed and analyzed separately by respective related parties. These two factors make manually analyzing such event-related information in Twitter a cumbersome or even impossible task.

Pohl(Pohl, Bouchachia, and Hellwagner 2012a) firstly defined the task of sub-event discovery to solve this problem in emergency management domain. Sub-event discovery aims to divide the large number of event-related information into several sub-sets through clustering and identifies these subsets as sub-events. Thus the sub-event discovery task is able to automatically organize the huge amount of complex and unstructured event-related data in Twitter. It separates a social event into significant and distinct parts and semantically represents them for further analysis and is recently considered as an effective method for event analysis.

Several clustering-based methods(Pohl, Bouchachia, and Hellwagner 2012a)(Pohl, Bouchachia, and Hellwagner 2012b)(Abhik and Toshniwal 2013) have been proposed to fulfill the sub-event discovery task. Pohl(Pohl, Bouchachia, and Hellwagner 2012a) firstly represented texts using tf-idf features and implemented Self Organizing Map(SOM) to cluster texts. Each cluster represented a specific sub-event. Pohl(Pohl, Bouchachia, and Hellwagner 2012b) and Abhik(Abhik and Toshniwal 2013) further employed more complex text representation methods that took different structural information, such as article titles, tags and geography coordinates, into consideration. However, these methods merely employed shallow features such as tf-idf for social media messages. These features contain basic and limited statistical information for the messages and take seldom advantage of the deep semantic information of social media texts. While, the semantic meanings of texts are of great importance in both the discovery and representation process of sub-events.

In order to fill this gap, we seek for topic model related methods. Topic models have long been considered as a powerful method for learning semantic representations of documents, with the topic distributions of which be considered as representation vectors. Latent Dirichlet Allocation(LDA)(Blei, Ng, and Jordan 2003) is one of the most popular topic models and has been followed by many others. However, the severe sparsity problem of short text poses a new challenge to traditional topic models like LDA. This is because LDA reveals latent topics by capturing document-level word co-occurrence patterns(Wang et al. 2014). Moreover, the informality and typos of the tweet language make

it even harder for traditional topic models to capture the semantic information of tweet data.

Fortunately, we find that hashtags in Twitter can give some help to leverage the two challenges for traditional topic models mentioned above. In tweet data, hashtags which are usually added by users to highlight topics or categorize messages(Wang et al. 2014), contain many locations and short event descriptions related to sub-events. For example, among the tweets for egypt revolution, most of tweets tagged "suez" are about the arrest of activists in Suez city and tweets with "humanrights" are about the protesters' specific activities to fight for human rights. Hence, hashtags can be viewed as a strong and concise semantic sign that could help traditional topic models to discover sub-events on semantic level. What's more, one hashtag normally tags multiple tweets. It indicates that hashtags can be used to reorganize and lengthen tweet texts. Thus this gives us a way to leverage the sparsity problem. Therefore, hashtags can not only be utilized to alleviate the sparsity of tweet data in sub-event discovery task, but also a strong semantic sign for sub-events that can help to overcome the informality and noise of tweet texts.

According to the facts mentioned above, we propose an extension of traiditional LDA, the mutually generative LDA model(MGe-LDA), which employs hashtag information during the generative process to help discover sub-events from a event-related tweet corpus. MGe-LDA is a generative topic model that could capture the semantic information of both tweet texts and hashtags, and learns the topic distributions of them. We then apply a clustering method on hashtags' topic representations learnt by MGe-LDA, to discover and represent sub-events. Compared to LDA, MGe-LDA adds a hashtag generation layer and models hashtags as distributions of topics. Moreover, MGe-LDA makes hashtags and topics of one tweet to be mutually generated. This process emphasizes hashtags' representativeness for the semantic meanings(topics) of tweet texts. It fits the fact that most of hashtags in event-related data sets act as representative factors of sub-events. Furthermore, to leverage the typos and repetitions in hashtags, we apply a hashtag graph for the hashtag assignment process of MGe-LDA. In this modification, we include the hashtags in the graph which is highly-similar to hashtags tagged in a tweet as assigning candidates for the tweet. This modification leverages the noise and informality problem of tweet data one step further and ensures the quality of topic distributions of both popular and rare hashtags.

