Detecting Real-time Burst Topics in Microblog Streams: How Sentiment Can Help

Lumin Zhang, Yan Jia, Bin Zhou
National University of Defense Technology
Changsha, China
{zlm.nudt,bin.zhou.cn}@gmail.com

Yi Han
Peking University
National University of Defense Technology
Beijing, China
hanyi@nudt.edu.cn

ABSTRACT

Microblog has become an increasing valuable resource of up-to-date topics about what is happening in the world. In this paper, we propose a novel approach of detecting real-time events in microblog streams based on bursty sentiments detection. Instead of traditional sentiment orientation like positive, negative and neutral, we use sentiment vector as our sentiment model to abstract subjective messages which are then used to detect bursts and clustered into new events. Experimental evaluations show that our approach could perform effectively for online event detection. Although we worked with Chinese in our research, the technique can be used with any other language.

Categories and Subject Descriptors

H.2.8 [**Data Management**]: Database Applications—*Data Mining*

Keywords

Sentiment vector, Event detection, Burst, Microblog

1. INTRODUCTION

Millions of users are sharing their views and discussing current issues through microblog every day, which makes microblog become a new valuable social media for public opinion mining. Unlike traditional social media, microblog which messages emerge in high rate contains massive volume and too much noise, making it much more challengeable to detect online events in real-time streams.

An effective way to detect events is using bursty features in data streams and rich research has been conducted on this area. Early in 2002, for example, Kleinberg proposed a formal approach to extract meaningful documents based on modeling the stream using an infinite-state automaton in which bursts appear naturally as state transitions[2]. And in [1], a new temporal representation for text streams based on bursty features combining with TFIDF was proposed. All those methods, which are very useful on long documents like news or blogs, may encounter disadvantages in microblog which only contains 140 words. First, it will take a long time to detect bursty features in massive messages. Second, noisy messages contain many bursty variation of Chinese words which those approaches may not recognize effectively.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

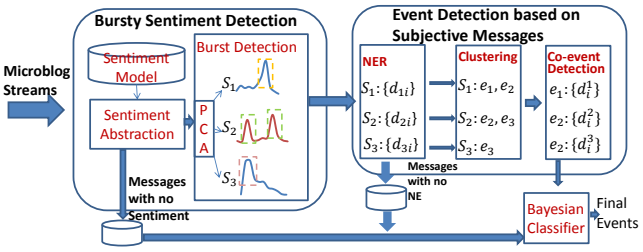

Figure 1: The framework of our model

Sentiment analysis(aka opinion mining) extracts subjective information from documents using natural language processing, computational linguistics and text analytics techniques. There is a strong correlation between bursty events and public moods, which means that there are always bursty sentiments with in the bursty events. For example, *scared* and *sad* emerge in large numbers if there is an earthquake. Inspired by this, we can detect new events by monitoring sentiment states in microblog streams. It is unnecessary to detect other bursty features, which makes online event detection available on massive data streams.

To the best of our knowledge, the work most related to ours is Nguyen's study[3], which also detected bursty events based on sentiments. The two methods, however, have some fundamental differences. Nguyen's method focused on long documents while we targets at microblog messages. Besides, we use a sentiment vector model in our early work in [4] with hierarchical structure to abstract subjective messages and the framework of our proposed model is different from theirs.

The rest of this paper is organized as follows. we present our model and event detection methods in Section 2, and report some selected empirical study results in Section 3.

2. MODEL AND METHODS

2.1 Problem Definitio

Let $D = \{d_1^{t_1}, d_2^{t_2}, d_3^{t_3}, \cdots\}$ is the microblog stream where t_i means the post time of message $d_i^{t_i}$, and $S =< S_1, S_2, \cdots S_m >$ is the sentiment vector proposed in [4] where S_j represents a sentiment. So the sentiment vector of a message d can be defined as $S_d =< \delta_d^1, \delta_d^2, \cdots \delta_d^m >$ where $\delta_d^k = 1$ if the message d contains sentiment S_k and $\delta_d^k = 0$ otherwise. For a period T and a certain sentimen S_i, we define $D_{S_k}^T = \bigcup_{d \in D^T} \delta_d^k$ as the messages sets which contains sentiment S_k, and $B^k =<$

$b_1^k, b_2^k, \cdots b_p^k >$ as the bursty periods of sentiment S_k where b_j^k means the j^{th} detected bursty period of sentiment S_k. Let $E = \bigcup e_i$ is the event set where $e_i = \{w_{i1}, w_{i2}, \cdots, w_{ip}\}$ is an event presenting by words. So our purpose is to detect new events in microblog stream D by using the features of bursty sentiments S.

2.2 Method Description

The framework of our method is illustrated in Figure 1. We first detect bursty sentiments in data streams. Then, we use the subjective messages to detect new events after three modules: Named Entity Recognition, Clustering and Co-event Detection. Finally, Bayesian Classification is performed to recycle the candidate messages into final events.

