# GEAM: A General and Event-Related Aspects Model for Twitter Event Detection

Yue You[1,2], Guangyan Huang[2], Jian Cao[1,⋆], Enhong Chen[3], Jing He[2], Yanchun Zhang[2], and Liang Hu[1]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[2] Centre for Applied Informatics, Victoria University, Australia
[3] School of Computer Science and Technology University of Science and Technology of China, China
{yy71107103,cao-jian,lianghu}@sjtu.edu.cn, {guangyan.huang,jing.he,yanchun.zhang}@vu.edu.au, cheneh@ustc.edu.cn

**Abstract.** Event detection on Twitter has become a promising research direction due to Twitter's popularity, up-to-date feature, free writing style and so on. Unfortunately, it's a challenge to analyze Twitter dataset for event detection, since the informal expressions of short messages comprise many abbreviations, Internet buzzwords, spelling mistakes, meaningless contents etc. Previous techniques proposed for Twitter event detection mainly focus on clustering bursty words related to the events, while ignoring that these words may not be easily interpreted to clear event descriptions. In this paper, we propose a General and Event-related Aspects Model (GEAM), a new topic model for event detection from Twitter that associates General topics and Event-related Aspects with events. We then introduce a collapsed Gibbs sampling algorithm to estimate the word distributions of General topics and Event-related Aspects in GEAM. Our experiments based on over 7 million tweets demonstrate that GEAM outperforms the state-of-the-art topic model in terms of both Precision and DERate (measuring Duplicated Events Rate detected). Particularly, GEAM can get better event representation by providing a 4-tuple ($Time, Locations, Entities, Keywords$) structure of the detected events. We show that GEAM not only can be used to effectively detect events but also can be used to analyze event trends.

**Keywords:** Event detection, Twitter, General and Event-Related Aspects Model (GEAM), Topic model, Gibbs sampling.

## 1 Introduction

Social network services such as Twitter[1], Facebook[2] have experienced a rapid growth in recent years. Millions of people turn from traditional news websites to

⋆ Corresponding author.

[1] https://twitter.com

[2] http://www.facebook.com/

X. Lin et al. (Eds.): WISE 2013, Part II, LNCS 8181, pp. 319–332, 2013.

these services to gather real-time news, share opinions, or read hot comments. As a real-time information network to share *"the latest stories, ideas, opinions and news*[3] *"*, Twitter allows users to express everything by writing a "*tweet*" up to 140 characters. According to the Paris-based analyst group Semiocast, as of July 2012, Twitter has more than 500 million users all over the world[4]. Twitter has some unique characteristics that make it a better source for event detection than traditional news articles, blogs, etc. Firstly, Twitter is the most up-to-date information channel of real events. Because of the length limit and the popularity of Twitter mobile applications, users can update information instantly at a pretty low cost, which makes Twitter a fresher resource than others. Secondly, Twitter covers a wider range of information than other sources. Millions of users in Twitter are not only information consumers but also news publishers, they can post tweets describing everything in their life. So, Twitter almost covers every aspect of the society, from breaking news to personal opinions. Thirdly, we can also analyze the opinions or sentiments related to the detected events in Twitter, which can help the organizations respond to the upcoming event quickly [1]. Thus, event detection on Twitter has attracted more and more interest recently.

Although event detection has been well studied in formal text such as news articles, blogs [2–4], Twitter dataset brings several challenges to us because it contains many abbreviations, Internet buzzwords, spelling mistakes, meaningless contents etc. In existing work on formal text, the underlying assumption is that all the documents in the corpus is related to some undiscovered events, but it is not reasonable in Twitter. According to Pear Analytics study[5], about 40% of tweets are "*pointless babbles*" like "I'm eating a sandwich". Such tweets are limited to users' personal feelings, and should not draw attention from the majority audience. Unlike well-written text, Twitter dataset also contains a lot of Internet buzzwords and misspelling due to the length limit and the free writing style[5]. All these unique characteristics make most existing techniques in formal text unavailable on Twitter dataset. Besides, most recently proposed methods for Twitter event detection are either limited to certain types of tweets (e.g. related to earthquakes or crimes and disasters) [6–8] or based on clustering bursty words [9–11], and provide results in the form of single terms, hashtags or $n$-grams. The outcome of these approaches may be uninformative or meaningless, and can't help the users obtain more fine-grained insights of the whole event.

To tackle the above challenges, we propose GEAM, a General and Event-Related Aspects Model for event detection in Twitter. GEAM is an unsupervised generative topic model that differentiates General words (describing general opinions) from Event-related Aspects words (describing different aspects of an event) in an event-related tweet. In this paper, we focus on detecting realistic events, which are discussed by a large number of users. We define a 4-tuple representation of a realistic event, e.g. *(Time, Locations, Entities, Keywords)*, which are corresponding to

[3] `https://twitter.com/about`

[4] `http://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/`

[5] `http://news.bbc.co.uk/2/hi/technology/8204842.stm`

4 Event-related Aspects in GEAM respectively. Here, Time and Locations represent when and where the event happened; Entities represent main subjects (persons, organizations, movies, etc.) in the event; Keywords describes the meaning of the event. Unlike formal news articles, users in Twitter also use General words in tweets. For example, during Obama's victory speech, there are 327k tweets per second posted/reposted to discuss Obama's re-election[6]. One of the typical tweets like "Thank The Lord for that!! Well done Obama in Chicago, Illinois!! That was a Great victory speech!". In this tweet: Time is *"November 7, 2012"*, *"Chicago"* and *"Illinois"* represent Locations, *"Obama"* represents Entity, Keywords are *"Great"*, *"victory"*, and *"speech"*, and the rest are General words. GEAM can provide a clearer description of events by separating General words from Event-related Aspects words.

