Clustering Based Topic Events Detection on Text Stream

Chunshan Li[1,2], Yunming Ye[1,2], Xiaofeng Zhang[1,2],
Dianhui Chu[3], Shengchun Deng[3], and Xiaofei Xu[3]

[1] Harbin Institute of Technology, Shenzhen Graduate School, China
[2] Shenzhen Key Laboratory of Internet Information Collaboration, China
[3] Department of Computer Science, Harbin Institute of Technology
lichunshan.hit@gmail.com,
{yeyunming,chudianhui,dengshengchun,xuxiaofei}@hit.edu.cn,
zhangxiaofeng@hitsz.edu.cn

Abstract. Detecting and tracking events from the text stream data is critical to social network society and thus attracts more and more research efforts. However, there exist two major limitations in the existing topic detection and tracking models, i.e. noise words and multiple sub-events. In this paper, a novel event detection and tracking algorithm, topic event detection and tracking (TEDT), was proposed to tackle these limitations by clustering the co-occurrent features of the underlying topics in the text stream data and then the evolution of events was analyzed for the event tracking purpose. The evaluation was performed on two real datasets with the promising results demonstrating that (1) the proposed TEDT algorithm is superior to the state-of-the-art topic model with respect to event detection; (2) the proposed TEDT algorithm can successfully track the event changes.

Keywords: social media, event detection, temporal analysis, topic model.

1 Introduction

With the social networking society blossomed into pervasive aspect of human beings, there exist a huge volume of text stream, such as blog posts and tweets. However, the emergence of the important events generally was quickly overwhelmed by the huge amount of non-important events due to the imbalance volume of posts. Therefore detecting these imbalanced but important events has become a hot research issue. Traditional event detection techniques generally involve two major approaches: topic detection and tracking (TDT) [1,2,3] and temporal text mining [4,5].

To detect and track events, text stream data is generally defined as a sequence of chronologically ordered documents, and topic event is a set of co-occurrent words within a short period of time. For example, the 2008 presidential campaign in the USA was reported in several news stories, which could be seen as the text stream data. Naturally, these documents were divided into several subgroups based on their timestamp. Then events detection and tracking algorithm groups words in each time window, e.g. words "obama, speech, iowa", appeared in t, and words "obama, hillari, support, vote" appeared in $t+1$. Subsequently, event words in different time windows could be grouped together into several specific events, i.e. (1) "Barack Obama speak at a campaign rally", (2)"Hillary Clinton leads the list of most admired women with 20%, according to the

N.T. Nguyen et al. (Eds.): ACIIDS 2014, Part I, LNAI 8397, pp. 42–52, 2014.

vote". In the literature, different methods have been proposed to detect events. The topic model [6] was adopted to detect topic by generating multiple topics from text corpus at the same time and each topic, consisting of several group of keywords, can be treated as one event. However, there exist two major drawbacks if we adopt topic model to detect events. First, topic model constructs an event by choosing top k topic keywords manually and thus brings the noisy words. Second, a topic generally includes more than one event. For example, topic model discovers an event containing words like "attack, soldier, militari, luzon, provinc, troop, southern, kill, area, libyan". In this event, it is easily to extract two sub-events: "Libyan people were killed" and "soldier were attacked at island in the southern Luzon".

In this paper, we investigated the problem of detecting and tracking topic event by extending topic model, and proposed topic event detection and tracking(TEDT) algorithm. The main difference between the proposed TEDT algorithm and topic model is how an event is generated. The topic model assumes that an event can be represented by a group of words having similar semantic topics. The proposed TEDT algorithm assumes that the co-occurrence relationship and the semantic topic of words work together to generate an event. In particular, for the TEDT algorithm, the topic features are clustered to generate probabilistic events using a revised topic based approach. The remaining of the paper is organized as follows. Section 2 reviews the related works. The TEDT algorithm is proposed in Section 3. Experiments and evaluation results are given in Section 4. Section 5 concludes the paper.

