A Unified Model for Topics, Events and Users on Twitter

Qiming Diao
Living Analytics Research Centre
School of Information System
Singapore Management University
qiming.diao.2010@smu.edu.sg

Jing Jiang
Living Analytics Research Centre
School of Information System
Singapore Management University
jingjiang@smu.edu.sg

Abstract

With the rapid growth of social media, Twitter has become one of the most widely adopted platforms for people to post short and instant message. On the one hand, people tweets about their daily lives, and on the other hand, when major events happen, people also follow and tweet about them. Moreover, people's posting behaviors on events are often closely tied to their personal interests. In this paper, we try to model topics, events and users on Twitter in a unified way. We propose a model which combines an LDA-like topic model and the Recurrent Chinese Restaurant Process to capture topics and events. We further propose a duration-based regularization component to find bursty events. We also propose to use event-topic affinity vectors to model the association between events and topics. Our experiments shows that our model can accurately identify meaningful events and the event-topic affinity vectors are effective for event recommendation and grouping events by topics.

1 Introduction

Twitter is arguably the most popular microblog site where people can post short, instant messages to share with families, friends and the rest of the world. For content analysis on Twitter, two important concepts have been repeatedly visited: (1) **Topics**. These are longstanding themes that many personal tweets revolve around. Example topics range from music and sports to more serious ones like politics and religion. Much work has been done to analyze topics on Twitter (Ramage et al., 2010; Hong and Davison, 2010; Zhao et al., 2011; Lau et al., 2012). (2) **Events**. These are things that take place at a certain time and attract many people's short-term attention in social media. Example events include concerts, sports games, scandals and elections. Event detection on Twitter has been a hot research topic in recent years (Petrović et al., 2010; Weng and Lee, 2011; Becker et al., 2011; Diao et al., 2012; Li et al., 2012).

The concepts of topics and events are orthogonal in that many events fall under certain topics. For example, concerts fall under the topic about music. Furthermore, being *social* media, Twitter users play important roles in forming topics and events on Twitter. Each user has her own topic interests, which influence the content of her tweets. Whether a user publishes a tweet related to an event also largely depends on whether her topic interests match the nature of the event. Modeling the interplay between topics, events and users can deepen our understanding of Twitter content and potentially aid many predication and recommendation tasks. In this paper, we aim to construct a unified model of topics, events and users on Twitter. Although there has been a number of recent studies on event detection on Twitter, to the best of our knowledge, ours is the first that links the topic interests of users to their tweeting behaviors on events.

Specifically, we propose a probabilistic latent variable model that identifies both topics and events on Twitter. To do so, we first separate tweets into *topic tweets* and *event tweets*. The former are related to a user's personal life, such as a tweet complaining about the traffic condition or wishing a friend

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1869–1879,
Seattle, Washington, USA, 18-21 October 2013.

happy birthday. The latter are about some major global event interesting to a large group of people, such as a tweet advertising a concert or commenting on an election result. Although considering only topic tweets and event tweets is a much simplified view of the diverse range of tweets, we find it effective in finding meaningful topics and events. We further use an LDA-like model (Blei et al., 2003) to discover topics and the Recurrent Chinese Restaurant Process (Ahmed and Xing,) to discover events. Details are given in Section 3.1.

Our major contributions lie in two novel modifications to the base model described above. The first is a duration-based regularization component that punishes long-term events (Section 3.2). Because events on Twitter tend to be bursty, this modification presumably can produce more meaningful events. More specifically, we borrow the idea of using pseudo-observed variables to regularize graphical models (Balasubramanyan and Cohen, 2013), and carefully design the pseudo-observed variable in our task to capture the burstiness of events. The second modification is adding event-topic affinity vectors inspired by PMF-based collaborative filtering (Salakhutdinov and Mnih, 2008) (Section 3.3). It uses the latent topics to explain users' preferences of events and subsequently infers the association between topics and events.

We use a real Twitter data set consisting of 500 users to evaluate our model (Section 4). We find that the model can discover meaningful topics and events. Comparison with our base model and with an existing model for event discovery on Twitter shows that the two modifications are both effective. The duration-based regularization helps find more meaningful events; the event-topic affinity vectors improve an event recommendation task and helps produce a meaningful organization of events by topics.

2 Related Work

Study of topics, events and users on Twitter is related to several branches of work. We review the most interesting and relevant work below.

Event detection on Twitter: There have been quite a few studies in this direction in recent years, including both online detection (Sakaki et al., 2010; Petrović et al., 2010; Weng and Lee, 2011; Becker et al., 2011; Li et al., 2012) and offline detection (Diao et al., 2012). Online detection is mostly concerned with early detection of major events, so efficiency of the algorithms is the main focus. These algorithms do not aim to identify *all* relevant tweets, nor do they analyze the association of events with topics. In comparison, our work focuses on modeling topics, events and users as well as their relation. Recently, Petrović et al. (2013) pointed out that Twitter stream does not lead news stream for major news events, but Twitter stream covers a much wider range of events than news stream. Our work helps better understand these additional events on Twitter and their relations with users' topic interests. Our model bears similarity to our earlier work (Diao et al., 2012), but we use a non-parametric model (RCRP) to discover events directly inside the probabilistic model.

