A probabilistic method for emerging topic tracking in Microblog stream

Article *in* World Wide Web · April 2016

DOI: 10.1007/s11280-016-0390-4

6 authors, including:

Jiajia Huang
Nanjing Audit University
14 PUBLICATIONS **169** CITATIONS
SEE PROFILE

Hua Wang
Victoria University Melbourne
257 PUBLICATIONS **2,623** CITATIONS
SEE PROFILE

Jinli Cao
La Trobe University
112 PUBLICATIONS **1,031** CITATIONS
SEE PROFILE

Xiuzhen Jenny Zhang
RMIT University
130 PUBLICATIONS **1,648** CITATIONS
SEE PROFILE

World Wide Web (2017) 20:325–350
DOI 10.1007/s11280-016-0390-4

A probabilistic method for emerging topic tracking in Microblog stream

Jiajia Huang[1] · **Min Peng**[1] · **Hua Wang**[2] · **Jinli Cao**[3] · **Wang Gao**[1] · **Xiuzhen Zhang**[4]

Received: 3 February 2016 / Revised: 21 March 2016 /
Accepted: 28 March 2016 / Published online: 21 April 2016

Abstract Microblog is a popular and opJiajia Huangen platform for discovering and sharing the latest news about social issues and daily life. The quickly-updated microblog streams make it urgent to develop an effective tool to monitor such streams. Emerging topic tracking is one of such tools to reveal what new events are attracting the most online attention at present. However, due to the fast changing, high noise and short length of the microblog

✉ Jiajia Huang
huangjiajia@whu.edu.cn

✉ Min Peng
pengm@whu.edu.cn

Hua Wang
hua.wang@vu.edu.au

Jinli Cao
j.cao@latrobe.edu.au

Wang Gao
gaowang@whu.edu.cn

Xiuzhen Zhang
xiuzhen.zhang@rmit.edu.au

[1] State Key Lab of Software Engineering, Wuhan University, Wuhan, 430072, China

[2] Centre for Applied Informatics, Victoria University, Melbourne, Victoria, 3001, Australia

[3] Computer Science and Computer Engineering, La Trobe University, Bundoora, Victoria, 3086, Australia

[4] School of CS&IT, RMIT University, GPO Box 2476, Melbourne, Victoria, 3001, Australia

feeds, two challenges should be addressed in emerging topic tracking. One is the problem of detecting emerging topics early, long before they become hot, and the other is how to effectively monitor evolving topics over time. In this study, we propose a novel emerging topics tracking method, which aligns emerging word detection from temporal perspective with coherent topic mining from spatial perspective. Specifically, we first design a metric to estimate word novelty and fading based on local weighted linear regression (LWLR), which can highlight the word novelty of expressing an emerging topic and suppress the word novelty of expressing an existing topic. We then track emerging topics by leveraging topic novelty and fading probabilities, which are learnt by designing and solving an optimization problem. We evaluate our method on a microblog stream containing over one million feeds. Experimental results show the promising performance of the proposed method in detecting emerging topic and tracking topic evolution over time on both effectiveness and efficiency.

Keywords Microblog stream · Emerging topic · LWLR · Topic evolution · Optimization problem

1 Introduction

It is true that microblog websites, such as Twitter and Sina Weibo, have developed into one of the most popular platforms of discussing social issues, ranging from politics, economics, sports, entertainments to people's daily life. These hot issues always attract wide discussions the first time they happen and evolve over time with sustained attention. The emerging trend always heralds an upcoming issue, which may be an emergency like *an earthquake* or *a terrorist attack*, or an anticipated event like *FIFA World Cup*. Tracking emerging topics not only can help us understand what things are drawing public attention, but also is beneficial for related organizations to draw up countermeasures earlier. For example, emerging topic tracking in microblog stream shows its advantage of high response speed in earthquake detection [27]. In addition, many organizations and brands are also keen on monitoring and tracking microblog streams to analyze user interests and detect emergencies [7, 40].

Two issues should be emphasized when tracking emerging topics from microblog streams. First is how to detect emerging topics as early and accurately as possible. Note that, in this study, the definition of *emerging* denotes that the topic not only should be novel, but also will evolve into a popular issue with more attention in the future instead of disappearing shortly. Many relevant methods have been proposed to address the problem of early detection, out of which emerging-feature-clustering methods are wide used [7, 18, 28, 33]. However, these methods mainly focus on detecting emerging topics as early as possible by emerging words clustering, but concern less on how these emerging topics evolve over time. Second is how to track topic evolution. Existing methods of topic evolution tracking focus either on capturing the evolution operations between all topics in two adjacent time windows [36, 42], or on generating latent topics over time under temporal topic models [17, 32]. The former usually defines operations, such as appearance (or emerging), growth (or absorption), merge (or join), split (or collapse), to figure out the life cycles of the evolving topics. The latter works on inferring the parameters of the temporal topic model when the latent topics evolve over time. However, in real-world microblog streams, large volumes of microblog feeds are posted by different people with very large vocabulary and discussing a broad range of issues, as a result, thousands of topics appear every day and some of them last for several days or weeks. These topics are also mixed with various noises, ranging from daily babbles to advertisements. Therefore, if an emerging topic can not be detected timely with coherent words, or its subsequent trend can not be correctly depicted with proper patterns, then it will perplex the relevant users of monitoring the streams or even mislead them.

In this study, we mainly address the following challenges: (1) detecting emerging topic as early as possible, (2) tracking emerging topic evolution and (3) presenting topic with coherent words. As probabilistic topic model is good at generating coherent topics whereas emerging-feature-based clustering methods excel at identifying emerging words or phrases, can we track emerging topics by combining the advantages of these two types of methods to tackle the above challenges? In fact, a study based on this idea, named BBTM (bursty biterm topic model) [35], is proposed to generate latent bursty topics in Twitter streams. However, the topic generation process is guided by the word bursty probability measured from the temporal perspective and it results in a low precision in identifying emerging topics. Therefore, we investigate two questions: Is there an approach of generating latent topics independently to word novelty evaluation process? Will the approach ensure all emerging topics and trends of before-detected topics can be dug out and distinguished from each other?

With this motivation, we propose a novel emerging topic tracking method, named ETT, which first generates latent topics from word co-occurrence space and estimates words novelty from timeline respectively. Furthermore, three kinds of operations, including *emerging*, *growing* and *fading* are defined based on the latent topic's significance and the word novelty to track the evolving topic. The foremost application of our approach is microblog streaming monitoring, including monitoring of both the global microblog streams that are posted on Twitter or Sina Weibo, and the local microblog streams that are related to specified originations, celebrities, brands or events. The advantage of this approach is highlighted when it is applied to the microblog stream that is composed of many topics and some of them last for a long time and evolve into a variety of trends. What's more, our approach can also discover and visualize a topic's temporal progression by displaying its word distribution, significance, trends (emerging, growing, etc.), best matching topic in previous time, and so on. Thus, the approach is feasible to be integrated into a practical monitoring system.

The main contributions of this study are summarized as follows.

First, we propose a novel probabilistic method along with three major topic evolution operations to track emerging topics from large volumes of microblog streams. This method can be used to detect emerging topics early from microblog streams and track their evolution at the same time.

Second, we propose metrics of novelty and fading to measure both words and topics from timeline and design a prediction method based on local weighted linear regression (LWLR) to estimate the words' novelty and fading probabilities. The LWLR-based method can detect emerging topics with early time and sound performance, because it can assign a high novelty to a new-coming word, and a much low novelty to an existing word even if the word has a significant increment in current stream.

Third, we formulate the learning of topic novelty and fading as an optimization problem that can be inferred from the novelty and fading probabilities of the words it contains. By defining operations based on this learning result, the ETT method can efficiently capture the topic evolution over time.

Finally, we demonstrate the effectiveness and efficiency of the ETT using a real-world microblog stream collected from Sina Weibo. A series of experimental results suggest that the method achieves better performance than baselines on emerging topics detection, including coherence, novelty and accuracy. In addition, this method also has promising performance on tracking topic evolution.

The rest of this paper is organized as follows. Section 2 briefly summarizes the related work of topic detection and tracking in microblog streams. Section 3 discusses the preliminaries and formulates the overview of emerging topic tracking task. Section 4 details the implementation of the ETT method. We evaluate the method in Section 5 and draw a conclusion in Section 6.

2 Related works

Research on topic detection and tracking (TDT) over microblogs has attracted great attention in recent years, due to its wide applications in emergency detection like earthquake [27], geo-location detection [31], political election outcomes prediction [30], and so on. In this section, we first briefly summarize the related work and then present the differences between existing work and ours.

2.1 Probabilistic topic model

Probabilistic topic models have been proposed to uncover the latent semantic structure (i.e., topic) in documents, which present a topic as a group of words with their probability distribution. Typical topic models, such as pLSA [12] and LDA [4], have achieved huge success in information retrieval [26], text mining [8, 19] and recommendation systems [21, 37]. In the last several years, LDA model has been extensively studied with many variants, such as addressing topic sparsity in short texts [19], automatically inferring topic numbers [38], generating more coherent topics by employing biterms [9] or knowledge [8]. These topic models usually perform well on generating coherent topics from various text corpus, such as news articles [9, 19, 38], Wikipedia webpages [19], Amazon reviews [8] and Twitter feeds [9, 38].

2.2 Topic evolution tracking

Research on topic evolution tracking can mainly be classified into two types according to tracking methods. One type of methods employ probabilistic topic model with considering temporal information to figure out the latent topic generation process in text streams. Fundamental work includes DTM (dynamic topic model) [3], TOT (topic over time) [32] and other temporal topic models [14, 23]. There are other temporal topic models that also work on detecting emerging topic at the same time. Online-LDA [2] is such a variant of LDA, which is able to identify bursty topics by inferring the current topic-term distribution with historical information. Similar to Online-LDA, Lau et.al. [17] proposed another Online-LDA model, which considers the characteristic of short texts when detecting emerging topic in Twitter streams. We call it OLDA to distinguish from the prior Online-LDA. In these methods, latent topics and their evolution are detected via inferring parameters changing in the probabilistic topic model with growing streams.