Related Work

In this section, we first give a brief overview of the studies on event/sub-event detection using social media data. Secondly, for semantic analysis, we survey the specific modification of LDA which add layers to involve factors like hashtags or authors in the generative model besides topics and words.

Event/sub-event detection using social media data. Event detection, which has drawn many interests in recent years, is a similar task to sub-event discovery and could offer some insights for the data representation during sub-event discovery. Most existing approaches of event detection can be classified as anomaly detection based approaches. They focus on detecting abnormal features, such as word/hashtag appearance or the number of messages published in a specific location, to monitor anomaly and represent new events. Fung and Yu(Fung et al. 2005) firstly proposed a parameter-free method to detect bursty words and entities in text streams, then detected events through clustering the detected words and entities. Similarly, EDCoW(Weng and Lee 2011) and Twevent(Li, Sun, and Datta 2012) are also based on word burst detection. Specifically, in order to deal with the short and noisy contents of tweets, Twevent(Li, Sun, and Datta 2012) split each tweet into non-overlapping segments then detected and clustered bursty segments to fulfill event detection. Besides texts, structural information is also utilized for event detection. Several works(Watanabe et al. 2011)(Chen and Neill 2014) utilized hashtags and geotags of social media data separately to detect bursts. Becker and Naanman(Becker, Naaman, and Gravano 2011) extracted temporal, social and topical features of message clusters in order to classify them as event or non-event clusters. Each of the clusters that was classified as event clusters represented one single detected event. Moreover, Aggarwal and Subbian(Aggarwal and Subbian 2012) implemented similar clustering method to group not only texts but also structural information of social media streams. Afterwards, they generated alarms to call attention to newly emerged events when the element-adding behavior of a particular cluster became abnormally frequent.

Comparing to event detection, sub-event discovery hasn't drawn enough attention it deserves. Pohl(Pohl, Bouchachia, and Hellwagner 2012a) firstly proposed a novel approach for sub-event detection in emergency management. Pohl represented texts using tf-idf features and implemented Self Organizing Map(SOM) to cluster texts. Each cluster in the result represented a specific sub-event. Furthermore, Pohl(Pohl, Bouchachia, and Hellwagner 2012b) and Abhik(Abhik and Toshniwal 2013) further applied two-phase clustering approaches that took structural information into consideration. In (Pohl, Bouchachia, and Hellwagner 2012b), the first phase of clustering used geo-referenced data to locate a sub-event, while the second phase used texts to describe sub-events. Thus these existing methods employed little semantic analysis of texts during detection.

Modifications of LDA by adding layers. Latent Dirichlet Allocation(LDA) has long been proved a very effective text understanding model. Many modifications of LDA have been proposed to fulfill various text-related tasks such as text mining and sentimental analysis. Rosen-Zvi and Griffiths introduced the Author Topic model(Rosen-Zvi et al. 2004), which is an extension of LDA to involve authorship information and then model the interests of authors. In Author Topic model, each author is associated with a distribution over topics and each topic is associated with a distribution over words. While it assumed that the author distribution of each document is uniform. HGTM(Wang et al. 2014) extended ATM to make it fit the characteristics of tweet data. HGTM treats hashtags in tweet data as authors and added a hashtag graph to leverage the sparsity problem of short tweet texts. Similarly, Lin and He(Lin and He 2009) in-

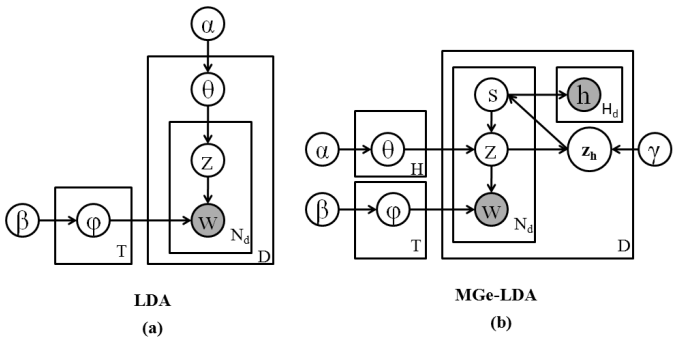

Figure 1: (a) Latent Dirichlet Allocation(LDA). (b) Muatually Generative LDA(MGe-LDA).

troduced joint sentiment/topic model(JST) for unsupervised sentimental analysis and opinion mining. JST added a sentiment label layer in front of the topic layer and assumed the sentiment distribution of each document is multinomial.