The most related work is the model proposed by Lau et al. [12]. They introduce an online processing variant of topic model Latent Dirichlet Allocation (LDA) [13, 14] to analyze tweets trend. Their graphic structure is the same as LDA, which models each tweet as a multinomial mixture of all topics or events. However, this assumption is obviously unreasonable due to Twitter's short text length, most users only discuss one event in a tweet. In this paper, we improve LDA for Twitter event detection by assigning only one event to an event-related tweet and differentiating General words from Event-related Aspects words.

To the best of our knowledge, we are the first to extend LDA to detect event in Twitter by simulating the generative process of the General words and Event-related Aspects words in a tweet. The advantages of our GEAM include: (1) GEAM outperforms the state-of-the-art topic models LDA in terms of both Precision and DERate (measuring Duplicated Events Rate detected); (2) the result is more informative than previous clustering algorithms which based on unigram text model, since we treat the words in a tweet differently based on different information corresponding words express (General words and Event-related Aspects words); (3) the event detection results of GEAM can also be easily utilized to perform event trend analysis.

The rest of the paper is organized as follows. Section 2 presents related work. Section 3 describes GEAM and its components in detail, and introduces the approximate inference algorithm for GEAM based on Gibbs sampling. In Section 4, we present the experimental results. Finally, we conclude the paper in Section 5.

## 2 Related Work

We present a survey of the state-of-the-art Twitter event detection methods and topic models in this section.

[6] `http://www.kansascity.com/2012/11/07/3903945/social-media-pickup-lines-illegal.html`

### 2.1 Twitter Event Detection

Although event detection from formal texts [2–4] has been studied for over a decade, event detection from Twitter is a relatively new topic which receives more interests in recent years. Most existing approaches for event detection from Twitter are limited to detect only certain types of events. For example, Sakaki et al. [6] devise a classifier to recognize tweets describing earthquakes in Japan, then they introduce a probabilistic spatiotemporal model to find the center and the trajectory of the earthquake. Rui et al. [7] focus on detecting Crime and Disaster related Events (CDEs), they also use a classifier to determine whether a crawled tweet is related to CDEs or not. Then, they utilize the author's network information to predict the location of a tweet. Benson et al. [8] utilize a graphic model to identify artists and venues mentioned with tweets posted by users in New York. Very recently, some open domain Twitter event detection approaches are proposed. Ritter et al. [9] extract an open-domain calendar of significant events from Twitter, they utilize a named entity tagger and sequence-labeling technology to extract event-related words. Chenliang et al. [10] propose a segment-based algorithm to detect events. They utilize Microsoft Web N-Gram service and Wikipedia to segment the tweets, detect the bursty event segments, and cluster the event segments using k-Nearest Neighbor Graph.

So, existing Twitter event detection approaches are either limited to certain types of tweets (e.g. related to earthquakes or CDEs, containing a predefined location) or based on detecting bursty words (single terms, hashtags or $n$-grams) which may be uninformative or meaningless.

### 2.2 Topic Models and Variants

Topic models (LDA) [13, 14] proposed by Blei et al. is popular for modeling latent topics in a corpus. Many variants of LDA has been proposed for different applications since it has been proven to be effective. Particularly, in social network or social media, Jonathan et al. [15] analyze documents on Wikipedia and infer descriptions of its entities and of relationships between those entities by proposing a probabilistic topic model. Hong et al. [16] study how the standard topic models can be trained on the microblogging dataset, they apply several schemes to train the model and compare their quality and effectiveness. Hu et al. [17] propose a Bayesian model called Event and Tweets LDA (ET-LDA) that performs topic modeling and event segmentation in one unified framework. Given transcripts of a known event from both New York Times and Twitter dataset, their work focus on how to jointly extract the topics covered by the two different datasets and segment the event.

The most related work is the model proposed by Lau et al. [12] to track emerging events in Twitter. They describe a topic model that processes documents in an online fashion. The model can update automatically based on time slices and can cope with dynamic vocabulary. However, they model each tweet as a multinomial mixture of all topics or events, which is obviously unreasonable due to short lengths of tweets. In this paper, we assign only one event to an event-related tweet.

## 3 General and Event-Related Aspects Model

We present an overview of our proposed system in Section 3.1. In Section 3.2, we develop the General and Event-Related Aspects Model (GEAM), and explain how to inference the model via collapsed Gibbs sampling.

