A Novel Method for Online Bursty Event Detection on Twitter

Yu Zhang
School of Information Science and Engineering, Lanzhou University
Lanzhou, Gansu Province, China
zhangyu901007@163.com

Zhiyi Qu
School of Information Science and Engineering, Lanzhou University
Lanzhou, Gansu Province, China
quzy@lzu.edu.cn

Abstract—**As one of the most popular social media platforms, twitter has become a tool that people widely used to share their contents, their interests and events with friends. Meanwhile, we are facing a big challenge to find the bursty events from the large volume of continuous text streams quickly and accurately due to millions of data produced every day. In this paper, we proposed a BBW (Basic-Burst Weight) method based on the Time Window to extract bursty words, then we exploit these bursty words to detect the meaningful bursty events combined with hierarchical clustering algorithm. Our experiments on a large twitter dataset show that our method can detect bursty events timely and precisely.**

Keywords-Twitter; time window; burst words; hierarchical clustering algorithm

I. Introduction

In recent years, social media has become an important source of real-time information on internet. The most typical and representative social media platform is online social network (OSN), such as Twitter, on which the post is limited to a small size and the social information are propagated rapidly. The information on Twitter has the following characteristics:

- brevity -- Twitter limits tweet length to 140 characters ;
- diversity-- In the view of information modality, tweet may consist of short textual sentences, individual images, or video links, and in the content, almost every topic, such as politics, economy and entertainment, can be propagated on twitter ;
- fast spreading -- contents are updated timely and can be retweet frequently.

In traditional media, only those professionals can report news with professional tools which has been proved to be a difficult way. While twitter is more convenient and efficient compared with traditional media, and anyone can deliver tweets on twitter no matter whether he/she is associated with the media industry or not. Twitter thus makes everyone like a self-media node connecting to the Internet, and presents a highly effective way to discover what is happening around the world or what happened in recent hours.

An important problem then arises: How to detect online bursty events from a large scale of information on twitter? A bursty event generally means an online event happen immediately and spread quickly in a short time. Bursty event may involve nature disaster, accidents disaster, public health incidents or social safety incidents that need to take emergency handling measures[1] .With the definition of bursty event, we can see that it is significant to detect bursty events from large amount of tweets. Due to the wide variety of purposes and topics, including daily chatter, conversations, sharing information, reporting news, a rumor, spam and advertisement. There are serious noises such as worthless words which has no contribution to discover online bursty events from the information of twitter. Moreover, to detect the bursty event need to analyze the change of retweet contents over time at a very small temporal scale.

In this paper, we propose a novel method called BBW for the online bursty event detection. BBW method aims to find the keywords with high contributions in the bursty event detection at a small temporal scale from the tweets of twitter. Therefore, we simplify the problem into extracting keywords from tweets based on time window and that is so called bursty words. In general, the bursty events are represented by a group of related bursty words with the bursty temporal information. But in fact, we need to use clustering method to form events. So we combine BBW method with hierarchical clustering algorithm to discover events. The experiments show that our method can present a better performance.

The rest of this paper is organized as follows: Section 2, related works will be introduced. Section 3, the method to extract bursty words. Section 4, introduce the clustering algorithm we used to detect bursty events. Section 5, show the experimental results. Section 6, give conclusions and future work.

II. ReLATED WORKS

Recent years, the widely used of social media makes huge changes, not only for users, but also for research area. Sharing and discussing conveniently over huge amount of users will makes those bursty events fermenting and spread quickly, so bursty event detection over social media platforms like Twitter and microblog becomes hot research topic. Whereas, this kind of new media that microblog/Twitter also brings some challenges to TDT(Topic Detection and Tracking) research area due to its variety of differences with traditional web-based

978-1-4799-8353-7 /15/$31.00 ©2015 IEEE

blog. While most of the research can be divided into two categories: one is based on microblog text, and the other is based on bursty features.

The methods based on microblog text will cluster bursty events in a time window, and the extract bursty features from the events. The problem of this method is that it will performs good on traditional blogs because it have longer contents than microblog or Twitter which have length limit on text content, another issue is that the microblog text or tweet is usually colloquial, so some researches tried to introduced some classical topic model like LDA(Latent Dirichlet Allocation)model, but they need to do some improvements because of those existing models cannot be used to detect bursty events that experience a sudden increase[2] in a short period. Meanwhile this kind of methods usually rely on the choice of hidden topic amount.

The methods based on bursty features need to extract bursty features first and then do the cluster work to get the bursty events features. This methods have high dependency on the extraction of bursty features words.

Some other researches want to find some new solution, utilizing the method of signal processing that translate the bursty words into a series of timing signals and then detected bursty events with frequency domain signal and threshold after carry on the Discrete Fourier Transform, but it is complex and needs more mathematical knowledge.

