Bursty Event Detection from Text Streams for Disaster Management

Sungjun Lee
Dept. of Industrial Engineering
Seoul National University
1 Gwanak-ro, Gwanak-gu
Seoul, 151-744, Korea
zaregn81@snu.ac.kr

Sangjin Lee
Dept. of Industrial Engineering
Seoul National University
1 Gwanak-ro, Gwanak-gu
Seoul, 151-744, Korea
sjinlee@snu.ac.kr

Kwanho Kim
Dept. of Industrial Engineering
Seoul National University
1 Gwanak-ro, Gwanak-gu
Seoul, 151-744, Korea
goalwisk@snu.ac.kr

Jonghun Park
Dept. of Industrial Engineering
Seoul National University
1 Gwanak-ro, Gwanak-gu
Seoul, 151-744, Korea
jonghun@snu.ac.kr

ABSTRACT

In this paper, an approach to automatically identifying bursty events from multiple text streams is presented. We investigate the characteristics of bursty terms that appear in the documents generated from text streams, and incorporate those characteristics into a term weighting scheme that distinguishes bursty terms from other non-bursty terms. Experimental results based on the news corpus show that our approach outperforms the existing alternatives in extracting bursty terms from multiple text streams. The proposed research is expected to contribute to increasing the situational awareness of ongoing events particularly when a natural or economic disaster occurs.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information search and retrieval—*information filtering, retrieval models*

General Terms

Algorithms, Experimentation, Performance

1. INTRODUCTION

Due to the recent explosion of social network service users and the emergence of real-time content delivery methods such as RSS feeds and XMPP, the number of situation aware text documents in the form of text streams have been steeply increased on the web. The ability to automatically detect bursty events in real-time from text streams on the web are crucial in disaster situation since it facilitates immediate reaction against disaster through increasing the awareness of the current situation. For effective detection of bursty events, it is necessary to identify the characteristics of bursty terms that are relevant to the events. In this paper, we represent a document by using the bag of words model, and investigate the characteristics of bursty terms to develop an effective term weighting scheme for identifying the bursty terms. Subsequently, we rank the documents in a certain period with the obtained term weights for retrieving bursty event related documents. The rest of the paper is organized as follows. In Section 2, related work is summarized. Section 3 describes our proposed approach. Experiment results are presented in Section 4, and the paper is concluded in Section 5.

2. RELATED WORK

In this section we briefly present some of the research literature related to bursty event detection. Fung et al. [2] modeled the probability of a term in a particular time window using a hyper-geometric distribution and approximated it with a binomial distribution for computational efficiency. They then defined the probability of a term's burstiness in a time window based on the binomial distribution. Chen et al. [1] constructed an energy function for each term to accommodate decay of term importance over time horizon. The energy function was built based on the χ^2 statistic that represents the difference of term occurrences in a specific time slot from those of the other time slots. That is, the χ^2 statistic captures anomaly state of a term in a text stream.

In [4], the authors integrated two statistics into one index to extract surprising terms from data set across several different time windows. χ^2 and Gaussian statistics were used as surprise statistics. There are also several other efforts to model term's periodicity for event detection. In [3], the authors detected the periodicity of terms by using a periodogram estimator, and grouped terms into five categories according to the term's periodicity and importance. Based on the research results from the previous studies, this paper attempts to further enhance the methods for bursty event detection through developing additional measures that can facilitate identification of bursty terms and combining them into into a single measure.

3. PROPOSED APPROACH

We consider a situation where there are multiple channels that produce respective text streams, and attempt to

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

derive features that indicate whether or not a term is in a bursty state. We call a term in a bursty state as a bursty term. The derived features are then employed as weights of a term for the purpose of identifying bursty terms from multiple text streams. Let $tf_{c,t,d}$ denote the term frequency of term t in channel c at day d, and c_i represent the i-th text stream in a given data set such that $c_i \in C$ where $C = \{c_1, c_2, ..., c_K\}$. We further denote the j-th day in the data set as d_j where $d_j \in D$ and $D = \{d_1, d_2, ..., d_L\}$. In the following, we present four features that have been empirically identified to be effective in extracting bursty terms from multiple text streams.

Skewness: It is usual that a bursty term appears intensively only in a specific time period during which the corresponding bursty event occurs. This contrasts to the general non-bursty terms that appear at rather constant rates throughout text streams. Accordingly, in terms of the term frequency distribution over time, the shape of the distribution for a bursty term will have thicker tail for highly frequent terms than that for a non-bursty term which has a relatively balanced bell shape.

