Real Time Event Detection in Twitter

Xun Wang[1], Feida Zhu[2], Jing Jiang[2], Sujian Li[1*]

[1]Key Laboratory of Computational Linguistics (Peking University), MOE, CHINA
[2]School of Information Systems, Singapore Management University
{xunwang,lisujian}@pku.edu.cn, {feidazhu,jingjiang}@smu.edu.sg

Abstract. Event detection has been an important task for a long time. When it comes to Twitter, new problems are presented. Twitter data is a huge temporal data flow with much noise and various kinds of topics. Traditional sophisticated methods with a high computational complexity aren't designed to handle such data flow efficiently. In this paper, we propose a mixture Gaussian model for bursty word extraction in Twitter and then employ a novel time-dependent HDP model for new topic detection. Our model can grasp new events, the location and the time an event becomes bursty promptly and accurately. Experiments show the effectiveness of our model in real time event detection in Twitter.

Keywords: HDP, Gaussian mixture, Twitter, event detection

1 Introduction

Events usually refer to something abnormal, that is, something that rarely happens in normal situation. Event detection aims to find such abnormal phenomenon from collected data. Various methods have been proposed to detect events such as disease outbreaks, criticism and so on [1, 2].

When it comes to Twitter, events are topics that suddenly draw public attention, for example, music concerts, football matches and so on. Some topics are not events, although they are popular. For example, "iphone" and "ipad" have always been popular according to the trendy topics provided by twitter.com officially [1]. They are talked about widely and repeatedly and show a high frequency everyday but are not bursty events. Neither are periodical topics events. For example, tweets on Friday usually talk about the coming of weekends while tweets on Monday usually complain about the beginning of a week of work. Such topics should not be regarded as events.

Several approaches have been proposed for event detection from tweets, such as wave analysis[3], topic model approach based on LDA[4], HDP[15] or text classification and clustering[5]. Existing approaches suffer from either failure in latent topics detection or inefficiency in involving topics models.

In this paper, we divide the event detection task into three steps. We firstly use a Gaussian mixture for bursty word candidate selection from huge tweets data. Our

* the corresponding author
[1] http://api.twitter.com/1/trends/

adfa, p. 1, 2011.

candidate selection model considers both weekly effects and monthly effects. Further to decide whether a bursty word represents a new topic, we borrow the ideas of evolutionary clustering which focuses on detecting the dynamics of a given topic such as appearance, disappearance and evolution for new topics. We develop a novel time dependent HDP(td-HDP) model based on HDP model[6] for new event detection. Our model is based on the assumption that the data of Twitter forms a Markovian chain and each day's topic distribution is affected by the topic distribution of previous time. The number of events would naturally increase or decrease with the appearance and disappearance of new topics which can be detected in td-HDP. Finally, the location of event is detected from a CRF algorithm[16]. Usually there're three compulsory components of an event in Twitter: topic, location, the time it becomes bursty. Our model can grasp the three points promptly and accurately at the same time.

The rest of paper is organized as follows. Section 2 describes related works in event detection in Twitter. Bursty words detection is introduced in Section 3. The td-HDP model which aims at detecting new topics from busty words is presented in Section 4. Location recognition is briefly described in Section 5. Experimental results are presented in Section 6 and we conclude the paper in Section 7.

Fig. 1. System architecture of our model

2 Related works

2.1 Event Detection

Existing event detection methods usually treat words as signals. Some researches deal with words in the time domain. Kleinberg[7] used an infinite-state automaton to detect events, with events represented by the transitions of state. Fung et al.[8] developed the idea to estimate the appearance of words from binomial distribution and bursty words are detected with a threshold-based heuristic. There are some other methods that deal with words in the frequency domain[8] where traditional Discrete Fourier Transformation (DFT) is applied to convert the signals from the time domain into the frequency domain.

When it comes to Twitter, Weng et al.[3] applied the wavelet analysis to tweets. Words are converted into signals and corresponding signal auto-correlations are used to find non-trivial words. These non-trivial words are then clustered to form events. The problem of such kind of work is that words are treated as signals and it is hard to capture the latent topics within text. Topic models such as LDA have gained great success in these days for their clear and rigorous probabilistic interpretations for latent topic modeling. LDA has also been used for event detection task in Twitter[4]. The problem is, in LDA, the number of topics should be fixed in advance. So it's not suitable for real time event detection where the amount of data gradually grows, especially when the data is huge.