2 Related Work

The event detection generally involves with several research domains, such as topic detection and tracking [7,8,9], text clustering [10,11,12,13,14] and temporal data analysis. The *topic detection and tracking based approach* is originally proposed to discover the topic hidden in the stream of news stories [7]. Then, Loulwah et al. employed KL-divergence to measure difference among topic over timeline and a threshold was used to detect the occurrence of events [15]. Mei et al. proposed mixture topic model to extract themes (hidden events) [8]. However, these topic-based approaches pay less attention to co-occurrence relation of corpus words and thus either make the generated topics contain noisy words or include fewer sub-topics.

The *clustering based approach* is developed for text mining in its own right and is now adopted for event detection and tracking. To cope with the huge amount of data, authors in [12] treated the words and their co-occurrence as a network and the graph partition algorithm was adopted to discover the densely connected subgraphs as the events. Lin et al. proposed a combined model for event detection from text and community features [10]. In [14], Twitter tags and spatio-temporal features worked together to discover events. Yao et al. [11] proposed to detect the burst of single word and their approach could be extended for the detection of multiple words but at much higher cost. Fund et al. [13] designed a parameter free word clustering approach, called HB-Event, to detect event, which clustered words by co-occurrence relation among words in corpus.

3 Topic Event Detection and Tracking

Although these existing approaches did not resolve aforementioned limitations, they motivate us to extend current topic model by filtering noisy features and unimportant sub-events. In this section, we first formulate the problem of topic event detection, then propose the detailed process to detect topic events. Finally, we discussed how to tract the change of detected events by topics of the events.

3.1 Problem Formulation

Given the text stream data $D = \{d_1, d_2, \ldots, d_i, \ldots\}$, where d_i is a document containing a unique timestamp t_i. All words in d_i are extracted out to form a vocabulary W, where $W = \{w_1 w_2 \ldots\}$. In D, if $i < j$, then $t_i \leq t_j$. The text stream D can be split into L non-overlapping time windows. The topic events detection is trying to find a set of topic events within each time slot, and a topic event in l_i consists of a set of topic features with the highest co-occurrence probability of extracted words. The tracking of topic events is to identify the related events in all timeslots. The Figure 1 explains the proposed word clustering based approach. We first proposed the probabilistic clustering approach to discover the topic events in text stream. Then, we utilized topics of events to track the change of events.

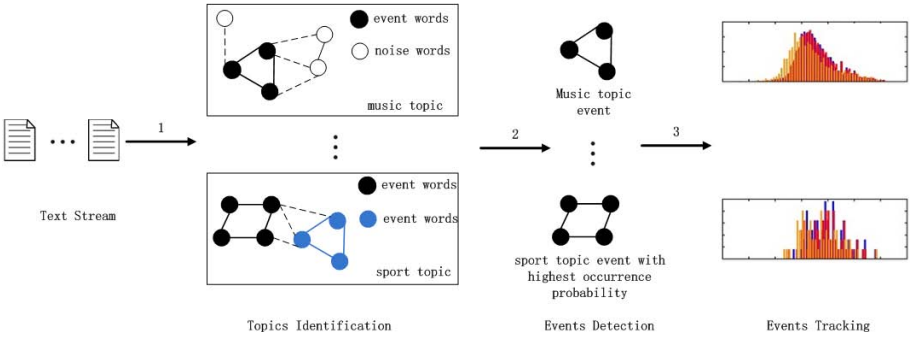


Fig. 1. The Overview of Topic Event Detection and Tracking

3.2 Topic Events Detection

After running online topic model, several sets of topic features were generated in each timeslot. A feature set of topic i can be defined as a word sequence $W_i = \{w_0, \ldots, w_m, w_n, \ldots\}$, in which, if $m < n$, then $probability(w_m|T_i) > probability(w_n|T_i)$. In traditional event detection method, the top k words in W_i are considered as a topic event. To filter the noisy features, these features in W_i will be clustered to form more accurate topic events. Details about the clustering procedure are described as follows.