To model this characteristic, we employ a skewness measure for the term frequency distribution function over time. Its definition is given as follows:

$$skew(t) = \frac{E(X(t) - \mu(X(t)))^3}{\sigma(X(t))^3} \tag{1}$$

where $X(t) = < \sum_{c \in C} tf_{c,t,1}, \sum_{c \in C} tf_{c,t,2}, \ldots, \sum_{c \in C} tf_{c,t,L} >$ and $\mu(X(t))$ and $\sigma(X(t))$ are mean and standard deviation of $X(t)$ respectively. This feature gives a high value when the distribution function is skewed left or right.

Consistency: We now examine the distinct characteristic of a bursty term across the multiple channels. Since we are interested in the bursty event in the context of disaster management which is of concern to almost all publishers, the frequency of a term related to a bursty event is expected to soar in all the streams. Furthermore, the occurrences of a bursty term will be in accordance with those of its corresponding event, resulting in more time synchronized behavior between the term and the event across the channels than for non-bursty terms. To quantify this nature, we adopt the mean square error among the time periods with respect to the term frequency. Since each channel publishes documents at different rates, we normalize the term frequencies for each channel. The feature is defined as follows:

$$cons(t) = \sum_{c \in C} \sqrt{\sum_{j=1}^{L} (tf_{c,t,j} - \sum_{d \in D} tf_{c,t,d})^2} \tag{2}$$

We call this feature as consistency, and its value increases when the term frequency changes over time become high.

Periodicity: As indicated in [3], the term periodicity is an important factor for detecting a bursty term. That is, periodic terms are less likely to be bursty terms. In order to capture this characteristic, we introduce a simple function that penalizes the terms exhibiting the periodicity. Periodicity detection is done by analyzing a periodogram. The feature is defined as follows:

$$peri(t) = \begin{cases} p & \text{if term } t \text{ is periodic} \\ 1 & \text{otherwise} \end{cases} \tag{3}$$

where p has the range of $[0, 1]$. For instance, the terms such as 'Monday' and 'Tuesday' are penalized by this feature since they appear weekly.

Table 1: Events on Feburary 01, 2010.

Events	Bursty Terms
E_1 (Suicide bomber)	female, suicide, bomber, kill, Baghdad
E_2 (Maxico killer)	gunman, Juarez, kill, Mexico
E_3 (Arms trade)	Taiwan, U.S., arms, sale, China
E_4 (Taliban)	Taliban, leader, die, Pakistan
E_5 (Haiti rescue)	Haiti, flight, resume, U.S., Medevac

Variation: Different channels have different writing styles and thereby exhibit different term usage patterns. For instance, some channels attach their own signature or template when posting a document to the channel. Without considering this characteristic, some terms are mistakenly regarded as bursty terms just because of their frequent channel specific term usages. To address this problem, we introduce the coefficient of variation to each channel for normalization at the channel level. The definition is given as follows:

$$vari(c, t) = \alpha + \frac{\sigma_c(c, t)}{\mu_c(c, t)} \tag{4}$$

An additional parameter α is introduced for smoothing purpose. This feature has an effect of reducing the possibility of identifying a term with high frequency only in a specific channel as a bursty term.

The four features introduced above are based on different rationales and different scales. Accordingly, we combine the scores of different features by utilizing additional parameters to fine tune their effects on the final term weighting scheme, $burst$, as follows:

$$\begin{aligned} burst(t, d) = skew(t)^{k_1} \times cons(t)^{k_2} \times peri(t)^{k_3} \\ \times \sum_{c \in C} \{vari(c, t) \times nf(c, t, d)\} \end{aligned} \tag{5}$$

where $nf(c, t, d) = \sum_{d \in D} (tf_{c,t,d} / \sum_{t \in T} tf_{c,t,d})$ and k_1, k_2 and k_3 are tuning parameters for the features, which are to be optimized through experimentation.

4. EXPERIMENT RESULT

In this section, we study the performance of our approach in comparison with the other existing alternatives. We first describe the dataset and experiment setup, and then we consider the news corpus on a single day to evaluate the proposed approach with the others in terms of the bursty term detection performance.

For our experiments, we constructed a corpus composed of six international news channels by collecting their news data from October 1, 2009 to March 15, 2010 through using Google Reader API, and applied a Porter stemmer. The number of news documents in the corpus was 23,515. For the performance evaluation purpose, we also manually labeled bursty events for the corpus.

February 1, 2010 was selected as a test day for evaluating the performance of bursty disaster event detection. Table 1 shows disaster related events occurred on that day. These events were manually crafted and the related bursty terms were also manually chosen. Four existing methods presented in [1, 2, 4, 3] were implemented and compared with the proposed approach. It is expected that the best method will produce the terms that are most overlapped with the bursty terms in Table 1.

Table 2: Top 10 terms extracted from the considered methods on Feburary 01, 2010.

