Time Series Topic Modeling and Bursty Topic Detection of Correlated News and Twitter

Daichi Koike
Yusuke Takahashi
Takehito Utsuro
Grad. Sch. Sys. & Inf. Eng.,
University of Tsukuba,
Tsukuba, 305-8573, JAPAN

Masaharu Yoshioka
Grad. Sch. Inf. Sci. & Tech.,
Hokkaido University,
Sapporo, 060-0808,
JAPAN

Noriko Kando
National Institute
of Informatics,
Tokyo, 101-8430,
JAPAN

Abstract

News and twitter are sometimes closely correlated, while sometimes each of them has quite independent flow of information, due to the difference of the concerns of their information sources. In order to effectively capture the nature of those two text streams, it is very important to model both their correlation and their difference. This paper first models their correlation by applying a time series topic model to the document stream of the mixture of time series news and twitter. Next, we divide news streams and twitter into distinct two series of document streams, and then we apply our model of bursty topic detection based on the Kleinberg's burst detection model. This approach successfully models the difference of the two time series topic models of news and twitter as each having independent information source and its own concern.

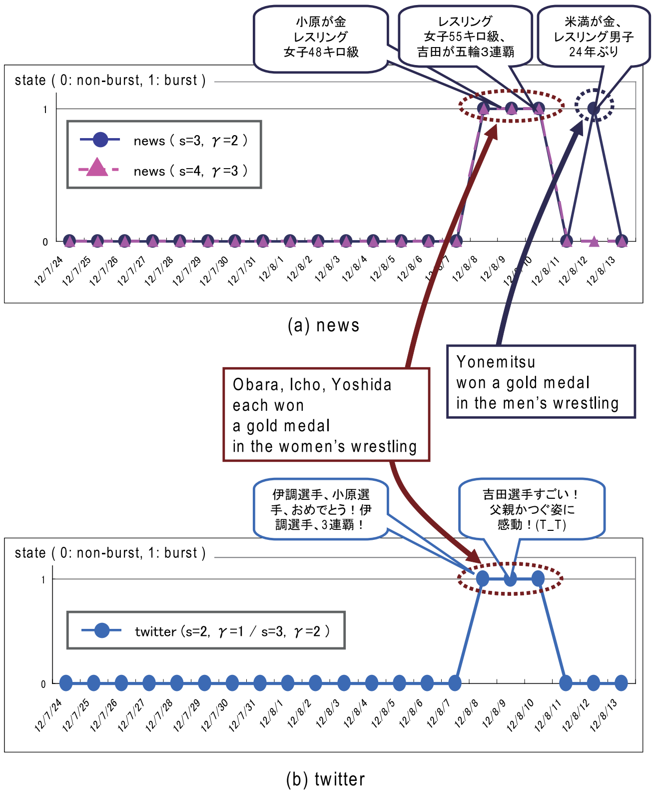


Figure 1: Optimal State Sequence for the Topic "wrestling"

1 Introduction

The background of this this paper is in two types of modeling of information flow in news stream, namely, burst analysis and topic modeling. Both types of modeling, to some extent, aim at aggregating information and reducing redundancy within the information flow in news stream.

First, when one wants to detect a kind of topics that are paid much more attention than usual, it is usually necessary for him/her to carefully watch every article in news stream at every moment. In such a situation, it is well known in the field of time series analysis that Kleinberg's modeling of bursts (Kleinberg, 2002) is quite effective in detecting burst of keywords. Second, topic models such as LDA (latent Dirichlet allocation) (Blei et al., 2003) and DTM (dynamic topic model) (Blei and Lafferty, 2006) are also quite effective in estimating distribution of topics over a document collection such as articles in news stream. Unlike LDA, in DTM, we suppose that the data is divided by time slice, for example by date. DTM models the documents (such as articles of news stream) of each slice with a K-component topic model, where the k-th topic at slice t smoothly evolves from the k-th topic at slice $t-1$.

Based on those arguments above, Takahashi et al. (2012) proposed how to integrate the two types of modeling of information flow in news stream. Here, it is important to note that Kleinberg's modeling of bursts is usually applied only to bursts of keywords but not to those of topics. Thus, Taka-

International Joint Conference on Natural Language Processing, pages 917–921,
Nagoya, Japan, 14-18 October 2013.

hashi et al. (2012) proposed how to apply Kleinberg's modeling of bursts to topics estimated by a topic model such as DTM. Typical results of applying the technique to time series news stream can be illustrated as in Figure 1 (a). In this example, we first estimate time series topics through DTM, among which is the one "wrestling" as shown in this figure. Then, we can detect the burst of the topic on the dates when those two Japanese wrestlers won the gold medals.