In this paper, MGe-LDA adds a hashtag generation layer to involve hashtags in generative process of a tweet corpus and models hashtags as distributions of topics. What's more, unlike other modifications of LDA that also add layers before the topic layer and model a one-way generation from the added layer to the topic layer, MGe-LDA makes hashtags and topics of one tweet to be mutually generated. More specifically, the model prefers generating hashtags with higher similarity in topic distributions to the tweet. This process emphasizes hashtags' representativeness for the semantic meanings of tweet texts. It fits the fact that most of hashtags in event-related data sets act as representative factors of sub-events such as locations and short descriptions. The high-quality hashtag topic distributions produced by the mutually generative process benefit both the discovery and representing process for sub-event discovery.

Hashtag-based Mutually generative LDA

We define the tweet corpus as $\mathbf{C} = \{\mathbf{d_1}, \mathbf{d_2}, ..., \mathbf{d_D}\}$; each document $\mathbf{d}$ in this corpus is composed of a word sequence $\mathbf{w_d} = \{w_{d1}, w_{d2}, ..., w_{dN_d}\}$ and a hashtag sequence $\mathbf{h_d} = \{h_{d1}, h_{d2}, ..., h_{dH_d}\}$. Note that the lengths of the hashtag sequence and the word sequence are different. Each word in documents is an item from the corpus vocabulary with the size of V. Also, let H be the total number of distinct hashtags in the corpus and T be the number of topics. Traditional LDA models document $\mathbf{d}$ as merely a word sequence $\mathbf{w_d}$. In LDA, documents are modeled as mixtures of topics and every topic is represented as a multinomial distribution over words. The generative process of a document $\mathbf{d}$ in LDA is divided into three steps as illustrated in Figure 1(a). First, one chooses a multinomial topic distribution θ for document $\mathbf{d}$ under the Dirichlet prior with parameter α. Then, for the ith word w_{di} in $\mathbf{w_d}$, one chooses a topic z_{di} according to multinomial topic distribution θ; finally, the ith word w_{di} is generated according to topic z_{di}'s word distribution Φ (which is also multinomial under a Dirichlet prior with parameter β). The general description of this generative process is as follows.

$$\theta \sim Dirichlet(\alpha)$$
$$\phi \sim Dirichlet(\beta)$$
$$z_{di}|\theta_{\mathbf{d}} \sim Multinomial(\theta_{\mathbf{d}})$$
$$w_{di}|\phi_{\mathbf{z_{di}}} \sim Multinomial(\phi_{\mathbf{z_{di}}})$$

In order to model the generative process of tweets with hashtags, we propose a hashtag-based mutually generative LDA(MGe-LDA) model which adds a hashtag layer between the document and topic layer and lets hashtags and topics of one tweet mutually generate each other. The graphic model of MGe-LDA is represented as Figure 1(b).

In MGe-LDA, each hashtag is associated with a multinomial distribution over topics and each topic is represented as a multinomial distribution over words. Different from LDA, the generation of one tweet includes not only generating the word sequence but also the hashtag sequence belonging to this tweet. Let $\mathbf{s_d} = \{s_{d1}, s_{d2}, ..., s_{dN_d}\}$ and $\mathbf{z_d} = \{z_{d1}, z_{d2}, ..., z_{dN_d}\}$ represent the hashtag and topic assignments separately for the word sequence of tweet d. Note that the distinct members of s_{di} form the hashtag sequence $\mathbf{h_d}$ of tweet d. Words of a tweet are generated in sequence. For word w_{di} in tweet d, one firstly chooses a hashtag s_{di} according to the $P(h|\mathbf{z_h})$, where $\mathbf{z_h}$ represents the topic assignments for the word sequence of hashtag h. This is because intuitively, the choice of hashtags when we pose tweets is tightly related to their topic distributions and the current topic assignments of hashtags directly reflect their topic distributions. Thus refereing to the topic assignments of hashtags gives a wiser choice of hashtags related to tweet texts. Then, one chooses a topic z_{di} with respect to the topic distribution of s_{di}, $\theta_{\mathbf{s_{di}}}$. Following this, w_{di} is chosen according to the word distribution $\phi_{z_{di}}$, which is the same as the last generating step of LDA. Through this process both the hashtag sequence $\mathbf{h_d}$ and word sequence $\mathbf{w_d}$ of tweet d are generated. Thus the formal description of MGe-LDA's generating process is as follows,