2.2 Evolutional Clustering and HDP

Evolutionary clustering is a relatively new research for topic detection, which aims to preserve the smoothness of clustering results over time, while fitting the data of each epoch. The work by Chakrabarti et al. [10] was probably considered as the first to address the problem of evolutionary clustering. They proposed a general framework of evolutionary clustering and extended two classical clustering algorithms to the evolutionary setting: (1) k-means clustering, and (2) agglomerative hierarchical clustering. The problem of evolutionary clustering is that the number of clusters stays the same over time. This assumption is obviously violated in many real applications.

Recently, HDP has been widely used in evolutionary clustering due to its capability of learning number of clusters automatically and sharing mixture components across different corpora. In HDP, each corpus is modeled by an infinite Dirichlet Process (DP) mixture model, and the infinite set of mixture clusters is shared among all corpora. Sethuraman [11] gave a stick-breaking constructive definition of DP for arbitrarily measurable base space and Blackwell and MacQueen[12] explained DP using the Polya urn scheme. The Polya urn scheme is closely related to the Chinese Restaurant Process (CRP) metaphor, which is applied on HDP demonstrating the 'clustering property' as the 'distribution on partition'. Based on HDP, some algorithms of evolutionary clustering are proposed by incorporating time dependencies, such as DPChain, HDP-EVO and dynamic HDP et al [13], [14], [15].

3 Bursty Words Extraction

As stated, we firstly try to find bursty words which serve as candidates for new event detection in tweets. Bursty words are those whose frequencies severely increase in a short time. Words in tweets can be regarded as being drawn from a certain distribution. According to the central limit theorem, the frequencies of a word in each day can be approximately modeled by a Gaussian distribution. The parameters of distribution can be estimated from historical data. We use the records of data in tweets from Jan 2011 to Dec 2011 as the historical data. The frequency of each word is recorded in a 1*365 dimensional vector, denoting the number of appearance of certain word in each day of the year. Let $F_{x_i,t}$ denote the frequency of word x_i at day t. If x_i is not a bursty word, we assume that the distribution of $F_{x_i,t}$ should be almost the same as the mean frequency distribution in the past whole year. In addition, we also have to get rid of the periodical effect such as weekly effect and monthly effect. For example, topics in weekdays and weekends would be different and topics in different months should also be different. For example, when December comes, it is normal to find words such as 'it is getting colder and colder' in tweets. And these topics should not be regarded as new topics. We propose a mixture Gaussian distribution which can consider both weekly and monthly effect. For word x_i at time t, if it is not a bursty word, we assume that:

$$F_{x_i t} \sim a_{x_i t} \frac{1}{\sqrt{2\pi\sigma_{x_i}^2}} \exp[-\frac{(F_{x_i t} - \mu_{x_i})^2}{\sigma_{x_i}^2}] + b_{x_i w_t} \frac{1}{\sqrt{2\pi\sigma_{x_i w_t}}} \exp[-\frac{(F_{x_i t} - \mu_{x_i w_t})^2}{\sigma_{x_i w_t}^2}] \quad (1)$$

$$+c_{x_i m_t} \frac{1}{\sqrt{2\pi\sigma_{x_i m_t}}} \exp[-\frac{(F_{x_i t} - \mu_{x_i m_t})^2}{\sigma_{x_i m_t}^2}] \qquad a_{x_i t} > 0 \quad b_{x_i w_t} > 0 \quad c_{x_i m_t} > 0 \qquad a_{x_i t} + b_{x_i w_t} + c_{x_i m_t} = 1$$

where w_t denotes whether time t is Monday, Tuesday,…Sunday, and m_t denotes whether time t is in Jan, Feb,…Dec. w_t=[1,2,…,7] and m_t=[1,2,..,12]. The first term in Eq(1) concerns about the overall effect within a year for word x_i. The second term concerns about the weekly effect and the third term concerns about the monthly effect. All parameters in Eq(1) can be obtained from EM algorithm by maximizing likelihood estimate. But in this paper, we adopt the following approximation which largely cuts off running time.

