# *BursT:* A Dynamic Term Weighting Scheme for Mining Microblogging Messages

Chung-Hong Lee, Chih-Hong Wu, and Tzan-Feng Chien

Dept of Electrical Engineering, National Kaohsiung University of Applied Sciences,
Kaohsiung, Taiwan
leechung@mail.ee.kuas.edu.tw, williamwu.tw@gmail.com,
tzanfeng@dml.ee.kuas.edu.tw

**Abstract.** One of the basic human needs is to exchange information and socialize with each other. Online microblogging services such as Twitter allow users to post very short messages related to everything ranging from mundane daily life routines to breaking news events. A key challenging issue of mining such social messages is how to analyze the real-time distributed messages and extract significant features of them in a dynamic environment. In this work, we propose a novel term weighting method, called *BursT,* using sliding window techniques for weighting message streams. The experimental results show that our weighting technique has an outstanding performance to reflect the shifts of concept drift. The result of this work can be extended to perform a periodic feature extraction, and also be able to integrate other sophisticated clustering methods to enhance the efficiency for real-time event mining in social networks.

**Keywords:** information retrieval, text mining, term weighting scheme, social networks, social mining

## 1 Introduction

### 1.1 Motivation

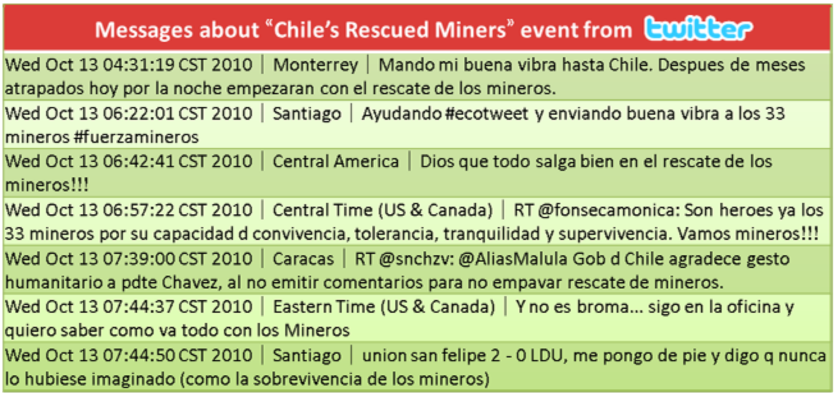
Recently, with a fast growing internet users keep up with newest information through text based microblogging streams like Twitter, searching for messages related to a hot news topic from the internet has been becoming an important activity for daily information acquisition. One of the key advantages of microblog is that it enables people to achieve a near real-time information awareness. For instance, Twitter has long offered users a list of real-time "trending topics" to explore--essentially all tweets containing the most-popular terms, such as "Haiti" or "Google Zeitgeist". This imposes an increasing need for utilizing automatic techniques to analyze and present correlative messages to a general reader in an efficient manner. Unfortunately, a key challenging issue of mining such social messages is how to analyze the real-time distributed messages and extract significant features of them in a dynamic environment. The view is taken, therefore, in this work we propose a novel term weighting method, called *BursT,* using sliding window techniques to tackle the challenges mentioned above.

D. Liu et al. (Eds.): ISNN 2011, Part III, LNCS 6677, pp. 548–557, 2011.

### 1.2 Problem Statements

In order to get insight into microblogging operations, the characteristics of microblogging messages are listed as follows:

- **Tremendous**: Microblogs have been becoming popular services with explosive subscribers in recent years. A statistic report[1] from Twitter reveals the message streams are over 1,000 TPS (Tweets/sec).
- **Text Un-integrality**: In order to shorten time of texting messages, most microblogging services limit the total length of a post. This leads to a problem with lack of semantic integrality in messages.
- **Time Sensitiveness**: Real-time operation is an essential factor for the use of microblogging messages. User posts a tweet with a timestamp indicating what someone says has happened (or will happen) at the specific time point.



