Knowledge-Based Systems 000 (2017) 1–11

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Detecting bursts in sentiment-aware topics from social media

Kang Xu[a,b], Guilin Qi[b,*], Junheng Huang[b], Tianxing Wu[b], Xuefeng Fu[c]

[a] School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing, China
[b] School of Computer Science and Engineering, Southeast University, Nanjing, China
[c] School of Information Engineering, Nanchang Institute of Technology, Nanchang, China

ARTICLE INFO

Article history:
Received 30 November 2016
Revised 6 November 2017
Accepted 8 November 2017
Available online xxx

Keywords:
Sentiment analysis
Burst detection
Sentiment topic model
Sina weibo

ABSTRACT

Nowadays plenty of user-generated posts, e.g., sina weibos, are published on the social media. The posts contain the public's sentiments (i.e., positive or negative) towards various topics. Bursty sentiment-aware topics from these posts reveal sentiment-aware events which have attracted much attention. To detect sentiment-aware topics, we attempt to utilize Joint Sentiment/Topic models, these models are achieved with Latent Dirichlet Allocation (LDA) based models. However, most of the existing sentiment/topic models cannot be directly utilized to detect sentiment-aware topics on the posts, since applying the models to the posts directly suffers from the context sparsity problem. In this paper, we propose a Time-User Sentiment/Topic Latent Dirichlet Allocation (TUS-LDA) which simultaneously models sentiments and topics for posts. Thereinto, TUS-LDA aggregates posts in the same timeslices or from the same users as pseudo-documents to alleviate the context sparsity problem. Based on TUS-LDA, we further design an approach to detect bursty sentiment-aware topics and these sentiment-ware topics can reflect bursty real-world events. Experiments on the Chinese sina weibos show that TUS-LDA outperforms previous models in the tasks of sentiment classification and burst detection in sentiment-aware topics. Finally, we visualize the bursty sentiment-aware topics discovered by TUS-LDA.

1. Introduction

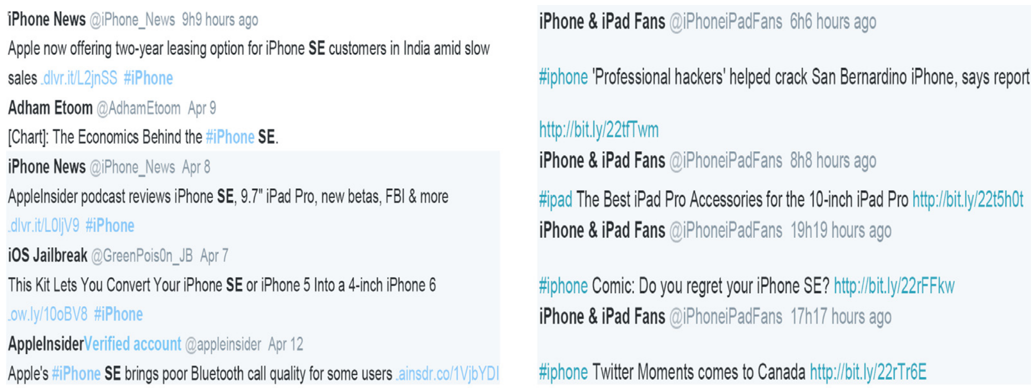
With the rapid growth of Web 2.0, a mass of user-generated posts, e.g., sina weibos, are published on the social media. These posts capture people's interests, thoughts, sentiments and actions, and have been accumulating on the social media over a long period. The topically similar posts, gathered at a certain time period with specific sentiment polarities, reveal the general public's interest. A sudden increase of these posts usually indicates a burst of interest about current events, e.g., the negative event, earthquake, or the positive event, the announcement of mobile phone. These events are always coupled with specific sentiments. Hence, finding bursty sentiment-aware topics can benefit us to monitor the most popular sentiment-aware topics which can affect the public. In this paper, we aim to work on mining sentiment-aware topics from posts. We further study to detect bursts from these sentiment-aware topics.

Topic modeling [1–3] and sentiment analysis [4,5] on the posts are complementary. Thereinto, sentiments on the posts are sensitive to topics, and topics in the posts often imply the sentiments of the public. Thus, jointly modeling topics and sentiments on the posts can reflect people's sentiments on different topics. For example, we can obtain a sentiment-aware topic, e.g., an topic about "Sinking of Dongfangeve Zhi Xing" {{东方星(Dongfangxing), 长江(Yangtze River), 客船(Passenger Ship), 沉船(Shipwrecks), 遇难者(Victims), 客轮(Passenger Ship), 翻沉(Sinking), 遗体(Remains), 遇难(Murdered) and 搜救(Rescue)} with the overall sentiment polarity "negative". However, unlike the normal documents (e.g., news and long reviews), the posts on the social media are short and informal. Thus, they often lack rich contextual information. However, conventional methods of modeling topics and sentiments mainly depend on the document-level contextual information. Hence, the task of topic modeling and sentiment analysis on the posts become more challenging than that of modeling topics and sentiments analysis on normal documents.