Temporal topic modeling: A number of models have been proposed for the temporal aspect of topics (Blei and Lafferty, 2006; Wang and McCallum, 2006; Wang et al., 2007; Hong et al., 2011), but most of them fix the number of topics. The Recurrent Chinese Restaurant Process (Ahmed and Xing,) was proposed to model the life cycles of topics and allows an infinite number of topics. It has later been combined with LDA to model both topics and events in news streams and social media streams (Ahmed et al., 2011; Tang and Yang, 2012). Our work also jointly models topics and events, but different from previous work, we do not assume that every document (tweet in our case) belongs to an event, which is important because Twitter contains many personal posts unrelated to major events.

Collaborative filtering with LDA: Part of our model is inspired by work on collaborative filtering based on probabilistic matrix factorization (PMF) (Salakhutdinov and Mnih, 2008). Recently there has been some work combining LDA with PMF to recommend items with textual content such as news articles and advertisements (Wang and Blei, 2011; Agarwal and Chen, 2010). They use topics to interpret the latent structure of users and items. We borrow their idea but our items are events, which are not known and have to be discovered by our model.

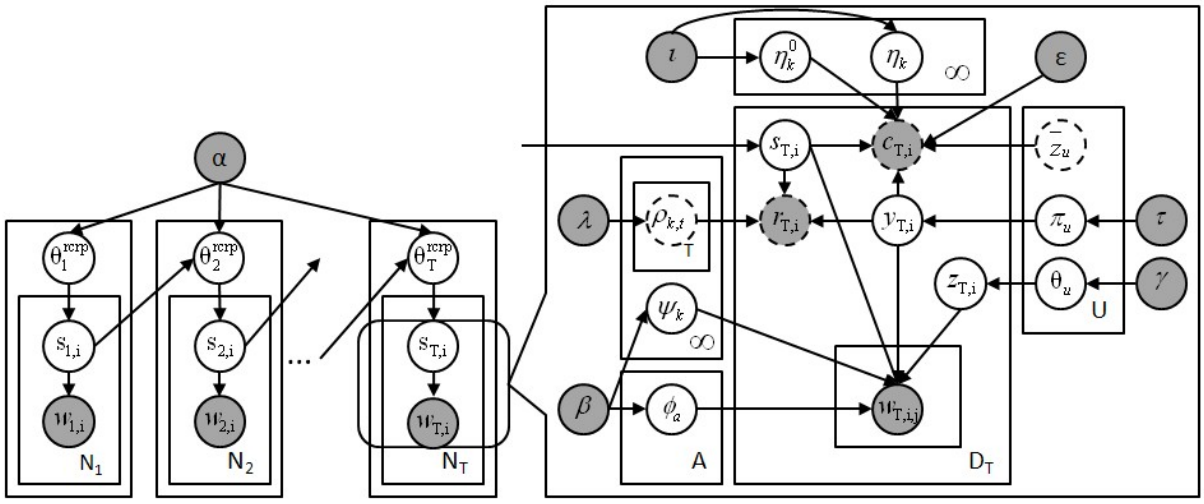

Figure 1: Plate notation for the whole model, in which pseudo-observed variables and distributions based on empirical counts are shown as dotted nodes.

3 Our Model

In this section, we present our model for topics, events and users on Twitter. We assume that we have a stream of tweets which are divided into T epoches. Let $t \in \{1, 2, \ldots, T\}$ be the index of an epoch. Each epoch contains a set of tweets and each tweet is a bag of words. We use $w_{t,i,j} \in \{1, 2, \ldots, V\}$ to denote the j-th word of the i-th tweet in the t-th epoch, where V is the vocabulary size. The author of the i-th tweet in the t-th epoch (i.e. the Twitter user who publishes the tweet) is denoted as $u_{t,i} \in \{1, 2, \ldots, U\}$, where U is the total number of Twitter users we consider.

We first present our base model in Section 3.1. We then introduce a duration-based regularization mechanism to ensure the burstiness of events in Section 3.2. In Section 3.3 we discuss how we model the relation between topics and events using event-topic affinity vectors. Finally we discuss model inference in Section 3.4.

3.1 The Base Model

Recall that our objective is to model topics, events, users and their relations. As in many topic models, our topic is a multinomial distribution over words, denoted as ϕ_a where a is a topic index. Each event is also a multinomial distribution over words, denoted as ψ_k where k is an event index. Because topics are long-standing and stable, we fix the number of topics to be A, where A can be tuned based on historical data. In contrast, events emerge and die along the timeline. We therefore use a non-parametric model called the Recurrent Chinese Restaurant Process (RCRP) (Ahmed and Xing,) to model the birth and death of events. To model the relation between users and topics, we assume each user u has a multinomial distribution over topics, denoted as θ_u.

As we have discussed, we separate tweets into two categories, topic tweets and event tweets. Separation of these two categories is done through a latent variable y sampled from a user-specific Bernoulli distribution π_u. For topic tweets, the topic is sampled from the corresponding user's topic distribution θ_u. For event tweets, the event is sampled according to RCRP. We now briefly review RCRP. Generally speaking, RCRP assumes a Chinese Restaurant Process (CRP) (Blackwell and MacQueen, 1973) for items within an epoch and chains up the CRPs in adjacent epochs along the timeline. Specifically, in our case, the generative process can be described as follows. Tweets come in according to their timestamps. In the t-th epoch, for the i-th tweet, we first flip a biased coin based on probability π_u to decide whether this tweet is event-related. If it is, then we need to decide which event it belongs to. It could be an existing event that has at least one related tweet in the

previous epoch or the current epoch, or it could be a new event. Let $n_{k,t-1}$ denote the number of tweets related to event k at the end of epoch $(t-1)$. Let $n_{k,t}^{(i)}$ denote the number of tweets related to event k in epoch t before the i-th tweet comes. Let N_{t-1} denote the total number of event-related tweets in epoch $(t-1)$ and $N_t^{(i)}$ denote the number of event-related tweets in epoch t before the i-th tweet. Then RCRP assumes that the probability for the i-th tweet to join event k is $\frac{n_{k,t-1}+n_{k,t}^{(i)}}{N_{t-1}+N_t^{(i)}+\alpha}$ and the probability to start a new event is $\frac{\alpha}{N_{t-1}+N_t^{(i)}+\alpha}$, where α is a parameter. As we can see, RCRP naturally captures the "rich-get-richer" phenomenon in social media.