Other types of methods focus on depicting the evolution operations of topics in two adjacent time windows. For example, Yang et.al. proposed [36] eight kinds of operations (appearance, disappearance, growth, shrinkage, etc.) to discover topic evolution in Twitter feeds. Zhu et.al. employed and improved Bayesian rose tree with considering three evolution operations (i.e., join, absorb and collapse) to capture the evolution structure among all detected latent topics in two adjacent microblog batches [42]. Cai et. al. [6] used four event operations (i.e., create, absorb, split and merge) and designed an event index structure of multi-layer inverted list to manage and discover evolving events over Twitter streams.

2.3 Emerging topic detection

Generally, research on emerging topic detection in microblog streams can be classified into three categories according to topic presentations.

Topic-model-based. In this type of methods, emerging topic is presented as a probability distribution over a vocabulary. Many variants of probabilistic topic models, such as Twitter-LDA [10], BBTM (bursty biterm topic model) [35], are proposed to detect latent bursty topics from microblog streams. Aside from Gibbs sampling, bursty topic can also be detected via solving an optimization problem. For example, Xie et.al. proposed TopicSketch [34], in which bursty topic detection task is formulated as an optimization problem of minimizing the square error between observed value and expectation of word acceleration. Yin et.al. [39] also constructed an optimization function to detect both bursty topics and stable topics. In this method, both user's relationship in social network and term bursty weight were considered for enhancing the model's performance.

Emerging-feature-clustering. Another type of emerging/bursty topic detection idea is to detect emerging/bursty features (e.g., terms, segments) and cluster them into topics. These methods usually explore more elaborated strategies for feature novelty estimation. For example, bursty/emerging features are selected via wavelet analysis of frequency-based term signal [33], frequency expectation of a segment over a time window [18], chi-square testing for term's foreground and background distributions [7], high utility pattern mining [13], and exponentially weighted average (EWMA) of terms and co-occur terms [28]. Then, clustering or classification algorithms, such as modularity-based partitioning [13, 33], kNN [13, 18] and SVM [7], are employed to gather these selected features into emerging/bursty topics. These methods are usually followed by a post-processing of filtering noisy topics.

Document-clustering. In this type of methods, documents are gathered together according to their similarity and a topic is presented as a cluster of relevant documents. Generally, incremental clustering is used for emerging topics detection [16, 25]. One typical study is first story detection [25], which tracked novel topics in Twitter stream by employing locality sensitive hashing (LSH) to search the neighbors for each new-coming text and then assigning this text to the cluster where its most similar neighbor is. Other similar study is to employ dictionary learning method [16] for emerging topic detection, in which a new-coming text is judged as an emerging topic if the reconstruction error of the text's sparse encoding under the dictionary trained by existing texts is below a specified threshold.

Our work in this study can be distinguished from existing studies in several ways. First, we not only detect emerging topics but also track their evolution at the same time. Unlike previous works that focus either on detecting emerging topics or on capturing topic evolution over time, this study deals with both tasks simultaneously. Second, we propose a LWLR-based word novelty and fading estimation method. Contrary to existing methods that evaluate word novelty based on average frequency of historical data [18, 28, 33, 35], this method ensures new-coming words stand out from existing words more significantly. Third, although only three major operations (i.e., emerging, growing and fading) are proposed to capture topic evolution, the proposed ETT method can be easily extended with more operations. As the topic trending tracking is based on the combination result of the topic novelty and fading evaluated from timeline and the latent topics inferred from word co-occurrence matrix, according to the definitions of the evolution operations proposed in [36], more operations can be introduced by specifying the thresholds of topic-pair similarity and rate between topic novelty and fading in a more elaborated manner.

3 Preliminaries

In this section, we formally discuss the microblog stream we are interested in, define relevant operations of topic evolution and formulate the overview of emerging topic tracking in microblog streams.

3.1 High quality microblog stream

Suppose that each microblog feed in the dataset consists of textual content, URL, number of forwards, number of reviews, and publishing timestamp. Next, we will release the definition of microblog stream as follows.

Definition 1 (**Microblog stream**) A microblog stream is a temporal sequence of microblog feeds $\mathcal{M} = \{M_1, M_2, ..., M_l, ...\}$, where each feed M_l in the stream contains following attributes:

- Textual content: a feed contains a short document, which may discuss something about a social issue, user's daily life or even an advertisement.
- URL: a group of URLs that link to news webpages, video webpages or shopping webpages.
- Forward and review: a feed may be forwarded or reviewed by user's followers.
- Text length: the number of words in a feed after stop-words removal.
- Posting time: the timestamp of posting this feed.

Conventionally, the microblog stream is divided into slices, where each slice collects all the feeds published in the same time interval (e.g., one hour or one day).

Note that, on one hand, similar to other real-world data streams, such as RFID data streams [15], and electrocardiogram streams [20], microblog streams are also full of large amount of noise. That is to say, microblog content does not always relate to a social, political or sport event. It may just be a bubble of someone's daily life, such as ”Good Morning”. According to a study of tweets [1], nearly 36 % of the rated tweets are worth reading, 25 % are not, and 39 % are middling. On the other hand, besides textual content, URLs sometimes can provide useful information for judging microblog feed quality. In our dataset, 36 % of feeds contain URLs. A feed may be associated with a meaningful event if its URL links to a news webpage. Therefore, we first try to extract a small number of high quality microblog feeds from the large scale of the raw stream for topic tracking.

Definition 2 (**High quality microblog stream**) A high quality microblog stream is the subset of the raw stream that is composed of meaningful microblog feeds and can represent the raw stream in a global manner.

Based on our previous research [24], we employ microblog feed attributes, including textual content, number of reviews, number of forwards, URL quality and text length to estimate the feed quality via an EM algorithm and then extract the top-N feeds in each slice to build a high quality microblog stream. Although the scale of the high quality stream is much smaller than the raw, it can reserve the feeds that express meaningful events and cause wide influence. This stream contains less noise, which is beneficial for enhancing the efficiency and effectiveness of topic tracking.

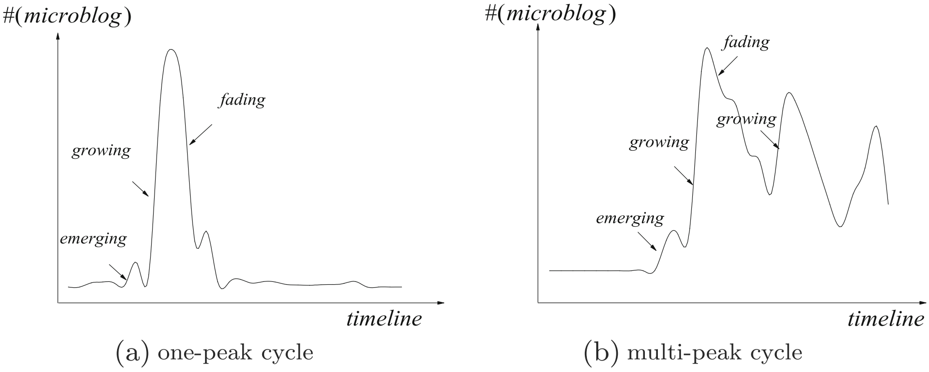
3.2 Topics evolution operations

Generally speaking, in mainstream topic detection studies [7, 10, 34, 35], a topic is considered as an emerging topic if it is (heavily) discussed in this time slice but not previously. Furthermore, emerging features, such as words or phrases, are usually utilized for detecting emerging topic. In this study, we propose a novelty probability and a fading probability to collectively identify emerging topics from the latent topics generated by a topic model. It is noted that although one of our goals is to detect emerging topics as early as possible, it is really hard to find a topic when there are only five or ten relevant microblog feeds mixed with thousands of irrelevant ones. Therefore, in this study, a topic is considered significant if it is discussed with over η percent (η is set to 5 % in our dataset) of feeds.

Definition 3 (**Emerging**) A topic in a time slice is called an emerging topic if it never appeared in previous slice and attracts a significant number of microblog feeds in current slice. Thus, an emerging topic should be a novel topic that has a high probability of novelty and low probability of fading.

In microblog streams, a topic usually presents some kind of life cycle, i.e., emerging at some time, building towards climax while gaining more and more attention, tending to fade and finally disappearing or restarting again. Figure 1 shows two types of topic life cycles, one-peak cycle and multi-peak cycle. In a topic life cycle, apart from emerging states, the other two evolution states are also worth noting, which we call growing and fading respectively. A major reason of dragging down the accuracy of emerging topic detection in previous work is that the growing trends are often treated as emerging topics, because some of the words they contain also show high novelty level. However, one significant difference between emerging topic and growing trend is whether it has appeared before.

Definition 4 (**Growing**) A topic shows a growing trend if it has appeared before and attracts more microblog feeds in current slice. Usually, a growing trend has a higher value of novelty than fading although it has appeared in previous slice.



(a) one-peak cycle (b) multi-peak cycle

Figure 1 Conceptual diagram of two types of life cycles and three topic evolution operations

Definition 5 (**Fading**) A topic shows a fading trend if it has appeared before but attracts fewer microblog feeds in current slice. Usually, a fading trend has a lower value of novelty than fading.

3.3 Overview of our solution

The three primary research challenges in this study are (I) how to detect an emerging topic as early and accurately as possible; (II) how to capture the topic evolution; and (III) how to generate coherent and meaningful topic.

