7-2012

# Finding Bursty Topics From Microblogs

Qiming Diao
*Singapore Management University*, qiming.diao.2010@smu.edu.sg

Jing JIANG
*Singapore Management University*, jingjiang@smu.edu.sg

Feida ZHU
*Singapore Management University*, fdzhu@smu.edu.sg

Ee Peng LIM
*Singapore Management University*, eplim@smu.edu.sg

## Citation

Diao, Qiming; JIANG, Jing; ZHU, Feida; and LIM, Ee Peng, "Finding Bursty Topics From Microblogs" (2012). *Research Collection School of Information Systems (Open Access)*. Paper 1547.
http://ink.library.smu.edu.sg/sis_research/1547

# Finding Bursty Topics from Microblogs

**Qiming Diao, Jing Jiang, Feida Zhu, Ee-Peng Lim**
Living Analytics Research Centre
School of Information Systems
Singapore Management University
{qiming.diao.2010, jingjiang, fdzhu, eplim}@smu.edu.sg

## Abstract

Microblogs such as Twitter reflect the general public's reactions to major events. Bursty topics from microblogs reveal what events have attracted the most online attention. Although bursty event detection from text streams has been studied before, previous work may not be suitable for microblogs because compared with other text streams such as news articles and scientific publications, microblog posts are particularly diverse and noisy. To find topics that have bursty patterns on microblogs, we propose a topic model that simultaneously captures two observations: (1) posts published around the same time are more likely to have the same topic, and (2) posts published by the same user are more likely to have the same topic. The former helps find event-driven posts while the latter helps identify and filter out "personal" posts. Our experiments on a large Twitter dataset show that there are more meaningful and unique bursty topics in the top-ranked results returned by our model than an LDA baseline and two degenerate variations of our model. We also show some case studies that demonstrate the importance of considering both the temporal information and users' personal interests for bursty topic detection from microblogs.

## 1 Introduction

With the fast growth of Web 2.0, a vast amount of user-generated content has accumulated on the social Web. In particular, microblogging sites such as Twitter allow users to easily publish short instant posts about any topic to be shared with the general public. The textual content coupled with the temporal patterns of these microblog posts provides important insight into the general public's interest. A sudden increase of topically similar posts usually indicates a burst of interest in some event that has happened offline (such as a product launch or a natural disaster) or online (such as the spread of a viral video). Finding bursty topics from microblogs therefore can help us identify the most popular events that have drawn the public's attention. In this paper, we study the problem of finding bursty topics from a stream of microblog posts generated by different users. We focus on retrospective detection, where the text stream within a certain period is analyzed in its entirety.

Retrospective bursty event detection from text streams is not new (Kleinberg, 2002; Fung et al., 2005; Wang et al., 2007), but finding bursty topics from microblog steams has not been well studied. In his seminal work, Kleinberg (2002) proposed a state machine to model the arrival times of documents in a stream in order to identify bursts. This model has been widely used. However, this model assumes that documents in the stream are all about a given topic. In contrast, discovering interesting topics that have drawn bursts of interest from a stream of topically diverse microblog posts is itself a challenge. To discover topics, we can certainly apply standard topic models such as LDA (Blei et al., 2003), but with standard LDA temporal information is lost during topic discovery. For microblogs, where posts are short and often event-driven, temporal information can sometimes be critical in determining the topic of a post. For example, typically a post containing the

*Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 536–544,
Jeju, Republic of Korea, 8-14 July 2012. ©2012 Association for Computational Linguistics

word “jobs” is likely to be about employment, but right after October 5, 2011, a post containing “jobs” is more likely to be related to Steve Jobs’ death. Essentially, we expect that on microblogs, posts published around the same time have a higher probability to belong to the same topic.

To capture this intuition, one solution is to assume that posts published within the same short time window follow the same topic distribution. Wang et al. (2007) proposed a PLSA-based topic model that exploits this idea to find correlated bursty patterns across multiple text streams. However, their model is not immediately applicable for our problem. First, their model assumes multiple text streams where word distributions for the same topic are different on different streams. More importantly, their model was applied to news articles and scientific publications, where most documents follow the global topical trends. On microblogs, besides talking about global popular events, users also often talk about their daily lives and personal interests. In order to detect global bursty events from microblog posts, it is important to filter out these “personal” posts.