$$\mu_{x_i} = \frac{1}{|T_1|} \sum_{t=1}^{t=|T_1|} F_{x_i t} \qquad \sigma_{x_i w_t}^2 = \frac{1}{|T_1|} \sum_{t=1}^{t=|T_1|} (F_{x_i t} - \frac{1}{T_1} \sum_{t=1}^{t=|T_1|} F_{x_i t})^2 \quad (2)$$

$$\mu_{x_i w_t} = \frac{1}{|T_{w_t}|} \sum_{t=1}^{t=|T_{w_t}|} F_{x_i t} \qquad \sigma_{x_i w_t}^2 = \frac{1}{|T_{w_t}|} \sum_{t=1}^{t=|T_{w_t}|} (F_{x_i t} - \frac{1}{T_1} \sum_{t=1}^{t=|T_{w_t}|} F_{x_i t})^2 \quad (3)$$

$$\mu_{x_i m_t} = \frac{1}{|T_{m_t}|} \sum_{t=1}^{t=|T_{m_t}|} F_{x_i t} \qquad \sigma_{x_i m_t}^2 = \frac{1}{|T_{m_t}|} \sum_{t=1}^{t=|T_{m_t}|} (F_{x_i t} - \frac{1}{|T_{m_t}|} \sum_{t=1}^{t=|T_{m_t}|} F_{x_i t})^2 \quad (4)$$

$|T_1|$ is 365. μ_{x_i} and $\sigma_{x_i w_t}^2$ are the mean and variance of frequency of x_i in the whole year. Similarly, $\mu_{x_i w_t}$ and $\sigma_{x_i w_t}^2$ are the mean and variance of frequency of x_i at all w_t in a year. w_t could be Monday, Tuesday… This is the same case with $\mu_{x_i m_t}$ and $\sigma_{x_i m_t}^2$. Then $a_{x_i t}$, $b_{x_i w_t}$ and $c_{x_i m_t}$ can be learned through a maximum likelihood estimation. Experiments show that this approximation works well.

Next we get down to the bursty word selection. At time t, word with a daily frequency higher than the upper boundary of its 99% confidence interval according to Eq(1) would be selected as a bursty word. At each time t, we selected all words whose frequencies exceed the thresholds as bursty words which serve as candidates for new event detection in Section 4.

4 New Events Detection

In this section, we use the bursty words extracted from Section 3 to detect novel events in Twitter. We firstly describe Dirichlet Process and Hierarchical Dirichlet process, and then introduce our time dependent HDP model (tdHDP).

4.1 DP and HDP

A DP[11] can be considered as a distribution of probability measure G. Suppose a finite partition $(T_1,\ldots,T_K)$ in the measure space Θ and a probability distribution G_0 on Θ, we write $G \sim DP(\alpha, G_0)$ if $(G(T_1),\ldots,G(T_K)) \sim Dir(\alpha G_0(T_1),\ldots,\alpha G_0(T_K))$, where α is a positive concentration parameter and G_0 is called a base measure. Sethuraman [11] showed that a measure G drawn from a DP is discrete by the stick-breaking construction:

$$\pi_k = \beta_k \prod_{l=1}^{k-1}(1-\beta_l) = \beta_k(1-\sum_{l=1}^{k-1}\pi_l),\ \ \{\phi_k\}_{k=1}^{\infty} \sim G_0,\ \ G=\sum_{k=1}^{\infty}\pi_k\delta_{\phi_k} \quad (5)$$

where δ_{ϕ_k} is a probability measure concentrated at ϕ_k. For convenience, we write $\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha)$ if $\boldsymbol{\pi}$ is a random probability measure defined by Eq. (5).