Our work is based on the second method, and our mainly work is the BBW method introduced below which could provide higher accuracy while extracting bursty features words, then we could gather more reliable bursty events with these features.

III. BBW

A. Notation Description

We first introduce the notations used in this paper and formally formulate our problem. First, we define **TW$_1$, TW$_2$,...TW$_n$** as the sliding Time Window list and n is the total number of time windows that we consider. During each time window **TWn**, we capture the dataset of tweets denoted by **D$_n$** , and then we have the text sets **D$_1$, D$_2$, ..., D$_n$** which correspond to the time windows, respectively.

B. Bursty Words Detection

In our method we use BBW to detect bursty events which means that the weight of a bursty keyword consist of two parts that is basic weight and burst weight.

TF-IDF is the classical method to calculate the term's weight in text representation. However, it cannot be applied to detect bursty words directly for the brevity characteristic of bursty event. So we use the method proposed in[3] to calculate the basic weight of each term and make several changes as the following equation[3] .

$$w_{t,n} = 0.5 + 0.5 \times \frac{tf_{t,n}}{tf_n^{max}} \qquad (1)$$

where $w_{t,n}$ is the basic weight of term t in **TW$_n$**, $tf_{t,n}$ is the frequency of term t in **TW$_n$**, tf_n^{max} is the maximum frequency of all the terms in **TW$_n$**.

Using $w_{t,n}$ we can give bigger weight to those terms which have high frequency and can distinguish from common words.

For the burst weight we adopt the method proposed in[4] [5] , it is defined as:

$$B_n(t) = \frac{f_n(t) - \mu_n(t)}{\mu_n(t)} \qquad (2)$$

Where $B_n(t)$ represent the burst extent of term t, $f_n(t)$ and $\mu_n(t)$ are the frequency and mean frequency of term t in **TW$_n$**, respectively. A common understanding of $B_n(t)$ is the incremental rate of term t. So if $B_n(t)$ is larger than zero, it means the term t is in grow stage which means the occurrence number of term t in **TW$_n$** is more than **TW$_{n-1}$**. On the contrary, if $B_n(t)$ is below zero, which means the term t is in descent stage. Therefor $B_n(t)$ is the crucial factor to detect bursty keywords timely.

In general, a typical bursty event needs a number of keywords with high frequency in the current time window and with large increase of frequency between two neighboring windows. Thus, to combine basic weight and burst weight is a reasonable strategy to detect the keywords with high contribution in discriminate the bursty events. We linearly combine both kinds of weights as follows by introducing a parameter λ

$$w_n(t) = \lambda w_{t,n} + (1 - \lambda)B_n(t) \qquad (3)$$

Here, λ is a factor to control the balance between basic weight and burst weight.

C. Hierarchical Clustering Algorithm

After the detection of bursty keywords, we use clustering algorithm to divide the bursty words into different groups, where each group is the set of keywords that are with high correlation and perhaps corresponds to a event. After that, we need to select the right group as a bursty event. The clustering results determine the accuracy of bursty event detection ultimately.

Hierarchical clustering algorithm is one of the most popular clustering algorithms in data mining and it is a bottom-up algorithm. The main idea of the algorithm can be describe as: treat each term a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all terms.

In this paper, we first divide all tweets into different time windows chronologically; In each time window we extract the bursty words accroding to Eq.3; Finally, we let every bursty word be the initial cluster and implement the hierarchical clustering algorithm[6] to obtain the final cluster results. The algorithm is summarized as follow:

Input: bursty words captured by BBW method

Output: bursty event cluster

Step1. Initialize each word as a cluster, and we will get N clusters for N bursty words input;

Step2. Build a N*N co-occurrence matrix ***C*** for all these words and then switch it to distance matrix, So ***C[i][j]*** represent the distance between term ***i*** and ***j***;

Step3. Get a cluster ***pair (r, s)*** by finding the nearest distance in the matrix C, which means we just need to find the minimum ***C[r][s]*** of ***C,*** then we merge this two clusters into one cluster ***(r, s)***;

Step4. Update the distance matrix by calculating the average distance between the new cluster and each other clusters based on ***C[k, (r, s)] = avg(C[k, r], C[k, s])***

Step5. Repeat step 3 and 4 until all the words are clustered into one cluster;

Step6. Traverse the cluster tree from the top, if the distance value is smaller than our predefined threshold, then set this cluster as a bursty event cluster, else traverse its subtrees until all bursty event cluster were found out.

The detected bursty event cluster can be represented by a number of bursty words, which can show the main semantics of the bursty events well. For example, the cluster {Sea, Russia, Okhotsk, trawler} represent the event of Russia trawler wrecked in the Sea of Okhotsk.