In this paper, we propose a topic model designed for finding bursty topics from microblogs. Our model is based on the following two assumptions: (1) If a post is about a global event, it is likely to follow a global topic distribution that is time-dependent. (2) If a post is about a personal topic, it is likely to follow a personal topic distribution that is more or less stable over time. Separation of “global” and “personal” posts is done in an unsupervised manner through hidden variables. Finally, we apply a state machine to detect bursts from the discovered topics.

We evaluate our model on a large Twitter dataset. We find that compared with bursty topics discovered by standard LDA and by two degenerate variations of our model, bursty topics discovered by our model are more accurate and less redundant within the top-ranked results. We also use some example bursty topics to explain the advantages of our model.

## 2 Related Work

To find bursty patterns from data streams, Kleinberg (2002) proposed a state machine to model the arrival times of documents in a stream. Different states generate time gaps according to exponential density functions with different expected values, and bursty intervals can be discovered from the underlying state sequence. A similar approach by Ihler et al. (2006) models a sequence of count data using Poisson distributions. To apply these methods to find bursty topics, the data stream used must represent a single topic.

Fung et al. (2005) proposed a method that identifies both topics and bursts from document streams. The method first finds individual words that have bursty patterns. It then finds groups of words that tend to share bursty periods and co-occur in the same documents to form topics. Weng and Lee (2011) proposed a similar method that first characterizes the temporal patterns of individual words using wavelets and then groups words into topics. A major problem with these methods is that the word clustering step can be expensive when the number of bursty words is large. We find that the method by Fung et al. (2005) cannot be applied to our dataset because their word clustering algorithm does not scale up. Weng and Lee (2011) applied word clustering to only the top bursty words within a single day, and subsequently their topics mostly consist of two or three words. In contrast, our method is scalable and each detected bursty topic is directly associated with a word distribution and a set of tweets (see Table 3), which makes it easier to interpret the topic.

Topic models provide a principled and elegant way to discover hidden topics from large document collections. Standard topic models do not consider temporal information. A number of temporal topic models have been proposed to consider topic changes over time. Some of these models focus on the change of topic composition, i.e. word distributions, which is not relevant to bursty topic detection (Blei and Lafferty, 2006; Nallapati et al., 2007; Wang et al., 2008). Some other work looks at the temporal evolution of topics, but the focus is not on bursty patterns (Wang and McCallum, 2006; Ahmed and Xing, 2008; Masada et al., 2009; Ahmed and Xing, 2010; Hong et al., 2011).

The model proposed by Wang et al. (2007) is the most relevant to ours. But as we have pointed out in Section 1, they do not need to handle the separation of “personal” documents from event-driven documents. As we will show later in our experiments, for microblogs it is critical to model users’ personal interests in addition to global topical trends.

To capture users' interests, Rosen-Zvi et al. (2004) expand topic distributions from document-level to user-level in order to capture users' specific interests. But on microblogs, posts are short and noisy, so Zhao et al. (2011) further assume that each post is assigned a single topic and some words can be background words. However, these studies do not aim to detect bursty patterns. Our work is novel in that it combines users' interests and temporal information to detect bursty topics.

## 3 Method

### 3.1 Preliminaries

We first introduce the notation used in this paper and formally formulate our problem. We assume that we have a stream of $D$ microblog posts, denoted as $d_1, d_2, \ldots, d_D$. Each post $d_i$ is generated by a user $u_i$, where $u_i$ is an index between 1 and $U$, and $U$ is the total number of users. Each $d_i$ is also associated with a discrete timestamp $t_i$, where $t_i$ is an index between 1 and $T$, and $T$ is the total number of time points we consider. Each $d_i$ contains a bag of words, denoted as $\{w_{i,1}, w_{i,2}, \ldots, w_{i,N_i}\}$, where $w_{i,j}$ is an index between 1 and $V$, and $V$ is the vocabulary size. $N_i$ is the number of words in $d_i$.

We define a bursty topic $b$ as a word distribution coupled with a bursty interval, denoted as $(\phi^b, t_s^b, t_e^b)$, where $\phi^b$ is a multinomial distribution over the vocabulary, and $t_s^b$ and $t_e^b$ $(1 \leq t_s^b \leq t_e^b \leq T)$ are the start and the end timestamps of the bursty interval, respectively. Our task is to find meaningful bursty topics from the input text stream.

