Topical Event Detection on Twitter

Lishan Cui(✉), Xiuzhen Zhang, Xiangmin Zhou, and Flora Salim

School of Science (Computer Science and Information Technology),
RMIT University, Melbourne 3001, Australia
{lishan.cui,xiuzhen.zhang,xiangmin.zhou,flora.salim}@rmit.edu.au

Abstract. Event detection on Twitter has attracted active research. Although existing work considers the semantic topic structure of documents for event detection, the topic dynamics and the semantic consistency are under-investigated. In this paper, we study the problem of topical event detection in tweet streams. We define topical events as the bursty occurrences of semantically consistent topics. We decompose the problem of topical event detection into two components: (1) We address the issue of the semantic incoherence of the evolution of topics. We propose to improve topic modelling to filter out semantically inconsistent dynamic topics. (2) We propose to perform burst detection on the time series of dynamic topics to detect bursty occurrences. We apply our proposed techniques to the real world application by detecting topical events in public transport tweets. Experiments demonstrate that our approach can detect the newsworthy events with high success rate.

Keywords: Dynamic topic modelling · Topic mutation · Event detection · Burst detection

1 Introduction

Recent years have seen an astonishing increase in the usage of the Twitter platform for various applications. On Twitter, users post messages, share information and communicate with friends. Twitter messages are often expressions by people with personal and public activities that occur worldwide, many of which describe real world facts and events. The ubiquitous use of Twitter has proven that the posts are updated more frequently than traditional news channel, and are distributed all over the world [18]. Therefore, detecting events over the Twitter platform is more effective and efficient in time-critical applications.

Existing event detection is mostly focused on detecting breaking news [19], query-based events [13], or monitoring disaster events [18]. A number of studies have been done for event summarisation [7,13], but focusing more on the description of events. Online detection of new events (sometimes called first story) over the tweet stream [4,17] very often results in noisy events like "7th Billionths Child Born", but the retrospective detection of semantically significant events is very important. For example, for public transports, discovering the prominent issue-related events are of utmost importance for the public transport administrative authority.

M.A. Cheema et al. (Eds.): ADC 2016, LNCS 9877, pp. 257–268, 2016.
DOI: 10.1007/978-3-319-46922-5_20

In this paper, we study the problem of retrospective analysis of tweets for discovering significant topics, and the hot periods of events that are associated. We focus on the detection of *topical events* in public transportation domain. Different from breaking events, topical events occur during a certain period and fluctuate over time. For example, the train delay recurrently happens on Monday and Friday. The cancellation always happens when the weather is unusual. Topical event analysis helps city planners to discover these "recurring" situations and provide more reliable public transport services. Frequent occurrences of semantically-related events form a topical event. Announcements on planned service disruptions or special concerns are not a topical event.

It is important to note that our study in this paper is different from TDT (topic detection and tracking) [1], which attempts to cluster documents as events using clustering techniques. Rather than clustering documents into topical events, our purpose is to discover the hidden topics in documents (tweets) based on analysing their dynamic semantic structures over streams of documents, along their time dimension. Topics are generally known, but we want to discover dynamic topics rather than a collection of independent topics. Under the same semantic structure, a dynamic topic evolves over time. The consistence of the topic is to measure the semantic consistency of the topic along the time series. When a topic is semantically consistent along the time series, the topic is defined as semantically consistent. To achieve effective topical event detection on Twitter, this paper is focused on the following research questions:

- How to detect the semantically consistent dynamic topics over time?
- How to detect the bursty periods of topical events?

Unlike the traditional event detection based on a given query topic, we do semantic summarisation and event detection at the same time. Therefore, conventional keywords based detection is not applicable to our work. In this paper, we propose to capture topics that evolve over Twitter content and epochs. We define a topical event as the bursty occurrences of a set of semantically consistent topics.