### 3.1 Overview of the System

We aim at detecting open domain realistic events that a large number of people talking about in Twitter. Figure 1 plots the main components and procedures in our system for Twitter event detection. Given a raw stream of tweets, we remove "*pointless babbles*", which may not attract the majority users' attention based on the Named Entity Tagger information. Then, we send tweets (labeled with named entity and time information) to the General and Event-Related Aspects Model (GEAM) to estimate the General topic or Event-related Aspect word distributions. Finally, we rank the detected events based on the number of tweets assigned to each event.

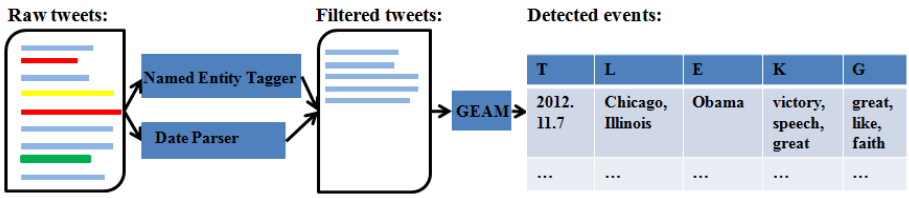



**Fig. 1.** Overview of our system

We manually investigate a dataset of one day's tweets on June 25, 2009. We observe that almost all the tweets that related to an realistic event contain named entities or hashtags. Table 1 shows 3 typical events on June 25, 2009, along with the number of whole tweets and tweets that contain named entities or hashtags corresponding to each event. From the numbers in Table 1, we can see that over 95% of tweets related to one realistic event contain named entities or hashtags. Based on this observation, we apply a named entity tagger for Twitter provided by Ritter et al. [5] to segment and label tweets. If no named entity or hashtag exists in a tweet, we ignore it. In other words, if a tweet contains some named entities or hashtags, it has great possibility related to a realistic event, and we will process it further. We also analyze the time information in tweets. There are several different ways to refer the same date in Twitter, for example "last night" and "the next Friday" may represent the same day depending on when the tweet was posted. So, we use natty[7] to extract the time information referred in a tweet. Natty is a natural language date parser that can recognize dates described in many ways, like *"the next Friday"*.

[7] `http://natty.joestelmach.com/`

**Table 1.** Number of whole tweets and tweets which contain named entities or hashtags corresponding to 3 typical events on June 25, 2009

| **Event** | **Iran election** | **"Transformer 2" released** | **Mark Sanford scandal** |
|---|---|---|---|
| all tweets | 24612 | 5644 | 3397 |
| tweets contain named entities or hashtags | 23583 | 5368 | 3234 |
| proportion | 95.8% | 95.1% | 95.3% |

After preprocessing and filtering, event-related tweets, which are labeled with named entities and time information, are sent to the General and Event-Related Aspects Model (GEAM), the kernel component of our system. GEAM is a probabilistic graphical model that simulates the generative process of a tweet related to a realistic event, and identifies General words and Event-related Aspects words in the tweet. A collapsed Gibbs sampling algorithm is designed to estimate word distributions of each event.

Finally, we rank the detected events, then present events in 4-tuple structures and associated General topics. For Event-related Aspect, *Time*, we provide the accurate date when the event occurred. For other three Event-related Aspects, we provide the top words according to the multinomial word distribution of each Event-related Aspect. We also provide the top General words associated with a event based on the underlying General topics.

### 3.2 General and Event-Related Aspects Model (GEAM)

We present how to model the realistic event and the corresponding general words associated with it by General and Event-related Aspects Model (GEAM), a hierarchical Bayesian model based on Latent Dirichlet Allocation (LDA). The inference of GEAM via collapsed Gibbs sampling will be introduced under the definition of the model.

**Model Description.** We use a generative process to model the tweets that describe a realistic event. GEAM models each tweet as a mixture of Event-related Aspects and General topics, then generates each word from the Event-related Aspects or General topics word distribution. We define a 4-tuple to represent an event in Twitter, i.e. *(Time, Locations, Entities, Keywords)*, each tuple reflects an aspect of the event. Due to informal expression of Twitter, people always use several words that doesn't belong to any Event-related Aspects mentioned before. We regard these part of words as General words in GEAM. For each tweet in the corpus, we assign a latent variable *event* to it, the *event* variable in GEAM is similar to a *topic* in LDA, and each *event* is characterized by 4 multinomial distributions over words, according to 4 Event-related Aspects. We assign an Event-related Aspect (i.e. *Time, Locations, Entities* or *Keywords*) to each named entity or hashtag word in a tweet, according to the named entity

tagger results and hashtags. For other words, we use a switch variable to indicate whether the word comes from General topic word distribution or Event-related Aspect word distribution. If the word is chosen from the General topic word distribution, then a *General topic* based on the whole corpus will be assigned to this word. Otherwise it will be assigned to the *Keyword Aspect* of an event. The graphical structure is shown in Figure 2.