Messages about "Chile's Rescued Miners" event from twitter

Wed Oct 13 04:31:19 CST 2010 | Monterrey | Mando mi buena vibra hasta Chile. Despues de meses atrapados hoy por la noche empezaran con el rescate de los mineros.

Wed Oct 13 06:22:01 CST 2010 | Santiago | Ayudando #ecotweet y enviando buena vibra a los 33 mineros #fuerzamineros

Wed Oct 13 06:42:41 CST 2010 | Central America | Dios que todo salga bien en el rescate de los mineros!!!

Wed Oct 13 06:57:22 CST 2010 | Central Time (US & Canada) | RT @fonsecamonica: Son heroes ya los 33 mineros por su capacidad d convivencia, tolerancia, tranquilidad y supervivencia. Vamos mineros!!!

Wed Oct 13 07:39:00 CST 2010 | Caracas | RT @snchzv: @AliasMalula Gob d Chile agradece gesto humanitario a pdte Chavez, al no emitir comentarios para no empavar rescate de mineros.

Wed Oct 13 07:44:37 CST 2010 | Eastern Time (US & Canada) | Y no es broma... sigo en la oficina y quiero saber como va todo con los Mineros

Wed Oct 13 07:44:50 CST 2010 | Santiago | union san felipe 2 - 0 LDU, me pongo de pie y digo q nunca lo hubiese imaginado (como la sobrevivencia de los mineros)

**Fig. 1.** A sample list of messages about the event of "Chile's Rescued Miners" from Twitter

In Figure 1, a sample list of messages regarding the event of "Chile's Rescued Miners" from Twitter is illustrated. In the message list, we found some characteristics within such short, timely and text-based user generated contents: a rapid growing amount of messages shared same topic words like "rescate", "mineros" within a short period of time, which can be regarded as a message burst. Our work suggests that the detection of a message burst should consider the factor of accumulated historical data. In our work, the term "burst" is defined as a situation that a large number of frequently posted messages suddenly happened in a short time. Due to the dynamic characteristics of microblogging messages, several issues are described as follows:

- The design of weighting scheme of microblogging messages should differ from traditional methods in respect of their topical dynamic characteristics. Under such

[1] " Twitter's New Search Architecture"— `http://engineering.twitter.com/2010/10/twitters-new-search-architecture.html`

- a circumstance, concepts are often not stable but change with time, which is also known as *concept drift* [1].
- For the purpose of instantly analyzing messages, an incremental approach for term weighting [2] is required to avoid re-calculating entire messages.
- Maintaining cost in managing microblogging messages would be higher than other static messages, due to the nature of timely dynamics in the microblogging streams.

In our previous work [3], we found that a crucial problem of mining streaming data is how to extract significant features in a dynamic environment. In this paper, the related work of term weighting techniques and burst detection algorithm are discussed in Section2. Section 3 presents a novel term weighting scheme for mining streaming messages. The experimental results are shown in Section 4. Finally, we conclude this paper in Section 5.

## 2 Related Work

For considering the timing factors in evaluation term significance, the incremental approach is regarded as an effective way to prevent re-calculating entire corpus once new messages being coming into the system. Thus, we discuss two popular weighting schemes associated with our method, say, *incremental TFIDF* and *TFPDF (i.e., Term Frequency*$\times$ *Proportional Document Frequency)* methods for streaming messages.

*Incremental TFIDF* term weighting scheme [4] is an extension to traditional *TFIDF* (*i.e., Term Frequency*$\times$ *Inverse Document Frequency*) method. The main idea of this method is to extend the document frequency *df* (see equation 3) and normalize term frequency *tf* (see equation 4) at each time *t* to fulfill the function of online calculation. It is now widely used in dynamic text retrieval and text mining systems. However, it is worth mentioning that almost all terms occur in each message only once due to the length limitation of messages. Therefore, the computation overhead of term frequency *tf* is strongly affected by the length of messages. In addition, the document frequency *df* conflicts the operation of topic mining since a higher *df* value of the words implies the terms occur in many documents, which might lead to the problem of missing topic words in messages to some extent.