Finally we place Dirichlet and Beta priors on the various parameters in our model. Formally, the generative process of our base model is outlined in Figure 2, excluding the lines in bold and blue. We also show the plate notation in Figure 1, in which the Recurrent Chinese Restaurant Process is represented as an infinite dynamic mixture model (Ahmed and Xing,) and θ_t^{rcrp} means the distribution on an infinite number of events in epoch t. D_t is the total number of tweets (both event-related and topic tweets), while N_t represents the number event-related tweets in epoch t.

3.2 Regularization on Event Durations

As we have pointed out, events on Twitter tend to be bursty, i.e. the duration of an event tends to be short, but this characteristic is not captured by RCRP. While there can be different ways to incorporate this intuition, here we adopt the idea of regularization using pseudo-observed variables proposed recently by Balasubramanyan and Cohen (2013). We introduce a pseudo-observed binary variable $r_{t,i}$ for each tweet, where the value of $r_{t,i}$ is set to 1 for all tweets. We assume that this variable is dependent on the hidden variables y and s. Specifically, if $y_{t,i}$ is 0, i.e. the tweet is topic-related, then $r_{t,i}$ gets a value of 1 with probability 1. If $y_{t,i}$ is 1, then we look at all the tweets that belong to event $s_{t,i}$. Our goal is to make sure that this tweet is temporally close to these other tweets. So we assume that $r_{t,i}$ gets a value of 1 with probability $\exp(-\sum_{t'=1,|t'-t|>1}^{T}\lambda|t-t'|n_{s_{t,i},t'})$, where $n_{s_{t,i},t'}$ is the number of tweets in epoch t' that belong to event $s_{t,i}$ and $\lambda > 0$ is a parameter. We can see that when we factor in the generation of these pseud-observed variables r, we penalize long-term events and favor events whose tweets are concentrated along the timeline. Generation of these variables r is shown in bold and blue in Figure 2.

- For each topic $a = 1, \ldots, A$
  - draw $\phi_a \sim \text{Dirichlet}(\beta)$
- For each user $u = 1, \ldots, U$
  - draw $\theta_u \sim \text{Dirichlet}(\gamma), \pi_u \sim \text{Beta}(\tau)$
- For each epoch t and tweet i
  - draw $y_{t,i} \sim \text{Bernoulli}(\pi_{u_{t,i}})$
  - If $y_{t,i} = 0$
    - draw $z_{t,i} \sim \text{Multinomial}(\theta_{u_{t,i}})$
    - For each j, draw $w_{t,i,j} \sim \text{Multinomial}(\phi_{z_{t,i}})$
  - If $y_{t,i} = 1$
    - draw $s_{t,i}$ from RCRP
    - If $s_{t,i}$ is a new event
      - draw $\psi_{s_{t,i}} \sim \text{Dirichlet}(\beta)$
      - **draw $\eta^0_{s_{t,i}} \sim$ Gaussian$(0, \iota^{-1})$**
      - **draw $\boldsymbol{\eta}_{s_{t,i}} \sim$ Gaussian$(0, \iota^{-1} I_A)$**
    - **draw $r_{t,i} \sim$ Bernoulli$(\rho_{s_{t,i},t})$, where $\rho_{s_{t,i},t} = \exp(-\sum_{t'=1,|t'-t|>1}^{T}\lambda|t'-t|n_{s_{t,i},t'})$**
    - **draw $c_{t,i} \sim$ Gaussian$(\eta^0_{s_{t,i}} + \boldsymbol{\eta}^T_{s_{t,i}} \cdot \bar{\boldsymbol{z}}_{u_{t,i}}, \epsilon^{-1})$**
    - For each j, draw $w_{t,i,j} \sim \text{Multinomial}(\psi_{s_{t,i}})$

Figure 2: The generative process of our model, in which the duration-based regularization (section 3.2) and the event-topic affinity vector (section 3.3) are in blue and bold lines.

3.3 Event-Topic Affinity Vectors

So far in our model topics and events are not related. However, many events are highly related to certain topics. For example, a concert is related to music while a football match is related to sports. We would like to capture these relations between topics and events. One way to do it is to assume that event tweets also have topical words sampled from the event's topic distribution, something similar to the models by Ahmed et al. (2011) and by Tang and Yang (2012). However, our prelimiary experiments show that this idea does not work well on Twitter, mainly because tweets are too short. Here we explore another approach inspired by recommendation methods based on probabilistic matrix factorization (Salakhutdinov and Mnih, 2008). The idea is that when a user posts a tweet about an event, we can treat the event as an item and this posting be-

havior as adoption of the item. If we assume that the adoption behavior is influenced by some latent factors, i.e. the latent topics, then basically we would like the topic distribution of this user to be close to that of the event.