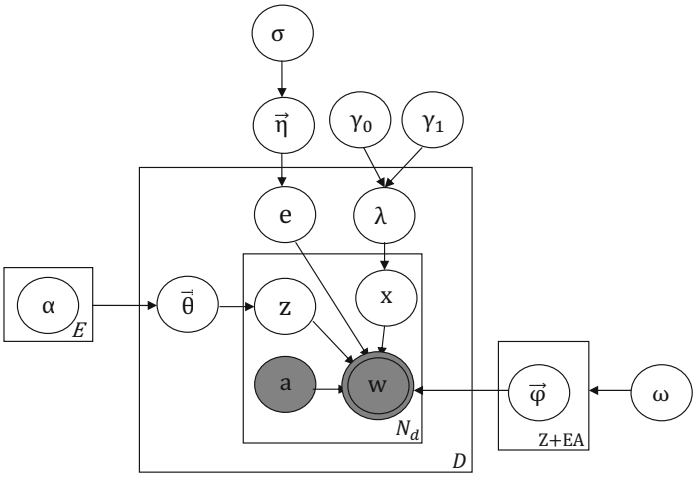



**Fig. 2.** Graphical representation of GEAM

Let us briefly introduce notations in Figure 2. $e$ is an latent variable assigned to each tweet specifying which event the tweets are describing, for example, a typhoon or a sports event. Each token in a tweet is associated with 4 variables (only 3 of them are meaningful at one time): word, $w$, switch variable, $x \in \{0, 1\}$, to indicate whether this word is an Event-related Aspect word or a General word, Event-related Aspect, $a$, to reflect which aspect (i.e. *Time, Locations, Entities* or *Keywords*) this word belongs to (in this case $x = 1$), and General topic, $z$, to specify which general topic the word corresponds to (in this case $x = 0$). Here we assume that there are $E$ events and $Z$ general topics in the whole corpus, each event has $A$ Event-related Aspects, and the size of vocabulary is $V$. $\overrightarrow{\eta}$ is the whole corpus' multinomial distribution over events, which is $E$ dimensional and drawn from a Dirichlet distribution parameterized by $\sigma$. $\overrightarrow{\theta}$ is a multinomial distribution over General topics for each tweet, which is $Z$ dimensional and drawn from a Dirichlet distribution parameterized by $\alpha_e$. $\overrightarrow{\varphi}$ is a multinomial distribution over words specific to topic $z$ or event $e$, which is $V$ dimensional and drawn from a Dirichlet distribution parameterized by $\omega$. $\lambda$ is the parameter for sampling the binary variable, $x$, and both $\gamma_0$ and $\gamma_1$ are beta parameters to generate $\lambda$. We summarizes the notations used in the GEAM in Table 2. Formally, the generative process is described in Algorithm 1: For the whole

**Table 2.** Notations in the model

| Symbol | Description |
|---|---|
| $w$ | a word in a tweet |
| $x$ | indicate if word $w$ is from a general topic $z$ or an event-aspect pair $ea$ |
| $e$ | the event described by the tweet |
| $z$ | the topic assigned to $w$, $x = 0$ |
| $a$ | the event aspect assigned to $w$, $x = 1$ |
| $\overrightarrow{\eta}$ | multinomial distribution over events |
| $\overrightarrow{\theta}$ | multinomial distribution over topics |
| $\overrightarrow{\varphi}_z$, $\overrightarrow{\varphi}_{ea}$ | multinomial distribution over words specific to topic $z$ or event $e$'s aspect $a$ |
| $\sigma, \omega, \alpha_e$ | Dirichlet priors to multinomial distributions $\overrightarrow{\eta}$, $\overrightarrow{\varphi}$, $\overrightarrow{\theta}$ |
| $\lambda$ | parameter for sampling the binary variable $x$ |
| $\gamma_0, \gamma_1$ | Beta parameter to generate $\lambda$ |
| $E$ | number of events |
| $A$ | number of aspects |
| $Z$ | number of general topics |
| $V$ | number of whole words, vocabulary size |

corpus, Lines 1-2 draw event distribution $\overrightarrow{\eta}$ and word distribution $\overrightarrow{\varphi}$. Then, for each tweet, draw 3 variables associated with it (Lines 4-6): a latent variable $e$, a general topic distribution $\overrightarrow{\theta}$, and a parameter $\lambda$ for sampling the switch variable $x$. Finally, for each token $w_i$ in the tweet: if the token is a named entity or hashtag word (Lines 8-11), determine the Event-related Aspect, $a_i$, based on the named entity tagger and date time parser information, then generate the word $w_i$ from $\overrightarrow{\varphi}_{ea_i}$. Otherwise, a switching variable, $x_i \in \{0, 1\}$, is drawn from a Binomial distribution (Line 13) to determine whether this word is chosen from the *Keywords* aspect distribution or from the general topic word distribution. If $x_i = 0$, choose a non-event (general) topic, $z_i$, and draw the word $w_i$ from $\overrightarrow{\varphi}_{z_i}$ (Lines 14-17). If $x_i = 1$, draw the word $w_i$ from event-relates *Keywords* aspect $\overrightarrow{\varphi}_{e3}$ (Lines 18-20).