A HDP[4] defines a distribution over a set of DPs. In HDP, a global measure G_0 is distributed as a DP with concentration parameter γ and base measure H. Then a set of measures $\{G_j\}_{j=1}^{J}$ is drawn independently from a DP with base measure G_0. Such a process is described as:

$$G_0 \sim DP(\gamma, H),\quad G_j \mid \alpha_0, G_0 \sim DP(\alpha_0, G_0) \quad (6)$$

For each j, let $\{\theta_{ji}\}_{i=1}^{n_j}$ be independent and identically distributed (i.i.d.) random variables drawn from G_j. n_j observations $\{x_{ji}\}_{i=1}^{n_j}$ are drawn from the mixture model:

$$\theta_{ji} \sim G_j,\quad x_{ji} \sim F(x \mid \theta_{ji}) \quad (7)$$

where $F(x|\theta_{ji})$ denotes the distribution of generating x_{ji}. Equations (6) and (7) complete the definition of a HDP mixture model, whose graphical representation is shown in Figure 2(a).

According to Eq. (5), the stick-breaking construction of HDP can be represented as:

$$(\beta_k)_{k=1}^{\infty} \sim GEM(\gamma),\ \ G_0=\sum_{k=1}^{\infty}\beta_k\delta_{\phi_k},\ \ (\pi_{jk})_{k=1}^{\infty} \sim \mathrm{DP}(\alpha 0,\ \beta),\ \ G_j=\sum_{k=1}^{\infty}\pi_{jk}\delta_{\phi_k} \quad (8)$$

The corresponding graphical model is shown in Figure 2(b).

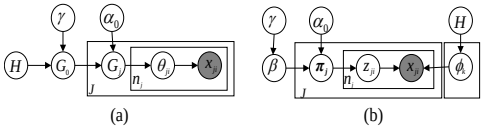


Fig. 2. HDP. (a) the original representation. (b)the stick-breaking construction

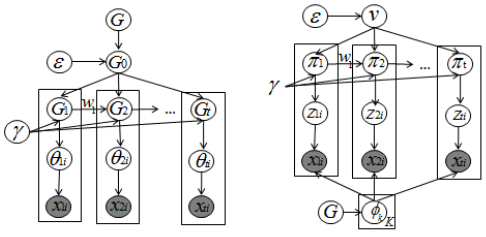
4.2 td-HDP

We model the multiple correlated corpus with Dirichlet Processes with Markovian time dependency. Specially, at each time t, we use a DP to model the data from different time and then put the time dependency between different epochs based on the Markovian dependency. All DPs at different time G_t would share an identical base measure G, and G is drawn from $DP(\xi, H)$ where H is the base measure. According to Markovian assumption, G_t is not only affected by the overall base measure G, but also affected by G_{t-1}. The generation process of td-HDP is as follows:

1. Draw an overall base measure $G \sim DP(\varepsilon, G_0)$, which denotes the base distribution for all time data.
2. If t is the start point, draw the local measure $G_t \sim DP(\gamma^t, G)$ according to the overall measure G, else draw $G_t \sim DP(\gamma^t, (1-w^t)G + w^t G_{t-1})$, where $w^t = \exp(-c\Delta_{t,t-1})$. $\Delta_{t,t-1}$ denotes the exact time difference between epoch t and epoch t-1. In this paper $\Delta_{t,t-1}$ is set to one day. w^t is the factor that controls influence of topic distribution from previous time. c is the time controlling factor.
3. For each word $x_{t,i} \in D_t$ draw $\theta_{t,i} \sim g(\theta \mid G_t)$, $x_{t,i} \sim f(x \mid \theta_{t,i})$

According to the stick-breaking construction of DP, the overall base measure G can be represented with the following form:

$$G = \sum_{k=1}^{\infty} v_k \delta_{\phi_k} \ , \boldsymbol{v} \sim GEM(\varepsilon) \ G_t = \sum_{k=1}^{\infty} \pi_k^e \delta_{\phi_k} \ , \ \ \boldsymbol{\pi}^t \sim DP(\gamma^e, (1-w^t)v + w^t \boldsymbol{\pi}^{t-1}) \quad (9)$$



(a) (b)

Fig. 3. (a)Graphical representation of our model (b) stick construction of our model

γ is sampled from a vague gamma prior which is set to be $Ga(10.0, 1.0)$.