Let the feature set of topic i be $W_i^l = \{w_0, w_1, ..., w_n\}$, l represents the current time, n is the number of words. A topic event E_i^l in l consists of several topic features and can be defined as $E_i^l = \{\ldots, w_i, w_j, \ldots\}$. E^l represents the set of detected events in l. For example, suppose $W_i^l = \{moive, music, titanic\}$, $E_i = \{moive, titanic\}$ implies that the words in event $\{moive, titanic\}$ have higher co-occurrence probability than $\{music\}$, $\{moive\}$ and $\{titanic\}$ in topic i. In order to find the optimal feature group in W_i^l, we formalize the problem of determining the cluster of topic features under a probabilistic framework, which is to discover occurrent event E_i^l with the highest probability. The objective function is defined as: (see Table 1 for the adopted notations).

$$\begin{aligned} &\max P(E_i|D,T_i) = \max\{P(\mathrm{w}_0,\mathrm{w}_1,...,\mathrm{w}_k|D,T_i)\} \\ &= \max\{\textstyle\prod_{m=1}^{M}\prod_{n}^{|E_i|} f_{m,n} p(\mathrm{w}_{m,n}|z_{m,n,i}) p(z_{m,n,i}|d_m) p(d_m)\} \end{aligned} \tag{1}$$

where $z_{m,n,i}$ represents the topic of word $w_{m,n}$ and $f_{m,n}$ is the frequency of $w_{m,n}$.

Table 1. Notations used in the TEDT algorithm

Parameter	description	Parameter	description
T_i	topic i	W_i^l	feature set of topic i in timeslot l
E_i	event i	D	the text corpus
d_m	a document in D	$w_{m,n}$	the word in document d_m for event n
$z_{m,n,i}$	the topic of word $w_{m,n}$		

Taking the logarithm, then the maximization of Eq. 1 is equivalent to maximize:

$$\begin{aligned} &\max\{\log P(E_i|D,T_i)\} \\ &= \max\{\log(\textstyle\prod_{m=1}^{M}\prod_{n}^{|E_i|} f_{m,n} p(w_{m,n}|z_{m,n,i}) p(z_{m,n,i}|d_m) p(d_m))\} \\ &= \max\{\textstyle\sum_{m}^{M}\sum_{n}^{|E_i|} \log(f_{m,n} p(w_{m,n}|z_{m,n,i}) p(z_{m,n,i}|d_m) p(d_m))\} \\ &= \max\{\textstyle\sum_{m}^{M}\sum_{n}^{|E_i|} (\log f_{m,n} + \log p(w_{m,n}|z_{m,n,i}) + \log p(z_{m,n,i}|d_m) \\ &\quad + \log p(d_m))\} \end{aligned} \tag{2}$$

Eq. 2 defines the priority of events. In Eq. 2, the document-topic distribution $p(z_{m,n,k}|d_m)$ and topic-word distribution $p(bw_{m,n}|z_{m,n,k})$ will be computed by online topic model [16], which is guaranteed to converge and is able to find comparably precise topics as the original topic model does. Through the preliminary experiments, the number of the topic in our experiments is set to be 20 to get the optimal results.

The proposed TEDT is illustrated in Algorithm 1. The algorithm returns a list of events $\{E_1, \ldots, E_k\}$, where k is the number of topic. The following example will show how the TEDT works with event detection. Suppose $W_i = \{moive, music, titanic\}$, then $p1 = probability(moive|D,T_i)$, $p2 = probability(music|D,T_i)$, $p3 = probability(moive, music|D,T_i)$ is calculated according to the Eq. 2. If $p1 > p3$ and $p2 > p3$, then $\{moive\}$ and $\{music\}$ are considered as two independent events.

If $p3 = probability(moive, titanic|D, T_i)$, which means term "titanic" is considered, then we have $p3 > p1$, $p3 > probability(music, titanic|D, T_i)$ and $p3 > probability(titanic|D, T_i)$. Obviously, the event $\{moive, titanic\}$ has the highest co-occurrence probability in topic i, which is the target event to be extracted out.