**Model Inference.** We use Gibbs sampling to estimate unknown parameters $\overrightarrow{\eta}$, $\overrightarrow{\theta}$, $\overrightarrow{\varphi}$ and $\lambda$ in GEAM. Gibbs sampling allows the learning of a model by iteratively updating each latent variable given the remaining variables. In particular, we follow the idea of collapsed Gibbs sampling to approximate the posterior distribution of $e, x, a$, and $z$. We alternately sample the document-level variable $e$ and the token-level variables $x, a$, and $z$. Then, given the sampling results of $e, x, a$, and $z$, we can easily infer $\overrightarrow{\eta}$, $\overrightarrow{\theta}$, $\overrightarrow{\varphi}$ and $\lambda$.

Firstly, **sampling tweet level event** $e$ according to:

$$p(e_d|\mathbf{e_{-d}}, \mathbf{w}, \mathbf{x}, \mathbf{z}, \mathbf{a}) \propto \frac{n_{e_d,-\mathbf{d}} + \sigma_{e_d}}{\sum_{e=1}^{E} n_{e,-\mathbf{d}} + \sigma_e} \times \prod_{k=1}^{K} \frac{\prod_{t=1}^{V} \prod_{i=0}^{n_{k,\mathbf{d}}^t - 1} (n_{k,-\mathbf{d}}^t + \omega^t + i)}{\prod_{i=0}^{n_{k,\mathbf{d}}^* - 1} (n_{k,-\mathbf{d}}^* + \omega^* + i)}$$
$$\times \prod_{a=1}^{A} \frac{\prod_{t=1}^{V} \prod_{i=0}^{n_{e_d a,\mathbf{d}}^t - 1} (n_{e_d a,-\mathbf{d}}^t + \omega^t + i)}{\prod_{i=0}^{n_{e_d a,\mathbf{d}}^* - 1} (n_{e_d a,-\mathbf{d}}^* + \omega^* + i)} \quad (1)$$

where $\mathbf{e_{-d}}$ is the *event* vector associated with each tweet in corpus $D$ excluding tweet $d$, $n_{e,-\mathbf{d}}$ is the number of tweets in the whole corpus assigned to event $e$

Draw an event distribution $\overrightarrow{\eta} \sim Dir\,(\sigma)$ ;
Draw multinomial word distribution $\overrightarrow{\varphi} \sim Dir\,(\omega)$ for General topics (Z) and Event-related Aspects (EA);
**foreach** *tweet* $d \in [1, D]$ **do**
  Draw an event $e \sim Mult\,(\overrightarrow{\eta})$ ;
  Draw a multinomial topic distribution $\overrightarrow{\theta} \sim Dir\,(\alpha_e)$ ;
  Draw a switching distribution $\lambda \sim Beta\,(\gamma_0, \gamma_1)$ ;
  **foreach** *word* $w_i \in N_d$ **do**
    **if** $w_i \in \{$*named entities or hashtag words*$\}$ **then**
      Determine aspect $a_i \in \{0, 1, 2, 3\}$ (Time, Locations, Entities, Keywords aspect);
      Draw word $w_i \sim Mult(\overrightarrow{\varphi}_{ea_i})$;
    **end**
    **else**
      Draw $x_i \in \{0, 1\} \sim Bi(\lambda)$ ;
      **if** $x_i = 0$ **then**
        Draw topic $z_i \sim Mult(\overrightarrow{\theta})$ ;
        Draw word $w_i \sim Mult(\overrightarrow{\varphi}_{z_i})$ ;
      **end**
      **else**
        Draw word $w_i \sim Mult(\overrightarrow{\varphi}_{e3})$ (Keywords aspect);
      **end**
    **end**
  **end**
**end**

**Algorithm 1.** Probabilistic generative process in GEAM

expect tweet $d$, $n^t_{k,\mathbf{d}}$ is the number of word $t$ assigned to general topic $k$ ($x = 0$) in tweet $d$ and $n^*_{k,\mathbf{d}} = \sum_{t=1}^{V} n^t_{k,\mathbf{d}}$, $n^t_{e_d a,\mathbf{d}}$ is the number of word $t$ assigned to event-related aspect $e_d a$ ($x = 1$) in tweet $d$ and $n^*_{e_d a,\mathbf{d}} = \sum_{t=1}^{V} n^t_{e_d a,\mathbf{d}}$.

Secondly, for each token in a tweet, if the token is recognized as a named entity or a hashtag word, then the aspect is observed based on the preprocessing result. Otherwise, we jointly **sample the token-level variables** $x, a,$ **and** $z$. For $x_i = 0$:

$$p(x_i = 0, z_i = k | e_d, \mathbf{w}, \mathbf{x}_{-\mathbf{i}}, \mathbf{z}_{-\mathbf{i}}, \mathbf{a}) \propto \frac{N_{d,0,-\mathbf{i}} + \gamma_0}{N_{d,-\mathbf{i}} + \gamma_0 + \gamma_1} \times \frac{N^k_{d,-\mathbf{i}} + \alpha^k_{e_d}}{\sum_{k=1}^{Z} N^k_{d,-\mathbf{i}} + \alpha^k_{e_d}} \times \frac{N^t_{k,-\mathbf{i}} + \omega^t}{\sum_{t=1}^{V} N^t_{k,-\mathbf{i}} + \omega^t} \tag{2}$$

where $N_{d,0,-\mathbf{i}}$ is the number of words in tweet $d$ associated with $x = 0$ excepting word $i$, $N^k_{d,-\mathbf{i}}$ is the number of words in tweet $d$ associated with general topic $k$ excepting the $i$th token, $N^t_{k,-\mathbf{i}}$ reflects the number of term $t$ associated with general topic $k$ excluding the $i$th token. For $x_i = 1$, the derivation formula is very similar to Eq.(2) ($x_i = 0$). Due to the length limit, thus, we omit the detail equation here. Finally, we can easily obtain the multinomial or binomial parameters $\overrightarrow{\eta}$, $\overrightarrow{\theta}$, $\overrightarrow{\varphi}$ and $\lambda$ based on the previous sampling results.