- For each hashtag h, choose a topic distribution $\theta_h \sim Dirichlet(\alpha)$;
- For each topic t, choose a word distribution $\phi_t \sim Dirichlet(\beta)$;

- For each tweet d in the corpus,
  - For each word w_{di} in tweet d,
    - choose a hashtag $s_{di} \sim P(h|\mathbf{z_h})$;
    - choose a topic $z_{di} \sim \theta_{s_{di}}$;
    - choose a word $w_{di} \sim \phi_{z_{di}}$

where $P(h|\mathbf{z_h}) \propto P(h) \cdot P(\mathbf{z}_h|h)$. $P(h)$ can be defined as a multinomial distribution γ that γ_h is the appearance frequency of hashtag h in the whole corpus. $\mathbf{z_h}$ is the current topic assignments for the word sequence of hashtag h. The first state of these topic assignments is randomly initialized.

In order to get the topic distributions of hashtags and mine semantic information from event-related tweets, we need to infer the latent variables θ and ϕ. Let $\mathbf{w}$ denotes the whole word sequence of the entire corpus and $\mathbf{z}$, $\mathbf{s}$ denotes the topic and hashtag assignments for $\mathbf{w}$ separately. To infer θ and ϕ, we firstly estimate the posterior probability of $\mathbf{z}$. By integrating out θ we can get,

$$p(\mathbf{z}|\mathbf{s}) = (\frac{\Gamma(T\alpha)}{(\Gamma(\alpha))^T})^H \prod_{h=1}^{H} \frac{\prod_j \Gamma(C_{jh}^{TH} + \alpha)}{\Gamma(\sum_{j'} C_{j'h}^{TH} + T\alpha)} \quad (1)$$

where C_{jh}^{TH} is the number of times that topic j has been assigned to words whose hashtag assignment is hashtag h. Next, by integrating out ϕ we could get,

$$p(\mathbf{w}|\mathbf{z}) = (\frac{\Gamma(W\beta)}{(\Gamma(\beta))^W})^T \prod_{t=1}^{T} \frac{\prod_w \Gamma(C_{wt}^{WT} + \beta)}{\Gamma(\sum_{w'} C_{w't}^{WT} + W\beta)} \quad (2)$$

where C_{wt}^{WT} is the number of times that word w is assigned to topic t. Then we could use Gibbs Sampling(Griffiths and Steyvers 2004) to construct $\mathbf{z}$ according to,

$$P(z|w,s) \propto P(z,w|s) = P(z|s)P(w|z) \quad (3)$$

Gibbs Sampling will sample every element z_{di} in $\mathbf{z}$ from the popularity of topics given the status of all other elements in $\mathbf{z}$. The probability that topic t is assigned to z_{di} can be calculated as follows,

$$\begin{aligned} p(z_{di} = t|\mathbf{z_{-di}}, \mathbf{w}, \mathbf{s}) \propto & \frac{C_{wt,-di}^{WT} + \beta}{\sum_{w'} C_{w't,-di}^{WT} + W\beta} \\ & \cdot \frac{C_{th,-di}^{TH} + \alpha}{\sum_{t'} C_{t'h,-di}^{TH} + T\alpha} \end{aligned} \quad (4)$$

where h is the hashtag assignment of w_{di}. Obviously, the assignment process of $\mathbf{z}$ relies on $\mathbf{s}$, the hashtag assignment sequence of $\mathbf{w}$. Thus next we discuss about the sampling process of $\mathbf{s}$.