Algorithm 1. TEDT algorithm

Require: a set of topic feature TW_i, the text corpus D,
a document-topic distribution matrix $\theta_{M\times T}$,
a topic-word distribution matrix $\beta_{T\times W}$
Ensure: topic event E_i
1: **for** each $w_j \in W$ **do**
2: $p_1 = log(p(E_i|D, T_i))$
3: $p_2 = log(p(w_j|D, T_i))$
4: $E_i.add(w_j)$
5: $p_3 = log(p(E_i|D))$
6: $p = max\{p_1, p_2, p_3\}$
7: **if** $p_1 == p$ **then**
8: $E_i.remove(w_j)$
9: **else if** $p_2 == p$ **then**
10: $E_i.clear()$
11: $E_i.add(w_j)$
12: **end if**
13: **end for**

In section 3.2, k topic events were already extracted out which verifies the assumed existence of one to one mapping relationship between topic events and topic. In [16], Blei et al. demonstrated that topics in the online topic model is consistent over the timeline, i.e. topic i in l contains correlated content to topic i in $l + 1$. Therefore, correlated topic events can be tracked by the consistency of topics.

4 Empirical Study

To evaluate the performance of the proposed TEDT, the online topic model and the Hot-Bursty-Event detection (HBE) approach [13], are implemented for performance comparison. In online topic model, the top 10 words in topic feature set are considered as an event. The HBE algorithm only considers the co-occurrence relation between topic features to generate topic event. Two real data sets, named *reuters* and *blog*, are used in the experiments. The *reuters* data set contains all news stories collected from 1987-2-26 to 1987-10-20. The general preprocessing steps are performed on *reuters*, such as stopword removal and stemming. Words whose term frequency is less than 3 were filtered out. After all these steps, the data set consists of 19,065 documents and 19,644 distinct words. The *blog* data set contains blog posts collected from 2008-1-1 to 2008-3-5. There are 2 categories in the *blog* which are "technologies" and "politics". After similar pre-processing steps, the date set used in the experiments is composed of 31,831 documents and 48,747 distinct words.

4.1 Performance Evaluation for Event Cohesiveness

Measuring the quality of detected events is a difficult task. Many works employed manual judgement to determine the performance of event detection approach [11]. However, such judgement is extremely expensive and time-consuming, thus is infeasible in reality. Yao et al. [12] proposed a criterion called "bursty cohesiveness", which estimates the performance of approach by quantifying the co-occurrent event words. Similar to Yao's approach, a new metric, called event cohesiveness, is designed to evaluate the quality of detected events. The event cohesiveness assumes event words are generally co-occurrent in the same document and their document frequency is also high enough. By removing these highly co-occurrent words, the event frequency drops quickly. If the co-occurrence of words in an event is higher, then the event is more cohesive. The event cohesiveness is therefore defined as:

$$conhesive(E_i) = \sum_{k}^{|E_i|} (1 - \frac{\text{independence}(bw_k)}{freqd(bw_k)}) \tag{3}$$

where E_i represent a specific event, w_k is an event word in E_i. $\text{independence}(w_k)$ indicates the frequency of w_k after removing the co-occurrent words, and $freq(w_k)$ is the frequency of w_k. Subsequently, the cohesiveness of detected event within a timeslot can be estimated as

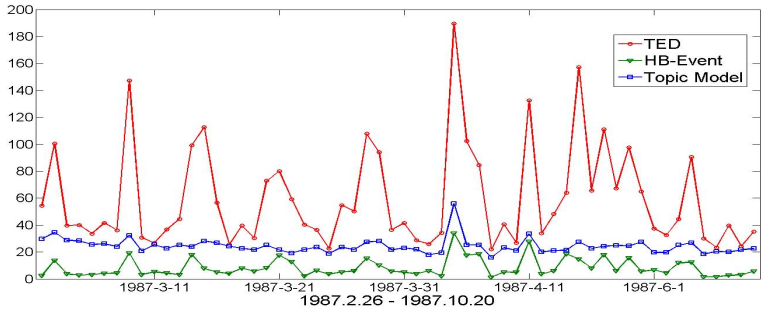
$$average\_Conhes(E) = \sum_{E_i \in E} \frac{conhesive(E_i)}{|E|}. \tag{4}$$