Our method consists of a topic discovery step and a burst detection step. At the topic discovery step, we propose a topic model that considers both users' topical interests and the global topic trends. Burst detection is done through a standard state machine method.

### 3.2 Our Topic Model

We assume that there are $C$ (latent) topics in the text stream, where each topic $c$ has a word distribution $\phi^c$. Note that not every topic has a bursty interval. On the other hand, a topic may have multiple bursty intervals and hence leads to multiple bursty topics. We also assume a background word distribution $\phi^B$ that captures common words. All posts are assumed to be generated from some mixture of these $C + 1$ underlying topics.

In standard LDA, a document contains a mixture of topics, represented by a topic distribution, and each word has a hidden topic label. While this is a reasonable assumption for long documents, for short microblog posts, a single post is most likely to be about a single topic. We therefore associate a single hidden variable with each post to indicate its topic. Similar idea of assigning a single topic to a short sequence of words has been used before (Gruber et al., 2007; Zhao et al., 2011). As we will see very soon, this treatment also allows us to model topic distributions at time window level and user level.

As we have discussed in Section 1, an important observation we have is that when everything else is equal, a pair of posts published around the same time is more likely to be about the same topic than a random pair of posts. To model this observation, we assume that there is a global topic distribution $\theta^t$ for each time point $t$. Presumably $\theta^t$ has a high probability for a topic that is popular in the microblogsphere at time $t$.

Unlike news articles from traditional media, which are mostly about current affairs, an important property of microblog posts is that many posts are about users' personal encounters and interests rather than global events. Since our focus is to find popular global events, we need to separate out these "personal" posts. To do this, an intuitive idea is to compare a post with its publisher's general topical interests observed over a long time. If a post does not match the user's long term interests, it is more likely related to a global event. We therefore introduce a time-independent topic distribution $\eta^u$ for each user to capture her long term topical interests.

We assume the following generation process for all the posts in the stream. When user $u$ publishes a post at time point $t$, she first decides whether to write about a global trendy topic or a personal topic. If she chooses the former, she then selects a topic according to $\theta^t$. Otherwise, she selects a topic according to her own topic distribution $\eta^u$. With the chosen topic, words in the post are generated from the word distribution for that topic or from the background word distribution that captures white noise.

```
1. Draw $\phi^B \sim \text{Dirichlet}(\beta), \pi \sim \text{Beta}(\gamma), \rho \sim \text{Beta}(\lambda)$
2. For each time point $t = 1, \ldots, T$
    (a) draw $\theta^t \sim \text{Dirichlet}(\alpha)$
3. For each user $u = 1, \ldots, U$
    (a) draw $\eta^u \sim \text{Dirichlet}(\alpha)$
4. For each topic $c = 1, \ldots, C$,
    (a) draw $\phi^c \sim \text{Dirichlet}(\beta)$
5. For each post $i = 1, \ldots, D$,
    (a) draw $y_i \sim \text{Bernoulli}(\pi)$
    (b) draw $z_i \sim \text{Multinomial}(\eta^{u_i})$ if $y_i = 0$ or
        $z_i \sim \text{Multinomial}(\theta^{t_i})$ if $y_i = 1$
    (c) for each word $j = 1, \ldots, N_i$
        i. draw $x_{i,j} \sim \text{Bernoulli}(\rho)$
        ii. draw $w_{i,j} \sim \text{Multinomial}(\phi^B)$ if
            $x_{i,j} = 0$ or $w_{i,j} \sim \text{Multinomial}(\phi^{z_i})$
            if $x_{i,j} = 1$
```

Figure 2: The generation process for all posts.

We use $\pi$ to denote the probability of choosing to talk about a global topic rather than a personal topic.

Formally, the generation process is summarized in Figure 2. The model is also depicted in Figure 1(a).

There are two degenerate variations of our model that we also consider in our experiments. The first one is depicted in Figure 1(b). In this model, we only consider the time-dependent topic distributions that capture the global topical trends. This model can be seen as a direct application of the model by Wang et al. (2007). The second one is depicted in Figure 1(c). In this model, we only consider the users' personal interests but not the global topical trends, and therefore temporal information is not used. We refer to our complete model as *TimeUserLDA*, the model in Figure 1(b) as *TimeLDA* and the model in Figure 1(c) as *UserLDA*. We also consider a standard LDA model in our experiments, where each word is associated with a hidden topic.