According to the generative process of MGe-LDA, the posterior popularity of $\mathbf{s}$ is,

$$P(\mathbf{s}|\mathbf{z}) \propto P(\mathbf{z},\mathbf{s}) = P(\mathbf{z}|\mathbf{s}) \cdot P(\mathbf{s}) \quad (5)$$

We have calculated $P(\mathbf{z}|\mathbf{s})$ in equation (1). Thus the probability of hashtag j assigned to hw_{di} that is required in Gibbs Sampling is as follows,

$$p(s_{di} = j|\mathbf{s_{-di}}, \mathbf{z}) \propto \gamma_j \cdot \frac{C_{th,-di}^{TH} + \alpha}{\sum_{t'} C_{t'h,-di}^{TH} + T\alpha} \quad (6)$$

where γ_j is the appearance frequency of hashtag j in the whole corpus. Moreover, it should be noted that during the sampling of s_{di}, we don't have to traverse all hashtags in hashtag list of the corpus since we've known the hashtag sequence $\mathbf{h}_d$ for tweet d. We only need to calculate all $p(s_{di} = j|\mathbf{s_{-di}}, \mathbf{z})$ when hashtag j belongs to $\mathbf{h}_d$.

After iterative sampling, the algorithm reaches its convergence. The final result of θ and ϕ are,

$$\theta_h \propto \frac{C_{th}^{TH} + \alpha}{\sum_{t'} C_{t'h}^{TH} + T\alpha} \quad (7)$$

$$\phi_t \propto \frac{C_{wt}^{WT} + \beta}{\sum_{w'} C_{w't}^{WT} + W\beta}. \quad (8)$$

The hashtag-topic distributions θ are implemented as features for clustering in the sub-event detection task.

MGe-LDA with hashtag graphs

Table 1: Statistical information of 3 event data sets extracted from TREC 2011

Name	#tweets	Vocabulary size	#hashtags
ER	6628	4433	1200
SB	2948	3104	614
SOTU	1840	2636	389

In this section we make a modification for MGe-LDA by applying a hashtag graph for the hashtag assignment process. This hashtag graph models the similarity and relation of all hashtags in the event-related tweet corpus. As illustrated above, in MGe-LDA, the hashtag candidates for hashtag assignments of tweet d are only those in the hashtag sequence $\mathbf{h}_d$. In this modification, we tend to loosen this limitation by including hashtags highly-related to those in the hashtag sequence $\mathbf{h}_d$ as candidates. The reason of this is that there are many typos and extremely similar hashtags in the whole tweet dataset since most of hashtags are added manually by users. For example, hashtag "jna25" is a typo of hashtag "jan25", and "jan25", "egypt" are both representing the Egypt revolution at Jan 25th in our data set. Thus even though these hashtags do not belong to the hashtag sequence of tweet d, they are reasonable candidates of hashtag assignments for the current tweet. Therefore, this modification leverages the noise and informality problem of tweet data one step further and ensures the quality of topic distributions of both popular and rare hashtags.

We take $\mathscr{G} = (V, E)$ to represent the hashtag graph, the nodes $V = \{h_1, h_2, ..., h_H\}$ in which are hashtags and each e_{ij} in $E = \{(e_{ij})\}_{i,j \in V, i \neq j}$ represents the number of times hashtag i and j have been added with the same tweets since appearing in the same tweets is a strong sign of semantic similarity for two hashtags. We introduce τ to represent the probability that we have to choose a hashtag in the graph. Thus the hashtag assigning process for every word w_{di} of MGe-LDA with graph can be divided into three sub-steps,

1. sample a hashtag t from $\mathbf{h_d}$ according to equation (6);

2. sample r, $r \sim Bernoulli(\tau)$;

3. if $r = 0$, w_{di}'s hashtag assignment s_{di}=t; else sample s_{di} from $\mathbf{g_t}$ according to equation(6).

where $\mathbf{g_t}$ is the set of hashtags that have direct edges to hashtag t.