4.2 Experiments on Event Cohesiveness

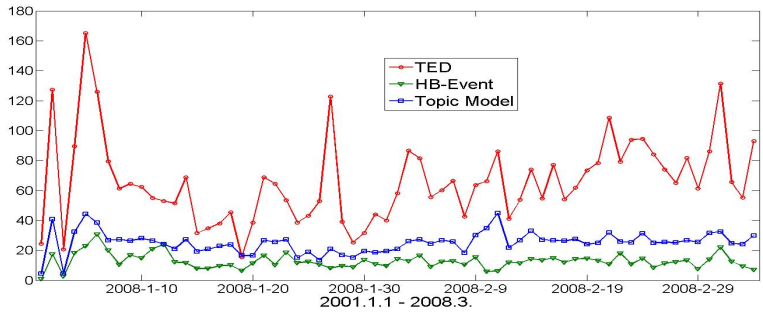
In this section, the event cohesiveness is used to evaluate the quality of detected event topics. The cohesiveness of both online topic model and HBE algorithm [13] will be calculated for the comparison. The experimental results are plotted in Figure 2. As can be seen, the value of the TEDT is always higher than that of the HBE and online topic model, which indicates that the TEDT is more cohesive. In general, the TEDT is more sensitive to filter the noisy features and non-important events in a topic. For example, the TEDT achieves a more bigger cohesiveness value on 1987-3-20, and event "the attack in Luzon" is detected with its event words "attack, armi, luzon, davao, rebel", whereas the online topic model detects two events "the attack in island" and "soldier was killed in Libyan". The reason why the TEDT can achieve a much better performance is that the online topic model only utilized semantic topic to find events, whereas the HBE which only consider the co-occurrence of words detected "attack, island, province".

4.3 Experiments on Visualization

To better understand the performance of the algorithms in detecting topic events, the detected topic events are visualized for the performance comparison with node represents event word, edge represents the co-occurrence between words, the size of node represents the word's frequency, and the width of the lines represents the strength of

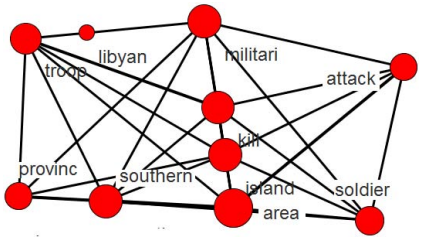


(a) event cohesiveness on reuters dataset

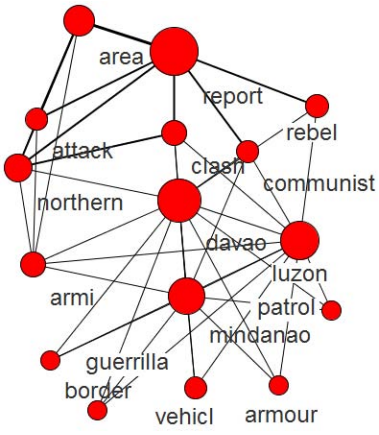


(b) event cohesiveness on blog dataset

Fig. 2. The event cohesiveness of two real datasets

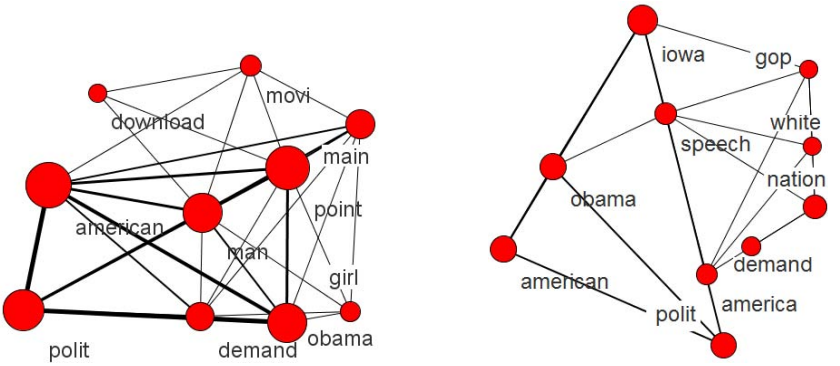


(a) topic model



(b) TEDT algorithm

Fig. 3. Event releasing fight between Philippine government and rebels



(a) topic model (b) TEDT algorithm

Fig. 4. Event releasing fight between Philippine government and rebels

co-occurrent words. Figure 3 reports the topic event "fight between Philippine government and rebels in the southern Philippine island at 1987". The proposed TEDT detects location of event, "luzon, davao", as well as key words of event "rebel, attack, armi", whereas the online topic model finds two events in one topic, which are "the attack in island" and "soldier was killed in Libyan". The result shows that the TEDT can achieve more accurate event, whereas the online topic model may uncover multiple events in one topic.