We decompose the problem of topical event detection into two components: (1) We address the issue of semantic mutation of dynamic topics during evolution over time. We propose to improve topic modelling to filter out semantically inconsistent dynamic topics. (2) We propose to perform state-based burst detection technique [3] to identify bursty occurrences of topical events in a discrete temporal sequence of dynamic topics. Experiments on the dataset of Melbourne public transport tweets show that our approaches can accurately detect and describe significant events over epochs. Our discovered events include persistent issues like "delay" as well hidden recurrent hot topics such as "accident", "cancelled", etc. In addition to the topical labels for semantically summarising events, we also pinpoint the timestamps for events.

2 Related Work

Event detection on Twitter has been discussed in the literature. The assumption is that all relevant documents for a topic contain some old or new events of

interest [2]. First story detection (FSD) was introduced to examine the Topic Detection and Tracking (TDT) task that ran as part of the TREC conference [1]. The objective of FSD was to detect the first occurrence of an article that was related to a given topic. The feature-pivot techniques model an event in text streams by grouping words together while rising sharply in frequency [8–10]. An event is represented by a number of keywords showing burst in frequency count [10]. The underlying assumption is that some of the related words about an event would show a spike in the usage when the event is happening.

Using topic distributions rather than the bag of words representing documents reduces the effect of lexical variability while retaining the overall semantic structure of the corpus [22]. Pan *et al.* [16] proposed event detection approaches by combining the LDA model with temporal segmentation and spatial clustering models. The Space-Time LDA is a Spatial Latent Dirichlet allocation (SLDA) [20] adapted from the detection of segments in images to the detection events in the text corpus. The Location-Time constrained Topic (LTT) [23] represents a social message as a probability distribution over a set of topics and captures the unknown composite social events by measuring the similarity between two messages over social media streams. These existing studies focus on discovering the hidden semantic topics for given breaking news events like natural disasters, rather than discovering events from the dynamics of hidden topics.

Our study aims to detect topical events from the tweet stream. The problem can be viewed as a retrospective event detection problem by analysing the Twitter archive. Different from existing event detection research on specific events [4,7,17,21], we aim for detecting significant unspecified events. Rather than relying solely on the word frequency count for event detection [8–10], we focus on the semantic structure of the document collection, which reduces discovering semantically insignificant events. Different from existing work on the topic-related event detection [16,20,23], we model the temporal dynamics when forming topics and our topical events are defined regarding the topic dynamics and the semantic consistence.

3 Problem Formulation

In this section, we formally define the problem of topical event detection on Twitter streams.

Given a text stream $D = \{d_1, d_2, \ldots, d_i\}$ where d_i is a tweet message with time stamp t_i and $t_i \leq t_j$ if $i < j$, and given a fixed time window (for example a week), a Twitter stream D can be divided into epochs of tweets. A *dynamic topic model* comprises a set of dynamic topics that evolve over the epoches under the same term distributional structure. It is well recognised that topic modelling can produce topics lacking semantic coherence [14,15]. The semantics for dynamic topics not only evolves but also can mutate over time, and the mutant topics consist of words that are statistically important for the dynamic topic in epochs but completely lack semantic coherence. Our experiments (Sect. 5.2) show that some topics for public transport can mutate completely, and it is hard for human annotators to discern their semantics for each epoch.

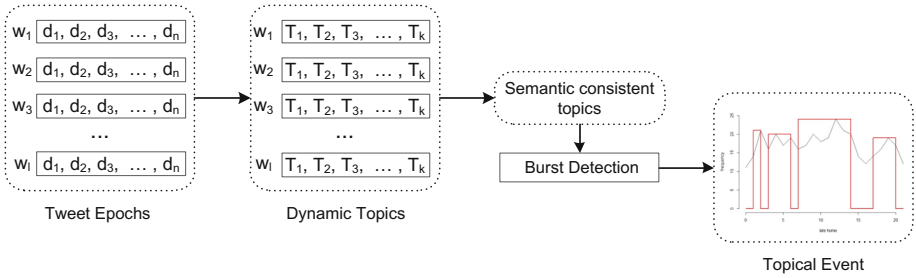

Fig. 1. The framework for topical event detection

The problem of topical event detection is to find a set of semantically consistent topics for the text stream that comprises occurrences of events. Figure 1 shows the framework of the topical event detection. The Twitter stream divided to fixed time window w_i. Let K be the number of topics in each epoch, N_l be the number of documents for each epoch w_i. Applying topic models on tweet epoch, the documents on each epoch modeled as topics under the same semantic structure. The problem of topical event detection can be decomposed into two tasks:

1. Discovery of semantically consistent dynamic topics over the tweet stream along epochs.
2. Detection of bursty occurrences events associated with the topics discovered in the first task.