Proposed	Whitney	Fung	Chen	He
haiti	reuter	**mexico**	ap	presid
haitian	govern*	**arm**	fridai	**kill**
davo	**kill**	test*	brief	ap
sri	**u.s**	school	**kill**	new
missil	world*	moon	thursdai	year
african	offici*	hl	presid	govern
earthquak	report*	food	wednesdai	offici
lanka	presid*	yesterdai	new	countri
qaddafi	leader*	report*	mondai	sai*
hakimullah	ap	mehsud	tuesdai	peopl

Table 3: Event related document retrieval performances. Bold indicates the best performance. Proposed1 is a proposed model without periodicity and Proposed2 is with periodicity.

Top-N	Method	Recall	Precision	F-measure
Top-5	Chen	0.0583	0.2272	0.0928
	Whitney	0.0530	0.2065	0.0844
	Fung	0.0781	0.3043	0.1243
	Proposed1	0.0884	0.3446	0.1408
	Proposed2	**0.1110**	**0.4326**	**0.1767**
Top-10	Chen	0.1155	0.2671	0.1613
	Whitney	0.1088	0.2516	0.1519
	Fung	0.1529	0.3535	0.2135
	Proposed1	0.1649	0.3813	0.2302
	Proposed2	**0.2098**	**0.4852**	**0.2929**
Top-20	Chen	0.2321	0.2261	0.2291
	Whitney	0.2009	0.1957	0.1982
	Fung	0.2919	0.2842	0.2880
	Proposed1	0.2907	0.2832	0.2869
	Proposed2	**0.3686**	**0.3590**	**0.3637**
Top-30	Chen	0.3387	0.2199	0.2667
	Whitney	0.2829	0.1837	0.2228
	Fung	0.4149	0.2694	0.3267
	Proposed1	0.3934	0.2554	0.3098
	Proposed2	**0.4886**	**0.3172**	**0.3847**

The top 10 bursty terms identified by the considered methods are shown in Table 2. Bold terms in Table 2 represent overlaps, indicating that the proposed method as well as Whitney's and Fung's models achieved the same performance. For further comparison, we underlined the terms that are topically related to the selected events and Table 2 shows that our model outperformed the others in extracting the bursty terms related to the events. The starred terms represent general terms that should not be extracted as bursty terms. The results also tell that Whitney's algorithm tends to identify verbs as bursty terms which would be less informative than nouns to describe the events. Furthermore, the highly ranked terms from Chen's and He's methods included many periodic terms or the terms specific to individual channels. For instance, 'fridai' is a stemmed form of 'Friday' and 'ap' is a term related to a specific news channel. In contrast, the proposed approach was able to filter out those non-bursty terms.

We also compared the proposed approach with the other three models in terms of the performance of retrieving documents relevant to the bursty events. We scored each document by summing up the weights of the terms contained in the document and rank them by the resulting scores. The performances were measured based on precision, recall, and F measure which are popular in information retrieval literature. Experimentation results under the cases of Top-5, Top-10, Top-20, and Top-30 documents are shown in Table 3, which indicates that the proposed method outperformed the others and the Fung's model was the second best. In addition, to examine the effect of the periodicity feature, we also tested the proposed method without it. Table 3 shows that the periodicity contributes positively to the event related document retrieval performance.

5. CONCLUSION

In this paper, we studied a bursty term identification problem from multiple text streams for detecting disaster related events. The considered problem is important as it provides valuable information with which people can react properly in decision critical situations. Specifically, we empirically developed a term weighting scheme that assesses the term's burstiness from the perspectives of skewness, consistency, periodicity, and variation. The proposed scoring function was compared with the other existing alternatives, and the result showed that our approach outperformed the others. As future work, we consider application of the proposed method to social media data that possess different characteristics from those of the news media considered in this paper.

6. ACKNOWLEDGMENTS

This research was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea Government (MEST) (No. 2011-0004423 and 2011-0030814).

7. REFERENCES

[1] K. Chen, L. Luesukprasert, and S. c. Chou. Hot topic extraction based on timeline analysis and multidimensional sentence modeling. *IEEE Transactions on Knowledge and Data Engineering*, 19(8):1016–1025, Aug. 2007.

[2] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu. Parameter free bursty events detection in text streams. In *Proceedings of the 31st international conference on Very large data bases*, VLDB '05, pages 181–192. VLDB Endowment, 2005.

[3] Q. He, K. Chang, and E. Lim. Analyzing feature trajectories for event detection. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 207–214, New York, NY, USA, 2007. ACM.

[4] P. D. Whitney, D. W. Engel, and N. O. Cramer. Mining for surprise events within text streams. In *Proceedings of the SIAM International Conference on Data Mining (SDM 2009)*, pages 617–627, Apr. 2009.