$$tfidf_t(w,d) = \frac{1}{Z_t(d)} f(d,w) * \log(\frac{N_t}{df_t(w)}) \tag{1}$$

$$Z_t(d) = \sum\nolimits_w f(d,w) * \log(\frac{N_t}{df(w)}) \tag{2}$$

$$df_t(w) = df_{t-1}(w) + df_{Ct}(w) \tag{3}$$

The same view is taken by [5], *TFPDF* algorithm tries to give higher weight to the term that was occurred frequently in many documents from the newswire sources for finding emerging topics. As shown in equation 4, $F_{jc}$ is the term frequency of *j-th* word in *c* channel (i.e., newswire source) and it will be normalized by the total words in the channel (see equation 5). *PDF* weighting represents an exponential distribution of the number of documents containing the term to the total number of documents in the channel. The *TFPDF* scheme benefits topic detection tasks particularly for finding significant words, but it is slightly inconvenient that extracting frequent features may enable oral words have heavier weights. These will be discussed in more detail later.

$$W_j = \sum_{c=1}^{c=D} F_{jc} \exp\left(\frac{n_{jc}}{N_c}\right) \tag{4}$$

$$F_{jc} = \frac{F_{jc}}{\sqrt{\sum_{k=1}^{k=K} F_{kc}^{\;2}}} \tag{5}$$

To process texts with a chronological order, a fundamental problem we concerned is how to find the significant features in text streams. In classic text retrieval systems, the most common method for feature extraction is to deal with each document as a bag-of-words representation. Such an approach is not suitable for our dynamic system. It is observed that in microblogging text streams, some words are "born" when they appear first time, and then their intensity "grow" in a period of time till reach a peak. These words are called *burst words*. As time goes by, the topics are no longer discussed by people, they "fade away" with power law and eventually become "death" (disappear), or change to a normal state. Such a phenomena is so-called life-cycle of the feature. Related idea of the state-based method was proposed by Kleinberg [6]. Their work considers the arrival rate of messages as the main factor to feature extraction, and the result yields a nested representation of the set of bursts that impose a hierarchical structure on the overall stream. Alternatively, a trend-based method [7] treats a word as falling or rising word depending on a measuring on absolute or relative change. In this paper, we do not compare two event time (current and the last record), but make a consideration of long term expectation value. The concepts would be described in detail in Section 3 and Section 4.

To the best of our knowledge, although some research work have been carried out on exploring the feature distribution on microblogging [8-10], little effort has been paid on utilizing sliding window techniques to developing a term weighting scheme for microblogging message streams. As a result, in this work we propose our novel term weighting scheme *BursT* to tackle the problem in the domain.

## 3 The Proposed Term Weighting Method for Microblogging Messages

For microblogs, the corpus tend to be dynamic as new items always being added and old items being changed or deleted. Therefore, an ideal term weighting scheme for mining microblogging messages should subtly reflect the changes over time and quickly assign proper weights in such a dynamic environment. In this section, we firstly describe the sliding window model, which is associated with the development of our weighting method, and then explain the weighting mechanism.

### 3.1 Fundamental Concepts of Sliding Window Model

The microblogging messages continuously posted by users around the world, and it is almost impossible to store all messages at one time due to the restrictions of memory limitation and constantly time lapsing. As a result, the *sliding window model* [11] is adopted for tackling the issues in this work, as shown in Figure 2. Briefly speaking, the steps of sliding window technique include: (i) the insertion operation in which a new index entry is built when a message comes in, (ii) the message is reserved until its lifetime exceeds the fixed length of time window *tw*, and (iii) the deletion operation in which the message will be removed from memory.