In the first task of summarising epochs of tweets into dynamic topics, we need to model the temporal dynamics of topics and also address the issue of semantic mutation of dynamic topics. Topic modelling has been widely used to discover hidden topics in a collection of documents. In our problem setting, the discovered topics should be semantically consistent over epochs. In this regard, we propose to measure the semantic coherence of topics across epochs and filter out semantically inconsistent topics. The temporal dynamics for the frequency of topics over epochs form a time series, and our second task of event detection is to detect the bursty occurrences of topics from the time series. To detect significant topical events from tweet posts associated with the topic, it is important to filter out the trivial or gibberish discussions of the topic. We propose to apply a state-based burst detection algorithm to detect topical events in a discrete temporal sequence.

4 Methodology

We will first describe our approach based on non-parametric topic model [6,12] to discover semantically consistent dynamic topics, then demonstrate the detection of bursty occurrences associated topics.

4.1 Discovering Semantically Consistent Dynamic Topics

We aim for discovering hidden topics for a stream of tweets with timestamps, specifically the semantically consistent topics overt time. When the tweet stream is divided into epochs, the semantics of topics summarised for epochs evolve over time. We propose to apply dynamic topic modelling to discover hidden topics from the tweet stream over epochs and then measure the semantic coherence to filter out inconsistent dynamic topics.

We applied the non-parametric dynamic topic model [6,12] that introduced components evolution as a chain, extended the standard topic model LDA [5] to identify semantically consistent latent topics over epochs.

The non-parametric dynamic topic models apply the hierarchical Pitman-Yor process (PYP) to model both the document-topic proportions and the topic-word distributions evolving over epochs. In each epoch, the process is similar to the standard LDA except for the PYP process. The posterior distribution of topics depends on the information from both word and time slice.

The non-parametric dynamic topic models identify and describe the topics over epochs. The topic consistency is ensured by maximizing the estimation. However, the topic sometimes may contain the mutation, the variation that happened during the topic evolution over epochs. The topical event is detected from a semantically consistent dynamic topic.

The Kullback-Leibler divergence [11] is a measure of the difference between two probability distributions. The semantic mutation of dynamic topics during evolution can be measured by KL divergence. Given dynamic topic T and l epochs, let T_x and T_y denote respectively the word probability distribution for topic T with epoch x and y, which $x, y \in l$. The semantic distance of T_x and T_y definded as:

$$s(T_x, T_y) = \sum_i T_x(i) * ln \frac{T_x(i)}{T_y(i)}. \tag{1}$$

where i denotes the words under consideration. For each dynamic topic, we measured the semantic distance between epoch pairs by counting word frequency.

$s(T_x, T_y)$ measures the semantic distance between time slices. To measure the consistency level across epochs for a dynamic topic, we define a threshold θ for the distance metric of a time series against the first epoch to measure the topic semantic consistency. In this paper, we use the mean of the semantic distance over epochs as the threshold θ. When the semantic distances over epochs are less than θ, we consider the topic as a semantically consistent topic over the epochs.

4.2 Topical Event Detection

The composition of topics for epochs changes over time. The frequency for a dynamic topic fluctuates and forms a time series. We aim to detect significant bursts of a dynamic topic as the event associated with the topic. We propose to apply state-based burst detection algorithm to detect the topical events in a discrete temporal sequence.