## Learning

We use collapsed Gibbs sampling to obtain samples of the hidden variable assignment and to estimate the model parameters from these samples. Due to space limit, we only show the derived Gibbs sampling formulas as follows.

First, for the $i$-th post, we know its publisher $u_i$ and timestamp $t_i$. We can jointly sample $y_i$ and $z_i$ based on the values of all other hidden variables. Let us use $\mathbf{y}$ to denote the set of all hidden variables $y$ and $\mathbf{y}_{\neg i}$ to denote all $y$ except $y_i$. We use similar symbols for other variables. We then have

$$p(y_i = p, z_i = c | \mathbf{z}_{\neg i}, \mathbf{y}_{\neg i}, \mathbf{x}, \mathbf{w}) \propto \frac{M^{\pi}_{(p)} + \gamma}{M^{\pi}_{(\cdot)} + 2\gamma} \cdot \frac{M^{l}_{(c)} + \alpha}{M^{l}_{(\cdot)} + C\alpha} \cdot \frac{\prod_{v=1}^{V} \prod_{k=0}^{E_{(v)}-1} (M^{c}_{(v)} + k + \beta)}{\prod_{k=0}^{E_{(\cdot)}-1} (M^{c}_{(\cdot)} + k + V\beta)}, \quad (1)$$

where $l = u_i$ when $p = 0$ and $l = t_i$ when $p = 1$. Here every $M$ is a counter. $M^{\pi}_{(0)}$ is the number of posts generated by personal interests, while $M^{\pi}_{(1)}$ is the number of posts coming from global topical trends. $M^{\pi}_{(\cdot)} = M^{\pi}_{0} + M^{\pi}_{1}$. $M^{u_i}_{(c)}$ is the number of posts by user $u_i$ and assigned to topic $c$, and $M^{u_i}_{(\cdot)}$ is the total number of posts by $u_i$. $M^{t_i}_{(c)}$ is the number of posts assigned to topic $c$ at time point $t_i$, and $M^{t_i}_{(\cdot)}$ is the total number of posts at $t_i$. $E_{(v)}$ is the number of times word $v$ occurs in the $i$-th post and is labeled as a topic word, while $E_{(\cdot)}$ is the total number of topic words in the $i$-th post. Here, topic words refer to words whose latent variable $x$ equals 1. $M^{c}_{(v)}$ is the number of times word $v$ is assigned to topic $c$, and $M^{c}_{(\cdot)}$ is the total number of words assigned to topic $c$. All the counters $M$ mentioned above are calculated with the $i$-th post excluded.

We sample $x_{i,j}$ for each word $w_{i,j}$ in the $i$-th post using

$$p(x_{i,j} = q | \mathbf{y}, \mathbf{z}, \mathbf{x}_{\neg\{i,j\}}, \mathbf{w}) \propto \frac{M^{\rho}_{(q)} + \gamma}{M^{\rho}_{(\cdot)} + 2\gamma} \cdot \frac{M^{l}_{(w_{i,j})} + \beta}{M^{l}_{(\cdot)} + V\beta}, \quad (2)$$

where $l = B$ when $q = 0$ and $l = z_i$ when $q = 1$. $M^{\rho}_{(0)}$ and $M^{\rho}_{(1)}$ are counters to record the numbers of words assigned to the background model and any topic, respectively, and $M^{\rho}_{(\cdot)} = M^{\rho}_{(0)} + M^{\rho}_{(1)}$. $M^{B}_{(w_{i,j})}$ is the number of times word $w_{i,j}$ occurs as a background word. $M^{z_i}_{(w_{i,j})}$ counts the number of times word $w_{i,j}$ is assigned to topic $z_i$, and $M^{z_i}_{(\cdot)}$ is the total number of words assigned to topic $z_i$. Again, all counters are calculated with the current word $w_{i,j}$ excluded.