Specifically, we assume that each event k has associated with it an A-dimensional vector $\boldsymbol{\eta}_k$ and a parameter η_k^0. The vector $\boldsymbol{\eta}_k$ represents the event's affinity to topics. η_k^0 is a bias term that represents the inner popularity of an event regardless of its affinity to any topic. We further assume that each tweet has another pseudo-observed variable $c_{t,i}$ that is set to 1. For topic tweets, $c_{t,i}$ gets a value of 1 with probability 1. For event tweets, $c_{t,i}$ is generated by a Gaussian distribution with mean equal to $\eta^0_{s_{t,i}} + \boldsymbol{\eta}_{s_{t,i}} \cdot \bar{\boldsymbol{z}}_{u_{t,i}}$, where $\bar{\boldsymbol{z}}_u$ is an A-dimensional vector denoting the empirical topic distribution of user u's tweets. This treatment follows the practice of fLDA by Agarwal and Chen (2010). Let $\bar{C}_{u,a}$ be the number of tweets by user u assigned to topic a, based on the values of the latent variables y and z. Then

$$\bar{\boldsymbol{z}}_{u,a} = \frac{\bar{C}_{u,a}}{\sum_{a'=1}^{A} \bar{C}_{u,a'}},$$
$$c_{t,i} \sim \text{Gaussian}(\eta^0_{s_{t,i}} + \boldsymbol{\eta}_{s_{t,i}} \cdot \bar{\boldsymbol{z}}_{u_{t,i}}, \epsilon^{-1}),$$

where ϵ is a parameter. We generate $\boldsymbol{\eta}_k$ and η_k^0 using Gaussian priors once event k emerges. The generation of the variables c is shown in bold and blue in Figure 2.

3.4 Inference

We train the model using a stochastic EM sampling scheme. In this scheme, we alternate between Gibbs sampling and gradient descent. In the Gibbs sampling part, we fix the values of η_k^0 and $\boldsymbol{\eta_k}$ for each event k, and then we sample the latent variables $y_{t,i}$,$z_{t,i}$ and $s_{t,i}$ for each tweet. In the gradient descent part, we update the event-topic affinity vectors $\boldsymbol{\eta_k}$ and the bias term η_k^0 of each event k by keeping the assignment of the variables $y_{t,i}$,$z_{t,i}$ and $s_{t,i}$ fixed.

For the Gibbs sampling part, we jointly sample $y_{t,i} = 0, z_{t,i} = a$ (topic tweet) and $y_{t,i} = 1, s_{t,i} = k$ (event tweet) as follows:

Topic tweet:

$$p(y_{t,i} = 0, z_{t,i} = a | \mathbf{y}_{\neg t,i}, \mathbf{z}_{\neg t,i}, \mathbf{w}, \mathbf{r}, \mathbf{c}, u_{t,i})$$
$$\propto \frac{n^{(\pi)}_{u,0} + \tau}{n^{(\pi)}_{u,(.)} + 2\tau} \frac{n^{(\theta)}_{u,a} + \gamma}{n^{(\theta)}_{u,(.)} + A\gamma} \frac{\prod_{v=1}^{V} \prod_{i=0}^{E_{(v)}-1} (n^{(\phi)}_{a,v} + i + \beta)}{\prod_{i=0}^{E_{(.)}-1} (n^{(\phi)}_{a,(.)} + i + V\beta)}$$
$$\prod_{t',i' \in I_u} \frac{\mathcal{N}(c_{t',i'} | \eta^0_{s_{t',i'}} + \boldsymbol{\eta}_{s_{t',i'}} \cdot \bar{\boldsymbol{z}}^*_u, \epsilon^{-1})}{\mathcal{N}(c_{t',i'} | \eta^0_{s_{t',i'}} + \boldsymbol{\eta}_{s_{t',i'}} \cdot \bar{\boldsymbol{z}}_u, \epsilon^{-1})}$$

Event tweet:

$$p(y_{t,i} = 1, s_{t,i} = k | \mathbf{y}_{\neg t,i}, \mathbf{z}_{\neg t,i}, \mathbf{w}, \mathbf{r}, \mathbf{c}, u_{t,i})$$
$$\propto \frac{n^{(\pi)}_{u,1} + \tau}{n^{(\pi)}_{u,(.)} + 2\tau} \frac{1}{\boldsymbol{N}} \Bigg(n^{\text{RCRP}}_{k,t} \mathcal{N}(c_{t,i} | \eta^0_{s_{t,i}} + \boldsymbol{\eta}_{s_{t,i}} \cdot \bar{\boldsymbol{z}}_u, \epsilon^{-1})$$
$$\cdot \exp(- \sum_{\substack{t'=1 \\ |t'-t|>1}}^{T} \lambda |t - t'| n_{k,t'}) \Bigg) \frac{\prod_{v=1}^{V} \prod_{i=0}^{E_{(v)}-1} (n^{(\psi)}_{k,v} + i + \beta)}{\prod_{i=0}^{E_{(.)}-1} (n^{(\psi)}_{k,(.)} + i + V\beta)}$$

in which,

$$n^{\text{RCRP}}_{k,t} = \begin{cases} (n_{k,t-1} + n_{k,t}) & \\ \quad \cdot \frac{n_{k,t} + n_{k,t+1}}{n_{k,t}} & \text{if } n_{k,t-1} > 0, n_{k,t} > 0, \\ n_{k,t-1} & \text{if } n_{k,t-1} > 0, n_{k,t} = 0, \\ n_{k,t+1} & \text{if } n_{k,t+1} > 0, n_{k,t} = 0, \\ \alpha & \text{if } k \text{ is a new event,} \end{cases}$$