In the area of Topic Detection and Tracking [1], the sequence of documents for a distinct topic is analysed as time series. The analysis of time series comprises of methods for extracting meaningful dynamic characteristics of the data. Among them, detecting and modelling bursts is a common task. We propose to apply state-based burst detection algorithm proposed by Araujo *et al.* [3] to detect the bursts of topical events in a discrete temporal sequence.

Events can happen at any time instant. Moreover, the timescale of some events is varied. For instance, low-intensity earthquakes have a timescale of days, and car crash accidents have a timescale of hours. Topical events detection need detect long-term vibrating bursts and short-term sharp bursts simultaneously. In the state-based burst detection model, the transit probability refers to the probability of changing the state. Differencing the transit probability can capture differences in the frequencies.

Under the state-based burst detection model, time intervals with different event frequencies are modeled as different states. The whole sequence of the event is regarded as a Hidden Markov Model of the states. The Poisson process is used to model arrivals. The detection of bursts is then achieved by applying the dynamic programming to find sequences of states that best fit the time series. Moreover, the fitness function generalizes this model to time series data that consists of sequences of events obtained from repeated measurements of time.

In our work, we model binary states model, which includes a non-bursty state and a bursty state. The states are pre-defined with different mean values according to the probabilistic distribution. The desired estimation for the sequence of frequencies $\{\lambda_1, \ldots, \lambda_T\}$ can be obtained by maximizing this probability, namely

$$f(\lambda_1, \ldots, \lambda_T) = \sum_{t=1}^{T} (x_t \ln \lambda_t - \lambda_t) + K \sum_{t=2}^{T} \delta_{\lambda_t, \lambda_{t-1}}. \tag{2}$$

where $\delta_{\lambda',\lambda} = 1$ if $\lambda' = \lambda$ and 0 otherwise. The parameter $K = \log[p(E-1)/(1-p)]$, with E is the number of frequencies in a discrete temporal sequence.

5 Experiments

We detected topical events from a collection of Twitter dataset gathered over a period of five months (1st January 2014 to 31st May 2014). The dataset was collected through Twitter's streaming API.

5.1 Datasets

The focus of our study is Melbourne public transport related event detection. To get Melbourne transport related tweets, we investigated three location related attributes, such as GPS coordinates, place name, and author's location indicated as Melbourne. The statistics of the Melbourne-based dataset are shown in Table 1. Then we aggregated the tweet contents by day.

Table 1. Melbourne-based tweets dataset

Month	# GeoTag	# Place	# User	# Combined
January	2,322	298	60,550	61,780
Feburay	1,335	241	41,574	42,338
March	1,918	335	56,281	57,307
April	1,652	687	57,992	58,936
May	1,877	1,375	61,174	62,214
Total	9,104	2,936	277,571	**282,575**

Table 2. The sample summarisations of Week 1 news articles

Week	Date	Summary	Topic Label
1	6 Jan	Delay traffic delays return work	delay
1	6 Jan	Hit killed by train Frankston	accident
1	10 Jan	Travelling tram seat beside syringes	service

The ground truth public transport events. We need to establish a reliable ground truth for public transport events to evaluate the topical events detected by our approach. To this end, we crawled the online news articles from The Age, a well known daily Melbourne newspaper, published for the same period as our Twitter dataset from 1st January 2014 to 31st May 2014. There are 799 pieces of news in five months from The Age for that duration, out of which 50 news articles are related to the public transport according to the keyword search. The fifty news articles have information of the date, snippet and news title. Two annotators summarised the content for each piece of news using five

Table 3. The semantic distance and Jaccard results for ten topics over 22 weeks

Topic	Topic Label	Avg	$s(T_x, T_y) \leq \theta$ (%)	Jaccard Distance
Topic0	late home	2.003	**95.5**	0.673
Topic1	service	2.399	86.4	0.605
Topic2	delay	1.997	**95.5**	0.486
Topic3	accident	2.269	**95.5**	0.555
Topic4	service	2.267	**95.5**	0.586
Topic5	stop	2.017	86.4	0.682
Topic6	train	2.284	90.9	0.691
Topic7	roadwork	2.150	**95.5**	0.564
Topic8	public	2.237	90.9	0.582
Topic9	driver	2.133	90.9	0.527

keywords from the news snippets and titles. The agreement was reached by discussions. The sample summarisations for the first week of news articles are shown in Table 2. The corresponding labels for each news were done manually by annotators based on the topic labels in Table 3 (See Sect. 5.2).