We propose a novel emerging topic tracking method, named ETT, to track what topics are emerging in microblog streams and how they evolve over time. To meet challenge (I), metrics of novelty and fading are proposed to measure words and topics from temporal perspective. Novelty is used to measure the freshness of a word/topic, whereas fading is used to measure the staleness of a word/topic. Furthermore, to estimate a word's novelty and fading, a method based on local weighted linear regression (LWLR) is designed to predict the frequency for each of the words in current stream. Thus, the novelty and fading of a word can be measured by the difference between the word's predicted frequency and actual frequency in current slice. To meet challenge (III), a probabilistic topic model is employed to generate latent topics in the stream from spatial perspective. Finally, an optimization problem is formulated and solved to estimate the novelty and fading for all the latent topics. To meet challenge (II), three topic evolution operations are defined to track emerging topics and their evolution over time.

Figure 2 shows the overview of the proposed ETT solution, which contains five modules: (1) A collector which selects high quality microblog stream; (2) A LWLR-based estimator which estimates the novelty and fading probabilities of all the words in current high quality stream; (3) A topic-model-based generator which generates latent topics based on the document-term matrix of microblogs in current stream; (4) An optimization learner which learns topic novelty and fading with inputting word distribution, word novelty, word fading and topic significant probability; (5) A tracker which tracks topic evolution according to the topic's word distribution, novelty and fading probabilities. A brief discussion of modules (1) has been presented in Section 3.1 and the details can be referred to our previous research [24]. Modules (2)-(5) will be detailed in Section 4.

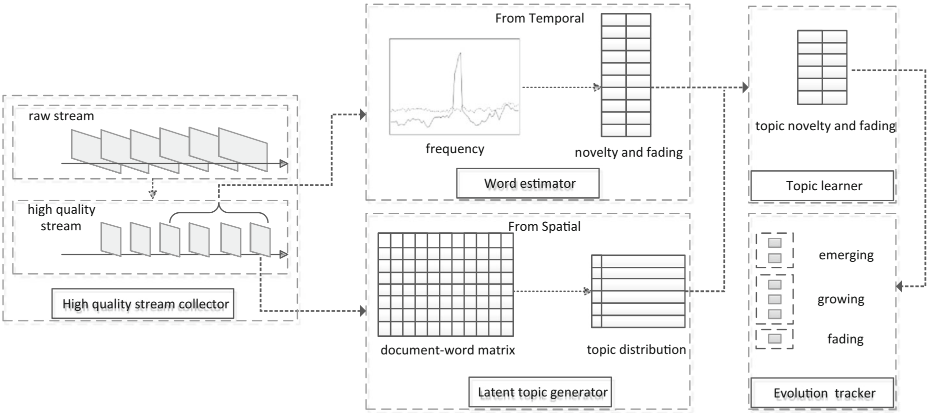


Figure 2 Overview of the ETT solution

4 Emerging topic tracking in microblog streams

We first discuss how to estimate the novelty and fading probabilities for terms, and then discuss how to learn these two probabilities for each latent topic in current microblog stream.

4.1 Term novelty and fading

In a high quality microblog stream, an emerging topic usually contains terms that their frequencies have sharp increments (i.e. emerging words) in current time slice. In some relevant studies, a primary metric of evaluating term novelty is to measure the difference between the term's predicted frequency and its actual frequency [6, 7, 18, 28, 33–35], which generally can be simplified as a z-score as follows:

$$z(x) = (y - \mu)/\sigma \tag{1}$$

where y denotes the term actual frequency in current time slice, μ denotes the average frequency in a historical time window and σ denotes the variance.

This kind of novelty measurement utilizes the term's average frequency in historical stream as its predicted frequency of current state. However, according to our analysis of the dataset, some terms also have significant increments (i.e., a high actual frequency) though their related topic has appeared before. This is because this topic is still very hot and receives more attention in current stage than in previous stage (i.e., presenting a growing trend). Therefore, the predicted frequencies of these terms will be low and thus their novelty levels are actually overvalued under (1).

In this study, we leverage local weighted linear regression (LWLR) to alleviate the problem of overvaluing the term's novelty of appearing in a growing trend.

Given an infinite high quality microblog stream, a term frequency table is first introduced to monitor term frequencies in a time window. Set TF to be the term frequency table, where each row records the frequencies $tf_t^i (i = 0, 1, ..., S - 1)$ of a term t appearing in the time window. Then the term frequency in current slice can be predicted based on its changing trend in the time window via LWLR. Suppose that the term frequency changing trend can be fitted by a linear function $h_t = ax + b$. Then the parameter a and b can be estimated by minimizing the errors between fitted value h_t^i and actual frequency $tf_t^i (i = 0, 2, ..., S - 1)$ as follows:

$$J(a, b) = \frac{1}{2} \sum_{i=0}^{S-1} w^i (h_t^i - tf_t^i)^2 \tag{2}$$

where the weight $w^i = e^{-\frac{(x^i - x^S)^2}{2\rho^2}}$ is used to control the impact of different slices in the time window on the estimation result. More historical data will have less impact on predicting the term frequency in the current state.

Through LWLR, a term's novelty can be calculated as follows:

$$nov_t = \begin{cases} \frac{w^S (tf_t^S - h(x^S))^2}{\sum_{i=0}^{S} w^i (tf_t^i - h(x^i))^2} & if \quad tf_t^S - h(x^S) > 0 \\ \varepsilon & otherwise \end{cases} \tag{3}$$

In LWLR, if a term's frequency shows an increasing trend in the time window, which usually appears in a growing trend, then the term's predicted frequency h_t^S will exceed the latest actual frequency tf_t^{S-1} in the time window. Thus, the term's novelty under the LWLR prediction will be much lower than the novelty under (1). If a term first appears in the window, it will achieve a comparable novelty to that under (1). Figure 3 shows an example of these two cases. If a term's predicted frequency h_t^S is lower than the actual frequency tf_t^S, then the frequency tends to decrease, as a result, its novelty will be very small(i.e., ε).

At last, we analyze the term novelty and fading from probabilistic theory. On one hand, suppose that all the terms in the microblog stream are independent and identically distributed, then, $\sum_{t \in V_S} p(f_t^S) = 1$, where V_S is the vocabulary of the current slice. On the other hand, the sum of a term's novelty probability and fading probability under the condition of term t should be 1, i.e., $p(n/t) + p(f/t) = 1$, where $p(n/t) = nov_t$. Thus, the joint probability can be estimated as:

$$p(n, t) = p(n/t) p(f_t^S) \tag{4}$$

$$p(f, t) = (1 - p(n/t)) p(f_t^S) \tag{5}$$

These two kinds of probabilities are used for topic novelty and fading probabilities learning.

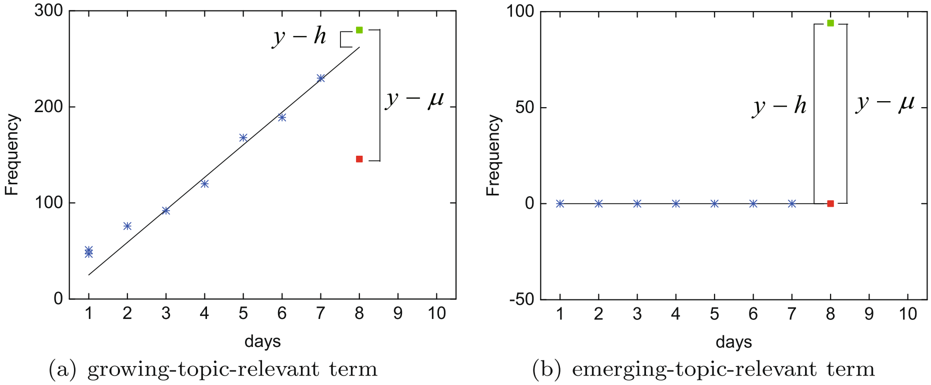
4.2 Topic generation

Note that all the topics appearing in current slice, including both emerging topics and others, are all generated from the microblog feeds of current stream. As the major task of this study is to track emerging topics by evaluating their novelty and fading, a topic model is employed as a module of the ETT method to generate coherent latent topics from the high quality microblog stream.

One of the most popular topic models is LDA, in which a document is modeled as a mixture of multiple topics that can be generated via Gibbs sampling [11] or variational inference [4]. Besides LDA, Yan *et. al.* proposed a Biterm Topic Model (BTM) [9], in which topics are learnt from short texts by directly modeling the generation of word co-occurrence patterns (i.e. biterms) in the text corpus. In both models, each latent topic is presented as a significance probability as well as a probability distribution over a vocabulary.

In this paper, both topic models are employed to generate latent topics from the high quality microblog stream with inferring a topic-term distribution matrix Φ and a topic significance probability vector θ by inputting the number of latent topics K.

There are two considerations as to why we prefer to employ a topic model rather than design a clustering method to gather these emerging terms into topics. First, probabilistic



(a) growing-topic-relevant term (b) emerging-topic-relevant term

Figure 3 An example of two types of novelty estimation methods

topic models are robust to the latent topic numbers K as compared with term clustering methods [7, 18, 28, 33]. It can avoid missing any possible topics by setting a higher value of K than the actual number of topics in the stream. Second, presenting a topic as a word probability distribution facilitates for displaying how significant the topic is and how important a term to a topic, which can help people better understand what a topic talks about. Therefore, we think it is feasible to employ a topic model to generate latent topics from microblog streams. Figure 6 shows the topic model's performance from coherence and F1 score with K varying from 20 to 70. The result demonstrates that both models achieve higher F1 score with higher K, which in essence, acquire higher recall and slightly lower precision. The decrease of precision is because more noisy topics are generated, whereas the increase of recall is because some less significant topics are also dug out. Overall, the result shows that both topic models are reliable in mining latent topics. In this study, a high value of K is specified to dig out as many latent topics as possible.