where we use u to represent $u_{t,i}$. $n^{(\pi)}_{u,0}$ is the number of topic tweets by user u while $n^{(\pi)}_{u,1}$ is the number of event tweets by user u. They stem from integrating out the user's Bernoulli distribution π_u. $n^{(\pi)}_{u,(.)}$ is the total number of tweets by user u. Similarly, $n^{(\theta)}_{u,a}$ is the number of tweets assigned to topic a for this user, resulting from integrating out the user's topic distribution θ_u. $n^{(\theta)}_{u,(.)}$ is the same as $n^{(\pi)}_{u,0}$. $E_{(v)}$ is the number of times word type v appears in the current tweet, and $E_{(.)}$ is the total number of words in the current tweet. $n^{(\phi)}_{a,v}$ is the number of times word type v is assigned to topic a, and $n^{(\phi)}_{a,(.)}$ is the number of words assigned to topic a. $n^{(\psi)}_{k,v}$ is the number of times word type v is assigned to event k, and $n^{(\psi)}_{k,(.)}$ is the total number of words assigned to event k. These word counters stem form integrating out each event's word distribution and are set to zero when k is a new event. $I_u = \{t', i' | y_{t',i'} = 1, u_{t',i'} = u\}$, which is the set of event tweets published by user u, and u represents $u_{t,i}$ for short. $\bar{\boldsymbol{z}}^*_u$ is the empirical

counting vector which considers the current tweet's topic assignment, while $\bar{\boldsymbol{z}}_u$ and all other counters do not consider the current tweet. Finally, $\boldsymbol{N}$ is a local normalization factor for event tweets, which includes the RCRP, event-topic affinity and regularization on event duration.

With the previous Gibbs sampling step, we can get the assignment of variables $y_{t,i}$,$z_{t,i}$ and $s_{t,i}$. Given the assignment, we use gradient descent to update the values of the bias term η_k^0 and the event-topic affinity vectors $\boldsymbol{\eta}_k$ for each current existing event k. First, we can get the logarithm of the posterior distribution:

$$\begin{aligned}&\ln P(\mathbf{y}, \mathbf{z}, \mathbf{s}, \mathbf{r}, \mathbf{c}|\mathbf{w}, \mathbf{u}, \text{all priors})\\ &= \text{constant} - \sum_{k=1}^{\infty}\{\frac{\iota}{2}({\eta_k^0}^2 + \boldsymbol{\eta}_k \cdot \boldsymbol{\eta}_k)\\ &+ \sum_{u=1}^{U} n_{u,k}\frac{\epsilon}{2}[1 - (\eta_k^0 + \boldsymbol{\eta}_k \cdot \bar{\boldsymbol{z}}_u)]^2\},\end{aligned}$$

where $n_{u,k}$ is the number of event tweets about event k published by user u. The derivative of the logarithm of the posterior distribution with respect to the bias term η_k^0 and the event-topic affinity vector $\boldsymbol{\eta}_k$ are as follows:

$$\begin{aligned}\frac{\partial \ln P}{\partial \eta_k^0} &= -\iota\eta_k^0 + \sum_{u=1}^{U}\epsilon n_{u,k}[1 - (\eta_k^0 + \boldsymbol{\eta}_k \cdot \bar{\boldsymbol{z}}_u)],\\ \frac{\partial \ln P}{\partial \boldsymbol{\eta}_k} &= -\iota\boldsymbol{\eta}_k + \sum_{u=1}^{U}\epsilon n_{u,k}[1 - (\eta_k^0 + \boldsymbol{\eta}_k \cdot \bar{\boldsymbol{z}}_u)]\bar{\boldsymbol{z}}_u.\end{aligned}$$

4 Experiment

4.1 Dataset and Experiment Setup

We evaluate our model on a Twitter dataset that contains 500 users. These users are randomly selected from a much larger pool of around 150K users based in Singapore. Selecting users from the same country/city ensures that we find coherent and meaningful topics and events. We use tweets published between April 1 and June 30, 2012 for our experiments. For preprocessing, we use the CMU Twitter POS Tagger[1] to tag these tweets and remove those non-standard words (i.e. words tagged as punctuation marks, emoticons, urls, at-mentions, pronouns, etc.) and stop words. We also remove tweets with less than three words. After preprocessing, the dataset contains 655,881 tweets in total.

Recall that our model is designed to identify topics, events and their relations with users. We therefore would like to evaluate the quality of the identified topics and events as well as the usefulness of the discovered topic distributions of users and event-topic affinity vectors. Because our topic discovery mechanism is fairly standard and a quick inspection shows that the discovered topics are generally meaningful and comparable to those discovered by standard LDA, here we do not focus on evaluation of topics. In Section 4.2 we evaluate the quality of the discovered events. In Section 4.3 we show how the discovered event-topic affinity vectors can be useful.

For comparison, we consider an existing method called **TimeUserLDA** introduced in our previous work (Diao et al., 2012). TimeUserLDA also models topics and events by separating topic tweets from event tweets. However, it groups event tweets into a *fixed* number of *bursty topics* and then uses a two-state machine in a postprocessing step to identify events from these bursty topics. Thus, events are not directly modeled within the generative process itself. In contrast, events are inherent in our generative model. We do not compare with other event detection methods because our objective is not online event detection.