Figure 4 reports the ability of the proposed TEDT algorithm to filter out the noisy features of topic. In Figure 4(a), online topic model detects several noisy words, such as "movi, download" and few event words, "obama, american". At the same time, the proposed TEDT algorithm finds "speech, american, iowa, obama", which are closely related to target event.

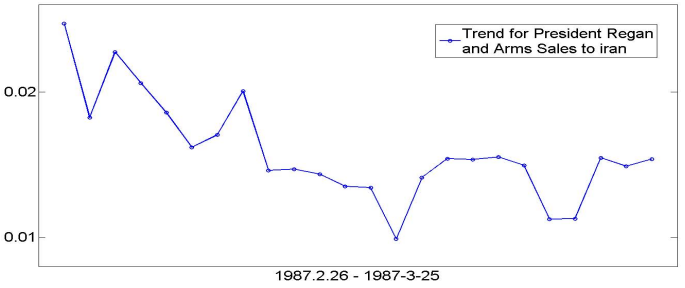


Fig. 5. Hotness for "Iran-Contra and President Regan"

Table 2. Topic Events Tracking On a given Topic

timeslot	events
1987-2-26	report,profit,control,republican,secretari,regan,iran,hous, analyst,arm, white,protect,washington,expect,shultz,person, polit,record,strategist,week
1987-3-1	poll,reagan,presid,conduct,person,three,disapprov,rate, pct,major,newsweek,magazin,iran
1987-3-2	presid,hous,white,gate,chief,staff,baker,senat, profit,offici,nomin,regan,nation,scandal,resign,tower, commiss,sale,record
1987-3-16	arm,presid,hous,iran,walsh,scandal,presidenti,secretari, rebel,white,probe,million,testimoni,tent,affair, chief,senat,wit,contra,georg,profit
1987-3-17	reagan,rebel,arm,hous,contra,iran,presid,white, report,novemb,aid,divers,profit
1987-3-18	reagan,presid,rebel,aid,hous,iran,contra,senat, militari,report,congress
1987-3-19	reagan,presid,aid,white,hous,contra,affair,georg, report,scandal,constitut,profit,told,john
1987-3-23	iran,presid,secretari,missil,radio,report,weinberg,tehran, ship,threat,militari,hawk,iranian
1987-3-24	iran,missil,reagan,strait,hormuz,silkworm,kuwaiti,chief, white,gulf,ship,report,warship,escort,defenc,radio, hous

4.4 Tracking on Topic Events

In this experiment, we will show the ability of the TEDT to track the event changes. Given a topic k, the TEDT first detects all topic events whose topic is k. We only show the results when the topic is "Iran-Contra and President Regan" and the event changes are tracked over all time slots. Part of the tracked events are recorded in Table 4.4.

As can be seen, all event changes related to "Iran-Contra and president Regan" have been detected in the period of ("1987-2-26"–"1987-3-24"). To measure the hotness of an event, the rate of topic k over all topics at each time slot is captured and is plotted in Figure 5. From the figure, it can be seen the tense situation of the Iran-Contra denoted by the rate of topic. Moreover, the evolutionary events can be well tracked. For example, the event "Reagan government sales arms to Iran" is found from the data extracted on "1987-2-26" which is the $1st$ row in the table. From the corresponding topic events on "1987-3-2", it is known that "the President Regan give a report about arms sales". After that on 1987-3-13, the White House investigate the arms sales affairs. In fact, the event changes report every details of the 'Iran-Contra" which shows the effectiveness of the proposed TEDT in tracking the event changes.