Unlike Takahashi et al. (2012), this paper studies the issue of time series topic modeling and bursty topic detection of possibly correlated news and twitter. News and twitter are sometimes closely correlated, while sometimes each of them has quite independent flow of information, due to the difference of the concerns of their information sources. In order to effectively capture the nature of those two text streams, it is very important to model both their correlation and their difference. This paper first models their correlation by applying a time series topic model to the document stream of the mixture of time series news and twitter. This approach successfully models the time series topic models of news and twitter as closely correlated to each other. Next, we divide news streams and twitter into distinct two series of document streams, and then we apply our model of bursty topic detection based on the Kleinberg's burst detection model. With this procedure, we show that, even though we estimate the time series topic model with the document stream of the mixture of news and twitter, we can detect bursty topics individually both in the news stream and in twitter. This approach again successfully models the difference of the two time series topic models of news and twitter as each having independent information source and its own concern.

2 Time Series Documents Set for Evaluation

In this paper, we collect time series news articles of a certain period as well as tweets texts of the same period that are closely related to the news articles. Then, we construct a time series document set consisting of the mixture of the news articles and tweets texts (Table 1) and use it for evaluation.

2.1 News

As the news stream documents set for evaluation, during the period from July 24th to August 13th, 2012, we collected 3,157 Yomiuri newspaper articles, 4,587 Nikkei newspaper articles, and 3,458 Asahi newspaper articles which amount to 11,202 newspaper articles in total[1]. Then, we select a subset of the whole 11,202 newspaper articles which are related to "the London Olympic game", where we collect 2,308 articles that contain at least one of 8 keywords[2] into the subset. The subset consists of 659 Yomiuri newspaper articles, 679 Nikkei newspaper articles, and 970 Asahi newspaper articles.

Table 1: Time Series Documents Set

news articles	tweets	total # of document
2,308	57,414	59,722

2.2 Twitter

As the tweet text data set for evaluation, during the period from July 24th to August 13th, 2012, we collected 9,509,774 tweets from the Twitter[3] with the Streaming API. Then, we removed tweets with official retweets and those including URLs, and 7,752,129 tweets remained. Finally, we select a subset which are related to "the London Olympic game". Here, we collect 57,414 tweets that contain at least one of the 8 keywords listed above, which are closely related to "the London Olympic game", into the subset.

3 Kleinberg's Bursts Modeling

Kleinberg (2002) proposed two types of frameworks for modeling bursts. The first type of modeling is based on considering a sequence of message arrival times, where a sequence of messages is regarded as bursty if their inter-arrival gaps are too small than usual. The second type of modeling is, on the other hand, based on the case where documents arrive in discrete *batches* and in each batch of documents, some are *relevant* (e.g., news text contains a particular word) and some are *irrelevant*. In this second type of bursts modeling, a sequence of batched arrivals could be considered bursty if the fraction of relevant documents alternates between reasonably long periods in which the fraction is small and other periods in which it is large. Out of the two modelings, this paper em-

[1] `http://www.yomiuri.co.jp/`, `http://www.nikkei.com/`, and `http://www.asahi.com/`.

[2] 五輪 (*Gorin* ("Olympic" in Chinese characters)), ロンドン (London), オリンピック (Olympic (in katakana characters)), 金メダル (gold medal), 銀メダル (silver medal), 銅メダル (bronze medal), 選手 (athlete), 日本代表 (Japanese national team)

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

ploys the latter, which is named as *enumerating bursts* in Kleinberg (Kleinberg, 2002).

4 Applying Time Series Topic Model

As a time series topic model, this paper employs DTM (dynamic topic model) (Blei and Lafferty, 2006). In this paper, in order to model time series news stream in terms of a time series topic model, we consider date as the time slice t. Given the number of topics K as well as time series sequence of batches each of which consists of documents represented by a sequence of words w, on each date t (i.e., time slice t), DTM estimated the distribution $p(w|z_n)$ ($w \in V$, the vocabulary set) of a word w given a topic z_n ($n = 1, \ldots, K$) as well as that $p(z_n|b)$ ($n = 1, \ldots, K$) of a topic z_n given a document b, where V is the set of words appearing in the whole document set. In this paper, we estimate the distributions $p(w|z_n)$ ($w \in V$) and $p(z_n|b)$ ($n = 1, \ldots, K$) by a Blei's toolkit[4], where the parameters are tuned through a preliminary evaluation as the number of topics $K = 50$ as well as $\alpha = 0.01$. The DTM topic modeling toolkit is applied to the time series document set shown in Table 1, which consists of the mixture of the news articles and tweets texts. Here, as a word w ($w \in V$) constituting each document, we extract Japanese Wikipedia[5] entry titles as well as their redirects.