## 4 Experiments

We demonstrate GEAM's performance from two aspects: effectiveness of event detection and capability for event trend analysis. For event detection effectiveness, the experiments show that GEAM outperforms the state-of-the-art topic models [13, 14] with better Precision and DERate (measuring Duplicated Events Rate detected). Also, 4-tuple structure *(Time, Locations, Entities, Keywords)* makes the detected events much easier to be understood by users. And majority people's reaction to the detected events can also be provided by the General topic words in GEAM. For event trend analysis, we demonstrate 3 typical events' trend based on detected results, and find different time patterns of these events.

### 4.1 Dataset and Experimental Setting

**Twitter Dataset.** We use one week data (from June 25, 2009 to July 1, 2009) from the tweets published by Stanford[8] to evaluate our system. There are a total of 7,088,229 tweets in the dataset. A number of realistic events happened in this period, such as "Iran Election 2009", "Micheal Jackson died", etc. Figure 3 shows the average of tweets published within each hour of a day. From Figure 3, we can see that most tweets were posted during midnight and afternoon. The average "*pointless babbles*" (tweets that contain neither named entities nor hashtags) that we filtered away in the dataset is about 47.32%.

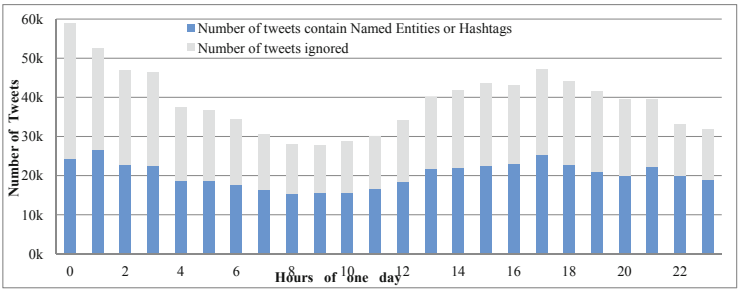



**Fig. 3.** Tweet volume against hour of day

**Parameter Setting.** There are several parameters in GEAM, and we use a validation set to find the optimal parameters. More specifically, GEAM archives good performance, when we set five parameters as follows: (1) $\gamma_0 = 0.8$ and $\gamma_1 = 0.2$ (to indicate that a word is more likely to be a General word), (2) $\alpha = 0.01$, $\omega = 0.01$, and $\sigma = 0.01$ (for hyperparameters of Dirichlet distribution). We can vary the number of events, $E$, and the number of general topics, $Z$, according to the data time-window (*e.g.*, one day or one hour). Note that, we fix time-window to one hour in this experiment, and GEAM is flexible enough to set the time-window to any time unit.

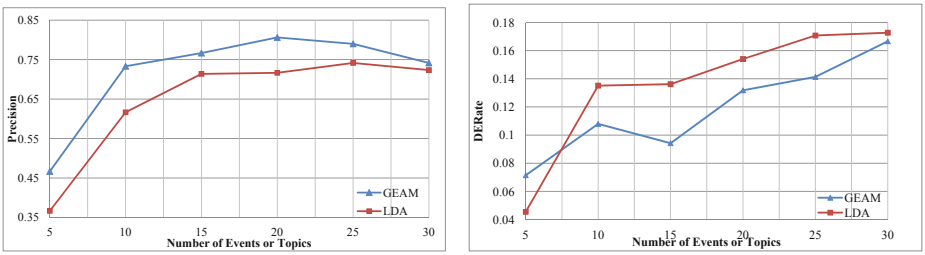
[8] `http://snap.stanford.edu/data/twitter7.html`

**Evaluation Metric.** A common problem existing in Twitter event detection is that we can't evaluate the quality of detected events by labeled ground truth, since it is infeasible to label over 7 million tweets manually. Alternatively, we manually search related real world facts to evaluate the detected events in terms of **Precision** and **DERate**. The **Precision** is defined as the fraction of detected events that are related to a realistic event. Note that, if there are several detected events related to the same realistic event, then all of them are regarded correct in **Precision**. Actually, the same realistic event may not be provided several times in the output. So, we use the metric **DERate** proposed in [10] to evaluate our GEAM, which is defined as the fraction of duplicate detected events among all all detected realistic events.

### 4.2 Event Detection Effectiveness

Our GEAM is different from LDA [13, 14]. LDA [13, 14] is a widely used statistical topic modeling technique, which aims at discovering the latent "topics" in a document corpus. For LDA, we set the number of topics K equals to the number of events E in GEAM; and we set the Dirichlet hyperparameters $\alpha = 0.01$, $\beta = 0.01$ similarly as $\alpha$, $\omega$ and $\sigma$ in GEAM.