IV. EXPERIMENTAL RESULTS

A. *Eeperimental Setting*

We build a dataset by collecting tweet messages from twitter using the Twitter API[7] . The dataset contains 91,8124 messages From April 1, 2015 to April 16, 2015. The tweets contain both Chinese and English.

Our experimental environment is a normal laptop, using Java language above windows 8(64-bit), 8GB memory and i5-4210M CPU, with the database management system MySQL. We use the Chinese Lexical Analysis System of Institute of Computing Technology (ICTCLAS)[8] to segment Chinese tweets and Stanford POS Tagger0to segment English tweets.

B. *Threshold Testing and Choose*

In our experiment, λ was empirically set to 0.5 and time window was set to six hours. In the hierarchical clustering algorithm, the threshold value θ make a great influence on the cluster result. Table 1 shows the precision and recall when the value of θ was set to 1.0, 1.1, 1.2, 1.3, 1.4, 1.5.

$$Precision = \frac{P}{S} \quad (4)$$

$$Recall = \frac{P}{T} \quad (5)$$

Where P represent the number of bursty events we detected correctly and S represent the total number of bursty events we detected. T is the actual number of bursty events of human judgment.

TABLE I. IMPACT ON THE RESULTS OF DIFFERENT THRESHOLDS OF θ

θ	Precision/%	Recall/%
1.0	55.2609	66.7
1.1	60.3416	75.0
1.2	68.1731	83.3
1.3	64.5885	83.3
1.4	58.9326	83.3
1.5	54.5575	83.3

From Table I we can see that when θ was too small or too big, it will lead to low precision for the reason that θ was too small the words of the cluster will too few to describe an event while θ was too big will lead to different bursty event clutering together. So both make low accuracy.

Finally, we set θ to 1.2 and get seven bursty events as shown in Table II. For the reason that the dataset contain both Chinese and English, so our result also contains two kinds of language. We use multiple bursty words to describe each event. We also extract bursty time of the bursty event at the same time. For example, 2015-04-02 00:00:00 represents that the bursty time of the corresponding event is between 00:00:00 and 06:00:00 on 2015-04-02, and similarly 06:00:00 represents the bursty time is between 06:00:00 and 12:00:00.

TABLE II. BURSTY KEYWORDS CLUSTERS OF EVENTS WITH BURSTY TIME

Event number	Bursty time	Bursty words description
1	2015-04-01 18:00:00	战斗机(fighter), 机场(airport), 台南(Tainan)
2	2015-04-02 00:00:00	平台(platform), 墨西哥(Mexico), 石油(oil)
3	2015-04-02 06:00:00	海域(sea), 俄罗斯(Russia), 霍次克(Okhotsk), 渔船(trawler)
4	2015-04-02 18:00:00	人质(hostage), 枪手(gunner), 肯尼亚(Kenya), 加里萨(garissa)
5	2015-04-04 12:00:00	农药(pesticide), 王府井(wangfujing), 北京(Beijing)
6	2015-04-0618:00:00	周边(round), 装置(device), 芳烃(aromatics), 项目(project)
7	2015-04-14 18:00:00	航空(Asiana), 客机(passenger plane), 广岛(Hiroshima), 跑道(runway)
8	2015-04-01 18:00:00	Fighter, jet, emergency, land, Tainan,
9	2015-04-02 00:00:00	Mexico, oil, platform
10	2015-04-02 06:00:00	Sea, Russia, Okhotsk, trawler
11	2015-04-14 18:00:00	Asiana, A320, skid

After all bursty events clusters were detected we will find out the tweet which is most relevant to the bursty event using a simple weighted matching algorithm. The tweets can describe the bursty event in detail. All the tweets of the detected events were shown in Table III.