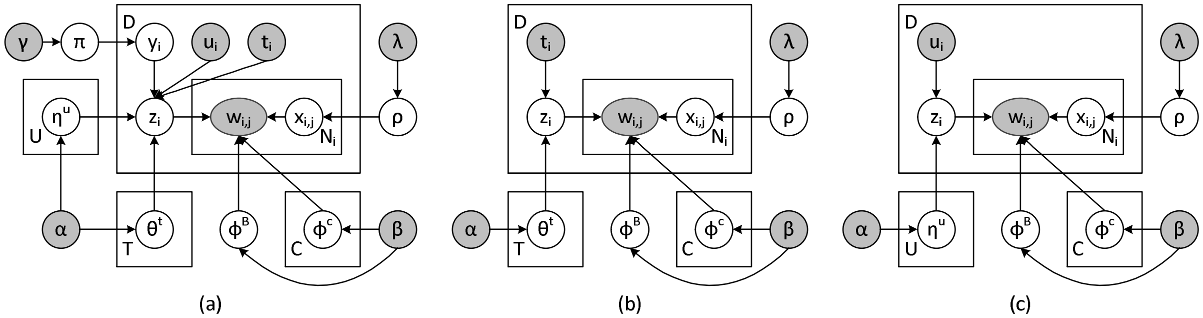

Figure 1: (a) Our topic model for burst detection. (b) A variation of our model where we only consider global topical trends. (c) A variation of our model where we only consider users' personal topical interests.

### 3.3 Burst Detection

Just like standard LDA, our topic model itself finds a set of topics represented by $\phi^c$ but does not directly generate bursty topics. To identify bursty topics, we use the following mechanism, which is based on the idea by Kleinberg (2002) and Ihler et al. (2006). In our experiments, when we compare different models, we also use the same burst detection mechanism for other models.

We assume that after topic modeling, for each discovered topic $c$, we can obtain a series of counts $(m_1^c, m_2^c, \ldots, m_T^c)$ representing the intensity of the topic at different time points. For LDA, these are the numbers of words assigned to topic $c$. For TimeUserLDA, these are the numbers of posts which are in topic $c$ and generated by the global topic distribution $\theta^{t_i}$, i.e whose hidden variable $y_i$ is 1. For other models, these are the numbers of posts in topic $c$.

We assume that these counts are generated by two Poisson distributions corresponding to a bursty state and a normal state, respectively. Let $\mu_0$ denote the expected count for the normal state and $\mu_1$ for the bursty state. Let $v_t$ denote the state for time point $t$, where $v_t = 0$ indicates the normal state and $v_t = 1$ indicates the bursty state. The probability of observing a count of $m_t^c$ is as follows:

$$p(m_t^c | v_t = l) = \frac{e^{-\mu_l} \mu_l^{m_t^c}}{m_t^c!},$$

where $l$ is either 0 or 1. The state sequence $(v_0, v_1, \ldots, v_T)$ is a Markov chain with the following transition probabilities:

$$p(v_t = l | v_{t-1} = l) = \sigma_l,$$

where $l$ is either 0 or 1.

| Method | P@5 | P@10 | P@20 | P@30 |
|---|---|---|---|---|
| LDA | 0.600 | 0.800 | 0.750 | N/A |
| TimeLDA | 0.800 | 0.700 | 0.600 | 0.633 |
| UserLDA | 0.800 | 0.700 | 0.850 | **0.833** |
| TimeUserLDA | **1.000** | **1.000** | **0.900** | 0.800 |

Table 1: Precision at $K$ for the various models.

| Method | P@5 | P@10 | P@20 | P@30 |
|---|---|---|---|---|
| LDA | 0.600 | 0.800 | 0.700 | N/A |
| TimeLDA | 0.400 | 0.500 | 0.500 | 0.567 |
| UserLDA | 0.800 | 0.500 | 0.500 | 0.600 |
| TimeUserLDA | **1.000** | **0.900** | **0.850** | **0.767** |

Table 2: Precision at $K$ for the various models after we remove redundant bursty topics.

$\mu_0$ and $\mu_1$ are topic specific. In our experiments, we set $\mu_0 = \frac{1}{T}\sum_t m_t^c$, that is, $\mu_0$ is the average count over time. We set $\mu_1 = 3\mu_0$. For transition probabilities, we empirically set $\sigma_0 = 0.9$ and $\sigma_1 = 0.6$ for all topics.

We can use dynamic programming to uncover the underlying state sequence for a series of counts. Finally, a burst is marked by a consecutive subsequence of bursty states.