4.3 Topic novelty and fading

Suppose that a high quality microblog stream is a mixture distribution of K latent topics $\mathcal{Z} = \{z_1, z_2, ..., z_K\}$ with their significance probabilities θ_k and term probability distributions $\Phi_k^T \in \mathbb{R}^V (k = 1, 2, .., K)$ over the entire vocabulary. These two groups of parameters have been inferred via a topic model, such as LDA or BTM. In addition, each term t in the vocabulary has already acquired a novelty probability $n_t = p(n, t)$ and a fading probability $f_t = p(f, t)$ in Section 4.1. The goal of this module is to learn the novelty probability $n_k = p(n, z_k)$ and fading probability $f_k = p(f, z_k)$ for each topic z_k.

On one hand, a topic's significance probability θ_k being shared by the topic novelty and fading, i.e. $n_k + f_k = \theta_k$. On the other hand, a term's novelty $p(n, t)$ and fading $p(f, t)$ should be shared by all the latent topics $z_k (k = 1, 2, ..., K)$ containing this term with probabilities $p(t/z_k) = \varphi_{kt}$. Thus, the expectation of a term's novelty probability is evaluated as:

$$Ep(n, t) = \sum_{k=1}^{K} p(n, z_k) p(t/z_k) = \sum_{k=1}^{K} n_k \varphi_{kt} \tag{6}$$

And the expectation of a term's fading probability is evaluated as:

$$Ep(f, t) = \sum_{k=1}^{K} p(f, z_k) p(t/z_k) = \sum_{k=1}^{K} f_k \varphi_{kt} \tag{7}$$

Therefore, the topic novelty n_k and fading f_k can be learnt by regularizing the square error between the observed values, n_t, f_t, and the expectations, $Ep(n, t)$, $Ep(f, t)$, with constraints as follows:

$$\min_{n_k, f_k} \sum_{t=1}^{|V|} (\sum_{k=1}^{K} n_k \varphi_{kt} - n_t)^2 + \sum_{t=1}^{|V|} (\sum_{k=1}^{K} f_k \varphi_{kt} - f_t)^2$$

$$s.t. \begin{cases} n_k + f_k = \theta_k \\ n_k \geq 0 \\ f_k \geq 0 \end{cases}, k = 1, 2, ..., K \tag{8}$$

This optimization function can be presented in a more simplified fashion as follows:

$$\min_{\mathbf{n}_Z,\mathbf{f}_Z} f(\mathbf{n}_Z,\mathbf{f}_Z) = \min_{\mathbf{n}_Z,\mathbf{f}_Z} \|\mathbf{n}_Z^T\Phi - \mathbf{n}_V\|_2^2 + \|\mathbf{f}_Z^T\Phi - \mathbf{f}_V\|_2^2$$
$$s.t. \begin{cases} \mathbf{n}_Z + \mathbf{f}_Z = \theta \\ n_k \geq 0 & , k = 1, 2, ..., K \\ f_k \geq 0 \end{cases} \tag{9}$$

where $\mathbf{n}_V$ and $\mathbf{f}_V$ are the novelty and fading vectors of the vocabulary V, Φ is the topic-term distribution probability matrix.

To solve this bi-convex problem (i.e., (9)), we employ an alternative direction method of multipliers (ADMM) [5] to optimize the objective function over the parameters $\mathbf{n}_Z$ and $\mathbf{f}_Z$ alternatively. With ADMM, the objective function has the augmented Lagrangian form:

$$L(\mathbf{n}_Z,\mathbf{f}_Z,\lambda) = f(\mathbf{n}_Z,\mathbf{f}_Z) + \frac{\rho}{2}\|\mathbf{n}_Z + \mathbf{f}_Z - \theta + \lambda\|_2^2$$
$$s.t., n_k \geq 0,\ f_k \geq 0, k = 1, 2, ..., K \tag{10}$$

Then, update $\mathbf{n}_Z$ with fixed $\mathbf{f}_Z$ and λ, and alternatively update another parameter with the rest fixed as follows:

$$\begin{cases} \mathbf{n}_Z^{t+1} & = \arg\min_{n_k\geq 0} f(\mathbf{n}_Z,\mathbf{f}_Z^t) + \frac{\rho}{2}\|\mathbf{n}_Z + \mathbf{f}_Z^t - \theta + \lambda^t\|_2^2 \\ \mathbf{f}_Z^{t+1} & = \arg\min_{f_k\geq 0} f(\mathbf{n}_Z^{t+1},\mathbf{f}_Z) + \frac{\rho}{2}\|\mathbf{n}_Z^{t+1} + \mathbf{f}_Z - \theta + \lambda^t\|_2^2 \\ \lambda^{t+1} & = \lambda^t + \mathbf{n}_Z^t + \mathbf{f}_Z^t - \theta \end{cases} \tag{11}$$

In the experimental implementation, the topic-term distribution Φ is set to a sparse matrix, i.e., a topic is presented only with a limited number of the most relevant terms (i.e., top50 terms are used in our experiment). Two benefits follow from this setting: 1) It can speed up the ADMM-based learning process when using a sparse matrix; 2) It can enhance the topic coherence, because the topic in real-world microblog stream relates to a few words rather than the entire vocabulary. Furthermore, a distributed solution of ADMM [5] can also speed up the learning process when the ETT method is applied to large-scale streams.

There are many convergence results for ADMM discussed in the literature. Here, we employ a practical termination criterion introduced in [5], i.e., both primal and dual residuals are very small or the maximum number of iterations is reached.

4.4 Topic evolution operations

Based on topic novelty and fading probabilities, each topic z_k can be evaluated by a triplet (n_k, f_k, θ_k) and a word probability distribution. According to our observation, a topic appearing in current time window at the first time usually acquires a high value of n_k and a much low value of f_k, and vice versa if the topic has evolved for a long time. To sum up, a higher value of θ_k indicates greater significance of the topic and a higher value of n_k/f_k means more freshness of the topic.

Next, we discuss how to define the topic evolution operations according to existing indicators about a topic. Following the definitions of the three operations described in Section 3.2, these operations can be quantitatively defined from topic similarity and gained attention. Generally, Kullback-Leibler divergence (KLD) is used to evaluate the similarity between topic pairs in many relevant work [17, 36, 42]. In this study, we also employ KLD to measure the similarity between each pair of the latent topics in two adjacent time slices. For any topic z_k^i in current slice, if there exists a topic z_l^{i-1} in previous slice such that $KLD(z_k^i, z_l^{i-1}) < \xi$, then it indicates that topic z_k^i is a continuation of topic z_l^{i-1}. However, if there is no topic in previous slice such that $KLD(z_k^i, z_l^{i-1}) < \xi$, it does not mean that topic z_k^i must be an emerging topic. This topic may also be noise that consists of many nonsensical words and fails to express a topic coherently. This phenomenon does happen, especially when the latent topic number K is specified to a value that is much higher than the actual number of topics in the stream.

Then, we discuss how to distinguish growing trend from fading one, as well as emerging topic from noise. On one hand, the major difference between growing and fading trends is whether the attention trend is increasing or not, which can be measured by the topic novelty and fading learnt from the words it contains. A high novelty level and a low fading level denote that the topic is attracting more and more attention, and vice versa when the topic is fading and gaining less and less attention. On the other hand, an emerging topic usually presents significant burstiness while it does not happen on noise. Therefore, we can define the three evolution operations by considering both the topic similarity measured by KLD and the rate between topic novelty and fading.

Finally, the topic evolution can be captured by quantitative defining these operations (including noise) as follows:

- **Emerging**: a topic z_k is emerging in current time slice *iff* $n_k/f_k \geq \zeta$, $(\zeta > 1)$ and there is no topic in previous slice such that $KLD(z_k^i, z_l^{i-1}) < \xi$.
- **Growing**: for a latent topic z_k in current time slice, *iff* $n_k/f_k > 1$ and there exists a topic z_l^{i-1} in previous slice such that $KLD(z_k^i, z_l^{i-1}) < \xi$, then topic z_k is called a growing trend of its most similar topic z_l^{i-1} .
- **Fading**: for a latent topic z_k in current time slice, *iff* $n_k/f_k \leq 1$ and there exists a topic z_l^{i-1} in previous slice such that $KLD(z_k^i, z_l^{i-1}) < \xi$, then topic z_k is called a fading trend of its most similar topic z_l^{i-1}.
- **Noisy**: a latent topic z_k in current time slice is treated as a noisy topic *iff* $n_k/f_k < \zeta$ and there is no topic in previous slice such that $KLD(z_k^i, z_l^{i-1}) < \xi$.

By applying these three evolution operations, one can capture the evolution path of topics over timeline. By defining a noisy topic, one can focus only on the meaningful topics and exclude the noisy disturbance in tracking topic trends.

In this study, we only define three major operations and do not perform in-depth analysis on more topic evolution operations for two reasons. First is this study focuses more on tackling the topic novelty and fading probabilities. Second is existing work, such as [36] has discussed and defined eight kinds of evolution operations in detail. Therefore, the ETT method can be extended with introducing more evolution operations (i.e., merge, split) feasibly by controlling the thresholds of the similarity between topic-pair and the rate between topic novelty and fading levels in a fine-grained fashion.

5 Experiments and evaluation

In this section, we will report the results of an extensive performance study implemented on a large-scale of real-world Sina Weibo microblog stream. The experiments are designed to verify the effectiveness and efficiency of the proposed method in tracking emerging topics in a microblog stream.

Table 1 Statistics on Weibo stream

Average	Text number	Text length	Event number	Biterm number	Word number
Raw stream	32308.9	12.0	13.5	119635	6697.1
High quality stream	5503.3	18.1	13.5	43720	3994

5.1 Generation of microblog stream

To empirically evaluate the performance of the proposed emerging topic tracking method ETT, we employ microblogs published on Sina Weibo[1] for testing. Nevertheless, it is hard to evaluate the method's performance with real-world microblog streams without any labelling information. For evaluation purpose, we are interested in building a synthetic stream by crawling microblog feeds of discussing social issues. With the aid of our microblog crawler system, it is convenient to collect a set of microblog feeds that all of them contain specified keywords (i.e., *NBA* or *XP*) and belong to specified time interval. Each feed in this collection is labelled as the event-relevant feed and the collection is called an ERF collection (event-relevant feed collection). We finally crawled 21 ERF collections, including *A Bite of China*, *2014 NBA Playoffs*, totally containing over one million feeds spanning from April 1st to 30th, 2014. All the feeds in these ERF collections are mixed together in chronological order to form a raw microblog stream.