5 Conclusion

Detecting events in text stream is a hot research issue in social network society, the conventional approaches only group words by the co-occurrence, which have two major

limitations: noisy words and multiple sub-events. In this paper, a novel event detection and tracking algorithm, called topic event detection and tracking (TEDT), was proposed to tackle these limitations. The TEDT extended the topic model and combined the co-occurrence and the topics of words simultaneously. The empirical experiments were performed on two real data sets and the results demonstrated that: (1) the TEDT is superior to the state-of-the-art topic model for event discovery; and (2) the TEDT can track the event changes. For the future work, we are investigating in detecting new events if one topic could appear in several events.

Acknowledgement. This work is supported in part by NSFC under Grant no.61073195, no.61073051, National Commonweal Technology R&D Program of AQSIQ China under Grant No.201310087, Shenzhen Strategic Emerging Industries Program under Grants No.JCYJ20130329142551746, Shenzhen Science and Technology Program under Grant No.CXY201107010163A and No.CXY201107010206A, National Key Technology R&D Program No.2012BAH10F03, 2013BAH17F00 and Science and Technology Development of Shandong Province No.2010GZX20126, 2010GGX10116.

References

1. He, T., Qu, G., Li, S., Tu, X., Zhang, Y., Ren, H.: Semi-automatic hot event detection. In: Li, X., Zaïane, O.R., Li, Z.-h. (eds.) ADMA 2006. LNCS (LNAI), vol. 4093, pp. 1008–1016. Springer, Heidelberg (2006)
2. Wang, C., Zhang, M., Ma, S., Ru, L.: Automatic online news issue construction in web environment. In: Proceedings of the 17th International Conference on World Wide Web, pp. 457–466 (2008)
3. Wang, Y., Xi, Y.H., Wang, L.: Mining the hottest topics on chinese webpage based on the improved k-means partitioning. In: 2009 International Conference on Proceedings of Machine Learning and Cybernetics, vol. 1, pp. 255–260 (2009)
4. Wang, X., Zhai, C., Hu, X., Sproat, R.: Mining correlated bursty topic patterns from coordinated text streams. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 784–793 (2007)
5. Hurst, M.F.: Temporal text mining. In: Proceedings of AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, pp. 73–77 (2006)
6. Eda, T., Yoshikawa, M., Uchiyama, T., Uchiyama, T.: The effectiveness of latent semantic analysis for building up a bottom-up taxonomy from folksonomy tags. In: Proceedings of World Wide Web, pp. 421–440 (2009)
7. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5), 513–523 (1988)
8. Mei, Q., Zhai, C.: Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 198–207 (2005)
9. Osborne, M., Petrovic, S., McCreadie, R., Macdonald, C., Ounis, I.: Bieber no more: First story detection using twitter and wikipedia. In: Proceedings of the SIGIR Workshop on Time-aware Information Access (2012)
10. Lin, C.X., Zhao, B., Mei, Q., Han, J.: Pet: A statistical model for popular events tracking in social communities. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 929–938 (2010)

11. Yao, J., Cui, B., Huang, Y., Jin, X.: Temporal and social context based burst detection from folksonomies. In: Proceedings of AAAI (2010)
12. Yao, J., Cui, B., Huang, Y., Zhou, Y.: Bursty event detection from collaborative tags. Proceedings of World Wide Web 15(2), 171–195 (2012)
13. Fung, G.P.C., Yu, J.X., Yu, P.S., Lu, H.: Parameter free bursty events detection in text streams. In: Proceedings of the 31st International Conference on Very Large Data Bases, VLDB Endowment, pp. 181–192 (2005)
14. Singh, V.K., Gao, M., Jain, R.: Social pixels: Genesis and evaluation. In: Proceedings of the International Conference on Multimedia, pp. 481–490 (2010)
15. AlSumait, L., Barbará, D., Domeniconi, C.: On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In: Proceedings of Eighth IEEE International Conference on Data Mining, pp. 3–12 (2008)
16. Hoffman, M., Bach, F.R., Blei, D.M.: Online learning for latent dirichlet allocation. In: Proceedings of Advances in Neural Information Processing Systems, pp. 856–864 (2010)