TABLE III. REPRESENTATIVE TWEETS CORRESPONDING TO BURSTY EVENTS

Event number	Representative tweets of the detected bursty event
1	美国空军的两架F-18战斗机周三午后降落在台湾的台南机场，美国在台协会说这是因其中一架飞机机械故障而迫降 (Two US F-18 fighter jets made an emergency landing at an air force base in the southern city of Tainan on Wednesday, with US authorities saying one of the planes had developed a mechanical failure)
2	墨西哥一石油钻井平台发生火灾致4人死亡… (An explosion and a fire erupted on an offshore oil platform operated by Mexico's Pemex, killing at least four workers)
3	俄罗斯大型渔船沉没已致54死、15失踪：鄂霍次克海域今天凌晨传出渔船沉没，已知造成54人死亡 (At least 54 sailors have been killed and 15 people missing after a Russian fishing trawler sank in the Sea of Okhotsk)
4	中国新闻网肯尼亚加里萨一所大学遭袭15人死亡65人受伤… (A university of Kenya garissa was attacked and 15 people were killed, 65 wounded which is reported by China news)
5	4月4日11时许，北京公安局东城分局执勤民警在王府井大街丹耀大厦门前发现30余人躺倒在地. . . (On April 4, 11 AM, dongcheng bureau of Beijing public security bureau police on duty in wangfujing street Dan yao found more than 30 people lying on the ground in front of the building)
6	4月6日18时58分，福建漳州消防支队指挥中心接到报警，古雷腾龙芳烃PX项目联合装置区发生爆. . . (At 6:58pm Apirl 6th , fujian zhangzhou fire detachment command center received a call, gu lei feng aromatics PX project joint device exploded)
7	【韓亞航空一架A320客機在日本沖出跑道致23人受傷】14日晚18時49分，一架從韓國仁川國際機場出發前往日本廣島的韓亞航空民航客機在著陸過程中沖出跑道… (Asiana A320 Skids After Landing at Hiroshima, Japan; 23 Hurt: At 6:49 p.m. April 14th, a passenger plane of South Korea's Asiana Airlines skidded off the runway when landing at the Hiroshima Airport in Japan...)
8	Two US F-18 fighter jets made an emergency landing at an air force base in the southern city of Tainan on Wednesday, with US authorities saying one of the planes had developed a mechanical failure
9	An explosion and a fire erupted on an offshore oil platform operated by Mexico's Pemex on Wednesday, killing at least four workers, injuring 16 and forcing 300 to be evacuated.
10	At least 56 sailors have been killed after a Russian fishing trawler sank in the Sea of Okhotsk
11	V. Asiana A320 Skids After Landing at Hiroshima, Japan; 23 Hurt

We calculate the change of the amount of tweets related to every bursty event over time. Figure 1 shows the trend of the related tweets of every hour from 2015-04-02 00:00:00 to 2015-04-02 23:00:00 for event 10. From the trend we can see that there is a sharply increasement between 10:00:00 and 11:00:00 and our method can also detect the bursty event during this time. In other words, our method can detect bursty events timely and effectively.

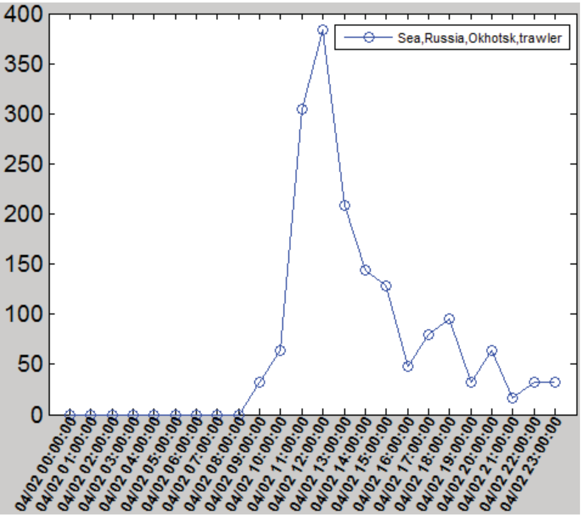

Figure 1. Trend of the number of related tweets of event 10

V. EXPERIMENTAL RESULTS

In this paper, we proposed a BBW method which can extract bursty words accurately, including some basic preprocessing above the microblog text with our design filtering rule, and then we made some improvements on traditional method TF-IDF. After that we could detect bursty events with our hierarchical clustering algorithm. What's more, we can detect the bursty time interval of the event.

With some experiments, we can find our method works well. In the future, we will try to find better way to reduce the time complexity and make our system arrive a higher level in TDT research area. Finally we want to make a real-time online detection with higher accuracy and lower time complexity.

ACKNOWLEDGEMENTS

The work is partly supported by Key Projects in the National Science & Technology Pillar Program (No. 2011BAK08B02).

REFERENCE

[1] http://baike.baidu.com/subview/39487/5131813.htm

[2] Qi X, Huang Y, Chen Z, et al. BURST-LDA : A NEW TOPIC MODEL FOR DETECTING BURSTY TOPICS FROM STREAM TEXT[J]. Journal of Electronics, 2014, 31(6):565-575.

[3] Gerard Salton, Chris Buckley. Term-Weighting Approaches in Automatic Text Retrieval.[J]. Information Processing & Management An International Journal, 1988, 24(88):513–523.

[4] Zhang Z, Xu M, Zheng N. Mining burst topical keywords from microblog stream[C]// Computer Science and Network Technology

(ICCSNT), 2012 2nd International Conference on. IEEE, 2012:1760 - 1765.

[5] Finch, “T. Incremental calculation of weighted mean and variance,” University of Cambridge, 2009.

[6] http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html

[7] http://www.codecademy.com/tracks/twitter

[8] http://ictclas.nlpir.org/

[9] http://nlp.stanford.edu/software/tagger.shtml