Experimental Analysis

In this section, we conduct sub-event discovery experiments on Twitter data to verify the effectiveness of MGe-LDA. The experiment procedures are as follows. After running MGe-LDA on an event-related tweet data set, we could get the topic distributions of hashtags in the data set. In our sub-event discovery experiment, we run K-means(Hartigan and Wong 1979) to cluster hashtags in their topic space, which takes hashtags' topic distributions as their feature vectors. The distance between two objects is measured by their cosine-similarity. To evaluate the sub-event discovery result, we use the H-score(Yan et al. 2013) of clustering results to evaluate the semantic compaction of one sub-event and the semantic distinction among all detected sub-events. What's more, we take top n hashtags nearest to the cluster centroids to represent discovered sub-events and analyze the semantic meanings of them.

We compare MGe-LDA and MGe-LDA with hashtag graphs with four other models: 1) tf-idf+SOM model(Pohl, Bouchachia, and Hellwagner 2012a), which builds tf-idf features for tweet texts and clusters them using Self Organizing Map(SOM); 2)LDA(Blei, Ng, and Jordan 2003), Latent Dirichlet Allocation, which takes each tweet as a document; 3)ATM(Rosen-Zvi et al. 2004), the Author Topic Model that we replace "author" as "hashtag" in our experiment; 4)HGTM(Wang et al. 2014), the hashtag graph topic model, which is an extension of ATM that implements related hashtags in hashtag graphs to leverage sparsity of tweet data. It should also be noted that since traditional LDA and tf-idf don't model the generation of hashtags in them, we aggregate all tweets tagged by the same hashtag to construct a pseudo hashtag document to infer the hashtag's topic distribution.

In our experiments, we set the number of topics T as 5 and the number of clusters K as 5. The hyperparameter β of all topic models is set as 0.01. As for the hyperparameter α and τ in topic models related to the hashtag graph, we tune them separately for each model to get its best performance.

Data Sets

We build three event data sets from the TREC 2011 microblog data set. It contains nearly 16 million tweets sampling from January 23rd to February 8th, 2011. There are three major events happening in this period which bring about heated discussions on Twitter, the Egyptian Revolution of 2011, Super Bowl 2011 and 2011 State of the Union Address. Thus we manually select tweets related to these three events separately and form three event data sets represented as ER, SB and SOTU.

H-score of sub-event discovery

We firstly measure the sub-event discovery result with H-score(Yan et al. 2013), which is the ratio of the average intra-cluster distance and the average inter-cluster distance of hashtag clustering results. The average inter-cluster distance can be calculated as follows,

$$IntraDis\,(C) = \frac{1}{K}\sum_{k=1}^{K}[\sum_{h_i,h_j \in C_k, i \neq j} \frac{2dis(h_i,h_j)}{\|C_k\|\|C_k - 1\|}] \tag{9}$$

where $dis(h_i, h_j)$ represents the cosine-similarity of hashtag i and j and $\|C_k\|$ represents the number of hashtags allocated to the kth cluster. The average inter-cluster distance is:

$$InterDis\,(C) = \frac{1}{K(K-1)} \sum_{C_k, C_{k'} \in C, k \neq k'} [\sum_{h_i \in C_k} \sum_{h_j \in C_{k'}} \frac{2dis(h_i,h_j)}{\|C_k\|\|C_{k'}\|}] \tag{10}$$

Thus the H-score of the clustering result is as follows,

$$H = \frac{IntraDis(C)}{InterDis(C)} \tag{11}$$

Since we cluster hashtags with their topic distributions as feature vectors, H-score of the clustering result measures the spatial distributions of sub-events in their topic space. From the equations above, we can easily see that smaller intra-cluster distances and larger inter-cluster distances result in a lower H-score. A lower H-score indicates that hashtags allocated to one sub-event are more semantically related and hashtags belonging to different sub-events are more semantically distinct. A decent H-score means that discovered sub-events are topically focused and independent. The H-scores for sub-event discovery in the 3 data sets are shown in Table 3.