4.3 Inference

We begin with the metaphor following Chinese Restaurant fanchise for inference. At time t, the word collection D_t corresponds to a restaurant, and each word $x_{t,i} \in D_t$ corresponds to a customer. Each table t in the restaurant would have a dish k, which can be interpreted as the topic. All customers sitting around table m would enjoy dish k, meaning that these words are assigned to topic k. In the Chinese Restaurant metaphor, customer x_{ti} sits at table p_{ti} while table p in restaurant D_t serves dish k_{tp}. MCMC sampling is used for td-HDP sampling.

We also need a notation for counts. $n_{tp\bullet}$ represents the number of customers in restaurant t at table p, and $n_{t\bullet k}$ represents the number of customers in restaurant t eating dish k. The notation m_{tk} denotes the number of tables in restaurant t serving dish k, $m_{t\bullet}$ denotes the number of tables in restaurant t, $m_{\bullet k}$ denotes the number of tables serving dish k in all restaurants, and $m_{\bullet\bullet}$ denotes the total number of tables in all restaurants. $n_{t\bullet k}$ denotes the number of customers in restaurant t having dish k and $m_{\bullet k}$ denotes the total number of tables serving dish k in all restaurants.

Sampling $\boldsymbol{p}$. Due to the space limit, we would just show the sampling formula without derivation. The likelihood due to $x_{t,i}$ given $p_{ti} = p$ for some previously p is $f_{k_{tp}}^{-x_{t,i}}(x_{t,i})$.

$$p(p_{ti} = p \mid p_{-ti}, \boldsymbol{k}) \propto \begin{cases} [(1-w_{t-1})n_{tp\bullet} + w_{t-1}n_{t-1\bullet k_{tp}} / m_{tk_{tp}}]\dfrac{E^{(k)}_{(x_{t,i})} + \beta}{n_{\bullet\bullet k} + V\beta} & \textbf{if } \boldsymbol{p} \textbf{ previously used} \\ \gamma\dfrac{1}{V} & \textbf{if } \boldsymbol{p} = p^{new} \end{cases} \tag{10}$$

w_{t-1} is the influence factor from restaurant t-1, $E^{(k)}_{(x_{t,i})}$ denotes the number of replicates of x_{ti} in topic k and β is the Dirichlet prior. If the sampled value of p_{ti} is p^{new}, then we can sample the topic of new table as follows $p_{kp^{new}}$

$$p(k_{tp^{new}} = k \mid p, \boldsymbol{k}_{-tp^{new}}) \propto \begin{cases} \dfrac{m_{\bullet k}}{m_{\bullet\bullet} + \varepsilon}\dfrac{E^{(k)}_{(x_{t,i})} + \beta}{n_{\bullet\bullet k} + V\beta} & \textbf{if } k \in K \\ \dfrac{\varepsilon}{m_{\bullet\bullet} + \varepsilon}\dfrac{1}{V} & \textbf{if } k = k^{new} \end{cases} \tag{11}$$

Sampling $\boldsymbol{k}$. Because the process of sampling $\boldsymbol{t}$ actually changes the component member of tables, we continue to sample k_{tp} for each table in the restaurant as follow:

$$p(k_{tp}=k \mid p, \boldsymbol{k}_{-tp}) \propto \begin{cases} \dfrac{m_{\bullet k}}{m_{\bullet\bullet}+\varepsilon} \dfrac{\Gamma(n_{\bullet\bullet k}+V\beta)}{\Gamma(n_{\bullet\bullet k}+n_{tp\bullet}+V\beta)} \displaystyle\prod_{x_{t,j}\in p_{tp}} \dfrac{\Gamma(E^{(k)}_{(x_{t,j})}+E^{(p_{tp})}_{(x_{t,j})}+\beta)}{\Gamma(E^{(k)}_{(x_{t,i})}+\beta)} & \textbf{if } k \in K \\ \dfrac{\varepsilon}{m_{\bullet\bullet}+\varepsilon} \dfrac{\Gamma(V\beta)}{\Gamma(n_{tp\bullet}+V\beta)} \displaystyle\prod_{x_{t,j}\in p_{tp}} \dfrac{\Gamma(E^{(p_{tp})}_{(x_{t,j})}+\beta)}{\Gamma(\beta)} & \textbf{if } k=k^{new} \end{cases} \tag{12}$$

$E^{(p_{tp})}_{(x_{t,j})}$ denotes the number of replicates x_{ti} at current table.