By convention, the time slice is set to one day. Due to the high noise of the raw stream, we first extract a small part of microblogs as a high quality stream. Here, 15 % of microblogs with the highest scores are selected from each slice. In pre-processing, we segment the sentences into word lists[2], remove meaningless words, including low frequency words (less than 5), stop-words and other characters not in Chinese or Latin, and exclude the texts with less than three words. Table 1 lists the statistical information of the dataset. Observed from Table 1, the average number of texts in each slice of the high quality stream is much lower than that of the raw stream, whereas the average event numbers are the same. It is because the representative microblog feeds of all events can be extracted from the raw stream. In addition, the average text length of the high quality stream is also much higher, as the feed with more words is more likely to be selected via the high quality microblog extraction method discussed in Section 3.1.

5.2 Experimental setting

To evaluate the proposed method, we introduce some indicators for evaluation and similar studies for comparison.

5.2.1 Performance indicators

Coherence We first introduce a coherence metric proposed in [22] for topic quality evaluation. Let f_l^s be the number of microblog feeds in the sth time slice of appearing term t_l, $f_{m,l}^s$ be the number of feeds that terms t_m and t_l co-occur, T be the total number of terms in

[1] http://weibo.com

[2] We use Jieba for Chinese word segmentation, which can be downloaded from https://github.com/fxsjy/jieba

each topic, then the topic coherence is defined as $C(z^s) = \sum_{t=1}^{T} \sum_{l=1}^{t} \log(f_{m,l}^s + 1)/f_l^s$. A *good* topic should have a high value of topic coherence. In the experiment, only top10 (i.e., $T = 10$) terms of each topic are used for calculating the indicators, including **Coherence**, F_1, **Novelty** and **F-measure**.

***F*₁** We hope a topic model can generate more meaningful topics that discuss meaningful events and fewer nonsensical topics that can't express any events. For this purpose, precision and recall are introduced here. Let $\{\mathcal{SW}^1, \mathcal{SW}^2, ..., \mathcal{SW}^\kappa\}$ be κ standard event-relevant term sets, which are selected from the κ labelled ERF collections based on their TF-IDF. Then, a detected topic z_k is judged as a real topic if its terms are also included by one of the standard event-relevant term sets. This topic is finally labelled as the event that shares the most number of terms. Thus, Precision is defined as the proportion of the detected topics that related to real events. Recall is defined as the proportion of the distinct real events detected by a method. As sometimes more than one detected topics are mapped with a same event, we only select the most relevant one when computing Recall. Finally, F_1 is defined as 2*Precision *Recall/(Precision+Recall).

Note that, F_1 is a rough indicator, as it judges a topic as true or false based on the standard event-relevant term sets. A topic that contains a part of the standard terms does not always express a significant and meaningful thing, because its keywords may come from different ERF collections. In this study, F_1 is used to evaluate the topic model's performance on both the raw stream and the high quality stream.

Novelty As one purpose of the proposed method is to detect emerging topics from microblog stream, we introduce Novelty indicator proposed in [35] to evaluate how the words in emerging topics vary over time. Let $\mathcal{W}^{s-1}$, $\mathcal{W}^s$ be two term sets of all detected emerging topics in two adjacent time slices $s - 1$ and s respectively, then the Novelty of the emerging topics is defined as $Novelty(z_k) = (|\mathcal{W}^s| - |\mathcal{W}^s - \mathcal{W}^{s-1}|)/(T * \widetilde{\kappa})$, where $\widetilde{\kappa}$ counts the emerging topics detected in current slice, and T denotes the number of words each topic contains.

F-measure F-measure [29] is introduced to evaluate topic tracking performance from text clustering, in which, each feed is assigned to the most similar topic cluster based on the cosine similarity between the feed and each of the tracked topics. Suppose that C_i is a feed cluster where all its elements are assigned to the same topic z_i, and K_j is a feed natural class where all its elements come from the same ERF collection. Then, F-measure for a pair of natural class K_j and cluster C_i are calculated as $F(K_j, C_i) = 2 * r(K_j, C_i) * p(K_j, C_i)/(r(K_j, C_i) + p(K_j, C_i))$, where $r(K_j, C_i) = n_{ij}/|K_j|$, $p(K_j, C_i) = n_{ij}/|C_i|$, and n_{ij} is the number of feeds of natural class K_j in cluster C_i. Finally, the overall F-measure is denoted as:

$$F = \sum_{j=1}^{\kappa} \frac{|K_j|}{N} \max_{C_i \in \mathcal{C}} \{F(K_j, C_i)\} \tag{12}$$

where $\mathcal{C}$ is the set of feed clusters, where each element is associated with one of the detected topics, and N denotes the total number of feeds in all the feed clusters.

5.2.2 *Baseline methods*

As the proposed emerging topic tracking method is on the basis of the latent topics inferred by the topic model, we first leverage other topic-model-based emerging topic detection/tracking methods for comparison.

- **OLDA** [17], where emerging topics are tracked as follows: 1) generates latent topics by Online-LDA, and 2) identifies emerging topics according to the Jensen-Shannon divergence (JSD) between the previous topic-word distribution and its subsequent update.
- **JSD-BTM**, a variant of BTM for emerging topic tracking, which is introduced as a baseline in [35]. JSD-BTM 1) generates latent topics by BTM model [9], 2) greedily matches the topic in current slice to the topics in previous slice according to cosine similarity, and 3) identifies emerging topics according to the JSD between a pair of topic-word distributions.
- **EWMA-BTM**, a method similar to ours, where the term novelty is estimated based on EWMA (Exponentially Weighted Moving Average) proposed in [28], and the topics are generated by BTM and finally the probabilities of topic novelty and fading are evaluated according to our proposed method (i.e., modules (3)-(5)). Here, the purpose of designing this method is to investigate how different term novelty estimation methods impact on the final detection performance.
- **BBTM** denotes bursty biterm topic model [35], which infers one background topic and K bursty topics along with their bursty probabilities.

In addition, we also introduce a feature-clustering method and a document-clustering method for comparison.

- **HUPC** [13], where top-k high utility patterns are mined from each time slice of the microblog stream and gathered into emerging topics and background topics via an incremental clustering method.
- **SFSD** [25], i.e., streaming first story detection, where new-coming microblog feed is always assigned to the topic cluster containing its most similar neighbor or a new cluster if the feed does not have such a similar neighbor.

Since HUPC and SFSD methods detect topics by clustering patterns and feeds, we selected top10 terms with the highest TF-IDF values from each of the detected clusters to represent this topic. As LDA and BTM are used in ETT solution respectively, our two methods are referred as E-LDA and E-BTM, where the latent topics are inferred by LDA and BTM respectively. Table 2 summarized the compositions of different methods, including

Table 2 Comparison of compositions of different methods

method	term novelty estimation	topic generation	detected topic types*
OLDA	none	LDA	e, b
JSD-BTM	none	BTM	e, b
EWMA-BTM	EWMA	BTM	e, g, f, n
BBTM	z-score	BBTM	e, b
E-LDA	LWLR	LDA	e, g, f, n
E-BTM	LWLR	BTM	e, g, f, n
HUPC	none	pattern clustering	e, b
SFSD	none	text clustering	e, b

* b denotes background topic, e denotes emerging topic, g denotes growing trend, f denotes fading trend, and n denotes noisy topic.

whether and how to estimate term novelty, how to generate topics and what type of topics are detected.

5.2.3 Parameter setting

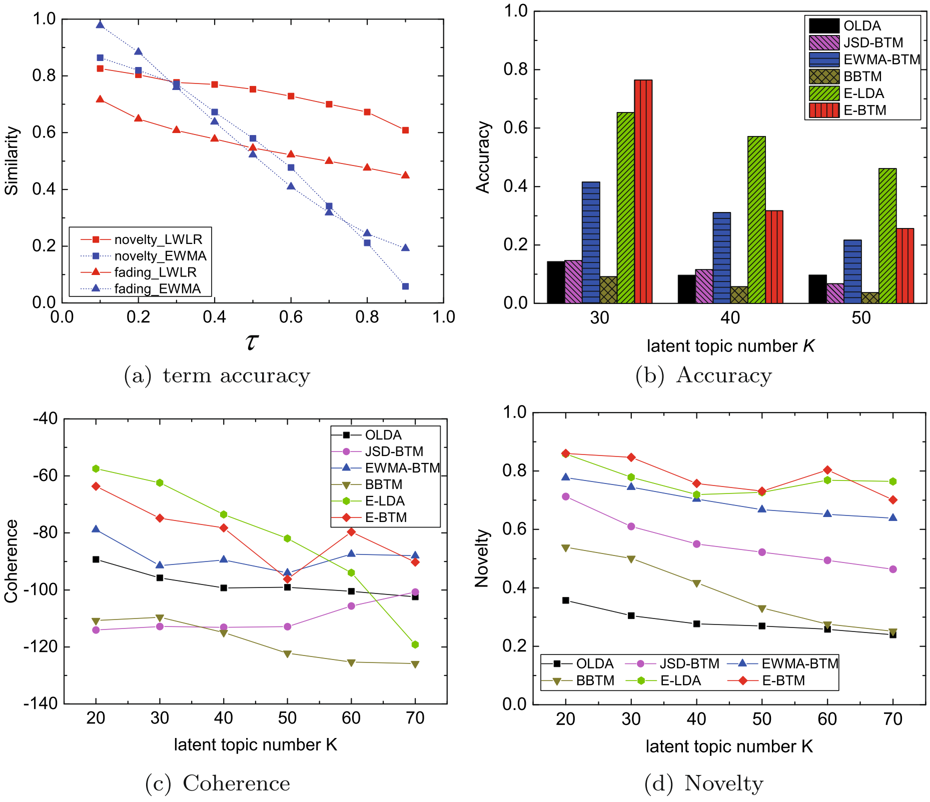
Similar to the references, the parameters α and β are set to 0.05 and 0.01 in LDA [9], 50/K and 0.01 in BTM [9] and BBTM [35], and 0.001 and 0.0l in OLDA [17]. The iterations are set to 50 for all topic models. For OLDA, and JSD-BTM, a topic is treated as an emerging topic if its JSD value exceeds 2.0. For BBTM and BTM, only the biterm with frequency exceeding 5 is used for topic inference. For HUPC, top-800 high utility patterns are mined for clustering and the cluster with less than 10 topic terms is discarded, because a cluster containing a few terms is difficult to express a meaningful thing. For SFSD, top-8000 terms are selected from the entire stream for building a unified vector space and the cluster with less than 10 feeds is discarded. Here, based the experimental result, we find it is more appropriated to set the minimum similarity between two feeds to 0.7 instead of 0.5 if they are assigned to the same topic. When the similarity threshold is too low, the method will produce too many clusters where most of them only contain a few feeds. The other parameters of the baselines are set to their default values in the papers.