From the H-score results, we could get the following conclusions. Firstly, all the topic model based methods, including MGe-LDA and MGe-LDA with hashtag graphs we proposed in this paper, outperform the current sub-event discovery method, tf-idf+SOM a great deal(with p-value¿0.6) on all the 3 event data sets. This proves that introducing topic model based methods benefits the sub-event discovery task a lot since topic models can better capture the semantic information embedded in tweet texts, which is a deeper level of hints that could help us splitting sub-events. Secondly, among all topic model based methods, MGe-LDA and MGe-LDA with hashtag graphs achieve the best performance on all the 3 events. This indicates that MGe-LDA methods(with and without hashtag graphs) produces more similar hashtag topic distributions for semantically similar hashtags and vice verse. In other words, hashtags form more **clear** clusters in their topical space. This is because in the generative process of MGe-LDA, we generate hashtags of a tweet according to the topic distribution of this tweet. Therefore, the tweet's topic distribution acts as a **bridge** for hashtags belonging to this tweet, or can be viewed as weak supervision information for hashtag assignment during the parameter estimation process.

Table 2: Semantic representations of sub-events of ER

MGe-LDA with graph					HGTM				
1	2	3	4	5	1	2	3	4	5
breakingnews	humanrights	google	socialmedia	palestine	feb01	freesherine	jan28	live	iraq
cnn	teaparty	tahrirsquare	islam	lebanon	freeegypt	europe	p2	blair	hamradio
egyptians	wikileaks	aje	fb	us	world	26jan	tcot	respect	pakistan
revolution	democracy	elbaradei	australia	jordan	justsaying	datalove	25jan	abc	haiti
jan28	egipto	freeayman	syria	freedom	hosnimub	chaos	tunisia	ac360	media
p2	usa	suez	censorship	protest	alarabiya	wef	cairo	tweet	fok
cairo	news	alexandria	twitter	iranelection	amman	anonops	tahrir	tunisie	tunis
tahrir	feb1	sidibouzid	uk	iran	london	feb4	mubarak	foxnews	teargas
jan25	obama	aljazeera	yemen	tcot	jan26	freeayman	jan25	jan	arab
egypt	mubarak	25jan	israel	tunisia	ff	uk	egypt	australia	aje

Table 3: H-Scores of sub-event discovery for the 3 events

Data set	Model	H-score
ER	tf-idf+SOM	0.9604
	LDA	0.1468
	ATM	0.1598
	HGTM	0.1389
	MGe-LDA	**0.0774**
	MGe-LDA with graph	**0.0232**
SB	tf-idf+SOM	0.9549
	LDA	0.1314
	ATM	0.3067
	HGTM	0.1329
	MGe-LDA	**0.0813**
	MGe-LDA with graph	**0.0364**
SOTU	tf-idf+SOM	0.9063
	LDA	0.1318
	ATM	0.0968
	HGTM	0.1526
	MGe-LDA	**0.0846**
	MGe-LDA with graph	**0.0124**

Semantic representations of sub-events

After the discovery of sub-events, we ought to figure out what the sub-events we discovered are about because the goal of this task is to understand the whole event in a finer-grained way. In this section, we choose n hashtags that are most popular in the cluster to semantically represent the sub-event. The semantic representations of sub-events of Egyptian Revolution 2011(ER) extracted by different models are shown in Table 2.

In Table 2, we list top ten most popular hashtags in the clusters separately to represent each of the five sub-events we discovered. Hashtags in black boxes are those who give clear clues about the contents of the sub-events they belong to. The result of MGe-LDA with hashtag graph gives the most explicable and rational representations for sub-events. From the representative hashtags for sub-event 1, we can easily tell that sub-event 1 our model discovered is about the protesters' occupation of the Tahrir Square during the first several days of the revolution. Representative hashtags gives clear time("jan25", "jan28"), location("tahrir", "cario", "egypt") and the progressiveness("breakingnews", "p2") for this movement. Sub-event 2 is about uncovering the deep reason for this revolution since the hashtag list contains the main general goal of this fight("humanrights", "democracy") and the major suspected back-stage planner("wikileaks", "usa", "obama"). As for sub-event 3, the representative hashtag list indicates that it is about the arrest of activists(mainly happened in "suez" and "alexandria"), especially the arrest of Al Jazeera English's("aje", "aljazeera") journalists such as Ayman Mohyeldin("freeayman"). Sub-event 4 is about the internet black out of Egypt and Syria("syria") ordered by their governments' "censorship" in Jan 27th and the whole Egypt almost disappeared in most of popular "socialmedia" sites("fb", "twitter"). The final sub-event is also quite clear according to its representative hashtag list. The list contains many countries affected by the Arab Spring such as "palestine", "lebanon", "jordan" and "iran", which indicates that this sub-event is mainly about the revolution's influence on other Arabian countries. In sum, the representative hashtag lists of sub-events extracted by MGe-LDA with graph is very semantically concentrated and explicit, which benefit the analysis of sub-events they represent.