## 4 Experiments

### 4.1 Data Set

We use a Twitter data set to evaluate our models. The original data set contains 151,055 Twitter users based in Singapore and their tweets. These Twitter users were obtained by starting from a set of seed Singapore users who are active online and tracing

| Bursty Period | Top Words | Example Tweets | Label |
|---|---|---|---|
| Nov 29 | vote, big, awards, bang, mama, win, 2ne1, award, won | (1) why didnt 2ne1 win this time!<br>(2) 2ne1. you deserved that urgh!<br>(3) watching mama. whoohoo | Mnet Asian Music Awards (MAMA) |
| Oct 5 ∼ Oct 8 | steve, jobs, apple, iphone, rip, world, changed, 4s, siri | (1) breaking: apple says steve jobs has passed away!<br>(2) google founders: steve jobs was an inspiration!<br>(3) apple 4 life thankyousteve | Steve Jobs death |
| Nov 1 ∼ Nov 3 | reservior, bedok, adlyn, slap, found, body, mom, singapore, steven | (1) this adelyn totally disgust me. slap her mum? queen of cine? joke please can.<br>(2) she slapped her mum and boasted about it on fb<br>(3) adelyn lives in woodlands , later she slap me how? | girl slapping mom |
| Nov 5 | reservior, bedok, adlyn, slap, found, body, mom, singapore, steven | (1) bedok = bodies either drowned or killed.<br>(2) another body found, in bedok reservoir?<br>(3) so many bodies found at bedok reservoir. alamak. | suicide near bedok reservoir |
| Oct 23 | man, arsenal, united, liverpool, chelsea, city, goal, game, match | (1) damn you man city! we will get you next time!<br>(2) wtf 90min goal!<br>(3) 6-1 to city. unbelievable. | football game |

Table 3: Top-5 bursty topics ranked by TimeUserLDA. The labels are manually given. The 3rd and the 4th bursty topics come from the same topic but have different bursty periods.

| Rank | LDA | UserLDA | TimeLDA |
|---|---|---|---|
| 1 | Steve Jobs' death | MAMA | MAMA |
| 2 | MAMA | football game | MAMA |
| 3 | N/A | #zamanprimaryschool | MAMA |
| 4 | girl slapping mom | N/A | girl slapping mom |
| 5 | N/A | iphone 4s | N/A |

Table 4: Top-5 bursty topics ranked by other models. N/A indicates a meaningless burst.

their follower/followee links by two hops. Because this data set is huge, we randomly sampled 2892 users from this data set and extracted their tweets between September 1 and November 30, 2011 (91 days in total). We use one day as our time window. Therefore our timestamps range from 1 to 91. We then removed stop words and words containing non-standard characters. Tweets containing less than 3 words were also discarded. After preprocessing, we obtained the final data set with 3,967,927 tweets and 24,280,638 tokens.

### 4.2 Ground Truth Generation

To compare our model with other alternative models, we perform both quantitative and qualitative evaluation. As we have explained in Section 3, each model gives us time series data for a number of topics, and by applying a Poisson-based state machine, we can obtain a set of bursty topics. For each method, we rank the obtained bursty topics by the number of tweets (or words in the case of the LDA model) assigned to the topics and take the top-30 bursty topics from each model. In the case of the LDA model, only 23 bursty topics were detected. We merged these topics and asked two human judges to judge their quality by assigning a score of either 0 or 1. The judges are graduate students living in Singapore and not involved in this project. The judges were given the bursty period and 100 randomly selected tweets for the given topic within that period for each bursty topic. They can consult external resources to help make judgment. A bursty topic was scored 1 if the 100 tweets coherently describe a bursty event based on the human judge's understanding. The inter-annotator agreement score is 0.649 using Cohen's kappa, showing substantial agreement. For ground truth, we consider a bursty topic to be correct if both human judges have scored it 1. Since some models gave redundant bursty topics, we also asked one of the judges to identify unique bursty topics from the ground truth bursty topics.

### 4.3 Evaluation

In this section, we show the quantitative evaluation of the four models we consider, namely, LDA, TimeLDA, UserLDA and TimeUserLDA. For each model, we set the number of topics $C$ to 80, $\alpha$ to $\frac{50}{C}$ and $\beta$ to 0.01 after some preliminary experiments. Each model was run for 500 iterations of Gibbs sampling. We take 40 samples with a gap of 5 iterations in the last 200 iterations to help us assign values to all the hidden variables.

Table 1 shows the comparison between these models in terms of the precision of the top-$K$ results. As we can see, our model outperforms all other models for $K <= 20$. For $K = 30$, the UserLDA model performs the best followed by our model.