**Event Detection Performance.** We compare GEAM and LDA in terms of Precision and DERate in Figure 4. We manually evaluate the top 5 words outputted by each model, if all the words are related to some realistic event, we regard it as true positive. We range the number of events (E) in GEAM and the number of topics (T) in LDA from 5 to 30. From the figure, we can see that GEAM outperforms LDA in terms of both Precision and DERate. Figure 4(a) demonstrates that based on one hour time-window dataset,the Precision of GEAM first increases with the increase of event number $E$, and archives highest of 80.6% at $E = 20$, then decreases when $E$ becomes larger. Similar behaviors are also observed in LDA, who receives highest Precision at $T = 25$. It is reasonable that the ptimal setting of $E$ based on one hour time-window Twitter dataset is about 20. When we set $E$ smaller, the GEAM may combine two similar realistic events into one detected event. This phenomenon is more obvious for LDA,




(a) (b)

**Fig. 4.** Comparison of GEAM and LDA. (a) shows the Precision of two models on different event numbers. (b) shows the DERate of two models on different event numbers.

which models each tweet as a mixture of $K$ latent topics. It shows in Figure 4(b) that generally the DERate of GEAM is smaller than LDA, which means that there are fewer duplicate events detected in GEAM. For each model, DERate increases with the increase of event number $E$ and topic number $K$. The reason for this phenomenon is that, when event number $E$ becomes larger, GEAM is more likely to assign different emphasis of one realistic event to different detected events. For example, when we set $E = 30$, the 2009 Iran election protest event will be divided into two events, one is for the video of Neda's death, the other is for an online activity to "Show support for democracy in Iran add green overlay to your Twitter avatar".

**Event Interpretation.** Five typical examples of the event detected by our GEAM in June 25, 2009 from 00:00 to 01:00 are demonstrated in Table 3, where T for *Time*, L for *Locations*, E for *Entities*, K for *Keywords*, and G for *General* words. For each event, we display top 10 Event-related Aspects words and top 3 General words. Due to the page limit, other events are not demonstrated. We illustrate that the output of GEAM is more informative, and can be easily understood with little background knowledge. From Table 3, we can see that the 4-tuple structure makes the detected events much clearer, and thus, users can easily access the different aspects of the event. GEAM can also provide reactions of majority people to the event by listing the General words. For example, $e_3$ expresses that the major audience enjoy the movie "Transformers 2". Among the 5 typical events listed in Table 3, we see that GEAM is an open-domain event detection model. It covers a wide range of events, from political events (Iran election), to entertainment events (Transformers 2), to sports (FIFA Confederations Cup), and to online hotspot (#lolquiz application).

### 4.3 Event Trend Analysis

By now, we evaluate GEAM on the basis of one hour time-window. It is more interseting to explore what can be found if we combine the detected events of continuous hours. As stated in Section 3.2, $\overrightarrow{\eta}$ in GEAM represents the multinomial distribution over events and $\overrightarrow{\eta}_e$ represents the fraction of tweets that are assigned to $event_e$. So, we can track the event in terms of hours. Figure 5 shows the trends of 3 events within a day. From Figure 5, we see that "*Iran Election*" was a stable event discussed on June 25, 2009, whose proportion didn't change dramatically during the whole day; "*Mark Sanford Scandal*" received less attention than "*Iran Election*", and was not discussed since 5:00; the "*Michael Jackson died*" event suddenly attracted majority users' attention from 20:00, and became much hotter than the other two events.

In summary, our GEAM outperforms the state-of-the-art topic models [13, 14] with better Precision and DERate. Particularly, 4-tuple structure makes the detected events much easier to be understood by users. Furthermore, event trend analysis can be performed easily based on the detection results.

**Table 3.** Examples of detected events

| Event | GEAM output | | Description |
|---|---|---|---|
| $e_1$ | $T$: | 2009-06-25 | Footage of the death of Neda drew international attention after she was shot dead during the 2009 Iranian election protests. |
| | $L$: | **Iran, Tehran** | |
| | $E$: | | |
| | $K$: | **#Iranelection, #Neda, #Iran, Mousavi, support, democracy, green, avatar** | |
| | $G$: | **video, crime, freedom** | |
| $e_2$ | $T$: | 2009-06-24 | The U.S. National Team won World No. 1 Spain, 2-0, in the semifinals of the FIFA Confederations Cup on June 24, advancing to next final against the winner of Brazil and South Africa. |
| | $L$: | | |
| | $E$: | **U.S., Spain, USA, Brazil, South Africa** | |
| | $K$: | **win, soccer, beat, team, school, 2-0** | |
| | $G$: | **great, watching, shocked** | |
| $e_3$ | $T$: | 2009-06-25 | "Transformers 2" released on June 24, 2009 in North America. Many people tweet about watching the film. |
| | $L$: | **America** | |
| | $E$: | **Transformers 2, Transformers** | |
| | $K$: | **good, movie, watch, tonight, wait, wish, lines** | |
| | $G$: | **tonight, join, happy** | |
| $e_4$ | $T$: | 2009-06-25 | South Carolina Governor Mark Sanford's disappearance and extramarital affair in June, 2009 was reported. He was in Argentina with his mistress for six days. |
| | $L$: | **Argentina, South Carolina** | |
| | $E$: | **Make Sanford, Sanford** | |
| | $K$: | **governor, mistress, e-mails, affair, exposed, bizarre** | |
| | $G$: | **shamed, news, dumbass** | |
| $e_5$ | $T$: | 2009-06-25 | A web entertainment application called #lolquiz gets popular. Many users try it to know which star available for them. |
| | $L$: | | |
| | $E$: | **twilight, Knotb Song, Jonas Brothers Song, Mcfly** | |
| | $K$: | **try, quiz, #lolquiz, song, fan, what** | |
| | $G$: | **took, lover, try** | |