As for other models, unlike MGe-LDA with hashtag graph, their results don't have obvious semantic focus for every sub-event. As shown in Table 3, the hashtag lists without black boxes are those that cannot be concluded a general description for their correspondent sub-events. Except MGe-LDA with graph, MGe-LDA performs the best and extracts four focused hashtag lists.

Conclusion and Future Work

In this paper, we focused on extending topic models for sub-event discovery in Twitter. We applied event-related hashtags in MGe-LDA model to capture semantic information from the short, informal and noisy tweet data. Moreover, to leverage the typos and repetitions in hashtags, we applied a hashtag graph for the hashtag assignment process of MGe-LDA. The hashtag graph benefits MGe-LDA in considering typos and repetition of hashtags in tweets, and then helps topic representation learning. In future work of this paper, except for hashtags, we hope to make use of other structural information such as geography coordinates, users' relationship and time series information in Twitter. Even though geography information is relatively rare, intuitively we suppose it can benefit sub-event discovery task very much since it is a significant attribute for sub-events.

References

Abhik, D., and Toshniwal, D. 2013. Sub-event detection during natural hazards using features of social media data. In *Proceedings of the 22nd international conference on World Wide Web companion*, 783–788. International World Wide Web Conferences Steering Committee.

Aggarwal, C. C., and Subbian, K. 2012. Event detection in social streams. In *SDM*, volume 12, 624–635. SIAM.

Becker, H.; Naaman, M.; and Gravano, L. 2011. Beyond trending topics: Real-world event identification on twitter. *ICWSM* 11:438–441.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.

Chen, F., and Neill, D. B. 2014. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1166–1175. ACM.

Fung, G. P. C.; Yu, J. X.; Yu, P. S.; and Lu, H. 2005. Parameter free bursty events detection in text streams. In *Proceedings of the 31st international conference on Very large data bases*, 181–192. VLDB Endowment.

Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101(suppl 1):5228–5235.

Hartigan, J. A., and Wong, M. A. 1979. Algorithm as 136: A k-means clustering algorithm. *Applied statistics* 100–108.

Li, C.; Sun, A.; and Datta, A. 2012. Twevent: segment-based event detection from tweets. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, 155–164. ACM.

Lin, C., and He, Y. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, 375–384. ACM.

Pohl, D.; Bouchachia, A.; and Hellwagner, H. 2012a. Automatic sub-event detection in emergency management using social media. In *Proceedings of the 21st international conference companion on World Wide Web*, 683–686. ACM.

Pohl, D.; Bouchachia, A.; and Hellwagner, H. 2012b. Supporting crisis management via sub-event detection in social networks. In *Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), 2012 IEEE 21st International Workshop on*, 373–378. IEEE.

Rosen-Zvi, M.; Griffiths, T.; Steyvers, M.; and Smyth, P. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, 487–494. AUAI Press.

Wang, Y.; Liu, J.; Qu, J.; Huang, Y.; Chen, J.; and Feng, X. 2014. Hashtag graph based topic model for tweet mining. In *Data Mining (ICDM), 2014 IEEE International Conference on*, 1025–1030.

Watanabe, K.; Ochi, M.; Okabe, M.; and Onai, R. 2011. Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2541–2544. ACM.

Weng, J., and Lee, B.-S. 2011. Event detection in twitter. *ICWSM* 11:401–408.

Yan, X.; Guo, J.; Lan, Y.; and Cheng, X. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, 1445–1456. International World Wide Web Conferences Steering Committee.