As we have pointed out, some of the bursty topics are redundant, i.e. they are about the same bursty event. We therefore also calculated precision at $K$ for unique topics, where for redundant topics the one ranked the highest is scored 1 and the other ones are scored 0. The comparison of the performance is shown in Table 2. As we can see, in this case, our model outperforms other models with all $K$. We will further discuss redundant bursty topics in the next section.

### 4.4 Sample Results and Discussions

In this section, we show some sample results from our experiments and discuss some case studies that illustrate the advantages of our model.

First, we show the top-5 bursty topics discovered by the TimeUserLDA model in Table 3. As we can see, all these bursty topics are meaningful. Some of these events are global major events such as Steve Jobs' death, while some others are related to online events such as the scandal of a girl boasting about slapping her mother on Facebook. For comparison, we also show the top-5 bursty topics discovered by other models in Table 4. As we can see, some of them are not meaningful events while some of them are redundant.

Next, we show two case studies to demonstrate the effectiveness of our model.

**Effectiveness of Temporal Models:** Both TimeLDA and TimeUserLDA tend to group posts published on the same day into the same topic. We find that this can help separate bursty topics from general ones. An example is the topic on the Circle Line. The Circle Line is one of the subway lines of Singapore's mass transit system. There were a few incidents of delays or breakdowns during the period between September and November, 2011. We show the time series data of the topic related to the Circle Line of UserLDA, TimeLDA and TimeUserLDA in Figure 3. As we can see, the UserLDA model detects a much large volume of tweets related to this topic. A close inspection tells us that the topic under UserLDA is actually related to the subway systems in Singapore in general, which include a few other subway lines, and the Circle Line topic is merged with this general topic. On the other hand, TimeLDA and TimeUserLDA are both able to separate the Circle Line topic from the general subway topic because the Circle Line has several bursts. What is shown in Figure 3 for TimeLDA and TimeUserLDA is only the topic on the Circle Line, therefore the volume is much smaller. We can see that TimeLDA and TimeUserLDA show clearer bursty patterns than UserLDA for this topic. The bursts around day 20, day 44 and day 85 are all real events based on our ground truth.

**Effectiveness of User Models:** We have stated that it is important to filter out users' "personal" posts in order to find meaningful global events. We find that our results also support this hypothesis. Let us look at the example of the topic on the Mnet Asian Music Awards, which is a major music award show that is held by Mnet Media annually. In 2011, this event took place in Singapore on November 29. Because Korean pop music is very popular in Singapore, many Twitter users often tweet about Korean pop music bands and singers in general. All our topic models give multiple topics related to Korean pop music, and many of them have a burst on November 29, 2011. Under the TimeLDA and UserLDA models, this leads to several redundant bursty topics for the MAMA event ranked within the top-30. For TimeUserLDA, however, although the MAMA event is also ranked the top, there is no redundant one within the top-30 results. We find that this is because with TimeUserLDA, we can remove tweets that are considered personal and therefore do not contribute to bursty topic ranking. We show the topic intensity of a topic about a Korean pop singer in

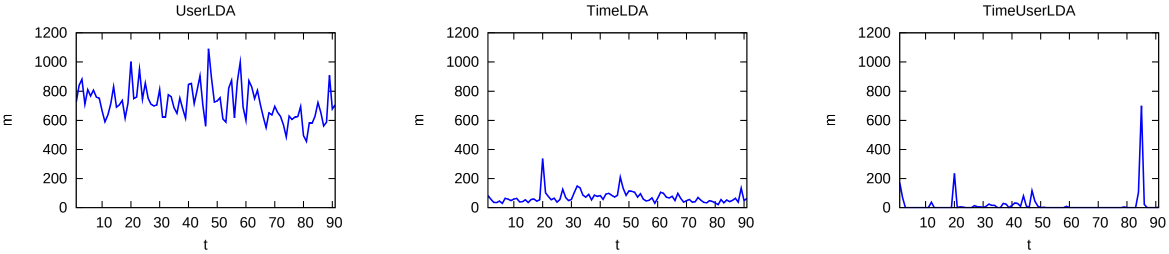

Figure 3: Topic intensity over time for the topic on the Circle Line.