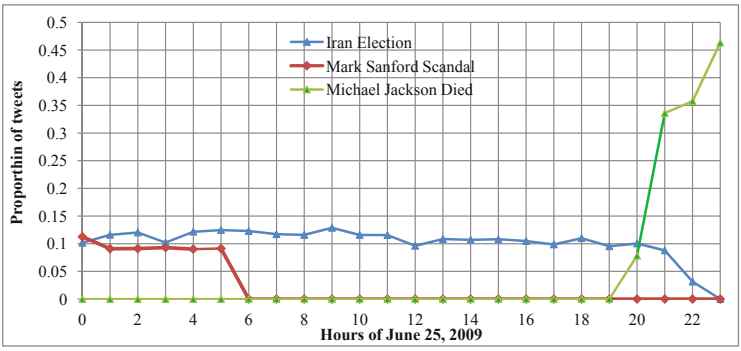



**Fig. 5.** Different proportion of tweets related to 3 events

## 5 Conclusions

In this work, we introduce a General and Event-Related Aspects Model (GEAM) for detect realistic event in Twitter. We design a collapsed Gibbs sampling algorithm to estimate the word distributions of an event. We also divide the words in an event-related tweet into General words or Event-related Aspect words, which matches the Twitter characteristic better than any unigram model. Our experiments demonstrate that GEAM outperforms the state-of-the-art topic model LDA in both Precision and DERate. Particularly, GEAM can get better event representation by providing a 4-tuple ($Time, Locations, Entities, Keywords$) of

the detected events and the associated General topics. Moreover, GEAM can be used to analyze event trends in continuous hours.

**Acknowledgments.** This work is partially supported by China National Science Foundation (Granted Number 61073021, 61272438), Research Funds of Science and Technology Commission of Shanghai Municipality (Granted Number 11511500102, 12511502704), Cross Research Fund of Biomedical Engineering of Shanghai Jiaotong University (YG2011MS38).

## References

1. Liu, K.L., Li, W.J., Guo, M.: Emoticon smoothed language models for twitter sentiment analysis. In: AAAI, pp. 1678–1684 (2012)
2. Fung, G.P.C., Yu, J.X., Liu, H., Yu, P.S.: Time-dependent event hierarchy construction. In: KDD, pp. 300–309 (2007)
3. Gabrilovich, E., Dumais, S.T., Horvitz, E.: Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In: WWW, pp. 482–490 (2004)
4. He, Q., Chang, K., Lim, E.P.: Analyzing feature trajectories for event detection. In: SIGIR, pp. 207–214 (2007)
5. Ritter, A., Sam, Clark, M., Etzioni, O.: Named entity recognition in tweets: an experimental study. In: EMNLP, pp. 1524–1534 (2011)
6. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors. In: WWW, pp. 851–860 (2010)
7. Li, R., Lei, K.H., Khadiwala, R., Chang, K.C.C.: Tedas: A twitter-based event detection and analysis system. In: ICDE, pp. 1273–1276 (2012)
8. Benson, E., Haghighi, A., Barzilay, R.: Event discovery in social media feeds. In: ACL, pp. 389–398 (2011)
9. Ritter, A., Mausam, Etzioni, O., Clark, S.: Open domain event extraction from twitter. In: KDD, pp. 1104–1112 (2012)
10. Li, C., Sun, A., Datta, A.: Twevent: segment-based event detection from tweets. In: CIKM, pp. 155–164 (2012)
11. Weng, J., Lee, B.S.: Event detection in twitter. In: ICWSM, pp. 401–408 (2011)
12. Lau, J.H., Collier, N., Baldwin, T.: On-line trend analysis with topic models: twitter trends detection topic model online. In: COLING, pp. 1519–1534 (2012)
13. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. In: NIPS, pp. 601–608 (2001)
14. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Science, 5228–5235 (2004)
15. Chang, J., Boyd-Graber, J.L., Blei, D.M.: Connections between the lines: augmenting social networks with text. In: KDD, pp. 169–178 (2009)
16. Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter. In: SOMA, pp. 80–88 (2010)
17. Hu, Y., John, A., Wang, F., Kambhampati, S.: Et-lda: Joint topic modeling for aligning events and their twitter feedback. In: AAAI, pp. 59–65 (2012)