5 Modeling Bursty Topics Independently from News and Twitter

In this section, we are given a time series document set which consists of the mixture of two types of documents originating from two distinct sources, e.g., news and tweets. In this situation, we assume that a time series topic model is estimated with the mixture of two types of time series documents, where the distinction of the two sources is ignored at the step of time series topic model estimation. Then, the following procedure presents how to model bursty topics for each of the two types of time series documents independently. This means, in the case of news and twitter, that, although the time series topic model is estimated with the mixture of time series news articles and tweets texts, bursty topics are detected independently from news and twitter.

[4]`http://www.cs.princeton.edu/~blei/topicmodeling.html`

[5]`http://ja.wikipedia.org/`

In this bursty topic modeling, first, we suppose that, on the date t (i.e., time slice t), we have two types of documents b_x and b_y each of which originates from the source x and y, respectively. Then, for the source x, we regard a document b_x as *relevant* to a certain topic z_n that are estimated through the DTM topic modeling procedure, to the degree of the amount of the probability $p(z_n|b_x)$. Similarly for the source y, we regard a document b_y as *relevant* to a certain topic z_n, to the degree of the amount of the probability $p(z_n|b_y)$. Next, for the source x, we estimate the number $r_{t,x}$ of relevant documents out of a total of $d_{t,x}$ simply by summing up the probability $p(z_n|b_x)$ over the whole document set (similarly for the source y):

$$r_{t,x} = \sum_{b_x} p(z_n|b_x) \quad r_{t,y} = \sum_{b_y} p(z_n|b_y)$$

Once we have the number $r_{t,x}$ and $r_{t,y}$ for the sources x and y, then we can estimate the total number of relevant documents throughout the whole batch sequence $\mathbf{B} = (B_1, \ldots, B_m)$ as $R_x = \sum_{t=1}^{m} r_{t,x}$ and $R_y = \sum_{t=1}^{m} r_{t,y}$. Denoting the total numbers of documents on the date t for the sources x and y as $d_{t,x}$ and $d_{t,y}$, respectively, we have the total numbers of documents throughout the whole batch sequence as $D_x = \sum_{t=1}^{m} d_{t,x}$ and $D_y = \sum_{t=1}^{m} d_{t,y}$, respectively. Finally, we can estimate the expected fraction of relevant documents as $p_{0,x} = R_x/D_x$ and $p_{0,y} = R_y/D_y$, respectively. Then, by simply following the formalization of bursty topics we proposed in Takahashi et al. (2012), it is quite straightforward to model bursty topics independently for each of the two sources x and y. In the following evaluation, we consider the sources x and y as time series news articles and tweet texts shown in Table 1. As the two parameters s and γ for bursty topic detection[6], we compare two pairs $s = 4$, $\gamma = 3$ and $s = 3$, $\gamma = 2$ for time series news articles, and $s = 3$, $\gamma = 2$ and $s = 2$, $\gamma = 1$ for tweets text.

6 Evaluation

6.1 The Procedure

As the evaluation of the proposed technique, we examine the correctness of the detected bursty top-

[6]s is a parameter for scaling expected fractions of relevant documents between burst / non-burst states. γ is a parameter for the cost of moving from the non-burst state to the burst sate. The details of the two parameters are described in Kleinberg (2002) and Takahashi et al. (2012).

Table 2: Evaluation Results: Precision of Detecting Bursty Topics (for 34 Topics relevant to “the London Olympic Games” out of the whole 50)

	bursts detected in both news and twitter	bursts detected only in one of news and twitter
news	per day: 87.5 % (14/16)	per day: 100 % (2/2), per topic: 100 % (1/1)
twitter	per topic 87.5 % (7/8)	per day: 100 % (32/32), per topic: 100 % (13/13)