4.4 New Events Detection

In the two following situations, the event is regarded as new:

1. $k_{x_{ti}}$ =new. This means that the topic has never appeared before, so it is a new event .
2. $k_{x_{ti}} \neq$ new and $k_{x_{ti}} \notin K_{t-i}$ i $\in [1,2,3]$. This means that even though $k_{x_{ti}}$ appeared before, but it did not appear in the past three days. So we can also regard $k_{x_{ti}}$ as new event.

Note each topic is represented by the top five words with largest probability.

5 Location Recognition

We also try to find the related locations for events. It is a traditional Named Entity Recognition task. In this paper, locations of an event is recognized through a CRF model[16]. The training data contains more than 1M tweets in Singapore, with several location names of events tagged. These events and their locations are manually selected and tagged using simple rules. For example, a car accident in Rochor Road is an event. Then "Rochor Road" in tweets of the same day such as "there's a car accident at Rochor Road" or "I saw an accident in Rochor Road" would be tagged as a location of event. The feature template is unigram, current word w and four adjacent words, which are w, w+/-1 and w+/-2. Bigram is also used as an important feature.

6 Experiments

We use tweets of Singapore in one year as history data to decide the normal situation of words and tweets in May 2012 as test data. All these tweets are from more than 15k users who follow the 5 most popular political accounts in Singapore. We use the data from Jan. 2011 to Dec. 2011 as training data. Specifically, data from Jan. 2011 to Dec.2011 is used as history data for training of parameters in Gaussian mixtures de-

scribed in Section 3 for bursty words extraction and data from Jan. 2011 to Mar.2011 for parameter tuning in 6.2. Data from Apr. 2012 to Jun. 2012 is used as test data.

6.1 Pre-processing

Words which contain too many replicates such as “hahahahha”, “mmmmmm” or “zzzzzz” or do not include valid characters such as “^_^” are deleted. Moreover stop words are also deleted. Tweets that contain less than two words are also ignored.

6.2 Parameter Tuning

Firstly the time controlling factor in time dependant HDP needs to be tuned. We use data from Jan 2011 to Mar 2011 as training data and firstly construct the new event collection by manually selecting new events detected by different models. We asked five undergraduates in Singapore to find true events detected from event collection detected by Trendy Topic in twitter (list of hot topics given by twitter in a certain period of time), LDA model, tf-idf model, tdHDP(C=0), tdHDP(c=1) and tdDHP(c=5). We collect 154 events in event collection. Then we experiment the td-HDP algorithm by setting c in the range from 0 to 5 with interval of 1. We compare the results with different c value with true event collection which was built previously according to manual evaluation and calculate accuracy for each experiment. We find that the accuracy drops sharply when c is set as a value larger than 2.0. Next, c is set in the range from 0 to 2.0 with interval of 0.2. Fig. 4 show the different value of accuracy with regard to different values of c. We find that the value of accuracy reaches their peak at around 1.2 and drops afterwards.

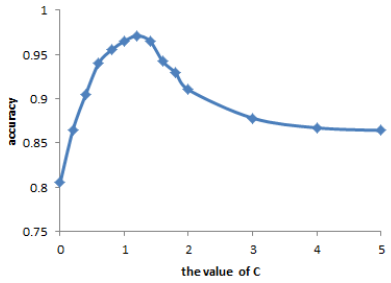


Fig. 4. Tuning the parameter C for tdHDP

6.3 Experimental Results

Due to the limitation of space, only events detected in May 2012 are shown in Table 1. We detect 33 events for May 2012 and 32 out of 33 are correct event detection according to manual evaluation. The wrongly detected events are marked in grey. As for Locaions, “Online” means it’s an online topic with no related locations in real world. All 8 locations are correct event locations according to human judgment. We check the related tweets of the 9th one which is not a valid event. There are two differ-

ent topics, “adobe cs6” and “cats in picasa”. They are mixed together because they share words such as “pics”.