We also compare our final model with two degenerate versions of it. We refer to the base model described in Section 3.1 as **Base** and the model with the duration-based regularization as **Base+Reg**. Comparison with these two degenerate models allows us to assess the effect of the two modifications we propose. We refer to the final model with both the duration-based regularization and the event-topic affinity vectors as **Base+Reg+Aff**.

For the parameter setting, we empirically set A to 40, γ to $\frac{50}{A}$, τ to 1, β to 0.01, α to 1, ι to 10, ϵ to 1, and the duration regularization parameter λ to 0.01. When a new event k is created, the inner popularity bias term η_k^0 is set to 1, and the factors in event-topic affinity vectors $\boldsymbol{\eta}_k$ are all set to 0. We run the stochastic EM sampling scheme for 300 iterations. After Gibbs sampling assigns each variable a value at the end of each iteration, we update the values of η_k^0 and $\boldsymbol{\eta}_k$ for the existing events using gradient descent.

[1] http://www.ark.cs.cmu.edu/TweetNLP/

Event	Top words	Duration	Inner popularity (η_k^0)
debate caused by Manda Swaggie	singapore, bieber, europe, amanda, justin, trending, manda, hates, swaggie, hate	17 June - 19 June	0.9457
Indonesia tsunami	tsunami, earthquake, indonesia, singapore, hit, warning, aceh, 8.9, safe, magnitude	10 April - 12 April	0.9439
SJ encore concert	#ss4encore, cr, #ss4encoreday2, hyuk, 120526, super, leader, changmin, fans, teuk	26 May - 28 May	0.8360
Mother's Day	day, happy, mother's, mothers, love, mom, mum, everyday, mother, moms	11 May - 14 May	0.9370
April Fools' Day	april, fools, day, fool, joke, prank, happy, today, trans, fool's	1 April - 3 April	0.9322

Table 1: The top-5 events identified by Base+Reg+Aff. We show the story name which is manually labeled, top ten ranking words, lasting duration and the inner popularity (η_k^0) for each event.

4.2 Events

First we quantitatively evaluate the quality of the detected events. Our model finds clusters of tweets that represent events. We first assess whether these events are meaningful. We then judge whether the detected event tweets are indeed related to the corresponding event.

Quality of Top Events

Method	P@5	P@10	P@20	P@30
Base+Reg+Aff	**1.000**	**1.000**	**0.950**	**0.900**
Base+Reg	**1.000**	**1.000**	**0.950**	0.867
Base	0.000	0.200	0.250	0.367
TimeUserLDA	**1.000**	0.800	0.750	0.600

Table 2: Precision@K for the various methods.

Usually we are interested in the most popular events on Twitter. We therefore assess whether the top events are meaningful. For each method, we rank the detected events based on the number of tweets assigned to them and then pick the top-30 events for each method. We randomly mix these events and ask two human judges to label them. The judges are given 100 randomly selected tweets for each event (or all tweets if an event contains less than 100 tweets). The judges can use external sources to help them. If an event is meaningful based on the 100 sample tweets, a score of 1 is given. Otherwise it is scored 0. The inter-annotator agreement score is 0.744 using Cohen's kappa, showing substantial agreement. Finally we treat an event as meaningful if both judges have scored it 1.

Table 2 shows the performance in terms of precision@K, and Table 1 shows the top 5 events of our model (i.e., Base+Reg+Aff). We have the following findings from the results: (1) Our base model performs quite poorly for the top events while Base+Reg and Base+Reg+Aff perform much better. This shows that the duration-based regularization is critical in finding meaningful events. A close examination shows that the base model clusters many general topic tweets as events, such as tweets about transportation and music and even foursquare tweets. (2) TimeUserLDA performs well for the very top events (P@5 and P@10) but its performance drops for lower-ranked events (P@20 and P@30), similar to what was reported by Diao et al. (2012). A close examination shows that this method is good at finding major events that do not have strong topic association and thus attract most people's attention, e.g. earthquakes, but not good at finding topic-oriented events such as some concerts and sports games. This is because this method mixes topics and events first and only detects events from bursty topics in a second stage of postprocessing. In contrast, our model performs well for topic-oriented events. (3) The difference between Base+Reg and Base+Reg+Aff is small, suggesting that the event-topic affinity vectors are not crucial for event detection.

Precision of Event Tweets

Next, we evaluate the relevance of the detected event tweets to each event. To make a fair comparison, we select only the common events identified by all the methods. We pick 3 out of 5 common events shared by all methods within top-30 events

Event	TimeUserLDA	Base	Base+Reg	Base+Reg+Aff
Father's Day	0.61	0.63	0.71	**0.72**
debate caused by Manda Swaggie	0.73	0.74	**0.84**	0.80
Indonesia tsunami	0.75	0.75	**0.82**	0.80
Super Junior album release	N/A	0.72	0.78	**0.81**

Table 3: Precision of the event tweets for the 4 common events.

(we pick "Fathers' day" to represent public festivals, and ignore the similar events "Mothers' day" and "April fools"). We also pick one event shared by three RCRP based models. We further ask one of the judges to score the 100 tweets as either 1 or 0 based on their relevance to the event. The precision of the 100 tweets for each event and each method is shown in Table 3. We can see that again Base+Ref and Base+Ref+Aff perform similarly, and both outperform the other two methods. We also take a close look at the tweets and find that the false positives mislabeled by Base is mainly due to the long-duration of the discovered events. For example, for the event "Super Junior album release," Base finds other music-related tweets surrounding the peak period of the event itself.

In summary, our evaluation on event quality shows that (1) Using the non-parametric RCRP model to identify events within the generative model itself is advantageous over TimeUserLDA, which identifies events by postprocessing. (2) The duration-based regularization is crucial for finding more meaningful events.