For E-LDA and E-BTM, the parameter ρ in LWLR is set to 5. The length of time window is set to 8 days, as there are some events appearing weekly. For example, the number of microblog feeds about *A Bite of China* weekly climbs to a peak on the day after the broadcast of the documentary. For E-LDA, E-BTM and EWMA-BTM, the maximum number of iterations in ADMM algorithm is set to 500. The two parameters in topic evolution are set to $\xi = 2.0$ and $\zeta = 5$ respectively.

5.3 Emerging topic tracking

First, we evaluate the performance of our method on correctly estimating term novelty and fading. Here, a term is identified as an emerging term if its novelty probability exceeds τ or is identified as a fading term if its fading probability exceeds τ. An event is labelled as an emerging event of current time slice if none of the microblog feeds in its ERF collection has appeared in previous slice. Otherwise, it is labelled as a background event. We calculate the proportion of correctly identified emerging/fading terms by comparing the terms of detected emerging/fading topics with the standard term set of emerging/background events. For comparison, we also employ a term significance measurement (i.e., EWMA) proposed in [28] for term novelty and fading evaluation. The result is shown in Figure 4a. It can be seen that our method is more competitive in correctly estimating term novelty and fading level, especially when τ exceeds 0.5. Nevertheless, EWMA performs poor on identifying high novelty and high fading level terms. It is because many growing-topics-relevant terms also have very high values of novelty.

Next, we evaluate the performance of the proposed method on detecting emerging topics by manual labelling. Six volunteers are invited to label a detected emerging topic as true or false. For each detected emerging topic, 20 most relevant terms as well as two standard keyword sets selected from two adjacent time slices are provided for all volunteers. Consulting these two keyword sets, a topic is labelled true if most of the words discuss an event that appears in the current slice but does not appear in the previous slice. In addition, if a topic contains words that come from different event sets or talks about an advertisement, it will not be judged ”true”. An emerging topic is labeled as ”true” if more than half of the volunteers label it ”true”.



(a) term accuracy (b) Accuracy

(c) Coherence (d) Novelty

Figure 4 Performance of emerging topic detection

Figure 4b compares the accuracy of all topic-model-based methods with different settings of latent topic number K and Table 3 compares our two methods (K set to 30) with HUPC and SFSD. From the results, we can see 1) our two methods and EWMA-BTM outperform the other baselines significantly in correctly identifying emerging topics. This is mainly because, in these methods, the term novelty is estimated from temporal perspective, whereas the latent topics are generated from term-occurrence space independently. Then, these two results work together to identify emerging topics from growing trends and noisy topics. On the contrary, other methods usually lack of effective strategy to estimate the term novelty and noisy topics are also not excluded. In HUPC and SFSD methods, a large number of nonsensical topics are treated as emerging ones, which is because a topic is treated as

Table 3 Topic tracking performance comparison

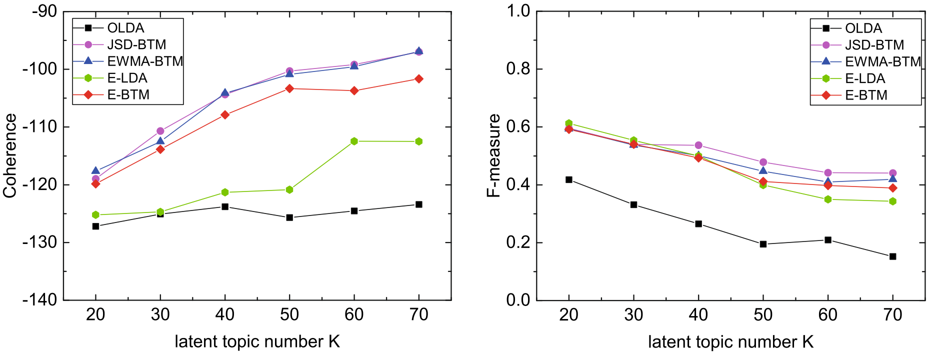
	Emerging topic detection performance			Topic tracking performance	
	Accuracy	Coherence	Novelty	Coherence	F-measure
E-LDA	0.50	-62.40	0.78	-124.69	0.55
E-BTM	0.77	-74.82	0.85	-113.87	0.54
HUPC	0.12	-124.81	0.18	-105.13	0.42
SFSD	0.18	-98.07	0.29	-117.77	0.43

emerging if it does not similar to any existing topics (i.e., background topics). As a result, they usually can obtain a group of meaningful and coherent background topics while a group of low quality emerging topics which are mixture with many noisy topics. In OLDA and JSD-BTM methods, growing trends are usually treated as emerging topics, as these methods only consider the difference between word distributions of topics in two consecutive time intervals while ignore the term novelty. 2) EWMA-BTM performs poorer than ours, because some growing trends are also treated as emerging topics. The further reason is that the terms referring to these growing trends usually acquire high novelty probabilities, which can also be verified from Figure 4a. 3) BBTM always performs the worst, because all the K latent topics are regarded as emerging topics, which inevitably introduces some background topics. 4) All the methods work poorer with increasing K due to more duplicate topics generated. At last, we note that E-BTM achieves higher accuracy than E-LDA when K=30 while lower accuracy when K is larger. The reason is that more noisy topics are generated by LDA with increasing K but not for BTM. These noisy topics can be identified by the proposed topic evolution operations and discarded when recommending emerging topics. On the contrary, BTM can generate coherent topics though K is larger, however, the duplicate emerging topics may be treated as emerging topics, thus leading to lower accuracy. This observation can also be seen in Figure 6a, which shows the increment of the topic coherence of BTM with K.

At last, we evaluate the performance of emerging topic tracking on topic coherence and novelty with increasing latent topic number K. Two observations can be drawn from the results displayed in Figures 4c, d and Table 3. First, the proposed method always achieves better performance than baseline methods on both coherence and novelty. It is because the method works better in tracking emerging topics by distinguishing them from growing trends and noisy topics, thus resulting in higher novelty and coherence. Second, the coherence of E-LDA decreases significantly with increasing K, while it does not happen on E-BTM, JSD-BTM and EWMA-BTM. The reason is the latent topics generated by different topic models (i.e., LDA and BTM) show different coherence with increasing K (see Figure 6a).

5.4 Topic evolution tracking

In this part, we investigate the performance of topic evolution tracking in terms of background topics. Here, a topic is called a background topic if it is just a continuation of some topic that has appeared in previous time slice. First, we evaluate the topic coherence and the performance of the microblog feed classification to background topics. A feed is classified into the most relevant background topic based on the similarity between the feed content and each of the topics. Here, F-measure is employed to evaluate the classification result. Figure 5 presents the results of coherence and F-measure with K varying from 20 to 70 and Table 3 presents the results of our two methods, HUPC and SFSD. From Figure 5, we can see 1) BTM-based methods always work better than the LDA-based methods, due to that fewer noisy topics are generated by BTM. 2) JSD-BTM performs best on both coherence and F-measure, which is because real emerging topics, noisy topics and some growing trends are all detected as emerging topics via this method. As a result, most of the rest topics are related to real events, show higher coherence and make the microblog feeds to be classified into corresponding topic cluster with higher F-measure. 3) For all these methods, topic coherence increases with increasing K whereas F-measure decreases with increasing K. We therefore analyze the reason why E-LDA and E-BTM achieve this result. On one hand, more latent topics are generated with increasing K, and more duplicate topics



(a) Coherence (b) F-measure

Figure 5 Performance of background topic tracking

are generated at the same time. These duplicate background topics disperse the feeds into more topic clusters and lead to a lower F-measure. On the other hand, despite of the fact that more topics are generated, the noisy topics still can be identified and discarded by the E-LDA and E-BTM, making the background topics with higher coherence to be retained. From Table 3, we can see HUPC achieves higher coherence but lower F-measure than ours. It is because it detects emerging topics via considering whether the new-coming pattern is similar enough to one of the background topics. Thus, the detected background topics are usually more coherent and meaningful whereas the emerging topics may contain noise. In addition, too much duplicated background topics also results in a low F-measure.

Next, we track the topic evolution by displaying its top5 terms in sequential time slices. Let $\mathcal{Z}_s = \{z^1, z^2, ..., z^s\}$ be a topic set of s days that evolves from emerging to fading. As each topic may evolve into several trends in next day, we select the best matching one for presentation. Table 4 displays the mainstream topic trends of *2014 NBA* from April 15th to 21th. Here, the latent topics are generated by BTM model with K setting to 40. As the relevant microblog feeds were collected from April 15th, this topic is successfully tracked as an emerging topic by E-BTM on that day. It shows significant difference from topics in previous day, which is manifested as a high KLD value (2.97) and a high rate between topic novelty and fading (5.01). Then its evolution can be tracked in the following days and presented as growing or fading trends, depending on how much attention it attracts. A significant increment of the feed counts is usually associated with a growing trend, whereas a decrement of the feed counts indicates that this topic gains less attention and behaves in a fading trend. Note that the topic also evolves into a fading trend when the feed counts increases slightly. It is because the term novelty is evaluated based on LWLR, which predicts a higher frequency for a term when its frequency shows an increasing trend in the historical time window.