5.2 Extraction of Consistent Dynamic Topics

As described in Sect. 4, we performed two steps to extract semantically consistent dynamic topics over five months. We first semantically summarised tweets topics to discovery semantically consistent dynamic topics. In our experiments, we set the parameter $Epoch = 22$, indicates the summarisation for the weekly in our corpus. We investigated the results that when $K = 10$, where the K is the number of topics. Figure 2 shows the trend of each topic by week for ten topics. The summaries of the meaning of each topic are shown in Table 3 (Topic Label). The labels for topics were manually assigned by annotators.

We can see from Fig. 2, during most of the period, the frequency of Topic6 is less than 15. Topic1 has a peak value of 32 at week 4, and Topic3 has another peak value of 31 at week 16. Topic7 has stable frequencies over all weeks compare to other topics.

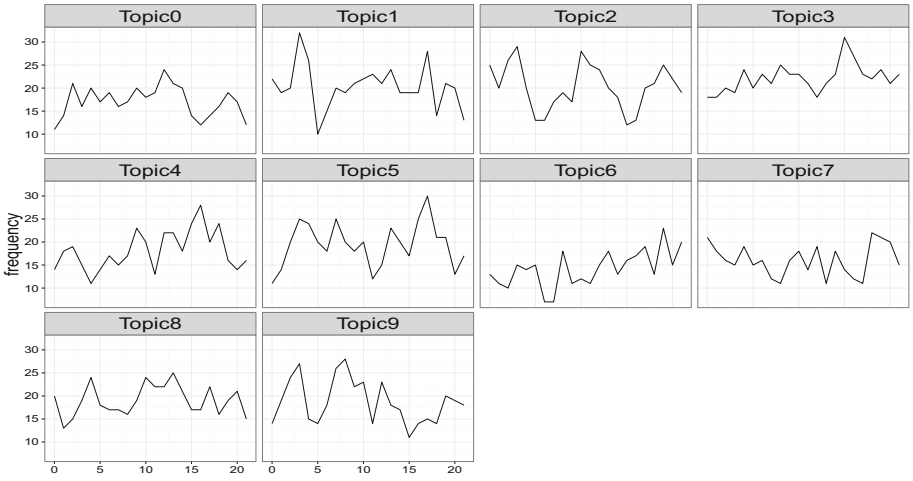

Fig. 2. The dynamic topics by week

The second step is to discover semantically consistent topics. We estimated the semantic consistence for topics over epochs. Table 3 lists the average semantic distance scores for each dynamic topic. Topic2 has the lowest average score while Topic1 has the highest score.

To discover the consistent topics, we applied the threshold θ to filter out the mutant topics. We set θ as the mean of the semantic distances over epochs. For

each topic, if there are more than 95 % of semantic distance scores ($s(T_x, T_y)$) less than the threshold, we treat this topic as a semantically consistent dynamic topic. The results are shown in Table 3 as the percentage of the consistency. The topics in bold are semantically consistent dynamic topics.

We evaluated the semantically consistent topics from the view of semantic topic evolution. We aggregated the word frequency for each topic over epochs, then selected the top ten high-frequency words for each topic as the topic representative words. These topic representative words lead the meaning of each topic. During the topic evolution, these words should have high probability occurrence. For each topic, we calculated the occurrence distributions of topic representative words over the epoch.