4.3 Event-Topic Association

Besides event identification, our model also finds the association between events and topics through the event-topic affinity vectors. The discovered event-topic association can potentially be used for various tasks. Here we conduct two experiments to demonstrate its usefulness.

Event Recommendation

Recall that to discover event-topic association, we treat an event as an item and a tweet about the event as indication of the user's adoption of the item. Following this analogy with item recommendation, we define an event recommendation task where the goal is to recommend an event to users who have not posted any tweet about the event but may potentially be interested in the event. Intuitively, if a user's topic distribution is similar to the event-topic affinity vector of the event, then the user is likely to be interested in the event.

Specifically, we use the first two months' data (April and May 2012) as training data to learn all the users' topic distributions. We then use a ransom subset of 250 training users and their tweets in June to identify events in June as well as the event-topic affinity vectors of these events. We pick 8 meaningful events that are ranked high by all methods for testing. For each event, we try to find among the remaining 250 users those who may be interested in the event and compare the results with ground truth obtained by human judgment. Because it is time consuming to obtain the ground truth for all 250 users, we randomly pick 100 of these 250 users for testing purpose. For each test user and each event, we manually inspect the user's tweets around the peak days of the event to judge whether she has commented on the event. This is used as ground truth.

With our complete model Base+Reg+Aff, we can simply rank the 100 test users in decreasing order of $\boldsymbol{\eta}_k \cdot \bar{\boldsymbol{z}}_u$. For the other methods, because we do not have any parameter that directly encodes event-topic association, we cannot rank users based on how similar their topic distributions are to the event's affinity to topics. We instead adopt a collaborative filtering strategy and rank the test users by their similarity with those training users who have tweeted about the event. Specifically, each of these methods produces a topic distribution θ_u for each user. In addition, for each test event these methods identify a list of training users who have tweeted on it. By taking the average topic distribution of these training users and compute its cosine similarity with a test user's topic distribution, we can rank the 100 test users.

Since we have turned the recommendation task into a ranking task, we use Average Precision, a commonly used metric in information retrieval, to compare the performance. Average Precision is the

Event	TimeUserLDA	Base	Base+Reg	Base+Reg+Aff	Inner popularity (η_k^0)
debate caused by Manda Swaggie	0.3533	0.3230	**0.3622**	0.2956	0.943
Father's Day	0.3811	0.3525	0.3596	**0.4362**	0.917
Big Bang album release	0.1406	0.1854	0.1533	**0.1902**	0.893
City Harvest Church scandal	N/A	0.2832	0.1874	**0.3347**	0.890
Alex Ong pushing an old lady	N/A	**0.1540**	0.1539	0.1113	0.876
final episode of Super Spontan (reality show)	N/A	0.0177	0.0331	**0.2900**	0.862
Super Junior album release	N/A	0.0398	0.0330	**0.5900**	0.792
LionsXII 9-0 Sabah FA (soccer)	0.0711	0.1207	0.2385	**0.3220**	0.773
MAP	N/A	0.1845	0.1901	**0.3213**	

Table 4: For the 8 test events that happened in June 2012, we compute the Average Precision for each event. We also show the Mean Average Precision (MAP) when applicable.

Topic	Top words of the topic	Related event	Top words of the event
Food	eat, food, eating, ice, hungry, dinner, cream, lunch, chicken, buy	Ben&Jerry free cone day	free, cone, day, ben, jerry's, today, b&j, zoo, #freeconeday, singapore
Korean Music	music, big, cr, super, bang, junior, love, concert, bank, album	Super Junior encore concert	#ss4encore, cr, #ss4encoreday2, hyuk, 120526, super, leader, changmin, fans, teuk
		Super Junior Shanghai concert	#ss4shanghai, cr, 120414, donghae, eunhyuk, giraffe, solo, hyuk, ryeowook, shanghai
		Super Junior Paris concert	#ss4paris, cr, paris, super, 120406, ss4, junior, siwon, show, update
Malay	aku, nak, tak, kau, ni, lah, tk, je, mcm, nk	final episode of Super Spontan	zizan, johan, friendship, jozan, #superspontan, skips, forever, real, juara, gonna
		LionsXII 9-0 Sabah FA	sabah, 9-0, #lionsxii, lions, singapore, 7-0, amet, sucks, sabar, goal
Soccer	win, game, man, chelsea, match, city, goal, good, united, team	Man City crowned English champions	man, city, united, qpr, fuck, bored, lah, love, glory, update

Table 5: Example topics and their corresponding correlated events.

average of the precision value obtained for the set of top items existing after each relevant item is retrieved (Manning et al., 2008). We also rank the 8 events in decreasing order of their inner popularity η_k^0 learned by our complete model. The results are shown in Table 4. We have the following findings from the table. (1) Our complete method outperforms the other methods for 6 out of the 8 test events, suggesting that with the inferred event-topic affinity vectors we can do better event recommendation. (2) The improvement brought by the event-topic affinity vectors, as reflected in the difference in Average Precision between Base+Reg+Aff and Base (or Base+Reg) is more pronounced for events with lower inner popularity. Recall that the inner popularity of an event shows the inherent popularity of an event regardless of its association with any topic, that is, an event with high inner popularity attracts attention of many people regardless of their topic interests, while an event with low inner popularity tends to attract attention of certain people with similar topic interests. The finding above suggests that the event-topic affinity vectors are especially useful for recommending events that attract only certain people's attention, such as those related to sports, music, etc.