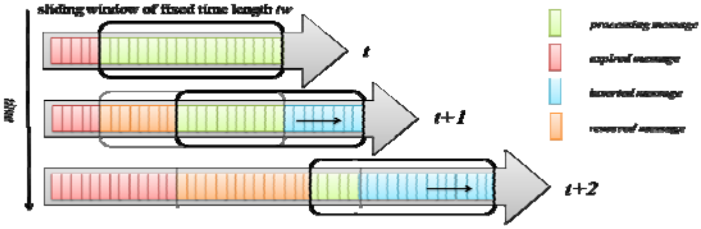


**Fig. 2.** A continuous message streams using sliding window model

### 3.2 The *BursT* Weighting Method

The primary goal of this work is to develop an incremental term weighting scheme for microblogging messages. The word *burst* is defined as an unusual number of frequently posted messages happened in a short time. Figure 3 shows a three-phase of word categories occurred in microblogging messages. In Figure 3, *uninformative word* means that a word rarely occurred in the sliding window, such as an oral word. If the word was occurring very frequently but with lower burstiness, they could be recognized as *common word* or *social word*, such as “haha” and “lol”. If a word has a higher burst than expectation within a certain range of document frequency, we will highlight its importance for weight design in the sliding window.

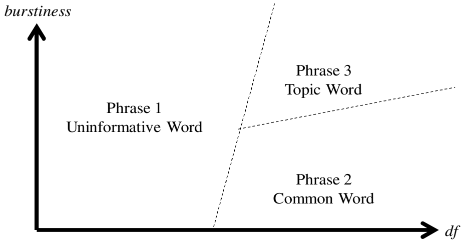


**Fig. 3.** Word categories in microblogging texts

Accordingly, our strategy in determining *BursT* value is that a heavier weight is achieved by a higher burstiness, in which some word occurs frequently in the window. Thus, the *BursT* weighting formula is shown in equation 6:

$$weight_{w,t} = BS_{w,t} * TOP_{w,t} \tag{6}$$

Where the weight of the word $w$ at time $t$ will be constituted by two factors: *BS (Burst Score)* and *TOR (Term Occurrence Probability)*. The detailed description of the weighting factors will be discussed in the following subsections.

### 3.2.1 The *BS* Weighting Factor

The interval of arrival time between messages can be transformed into arrival rate for many streaming data applications. If the feature of some message arrives with short intervals incessantly, the feature representing the importance of a message may be more useful. Suppose there is a feature word $w$ occurs in message sequence $\{m_{w,1}, m_{w,2}, m_{w,3}...m_{w,t}\}$, and each message has a specified arrival time $at_{w,t}$. We can then define the arrival rate $ar_{w,t}$ for current message $m_{w,t}$ by the formula shown in Equation 7:

$$ar_{w,t} = \frac{1}{at_{w,t} - at_{w,t-1} + 1} \tag{7}$$

In Equation 7, if $t=1$, the interval value becomes zero because $w$ is a brand new word in the system. The arrival rate $ar_{w,t}$ represents the reciprocal type of arrival gap($at_{w,t}$-$at_{w,t-1}$) which could be normalized between 0 to 1. In addition, there is an essential situation in microblogging messages according to our experimental statistics. The inter-arrival messages of the feature will accumulate a certain amount of arrival rate, particularly those words periodically appear daily, weekly and monthly. In order to reflect such a phenomenon and assign reasonable weights, an incremental mean method is adopted to compute the mean value of arrival rate of the word in this work.

$$\mu_n = \mu_{n-1} + (\frac{1}{n})(x_n - \mu_{n-1}) \tag{8}$$

$$E_i(ar_{iw,t}) = \mu_{iw,t} = \mu_{iw,t-1} + (\frac{1}{n_{iw,t}})(ar_{iw,t} - \mu_{iw,t-1}) \tag{9}$$

Equation 8 represents the calculation of an incremental mean in mathematics form. Finch [12] explained a numerically stable way to avoid accumulating too large sums. In this work we apply their approach in our weighting scheme to formulate equations of insertion (i.e. equation 9), where $ar_{iw,t}$ is the new arrival rate of the word. After computing the expectation values of arrival rate, the burst score is calculated as below:

$$BS_{w,t} = \max\{\frac{ar_{w,t} - E(ar_{w,t})}{E(ar_{w,t})}, 0\} \tag{10}$$

Therefore, we regard $ar_{w,t}$ as the current observation result, to compare with expected value $E(ar_{w,t})$of the word $w$ at *t-th* arrival. In addition, we derive a formula equation 10 in which residual is the deviation between observation and expectation values. It should be noted that the result of equation 10 would not always be positive if the observation result is less than expectation value. In such a case, we define the word as a "falling word" at that time, and enable *BS* factor to be zero.