We applied Jaccard distance to calculate the dissimilarity between each topic over 22 weeks using the top ten high-frequency topic words. The formula is as follow:

$$d_J(E_i, E_j) = 1 - J(E_i, E_j) = \frac{|E_i \bigcup E_j| - |E_i \bigcap E_j|}{|E_i \bigcup E_j|}. \tag{3}$$

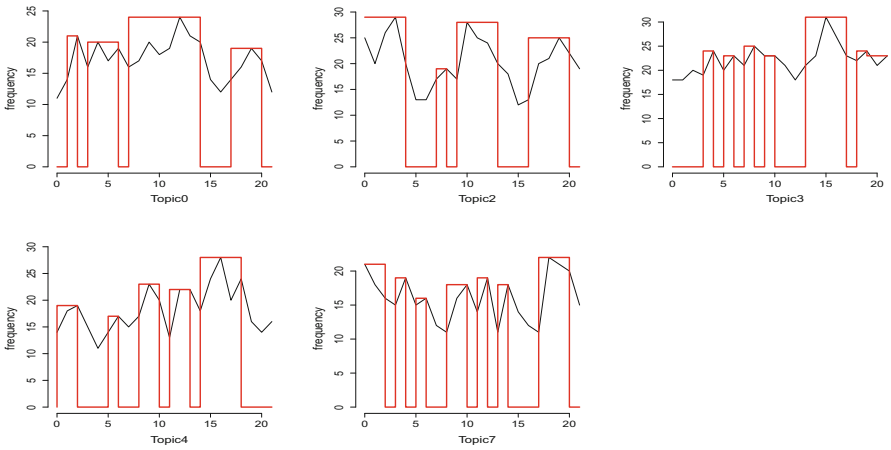

Fig. 3. The topical event detection

The results are shown in Table 3. Topic5 and Topic6 have the highest dissimilarity. In our approach, these two topics were filtered out as well. For Topic2 and Topic3, the Jaccard values are lower, means these two topics are semantically consistent over 22 weeks.

5.3 Detection of Event Occurrences

The state-based burst detection was applied to detect the period of topical events. We treat dynamic topics as temporal series, and the time scale is a week. In total, we ran the algorithm for burst detection on 22 weeks series.

The burst detection results for the five semantically consistent topics are shown in Fig. 3. Topic0 detected four occurrences, including week 2, from week 4 to week 6, from week 8 to week 14 and week 18 to week 20. Topic3 detected six bursts and Topic4 detected five occurrences.

By and large, the news of public transport is only about major accidents that caused very long delays and disruptions. Such accidents usually involved casualties or major injuries, e.g. a person being hit by a train. On the other hand, the tweets have more diverse topics when it comes to public transport. People send updates on the delays or their grievances on trivial matters.

Table 4. Event detection

	News	Event
January	9	7
Feburary	4	2
March	20	17
April	4	3
May	13	6
Sum	50	35

Table 4 shows the numbers of public transport related events from news articles and the number of events correctly detected on Twitter. In January, there is nine public transport news featured in the news. From our method, using five consistent topics, we detected seven events from our tweets dataset correspond to the same week as the news reported. However, there are two bursts not detected in the tweets dataset.

In comparison to the news content, the event of the delay, our approach can reach the recall of 73 %, and the event of the service, our approach can only get the recall of 70 % since the news reported the planned work or project annotated as service as well.

Accurately detecting bursts is not only related to the settings of burst detection parameters, but the numbers of consistent topics are also important. The larger number of consistent topics can increase the accuracy of event detection, but decrease the quality of the event summary.

6 Conclusions

In this paper, we studied the problem of topical event detection in a stream of tweet messages. We decomposed the problem of topical event detection into two components: (1) Semantically consistent dynamic topic discovery: We applied dynamic topic modelling to discover dynamic topics. More importantly, to address the issue of semantic mutation for the evolution of dynamic topics, we proposed to use the KL divergence measure to filter out semantically inconsistent

dynamic topics. (2) Detection of events burst occurrences: We applied state-based burst detection on the time series for dynamic topics to detect bursty occurrences of topical events. We applied our proposed technique to the real world by detecting topical events for a public transport Twitter dataset, for five months. Our results demonstrated that our approach can detect the newsworthy events for public transport with high success rate. For future work, we will focus on developing a unified model that combines dynamic topic modelling and mutation topic pruning. We will also investigate how to achieve online topical event detection for live Twitter streams.