Date & Time	Events	Event Description	Locs
1st 4:55:19	labour happy holiday work	Labour day	Online
2nd 9:26:3	khj fans session meeting	Fan meeting of KHJ, a pop star	Online
2nd 15:36:42	Infinite word album brand	An album called Infinite is released	Online
4th 12:58:2	Kelantan match fans football	Football match, Kelantan vs LionsXII	Online
8th 13:33:9	whatsapp working wrong problem	Whatsapp, one app on smartphone	Online
8th 14:59:13	opera jap super junior dance	Super Junior’s opera	Online
9th 7:34:19	ironman hulk captain avengers	Some popular movies	Online
9th 8:58:39	election hougang tony tan news	hougang election	Hougang
9th 5:29:34	cats cs6 faker picasa gallery	Not a valid event	N/A
12th 12:48:57	mothers day happy love mum	mothers’ day comes	Online
13th 13:51:49	Alonso win Spanish Maldonado	Maldonado beat Alonso in Spanish Grand Prix	Online
13th 14:40:0	united man city beat play	Man City beat Man United	Online
13th 14:47:22	ferrari taxi accident driver dead	A ferrari taxi accident	Bugis
15th 6:57:2	diablo Funan Diablo iii fans	Funan Digital Mall released the game Diablo III	Online
16th 2:58:29	hougang pappng work election	hougang election	Hougang
16th 3:54:27	Zeng Guoyuan parrot police	Zeng Guoyuan's bird abused the police	Online
16th 8:38:31	speeding red driver beating lights	A speeding driver drove through the red light	Online
16th 12:25:54	pixie lott live topshop she	A famous singer Pixie Lott went to TopShop	Bugis
16th 12:39:10	England squad euro Neville	Gary Neville became England coach	Online
19th 20:26:18	goal Bayern Chelsea Thomas	Thomas Müller in the match: Bayern vs Chelsea	Online
19th 21:17:3	Chelsea win germans Bayern hope	Football match: Chelsea beats Bayern	Online
20th 12:28:35	Phua Chu kang Denise politics	hougang election	Hougang
20th 23:59:12	Gibb Robin dies singer cancer	Famous singer, Gibb Robin dies of cancer	Online
22nd 3:50:7	MBC concert google korean music	MBC's 'Korean Music Wave in Google'	Online
22nd12:28:18	pxdkitty camera win blog world	To win a camera called Pxdkitty by blogging	Online
24th 15:7:5	thor thunder loki hammer rain	A movie called Thor	Online
26th 7:55:12	Joongki song today man running	Song Joongki, an actor in TV series	Marina
26th 14:38:1	worker party hougang partner win	hougang election	Hougang
27th 11:23:4	Taufik Rossa Batisah excited	Taufik & Rossa in Singapore TV show	Jalan
27th 13:58:54	Webber Mark Monaco wins	Mark Webber wins the game in Monaco	Online
28th 15:34:45	gaga lady concert stadium indoor	Lady GaGa’s concert in statium	Online
30th 9:48:33	sep 28th big bang coming	Pop band Big Bang in SG on Sept. 28th	Online

Table 1. Events in May 2012

6.4 Evaluations

The evaluation of our model is conducted in three aspects. The first one is Timeliness, which represents whether our model can detect new events quickly. In addition, we evaluate the precision and recall of our model, which respectively denote the model’s ability in detecting new events correctly and completely.

6.4.1 Evaluation of Timeliness

In tdHDP model, each word has a probability that is generated by a topic. Here we use the top word in the new topic to represent the event. The timeliness of an event is evaluated by the difference between the time when we detect the bursty word that represents the new topic and the peak appearance time of that word. Results of 32 events detected in May 2012 are shown in Fig. 4. Only 4 out of 32 events are detected after the frequency of bursty words arrives at its peak.

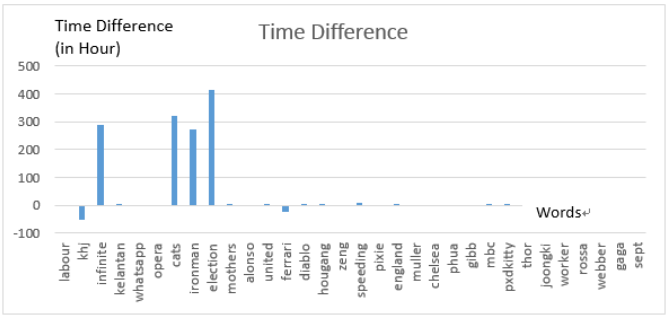


Figure. 4. The time difference between event detected time and peak time of bursty words

6.4.2 Evaluation of Precision and Recall

In this subsection, we evaluate the compare precision and recall of our model with existing models. The first baseline is standard HDP model which does not consider time dependent relations. LDA model and tf-idf model are also used as baselines.