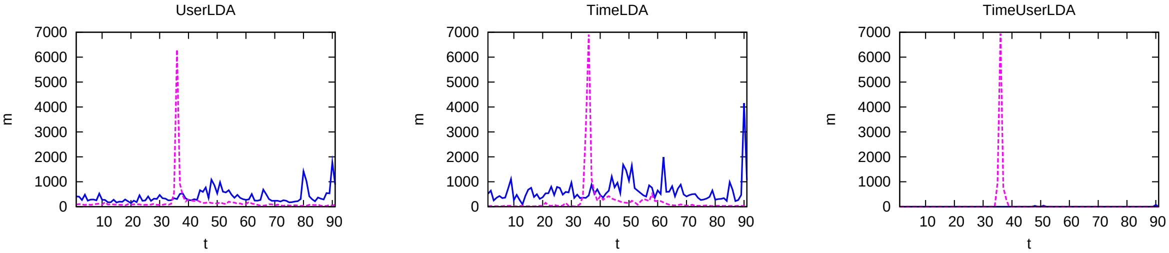

Figure 4: Topic intensity over time for the topic about a Korean pop singer. The dotted curves show the topic on Steve Jobs' death.

Figure 4. For reference, we also show the intensity of the topic on Steve Jobs' death under each model. We can see that because this topic is related to Korean pop music, it has a burst on day 90 (November 29). But if we consider the relative intensity of this burst compared with Steve Jobs' death, under TimeLDA and UserLDA, this topic is still strong but under TimeUserLDA its intensity can almost be ignored. This is why with TimeLDA and UserLDA this topic leads to a redundant burst within the top-30 results but with TimeUserLDA the burst is not ranked high.

## 5 Conclusions

In this paper, we studied the problem of finding bursty topics from the text streams on microblogs. Because existing work on burst detection from text streams may not be suitable for microblogs, we proposed a new topic model that considers both the temporal information of microblog posts and users' personal interests. We then applied a Poisson-based state machine to identify bursty periods from the topics discovered by our model. We compared our model with standard LDA as well as two degenerate variations of our model on a real Twitter dataset. Our quantitative evaluation showed that our model could more accurately detect unique bursty topics among the top ranked results. We also used two case studies to illustrate the effectiveness of the temporal factor and the user factor of our model.

Our method currently can only detect bursty topics in a retrospective and offline manner. A more interesting and useful task is to detect realtime bursts in an online fashion. This is one of the directions we plan to study in the future. Another limitation of the current method is that the number of topics is predetermined. We also plan to look into methods that allow appearance and disappearance of topics along the timeline, such as the model by Ahmed and Xing (2010).

## 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

Amr Ahmed and Eric P. Xing. 2008. 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*, pages 219–230.

Amr Ahmed and Eric P. Xing. 2010. Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, pages 20–29.

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*.

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

Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Philip S. Yu, and Hongjun Lu. 2005. Parameter free bursty events detection in text streams. In *Proceedings of the 31st International Conference on Very Large Data Bases*, pages 181–192.

Amit Gruber, Michal Rosen-Zvi, and Yair Weiss. 2007. Hidden topic Markov model. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*.

Liangjie Hong, Byron Dom, Siva Gurumurthy, and Kostas Tsioutsiouliklis. 2011. A time-dependent topic model for multiple text streams. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 832–840.

Alexander Ihler, Jon Hutchins, and Padhraic Smyth. 2006. Adaptive event detection with time-varying poisson processes. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 207–216.

Jon Kleinberg. 2002. Bursty and hierarchical structure in streams. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 91–101.

Tomonari Masada, Daiji Fukagawa, Atsuhiro Takasu, Tsuyoshi Hamada, Yuichiro Shibata, and Kiyoshi Oguri. 2009. Dynamic hyperparameter optimization for bayesian topical trend analysis. In *Proceedings of the 18th ACM Conference on Information and knowledge management*, pages 1831–1834.

Ramesh M. Nallapati, Susan Ditmore, John D. Lafferty, and Kin Ung. 2007. Multiscale topic tomography. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 520–529.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494.

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. Mining correlated bursty topic patterns from coordinated text streams. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 784–793.

Chong Wang, David M. Blei, and David Heckerman. 2008. Continuous time dynamic topic models. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 579–586.

Jianshu Weng and Francis Lee. 2011. Event detection in Twitter. In *Proceedings of the 5th International AAAI 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 Advances in information retrieval*, pages 338–349.