References

1. Allan, J.: Introduction to topic detection and tracking. In: Allan, J. (ed.) Topic Detection and Tracking, pp. 1–16. Springer, New York (2002)
2. Allan, J., Lavrenko, V., Jin, H.: First story detection in TDT is hard. In: Proceeding 19th International Conference on Information and Knowledge Management, pp. 374–381. ACM (2000)
3. Araujo, L., Cuesta, J.A., Merelo, J.J.: Genetic algorithm for burst detection and activity tracking in event streams. In: Runarsson, T.P., Beyer, H.-G., Burke, E.K., Merelo-Guervós, J.J., Whitley, L.D., Yao, X. (eds.) PPSN 2006. LNCS, vol. 4193, pp. 302–311. Springer, Heidelberg (2006)
4. Becker, H., Naaman, M., Gravano, L.: Beyond trending topics: real-world event identification on Twitter. ICWSM **11**, 438–441 (2011)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Mach. Learn. Res. **3**, 993–1022 (2003)
6. Buntine, W.L., Mishra, S.: Experiments with non-parametric topic models. In: Proceeding of 20th ACM SIGKDD, pp. 881–890. ACM (2014)
7. Cordeiro, M.: Twitter event detection: combining wavelet analysis and topic inference summarization. In: Proceeding Doctoral Symposium on Informatics Engineering, DSIE, p. 123–138 (2012)
8. Fung, G.P.C., Yu, J.X., Yu, P.S., Lu, H.: Parameter free bursty events detection in text streams. In: Proceeding of 31st International Conference on Very Large Data Bases, pp. 181–192. VLDB Endowment (2005)
9. He, Q., Chang, K., Lim, E.P.: Analyzing feature trajectories for event detection. In: Proceeding of 30th ACM SIGIR, pp. 207–214. ACM (2007)
10. Kleinberg, J.: Bursty and hierarchical structure in streams. Data Min. Knowl. Disc. **7**(4), 373–397 (2003)
11. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**(1), 79–86 (1951)
12. Li, J., Buntine, W.: Experiments with dynamic topic models. In: Proceeding of NewsKDD workshop on Data Science for News Publishing. ACM (2014)
13. Metzler, D., Cai, C., Hovy, E.: Structured event retrieval over microblog archives. In: Proceeding of the North American Chapter of the Association for Computational Linguistics, pp. 646–655. Association for Computational Linguistics (2012)
14. Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proc. Empirical Methods in Natural Language Processing. pp. 262–272. Association for Computational Linguistics (2011)

15. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Proceeding of the North American Chapter of the Association for Computational Linguistics, pp. 100–108. Association for Computational Linguistics (2010)
16. Pan, C.C., Mitra, P.: Event detection with spatial latent dirichlet allocation. In: Proceeding of 11th ACM/IEEE International joint Conference on Digital libraries, pp. 349–358. ACM (2011)
17. Petrović, S., Osborne, M., Lavrenko, V.: Streaming first story detection with application to Twitter. In: Proceeding of the North American Chapter of the Association for Computational Linguistics, pp. 181–189. Association for Computational Linguistics (2010)
18. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceeding of 19th International Conference WWW, pp. 851–860. ACM (2010)
19. Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D., Sperling, J.: Twitterstand: news in tweets. In: Proceeding 17th ACM SIGSPATIAL International Conference on advances in geographic information systems, pp. 42–51. ACM (2009)
20. Wang, X., Grimson, E.: Spatial latent Dirichlet allocation. In: Proceeding Advances in Neural Information Processing Systems, pp. 1577–1584 (2008)
21. Weng, J., Lee, B.S.: Event detection in Twitter. ICWSM **11**, 401–408 (2011)
22. Yao, L., Mimno, D., McCallum, A.: Efficient methods for topic model inference on streaming document collections. In: Proceeding of 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 937–946. ACM (2009)
23. Zhou, X., Chen, L.: Event detection over Twitter social media streams. Int. J. VLDB **23**(3), 381–400 (2014)