### 3.2.2 The *TOP* Weighting Factor

The second consideration in *BursT* weighting scheme is *TOP (term occurrence probability)* factor, which is formulized by the proportion of the term in the sliding window. For the operation of mining hot news topics from messages, if a word occurs in more messages, it is more likely to be a trending topic. Thus, the term occurrence probability corresponding to the word $w$ at *t-th* arrival is formulated as below:

$$TOP_{w,t} = P(w_t \mid c_t) = \frac{|\{m : w_t \in c_t\}|}{|c_t|} \tag{11}$$

Where *TOP* represents the probability of the word occurrence in the sliding window, and $c_t$ denotes the message collection in the corpus collected from the time *t-tw* to current time. This factor would enable the weight of the word to grow with its occurrence frequency in messages, for identification of trending topics.

## 4 Experiments and Results

### 4.1 Data Source

The goal of this work is to develop a term weighting method for mining microblogging messages using sliding window techniques. In this section, we inspected *BursT* by setting the length of sliding window *tw* as an hour, and a total number of 27,929,301 microblogging posts were collected from Twitter (dating from September 19, 2010 to October 27, 2010). The test samples were collected through Twitter Stream API. After filtering out non-ASCII tweets, 14,009,908 available tweets had been utilized as our data source.

Subsequently, we partitioned messages into unigrams and removed the substring "RT @username:". In this work, the stopword list contains stopwords in seven different languages since the collected tweets contain multilingual texts. Finally, all capital letters in each tweet were converted into lowercase for our experiments.

### 4.2 Preliminary Experiment

One of the challenges encountered is due to the limitation of memory space for processing the streaming messages. Furthermore, an elimination process is also required for updating the weights in the limited memory space. Thus it is essential to start with reducing workloads by removing the "redundant features" in the index table. In order to look deep into these features, we have collected tweets for three weeks, as shown in Table 1.

**Table 1.** Number of the term occurrence and the cumerative term occurrences

| | 2010/08/11-2010/08/17 | | 2010/08/18-2010/08/24 | | 2010/08/25-2010/08/31 | |
|---|---|---|---|---|---|---|
| | # terms | percentage | # terms | percentage | # terms | percentage |
| TO = 1 | 553210 | 64.553281% | 546361 | 64.682219% | 561381 | 64.803643% |
| TO ≤ 2 | 653148 | 76.214903% | 644370 | 76.285242% | 661501 | 76.361107% |
| TO ≤ 3 | 696744 | 81.302058% | 686965 | 81.327951% | 705489 | 81.438911% |
| TO ≤ 4 | 722651 | 84.325108% | 712591 | 84.361744% | 731402 | 84.430207% |
| TO ≤ 5 | 739801 | 86.326317% | 729611 | 86.376697% | 748713 | 86.428522% |
| TO ≤ 6 | 752510 | 87.809312% | 742143 | 87.860327% | 761579 | 87.913723% |
| TO ≤ 7 | 762210 | 88.941191% | 751718 | 88.993885% | 771401 | 89.047537% |
| TO ≤ 8 | 769904 | 89.838993% | 759160 | 89.874924% | 779131 | 89.939858% |
| TO ≤ 9 | 776251 | 90.579615% | 765379 | 90.611175% | 785512 | 90.676456% |
| TO ≤ 10 | 781414 | 91.182079% | 770604 | 91.229748% | 790885 | 91.296694% |
| total | 856982 | 100.000000% | 844685 | 100.000000% | 866280 | 100.000000% |

In table 1, we found that the most of features occurred only few times within a week so we removed these terms if they appear less than three times as long as the length of sliding window. According to the experimental result, the elimination can remove more than 81% redundant features to enhance the performance.