Topic models, e.g., LDA [1,6] and pLSA [7], originally focus on mining topics from lengthy texts, and they can further be extended to extract sentiments. Conventional sentiment-aware topic models, like Joint Sentiment/Topic Model (JST) [8] and Aspect/Sentiment Unification Model (ASUM) [9], are utilized for uncovering the hidden topics and sentiments from text corpus. In JST and ASUM, each document is a mixture of sentiment/topics and each sentiment/topic is a mixture of words. Thereinto, each sentiment label

* Corresponding author.
E-mail addresses: kxu@seu.edu.cn (K. Xu), gqi@seu.edu.cn (G. Qi), jhhuang@seu.edu.cn (J. Huang), wutianxing@seu.edu.cn (T. Wu), fxf@nit.edu.cn (X. Fu).

https://doi.org/10.1016/j.knosys.2017.11.007
0950-7051/



(a) Announcement of iPhone SE (b) Apple products

Fig. 1. (a) A temporal topic (b) A stable topic.

in the models is viewed as a special kind of topic, i.e., topics are unknown and data-driven while sentiments are known and specified. Since the posts are short and informal and they often lack rich contextual information, applying the models to the short posts on the social media directly suffers from the context sparsity problem.

One simple and effective way to alleviate the sparsity problem is to aggregate the short posts into lengthy pseudo-documents [10,11]. Hence, we assume that the posts on the social media is a mixture of two kinds of topics: (1) temporal topics which are related to current events (e.g., posts about a topic "Announcement of iphone SE" in Fig. 1(a) which are published in a short time), (2) stable topics which are related to personal interests (e.g., posts about a topic "Apple products" in Fig. 1(b) which are published by a user). In these posts, temporal topics are sensitive to time and related to specific sentiments. For example, when an event occurs, posts with specific sentiments about the event may burst in a short period within a large volume. Hence, if posts talk about temporal topics, all these posts in the same timeslice are aggregated as pseudo-documents and these pseudo-documents are mixtures of sentiment-aware topics. Moreover, stable topics are related to specific users, where each user focuses on several topics with specific sentiments. Hence, if posts talk about stable topics, all these posts published by the same users are aggregated as pseudo-documents and these pseudo-documents are also mixtures of sentiment-aware topics. We assume that posts on social media belong to two kinds of topics, temporal topics and stable topics. Temporal topics are sensitive to time and are generated from posts in the same timeslices. Stable topics are related to users and are generated from posts published by the same users. Thus, if a post belongs to a temporal topic, then it is assigned to a sentiment-aware topic in its corresponding timeslice; if a post belongs to a stable topic, it is assigned to a sentiment-aware topic in its corresponding user.

Based on the analysis of the characteristics of topics and sentiments, Zhao et al. [11] have an important observation of topics: A single post always talks about a single topic. Furthermore, according to Kiritchenko et al. [12] and Lu et al. [13], although a post usually talks about a single topic, it may talk about multiple fine-grained aspects of the topic with different sentiment polarities. For a post, it may express only a kind of sentiment, e.g., the post "It is a good day" is positive, and it can also express more than one kind of sentiments, e.g., the post "Lily looks nice, but Tom does not." contains both positive and negative sentiments [14]. Thus, to accurately model sentiment polarities for the posts, we follow the observations in [12,13] and assume that words in a single post can correspond to multiple sentiment polarities.

To model the association between each post's sentiments and topics, we further assign a sentiment label to each post. The sentiment label represents the overall sentiment polarities of the post and is determined by the sentiment polarities of words in the post. If the words in a post express both positive and negative sentiments, the overall sentiment polarities of the post should be judged based on the stronger one [14].

To handle the aforementioned problems, our work is based on a Time-User Sentiment/Topic Latent Dirichlet Allocation (TUS-LDA) [15], which utilizes user and timeslice information to aggregate posts to alleviate the context sparsity problem. Moreover, TUS-LDA models topics and sentiments based on the characteristics of topics and sentiments. Specifically, the sentiments of a post and the words in the post are all drawn from document-level sentiment distribution; within the chosen sentiment of the post, the topic of the post is drawn from a user-level or timeslice-level sentiment/topic distribution.

In this paper, we utilize TUS-LDA to mine sentiment-aware topics from posts and then detect bursts from sentiment-aware topics discovered by TUS-LDA. In TUS-LDA, we can not only capture the sentiment-aware topics from posts, but also monitor the variations of sentiment-aware topics over time. Our work focuses on the variations of sentiment-aware topics over time to detect and track bursty sentiment-aware topics. These topics are often triggered by real-world events. When a negative event "东方之星号客轮翻沉事件" ("Sinking of Dongfang zhi Xing")[1] occurred, the volume of weibos about the negative event sent spiked to more than 5000 times per second. An effective way to detect bursts is using bursty features, e.g., the bursty volumes in posts corresponding to the same topics [19], in data streams. In our work, we consider detecting bursts by monitoring the variations of sentiment-aware topics over time. Different from the work on detecting bursty topics in [10,16], we consider to detect bursts on sentiment-aware topics. Because there is a strong correlation between bursty topics and public moods, there are always bursty sentiments within the bursty events [17]. For example, negative words "害怕" ("scared") and "伤心" ("sad") emerge with a large volume after an earthquake occurs. Thus, we detect new bursty events by monitoring states of sentiments and topics in weibos. To detect the bursts in sentiment-aware topics, we propose to apply Kleinberg's [18,19] modeling of bursts to sentiment-aware topics discovered by TUS-LDA.

Compared with our previous work [15], there are three new contributions added in this work:

[1] "东方之星号客轮翻沉事件" ("Sinking of Dongfang zhi Xing"): The ship which was traveling on the Yangtze River in Jianli, Hubei Province with 454 people on board was capsized by a severe thunderstorm on June 1, 2015.

Please cite this article as: K. Xu et al., Detecting bursts in sentiment-aware topics from social media, Knowledge-Based Systems (2017), https://doi.org/10.1016/j.knosys.2017.11.007