One may wonder for the events with low inner popularity why we could not achieve the same effect by Base or Base+Reg where we consider the topic similarity of test users with training users who have tweeted about the event. Our close examination shows that for these events although Base and Base+Reg may identify relevant event tweets with decent precision, the users they identify who have

tweeted about the event may not share similar topic interests. As a result, when we average these users' topic interests, we cannot obtain a clear skewed topic distribution that explains the event's affinity to different topics. In contrast, Base+Reg+Aff explicitly models the event-topic affinity vector and prefers to assign a tweet to an event if its author's topic distribution is similar to the event's affinity vector. Through the training iterations, the users who have tweeted about an event as identified by Base+Reg+Aff will gradually converge to share similar topic distributions.

Grouping Events by Topics

Finally, we show that the event-topic affinity vectors can also be used to group events by topics. This can potentially be used to better organize and present popular events in social media. In Table 5 we show a few highly related events for a few popular topics in our Twitter data set. Specifically given a topic a we rank the meaningful events that contain at least 70 tweets based on $\eta_{k,a}$. We can see from the table that the events are indeed related to the corresponding topic. The event "LionsXII 9-0 Sabah FA" is particularly interesting in that it is highly related to both the topic on Malay and the topic on soccer. (LionsXII is a soccer team from Singapore and Sabah FA is a soccer team from Malaysia.)

5 Conclusion

In this paper, we propose a unified model to study topics, events and users jointly. The base of our method is a combination of an LDA-like model and the Recurrent Chinese Restaurant Process, which aims to model users' longstanding personal topic interests and events over time simultaneously. The Recurrent Chinese Restaurant Process is appealing in the sense that it provides a principled dynamic nonparametric model in which the number of events is not fixed overtime. We further use a time duration-based regularization to capture the fast emergence and disappearance of events on Twitter, which is effective to produce more meaningful events. Finally, we use an inner popularity bias parameter and event-topic affinity vectors to interpret an event's inherent popularity and its affinity to different topics. Our experiments quantitatively show that our proposed model can effectively identify meaningful events and accurately find relevant tweets for these events. Furthermore, the event-topic association inferred by our model can help an event recommendation task and organize events by topics.

6 Acknowledgments

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office. We thank the reviewers for their valuable comments.

References

Deepak Agarwal and Bee-Chung Chen. 2010. fLDA: matrix factorization through latent Dirichlet allocation. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 91–100.

Amr Ahmed and Eric P. Xing. Dynamic non-parametric mixture models and the recurrent Chinese restaurant process: with applications to evolutionary clustering. In *Proceedings of the SIAM International Conference on Data Mining, SDM 2008*.

Amr Ahmed, Qirong Ho, Jacob Eisenstein, Eric Xing, Alexander J. Smola, and Choon Hui Teo. 2011. Unified analysis of streaming news. In *In Proceedings of the 20th international conference on World wide web*, pages 267 – 276.

Ramnath Balasubramanyan and William W. Cohen. 2013. Regularization of latent variable models to obtain sparsity. In *Proceedings of SIAM Conference on Data Mining*.

Hila Becker, Mor Naaman, and Luis Gravano. 2011. Beyond trending topics: Real-world event identification on twitter. In *Fifth International AAAI Conference on Weblogs and Social Media*.

D. Blackwell and J. MacQueen. 1973. Ferguson distributions via Polya urn schemes. *The Annals of Statistics*, pages 353–355.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113 – 120.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Qiming Diao, Jing Jiang, Feida Zhu, and Ee-Peng Lim. 2012. Finding bursty topics from microblogs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 536 – 544.

Liangjie Hong and Brian D. Davison. 2010. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88.

Liangjie Hong, Byron Dom, Siva Gurumurthy, and Kostas Tsioutsiouliklis. 2011. A time-dependent topic model for multiple text streams. In *Proceedings of SIGKDD*.

Jey Han Lau, Nigel Collier, and Timothy Baldwin. 2012. On-line trend analysis with topic models: #twitter trends detection topic model online. In *Proceedings of the 24th International Conference of on Computational Linguistics*, pages 1519–1534.

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

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze, 2008. *Introduction to Information Retrieval*, chapter Evaluation in information retrieval. Cambridge University Press.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to Twitter. In *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 181 – 189.

Saša Petrović, Miles Osborne, Richard McCreadie, Richard Macdonald, Richard Ounis, and Luke Shrimpton. 2013. Can twitter replace newswire for breaking news? In *Proceedings of the 7th International Conference on Weblogs and Social Media*.

Daniel Ramage, Susan T. Dumais, and Daniel J. Liebling. 2010. Characterizing microblogs with topic models. In *Proceedings of the 4th International Conference on Weblogs and Social Media*.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, pages 851–860.

Ruslan Salakhutdinov and Andriy Mnih. 2008. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, volume 20.

Xuning Tang and Christopher C. Yang. 2012. TUT: a statistical model for detecting trends, topics and user interests in social media. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 972 – 981.

Chong Wang and David M. Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 448 – 456.

Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.*, pages 424 – 433.

Xuanhui Wang, ChengXiang Zhai, Xiao Hu, and Richard Sproat. 2007. Proceedings of the 13th acm sigkdd international conference on knowledge discovery and data mining. In *Proceedings of SIGKDD*, pages 784 – 793.

Jianshu Weng and Bu-Sung Lee. 2011. Event detection in Twitter. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing Twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on IR Research*, pages 338–349.