2.2.1 Bursty Sentiment Detection

We use the sentiment vector model proposed in our early work in [4] to perform sentiment abstraction. The model contains 284 Chinese words including new Internet words and common emoticons, and is automatically classified into 37 categories. For the messages containing no sentiments, like objective microblogs, we put them into a candidate set $setC_1$. For those subjective messages, we perform Principal Component Analysis to detect the main sentiments in time window T. For each main sentiments S_k, we detect its bursty periods B^k and the corresponding messages set $D_{S_k}^{b^k}$ using Kleinberg's methods proposed in [2].

2.2.2 Event Detection based on Sentiment Messages

For each $D_{S_k}^{b^k}$, Named Entity Recognition is performed, and the messages with no *time, location* will be put into candidate set $setC_2$. We then use spectral clustering to detect the events in each $D_{S_k}^{b^k}$. Here, we think the named entities weight higher than the other words , and use a parameter λ to adjust the weights while computing similarity between two messages as $Sim(d_i, d_j) = \lambda Sim_{NE} + (1-\lambda)Sim_{others}$. Finally, we integrate co-occurrence events. Event e_1 in sentiment S_i and event e_2 in S_j will be merged as one event only if $\exists k, s.t.(b_k^i \cap b_k^j)/(b_k^i \cup b_k^j) > \theta_1$ and $(e_1 \cdot e_2)/(||e_1|| \cdot ||e_2||) > \theta_2$ where θ_1 and θ_2 are thresholds.

2.2.3 Recycling

Messages in candidate $setC_1$ and $setC_2$ also contain some information about events, especially the objective messages in $setC_1$. In order to make precise event abstraction, we build a Naive Bayes classifier to absorb the information in the candidate sets. The event messages we have detected above could be considered as the training corpus. And a messages d in candidate sets belongs to an existing event $e_i$only if the $P(e_i|d) > \theta_3$.

3. EXPERIMENTAL RESULTS

3.1 Dataset

We worked with Chinese microblogs in SINA. By using API, we collected 3,923,641 messages from July 25th to Aug. 15, 2012 during the London Olympic Games.

3.2 Results and Discussions

In order to make real-time event detection, we set the time window $T = 1$ hour. Figure 2 shows the changes of two typical sentiments by hour. Here, we use the proportion of a certain sentiments in all subjective messages rather than the absolute number. We can see that the sentiment *happy* almost burst each day, yet *angry* burst only in certain times. In fact, each time China got a medal, there was a growth in sentiment *happy* and we really detected all the medal events during the Games.

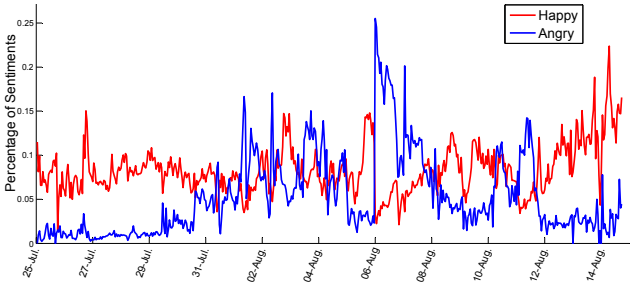

Figure 2: Changes of two typical sentiments by hour

We selected $top-5$ events for each time window as hot events. As manually examining all the events is prohibitively expensive, we chose 50 time periods randomly to evaluate the effects of our proposed model and the average precision is 86.3%

The framework of our model is quite efficient. In fact, almost 62% messages are objective messages, so most messages will be put into candidate set $setC_1$ until recycling in the last module. Besides, although there are 37 categories of sentiments in our sentiment vector model, only around 5 types are main sentiment after the process of principal component analysis on average. So it will not take a long time to detect bursts.

4. ACKNOWLEDGMENTS

This research is supported by the Major State Basic Research Development Program (No. 2013CB329600); the National Natural Science Foundation of China (No.60933005, 91124002); the National 863 Program (No.012505, 2011AA010702, 2012AA01A401, 2012AA01A402); 242 Program (No.2011A010); and the National Science and Technology Ministry (No.2012BAH38B04, 2012BAH38B06).

5. REFERENCES

[1] Q. He, K. Chang, and E.-P. Lim. Using burstiness to improve clustering of topics in news streams. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 493–498. IEEE, 2007.

[2] J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.

[3] T. Nguyen, D. Phung, B. Adams, and S. Venkatesh. Emotional reactions to real-world events in social networks. In *Proceedings of the 15th international conference on New Frontiers in Applied Data Mining*, PAKDD'11, pages 53–64, Berlin, Heidelberg, 2012. Springer-Verlag.

[4] L. Zhang, Y. Jia, B. Zhou, and Y. Han. Microblogging sentiment analysis using emotional vector. In *Cloud and Green Computing (CGC), 2012 Second International Conference on*, pages 430–433. IEEE, 2012.