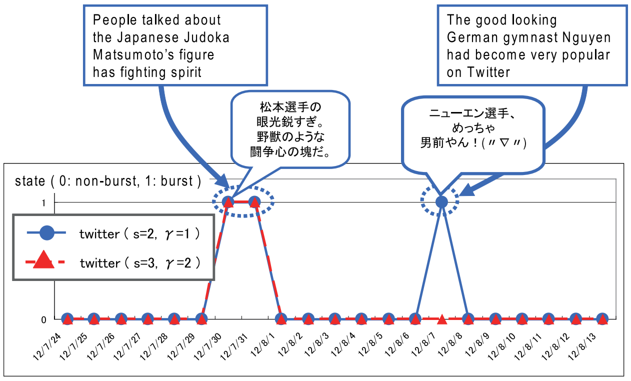
ics. For each topic z_n, collect the documents b which satisfies $z_n = \arg\max_{z'} p(z'|b)$ into the set $B_{1st}(z_n)$. Then, we first judge whether most of the collected documents (both news articles and tweets texts) $b \in B_{1st}(z_n)$ have relatively similar contents. If so, next we examine the correctness of the detected burst of that topic.

We evaluate the detected bursty topics per day or per topic. As for “per day evaluation”, we examine whether, on each day of the burst, the detected burst is appropriate or not. As for “per topic evaluation”, we examine whether, for each topic, all of the detected bursts are appropriate or not.

Out of the whole 50 topics, we manually select 34 that are relevant to “the London Olympic games”, and show the evaluation results of detecting bursty topics in Table 2. Here, as the two parameters s and γ for bursty topic detection, we show those with $s = 4$ and $\gamma = 3$ for news and $s = 3$ and $\gamma = 2$ for tweets, for which we have the highest precision in bursty topic detection. We also classify the detected bursts per day and detected bursty topics (i.e., *per topic*) into the following two types: (a) *the bursty topic is shared between news and twitter*, and (b) *the bursty topic is detected only in one of news and twitter*.

6.2 Evaluation Results

As shown in Table 2, for the bursty topic of type (b), precisions for both “per day” and “per topic” evaluation are 100% (both for news and twitter). The proposed technique is quite effective in detecting many bursty topics that are observed only in twitter. For the bursty topic of type (a), over detection of bursty topics is only for one topic, which is about “*politics*”. The reason why this over detection occurred is mainly because we observed fewer numbers of news articles and tweets on politics during the period of “the London Olympic games”, and then, the periods other than “the London Olympic games” are detected as bursty. Also



(b) twitter

Figure 2: Optimal State Sequence for the Topic “good looking athletes” (observed only in twitter)

for the bursty topic of type (a), reasons of bursts in news articles and tweets texts are almost the same as each other. This result clearly supports our claim that the proposed technique is quite effective in detecting closely related bursty topics in news and twitter.

Figure 1 plots the optimal state sequence for the topic “wrestling” for both news and twitter. For this topic, some of the bursts are shared between news and twitter, so we also show the results of aligning bursts between news and twitter. Figure 2 also plots the optimal state sequence for the topic “good looking athletes”, where for this topic, all the documents are from the source twitter and bursts are detected only for twitter[7].

7 Conclusion

This paper showed that, even though we estimate the time series topic model with the document stream of the mixture of news and twitter, we can detect bursty topics independently both in the news stream and in twitter. Among several related works, Diao et al. (2012) proposed a topic model for detecting bursty topics from microblogs. Compared with Diao et al. (2012), one of our major contributions is that we mainly focus on the modeling of correlation and difference between news and twitter.

[7] It is surprising that tweets that mentioned good looking athletes are collected altogether in this topic. Many tweets collected in this topic on the non-bursty days said that he/she likes a certain athlete. And, those tweets share the terms 選手 (athlete) and 好き (like). But, especially on the days when the bursts were observed, much more people posted that the Japanese judoka Matsumoto and the German gymnast Nguyen were so impressive because of their looking. This is why we observed bursts on those days.

References

D. M. Blei and J. D. Lafferty. 2006. Dynamic topic models. In *Proc. 23rd ICML*, pages 113–120.

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

Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim. 2012. Finding bursty topics from microblogs. In *Proc. 50th ACL*, pages 536–544.

J. Kleinberg. 2002. Bursty and hierarchical structure in streams. In *Proc. 8th SIGKDD*, pages 91–101.

Y. Takahashi, T. Utsuro, M. Yoshioka, N. Kando, T. Fukuhara, H. Nakagawa, and Y. Kiyota. 2012. Applying a burst model to detect bursty topics in a topic model. In *JapTAL 2012*, volume 7614 of *LNCS*, pages 239–249. Springer.