Model	Precision
td-HDP	0.968
Standard HDP	0.890
LDA	0.841
tf-idf	0.806

(a)

Model	Recall
td-HDP	0.931
Standard HDP	0.832
LDA	0.797
tf-idf	0.710

(b)

Table 2. (a)Precision comparison with other models.(b)Recall comparison with other models

As in 6.2, we build the event collection and manually judge the results of these methods. Then we compare the results of each model to the collection and get the Precision and Recall of each model. The results are shown in Table 2. Standard HDP achieves better results than LDA because HDP can model data more properly and topic number can increase and decrease naturally with the appearance and disappearance of topics. But in LDA, topic number has to be fixed in advance. Td-HDP achieves best result and this further illustrates the necessity of adding Markovian time relation in topic modeling. Our approach enjoys the precision of 0.968 and a Recall of 0.931, demonstrating the effectiveness of our model.

6. Conclusion

We propose a novel event detection model for Twitter. A Gaussian Mixture model is constructed for word candidate selection in regard with periodic effect and a time-dependent HDP model is developed for new event detection from word candidates. Our model can deal with large amount of data effectively and detect events efficiently. Experiments show the good performance of our model.

Acknowledgements This work is supported by NSFC programs (No: 61273278), NSSFC (No: 10CYY023), National Key Technology R&D Program (No: 2011BAH10B04-03), National High Technology R&D Program of China (No. 2012AA011101) and Singapore National Research Foundation under its IRC @ SG FI and administered by the IDM Programme Office.

References

1. M. Kulldorff, F. Mostashari, L. Duczmal, W. K. Yih, K. Kleinman, and R. Platt. Multivariate scan statistics for disease surveillance. *Statistics in Medicine*, 26: 1824–1833, 2007.
2. D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, 32(3): 962–994, 2004.
3. Weng, J., & Lee, B.-S. 2011. Event detection in twitter. ICWSM'11. Barcelona, Spain.
4. Qiming Diao, Jing Jiang, Feida Zhu and Ee-Peng Lim. 2012, Finding bursty topics from microblogs. ACL'12, Jeju, Korea.
5. Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. WWW '10. New York, USA.
6. Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566--1581, 2006.
7. J. Kleinberg. 2002 Bursty and hierarchical structure in streams. KDD'02, New York, USA.
8. Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Philip S. Yu, and Hongjun Lu. 2005 Parameter free bursty events detection in text streams. VLDB'05. Trondheim, Norway.
9. Qi He, Kuiyu Chang, and Ee-Peng Lim.2007. Analyzing feature trajectories for event detection. SIGIR'07, New York, USA.
10. Chakrabarti, D., Kumar, R. and Tomkins, A. 2006. Evolutionary clustering, KDD'06, Philadelphia, USA
11. Sethuraman J.1994. A constructive definition of Dirichlet priors. *Statistica Sinaca*, Vol 2 pages 639-650.
12. Blackwell, D. and MacQueen, J. B.1973. Ferguson distributions via Polya urn schemes. *The Annals of Statistics*, Vol 1, page 353-355.
13. Xu, T., Zhang, Z. M. and Yu, P. S. and Long. B. 2008. Dirichlet process based evolutionary clustering. ICDM'08. Pisa, Italy,
14. Ren, L, Dunson, D. B., and Carin, L. 2008. The dynamic hierarchical Dirichlet process. ICML'08, Helsinki, Finland.
15. Gao, Z. J., Song, Y., and Liu, S. 2011. Tracking and Connecting Topics via Incremental Hierarchical Dirichlet Processes, ICDE'11. Hannover, Germany.
16. John L., Andrew M. , Fernando P., 2001, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, ICML'01, Williamstown, USA