Table 5 displays the topic evolution process of *A bite of China (the second season)*, a documentary movie shown on television about the history of food, eating and cooking in China, which was also collected from April 15th. The first season of this documentary gained widespread popularity in 2012, making the second season attracts extensive attention before showtime (April 15th to17th). Different from the topic trend of *NBA*, this topic shows a sharp increment trend on April 18th, Friday, the day when the documentary began to broadcasting on CCTV. In the next several days, many people watched the documentary

Table 4 Topic evolution of *2014 NBA*

date	topic id	topic terms	θ_k	state	n_k/f_k	KLD	feed counts
4/15	39	NBA, team, basketball, sport, Rockets	0.014	emerging	5.01	2.97	286
4/16	33	NBA, Spurs, Rockets, Bull, James	0.025	fading	0.78	1.54	258
4/17	9	NBA, playoffs, Thunder, Heat, Spurs	0.034	growing	3.51	1.35	477
4/18	8	NBA, Spurs, east, champion, regular	0.036	fading	0.40	1.46	337
4/19	33	NBA, playoffs, today, Thunder, Rockets	0.017	fading	0	1.45	371
4/20	17	NBA, Thunder, playoffs, today, Pacers	0.044	growing	1.92	0.77	575
4/21	9	NBA, playoffs, Rockets, home, Heat	0.028	fading	0.89	1.41	601

and discussed it on Sina Weibo. Note that, although the rate between topic novelty and fading level is as high as 16.2 on April 18th, this topic still can be accurately tracked as a growing trend rather than an emerging topic. This success should be attributed to the KLD measurement used for topic evolution tracking. At last, this topic evolved into a noise due to that the core terms had changed from *tongue* and *documentary* to *documentary* and *global*.

Table 5 Topic evolution of*A bite of China*

date	topic id	topic terms	θ_k	state	n_k/f_k	KLD	feed counts
4/15	31	tongue, China, broadcast, food, CCTV	0.017	emerging	33.4	2.77	234
4/16	27	tongue, food, season, CCTV, broadcast	0.023	fading	0.54	1.28	277
4/17	30	tongue, food, season, broadcast, CCTV	0.025	fading	0.54	0.84	325
4/18	6	tongue, tonight, season, food, CCTV	0.078	growing	16.2	0.78	1526
4/19	3	tongue, season, footstep, China, food	0.060	fading	0.40	1.03	1600
4/20	7	China, documentary, tongue, food, flavour	0.044	fading	0.21	1.26	1335
4/21	37	documentary, works, global, hard, America	0.018	noisy	1.24	2.06	1362

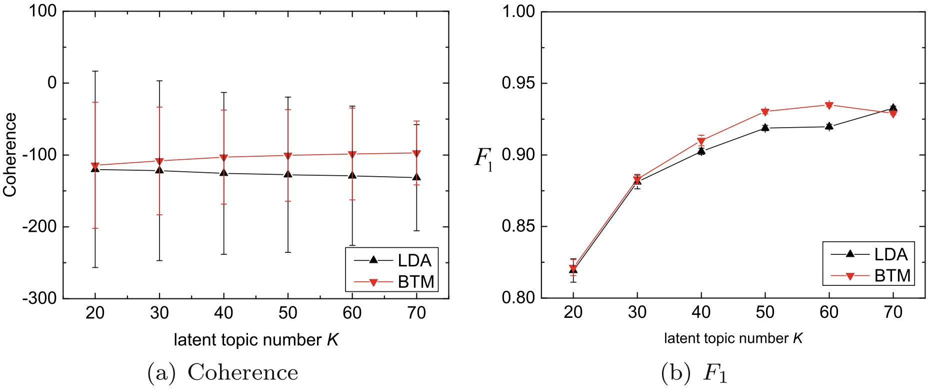
Table 6 Topic quality in raw stream and high quality stream generated via LDA and BTM

	Streams	Coherence	Precision	Recall	F1	Runtime(.sec.)
LDA	Raw	-144.712	0.997	0.828	0.901	759.690
	High quality	-125.527	0.973	0.845	0.903	197.293
BTM	Raw	-145.724	0.998	0.876	0.930	1769.694
	High quality	-102.971	0.978	0.857	0.910	475.512

5.5 Performance of topic models on different streams

High quality Stream In this study, emerging topics are tracked from the high quality microblog stream instead of the raw stream. It is because the high quality microblog stream contains less noise, which is amenable to generating more coherent latent topics. Here, we investigate the performance of LDA and BTM when they are employed to generate latent topics from the raw stream and the high quality stream respectively. Table 6 presents the average topic coherence, precision, recall, F_1 and runtime. Here, K is set to 40. Observed from Table 6, on one hand, the topics generated from the high quality stream are usually more coherent, and the F_1 is comparable to that from the raw stream. It is because most of the meaningless texts are excluded from the high quality stream and thus fewer nonsensical topics are generated. Meanwhile, it speeds up the topic generation process significantly when only 15 % of microblog feeds are selected to build the high quality stream. On the other hand, both LDA and BTM achieve promising performance in terms of F_1 score, whereas the topics inferred by BTM are more coherent, which is at the cost of more time consumption.

Topic model Next, we evaluate the performance of the two topic models when they are implemented on the high quality stream with K ranging from 20 to 70. The results are shown in Figure 6. From Figures 6a and b, we find LDA achieves competitive F_1 score but lower coherence as compared with BTM. In addition, both models can achieve promising F_1 score results with increasing K. It states that both models are credible for generating latent topics from the streams, which is consistent with our assumption. However, based on the error-bar shown in Figure 6a, it is obvious that BTM can generate more coherent



(a) Coherence (b) F_1

Figure 6 Performance of different topic models on high quality microblog stream

topics, which presents lower variance among different topics. At last, note that, both models achieve higher recall while lower precision with increasing K. It is because they generate more duplicate topics about the same event when K increases.

5.6 Parameter tuning

In this part, we investigate the impact of parameters ξ and ζ on the final result of emerging topic tracking, as the evolution operations are defined based on them. Here, we show their impact on emerging topic detection with latent topics generated by BTM. Figure 7 presents the average coherence and novelty of all emerging topics with K setting to 20, 40 and 60 respectively. Note that the parameter ξ is set to two different values, i.e., 1.5 and 2.0, because the variance of the minimum KLD among different topic pairs is small. A high value of ξ usually indicates the missing of some insignificant emerging topic, whereas a low value of ξ indicates some background topics with significant variations mistakenly judged as emerging topics.

From Figure 7, we find 1) ξ has less impact on the final result, especially on the Novelty indicator. This may be because emerging topics and noisy topics usually have much higher minimum KLD than the background trends. 2) A low value of ζ (e.g., ζ=1 or ζ= 3) has significant impact on the final result,which is because a low value of ζ makes some growing trends to be judged as emerging topics. However, such errors become less when ζ(e.g.,

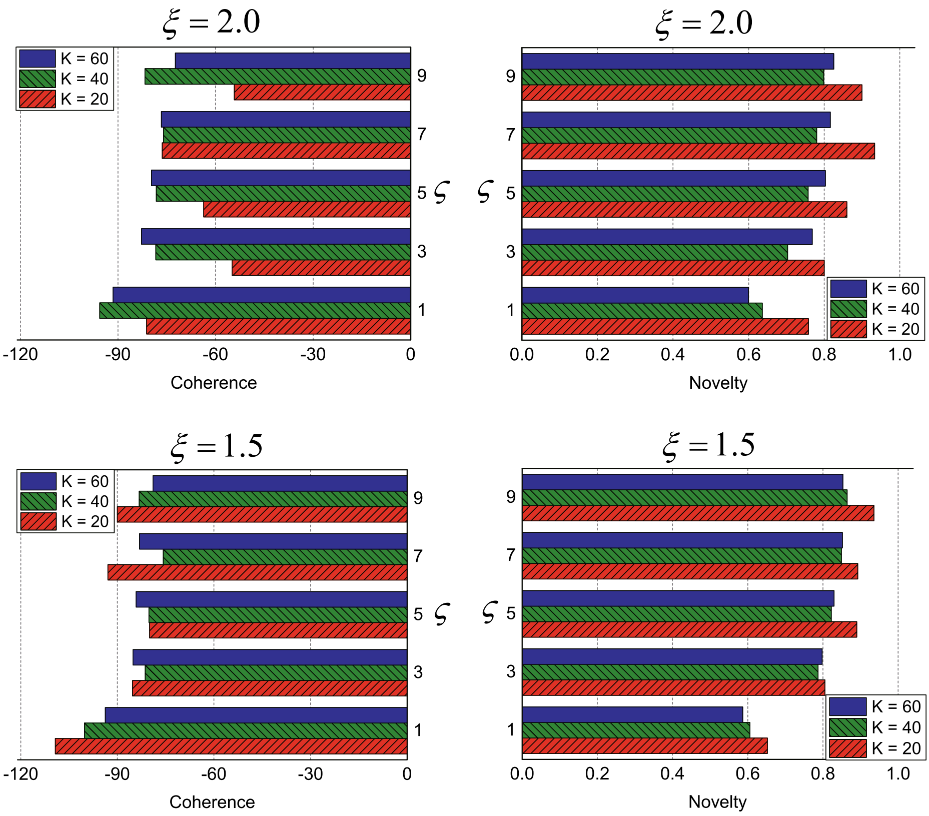


Figure 7 Parameter tuning of E-BTM

Table 7 Runtime of per day (sec.)