### 4.3 Concept Drift with Burst Features and Oral Features

In order to examine the system performance in reflecting the concept drift of words, we selected "Chile's Rescued Miners" event as our case study. Our experiment started with the first miner was rescued at 2010/10/13 11:11 (GMT +8:00), until all miners were rescued at 2010/10/14 20:56 (GMT +8:00). Figure 4 indicates that the intensity of inter-arrival gap the feature word "chile", and it suddenly dropped at Oct 13 (GMT +8:00) when the event was happening.

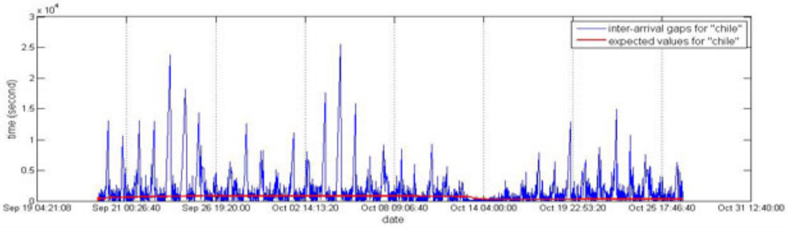
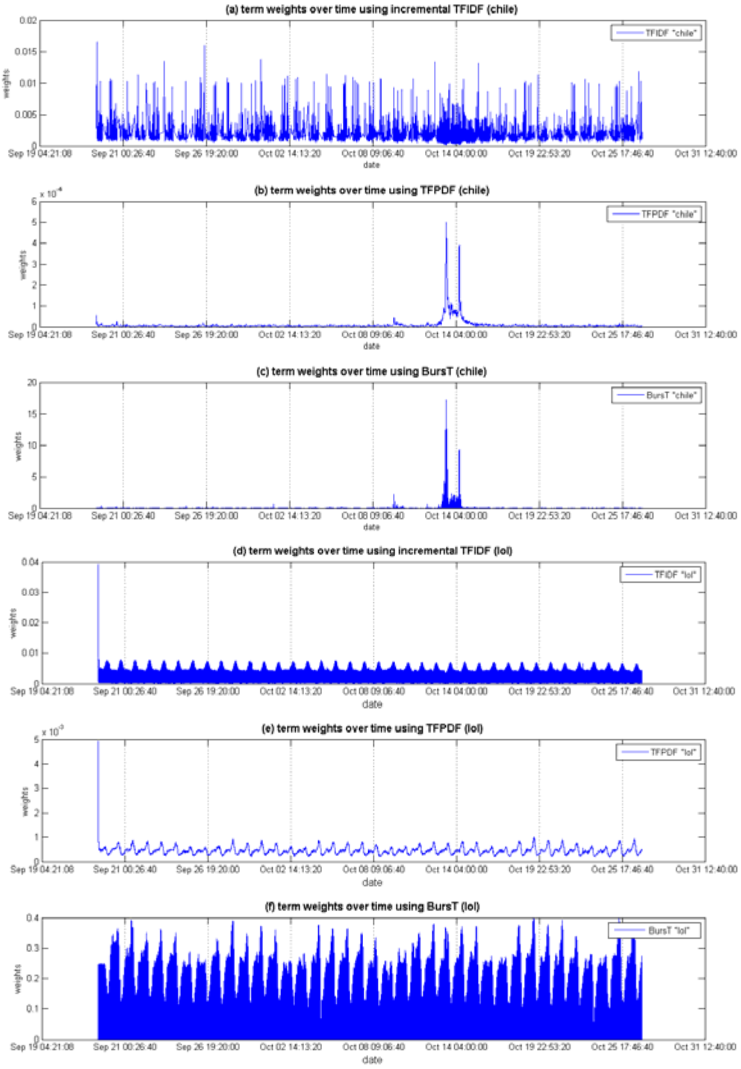