	Pre-processing	Topic inference	Post-processing	Total
OLDA	None	100.64	0.09	100.73
JSD-BTM	3.36	475.51	0.78	479.65
EWMA-BTM	1.93	475.51	0.92	478.36
BBTM	4.39	596.71	None	601.1
HUPC	66.06	78.65	0.46	145.17
SFSD	None	808.97	4.54	813.51
EB-BTM	4.7	475.51	1.33	481.54
EB-LDA	1.35	197.29	1.35	199.99

$\zeta \geq 5$) increases, and the result tends to become stable finally. 3) These two parameters' impact on final result is more significant when K is set to 20. Although the average number of events in each time slice is less than 20, the percentages of relevant feeds of different events are quite difference. Some popular background events may take up over 50 % of texts, whereas some emerging events take up less than 10 %. As a result, the emerging event with only a few feeds can not be inferred by the topic model, whereas those popular events can be inferred with the presentation of several trends. Therefore, emerging topic detection based on these latent topics is more sensitive to the parameters. However, when K is large enough, almost all events can be inferred by the topic model. In this case, the proposed ETT achieves much more stable performance on emerging topic detection.

5.7 Efficiency

The efficiency is also very important when ETT is applied to real-world microblog stream. In this part, we compare the runtime of these eight methods. To be fair, all methods are implemented in PYTHON27, which are carried out on a Window 7 server with Intel Core i5 3.20GH CPU and 4G memory. Collapsed Gibbs sampling is employed for all these methods. Table 7 lists the average execution time per time slice for emerging topic tracking. Here, K is set to 40 for topic-model-based methods. Since the proposed method contains three parts, i.e., a pre-processing of term novelty evaluation, latent topic generation and a post-processing of topic evolution tracking. We list the time cost of each of the three parts to explain which part is the most time-consuming. It can be seen that the major time consumption of all methods except HUPC is the topic inference process, which costs much time than the rest. In HUPC, high utility pattern mining costs comparable time to the pattern clustering. On the contrary, term novelty evaluation via LWLR and topic novelty learning via ADMM method cost only a few seconds which almost can be ignored. In addition, note that BTM is much more time-consuming than LDA, but generates more coherent topics in return.

6 Conclusions

In this study, we propose a novel method, ETT, to track emerging topics in the microblog stream. The goal of our method is to track emerging topics as early as possible and track their evolution at the same time. For this purpose, we first estimate term novelty by employing local weighted linear regression (LWLR). We find that LWLR has satisfying performance on estimating term novelty and fading level, i.e., it assigns high novelty to emerging-topic-relevant terms and low novelty to background-topic-relevant terms. We then formulate the learning process of topic novelty and fading probabilities as an optimization problem and solve it with alternative direction method of multipliers (ADMM). Based on these two probabilities, topic evolution operations are defined subsequently to identify emerging topics from the large amount of latent ones and track how these topics evolve over time.

Several future directions remain for us to explore. One interesting direction is to investigate how to depict topics evolution with more operations. As discussed in Section 5.6, two parameters (i.e., the rate between novelty and fading probabilities, and the minimum KLD) have significant impact on the tracking result, we therefore plan to investigate the relationship between them and design a group of more-refined evolution operations based on them.

Acknowledgments The work was supported by the National Science Foundation of China (NSFC, No. 61472291 and No. 61303115), and Interdisciplinary Foundation of Independent Scientific Research (No. 2042016kf0182).

References

1. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: WSDM, pp. 183–194 (2008)
2. AlSumait, L., Barbar, D., Domeniconi, C.: On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In: ICDM, pp. 3–12 (2008)
3. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: ICML, pp. 113–120 (2006)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. JMLR **3**, 993–1022 (2003)
5. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. FTML3(1), 1–122 (2011)
6. Cai, H., Huang, Z., Srivastava, D., Zhang, Q.: Indexing evolving events from tweet streams. TKDE **27**(11), 3001–3015 (2015)
7. Chen, Y., Amiri, H., Li, Z., Chua, T.S.: Emerging topic detection for organizations from microblogs. In: SIGIR, pp. 43–52 (2013)
8. Chen, Z., Liu, B.: Mining topics in documents: Standing on the shoulders of big data. In: SIGKDD, pp. 1116–1125 (2014)
9. Cheng, X., Yan, X., Lan, Y., Guo, J.: BTM: Topic model over short texts. TKDE **26**(12), 2928–2941 (2014)
10. Diao, Q., Jiang, J., Zhu, F., Lim, E.P.: Finding bursty topics from microblogs. In: ACL, pp. 536–544 (2012)
11. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci **101**, 5228–5235 (2004)
12. Hofmann, T.: Probabilistic latent semantic indexing. In: SIGIR, pp. 50–57 (1999)
13. Huang, J., Peng, M., Wang, H.: Topic detection from large scale of microblog stream with high utility pattern clustering. In: Proceedings of the 8th Workshop on Ph. D. Workshop in CIKM, pp. 3–10 (2015)
14. Iwata, T., Watanabe, S., Yamada, T., Ueda, N.: Topic tracking model for analyzing consumer purchase behavior. In: IJCAI, pp. 1427–1432 (2009)
15. Jeffery, S.R., Garofalakis, M., Franklin, M.J.: Adaptive cleaning for RFID data streams. In: VLDB, pp. 163–174 (2006)
16. Kasiviswanathan, S.P., Melville, P., Banerjee, A., Sindhwani, V.: Emerging topic detection using dictionary learning. In: CIKM, pp. 745–754 (2011)
17. Lau, J.H., Collier, N., Baldwin, T.: On-line trend analysis with topic models: Twitter trends detection topic model online. In: COLING, pp. 1519–1534 (2012)
18. Li, C., Sun, A., Datta, A.: Twevent: segment-based event detection from tweets. In: CIKM, pp. 155–164 (2012)
19. Lin, T., Tian, W., Mei, Q., Cheng, H.: The dual-sparse topic model: mining focused topics and focused terms in short text. In: WWW, pp. 539–550 (2014)

20. Ma, J., Sun, L., Wang, H., Zhang, Y., Aickelin, U.: Supervised anomaly detection in uncertain pseudoperiodic data streams. ACM Trans. Internet Technol. (TOIT) **16**(1), 4 (2016)
21. McAuley, J., Leskovec, J.: Hidden factors and hidden topics: understanding rating dimensions with review text. In: Recommender Systems, pp. 165–172 (2013)
22. Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: EMNLP, pp. 262–272 (2011)
23. Nallapati, R.M., Ditmore, S., Lafferty, J.D., Ung, K.: Multiscale topic tomography. In: SIGKDD, pp. 520–529 (2007)
24. Peng, M., Huang, J., Fu, H., Zhu, J., Zhou, L., He, Y., Li, F.: High quality microblog extraction based on multiple features fusion and time-frequency transformation. In: WISE, pp. 188–201 (2013)
25. Petrovi, S., Osborne, M., Lavrenko, V.: Streaming first story detection with application to Twitter. In: NAACL, pp. 181–189 (2010)
26. Pu, X., Jin, R., Wu, G., Han, D., Xue, G.: Topic modeling in semantic space with keywords. In: CIKM, pp. 1141–1150 (2015)
27. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: Real-time event detection by social sensors. In: WWW, pp. 851–860 (2010)
28. Schubert, E., Weiler, M., Kriegel, H.P.: Signitrend: Scalable detection of emerging topics in textual streams by hashed significance thresholds. In: SIGKDD, pp. 871–880 (2014)
29. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: KDD Workshop on text mining, pp. 525–526 (2000)
30. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with Twitter: What 140 characters reveal about political sentiment. In: ICWSM, pp. 178–185 (2010)
31. Unankard, S., Li, X., Sharaf, M.A.: Emerging event detection in social networks with location sensitivity. JWWW **18**(5), 1–25 (2014)
32. Wang, X., McCallum, A.: Topics over time: A non-Markov continuous-time model of topical trends. In: SIGKDD, pp. 424–433 (2006)
33. Weng, J., Lee, B.S.: Event detection in Twitter. In: ICWSM, pp. 401–408 (2011)
34. Xie, W., Zhu, F., Jiang, J., Lim, E.P., Wang, K.: Topicsketch: real-time bursty topic detection from Twitter. In: ICDM, pp. 837–846 (2013)
35. Yan, X., Guo, J., Lan, Y., Xu, J., Cheng, X.: A probabilistic model for bursty topic discovery in microblogs. In: AAAI Conference on artificial intelligence, pp. 353–359 (2015)
36. Yang, X., Ghoting, A., Ruan, Y., Parthasarathy, S.: A framework for summarizing and analyzing Twitter feeds. In: SIGKDD, pp. 370–378 (2012)
37. Yao, W., He, J., Wang, H., Zhang, Y., Cao, J.: Collaborative topic ranking: Leveraging item Meta-Data for sparsity reduction. In: AAAI, pp. 374–380 (2015)
38. Yin, J., Wang, J.: A dirichlet multinomial mixture model-based approach for short text clustering. In: SIGKDD, pp. 233–242 (2014)
39. Yin, H., Cui, B., Lu, H., Huang, Y., Yao, J.: A unified model for stable and temporal topic detection from social media data. In: ICDE, pp. 661–672 (2013)
40. Zhang, H., Kim, G., Xing, E.P.: Dynamic topic modeling for monitoring market competition from online text and image data. In: SIGKDD, pp. 1425–1434 (2015)
41. Zhu, J., Xing, E.P.: Sparse topical coding. In: UAI, pp. 831–838 (2011)
42. Zhu, J., Peng, M., Huang, J., Qian, T., Huang, J., Liu, J., Hong, R., Liu, P.: Coherent topic hierarchy: A strategy for topic evolutionary analysis on microblog feeds. In: WAIM, pp. 70–82 (2015)