**Fig. 4.** Inter-arrival gaps using the feature word "chile"

Subsequently, we compared the weighting values of *BursT* and *TFIDF* methods, as shown in Figure 5. We found that the incremental *TFIDF* can't reflect the actual trends in sliding window algorithm, but *TFPDF* and our approach performed well in topic words. However, in the outcome of oral word analysis, we demonstrate the word "lol", which both has a high density of collection and arrival rate, as an example, and obviously it might not be suitable to define "lol" as a valid feature. It is worth


(a) term weights over time using incremental TFIDF (chile)

(b) term weights over time using TFPDF (chile)

(c) term weights over time using BursT (chile)

(d) term weights over time using incremental TFIDF (lol)

(e) term weights over time using TFPDF (lol)

(f) term weights over time using BursT (lol)

**Fig. 5.** Evaluation of *BursT,* incremental *TFIDF* and *TFPDF* weighting techniques

mentioning that some popular oral words might be easily over weighted in *TFPDF* because it places too much emphasis on document frequency. As shown in Figure 5(d) and (e), the weighted number of "lol" in *TFPDF* is still higher than in incremental *TFIDF* even the event is still on the fly.

## 5 Conclusion

In this paper, we describe a novel algorithm so called *BursT,* using the sliding window technique for weighting message streams. *BursT* facilitates online burst analysis by adopting a long-term expectation of arrival rate as a global baseline. The experimental results show that our weighting technique has an outstanding performance to reflect the shifts of concept drift, especially in dealing with the issue of dilimishing ineffective oral phrases. The result of this work can be extended to perform a periodic feature extraction, and is able to integrate other sophisticated clustering methods to enhance the efficiency for real-time event mining in social networks.

## References

1. Tsymbal, A.: The Problem of Concept Drift: Definitions and Related Work. Trinity College, Dublin (2004)
2. Yang, Y., Pierce, T., Carbonell, J.: A Study of Retrospective and On-line Event Detection. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 28–36. ACM, Melbourne (1998)
3. Lee, C.H., Chien, T.F., Yang, H.C.: DBHTE: A Novel Algorithm for Extracting Real-time Microblogging Topics. In: The Conference Proceedings at the ISCA 23rd International Conference on Computer Applications in Industry and Engineering (CAINE 2010), Las Vegas, USA (2010)
4. Brants, T., Chen, F., Farahat, A.: A System for New Event Detection. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 330–337. ACM, Toronto (2003)
5. Bun, K.K., Ishizuka, M.: Topic Extraction from News Archive Using TFPDF Algorithm. In: Proceedings of the 3rd International Conference on Web Information Systems Engineering, pp. 73–82. IEEE Computer Society, Los Alamitos (2002)
6. Kleinberg, J.: Bursty and Hierarchical Structure in Streams. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 91–101. ACM, Edmonton (2002)
7. Kleinberg, J.: Temporal Dynamics of On-Line Information Streams. In: Conference Temporal Dynamics of On-Line Information Streams (2005)
8. Cataldi, M., Caro, L.D., Schifanella, C.: Emerging Topic Detection on Twitter Based on Temporal and Social Terms Evaluation. In: Proceedings of the Tenth International Workshop on Multimedia Data Mining, pp. 1–10. ACM, Washington, D.C (2010)
9. Cheong, M., Lee, V.: Integrating Web-based Intelligence Retrieval and Decision-making from the Twitter Trends Knowledge Base. In: Proceeding of the 2nd ACM Workshop on Social Web Search and Mining, pp. 1–8. ACM, Hong Kong (2009)
10. Chen, J., Nairn, R., Nelson, L., Bernstein, M., Chi, E.: Short and Tweet: Experiments on Recommending Content from Information Streams. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems, pp. 1185–1194. ACM, Atlanta (2010)
11. Bifet, A.: Adaptive Stream Mining: Pattern Learning and Mining from Evolving Data Streams. IOS Press, Amsterdam (2010)
12. Finch, T.: Incremental Calculation of Weighted Mean and Variance. University of Cambridge (2009)