

Identifying Relevant Event Content for Real-time Event Detection

Xinyue Wang, Laurissa Tokarchuk, and Stefan Poslad

School of Electronic Engineering and Computer Science

Queen Mary, University of London

London, UK

{xinyue.wang, laurissa.tokarchuk, stefan.poslad }@qmul.ac.uk

Abstract—A variety of event detection algorithms for microblog services have been proposed, but their accuracy relies on the microblog feeds they analyse. Existing research explores datasets that are collected using either a set of manually pre-defined terms or information from external sources. These methods fail to provide comprehensive and quality feeds for real-time event detection. In this paper, we present a novel adaptive keyword identification approach to retrieve a greater amount of event relevant content. This approach continuously monitors emerging hashtags and rates them by their similarity to specific pre-defined event hashtags using TF-IDF vectors. Top rated emerging hashtags are added as filter criteria in real time. By comparing our proposed approach, called CETRE (Content-based Event Tweet Retrieval) with an existing baseline approach applied to real-world events, we show that CETRE not only identifies event topics and contents, but also enables better event detection.

Keywords—Twitter, Hashtag, Contents Analysis, Query Expansion, Event Detection

I. INTRODUCTION

Online social networking services provide platforms for the public to share their observations and opinions about breaking news events through the Internet. Among all the emerging social network services, Twitter¹, a microblog service that supports multiple short updates for expressing opinions, can be used for event broadcasting [1]. Statistic shows that Twitter accumulates over hundreds of posts per second just because of the occurrence of a breaking news event [3].

In order to analyse real-world events through Twitter, event-relevant tweets need to be identified within a large, dynamic and noisy tweets stream in real-time (RT). Although the breaking news events arouse large volume of traffic, most of the other tweets are people's babblings. This great proportion of noise makes the use of the full amount of tweets traffic unrealistic. Most existing approaches exploit a fixed amount of manually pre-defined terms for the task [4]. The assumption here is that if content contains these selected terms, then it will contain information related to the event of interest. However, the set of predefined keywords is subjective and can easily lead to an incomplete dataset. Consequently, the problem we

address here is how to gather a more comprehensive set of event tweets in RT by solely relying on keyword searching.

We first deal with the sparse TF-IDF vector issue of tweets by formulating the document based upon hashtags (section III.A). We then update the keyword adaptation algorithm from our previous work [13] so as to allow the embedment of the hashtag-based TF-IDF vector (section III.B). We analyse the data feeds retrieved by our approach for the Sochi 2014 Winter Olympic Games (Sochi 2014) and the 2014 Malaysia Airlines missing Flight 370 (MH370 missing plane) event, and discuss our findings and future work in section IV.

II. RELATED WORK

Query expansion for extracting event related content from social websites has become a popular research topic, i.e., to use more than keywords as search criteria. Some of these rely on other websites' content for additional situational knowledge [5][6], while others try to make the most of the Tweets themselves [7][8]. However, their application is limited to pre-planned events. In addition, the extra cost in retrieving and processing additional data prohibits them from being used for RT services. The balance of retrieval accuracy and calculation efficiency is a key problem for the event detection and analysis of large-scale breaking news events, i.e. world-wide events that are of public concern and discussed widely online, especially from real-time social network data feeds.

Some researchers report success with the use of event detection on large Twitter datasets [8][10]. However these types of techniques rely on the collection of large datasets that represent the state of the Twitterverse at a particular point in time, which is not scalable enough to events that may last up to several months. Other research has focused on the detection of the event itself [9][2]. Namely, they delve more into the details about tracking event trends, but don't consider the temporal features. While tweet clustering runs in real-time, it requires the use of human annotated Tweets during a training stage [11], which is not feasible for live events. To summarise, existing methods are limited because they require the need for manual interactive operation, prohibiting their support for RT detection of both changing planned and unplanned events and are not oriented to large-scale events.

¹ Twitter: <https://twitter.com>

III. REAL-TIME EVENT-RELATED CONTENT RETRIEVAL

We are interested in detecting social feeds that relate to large-scale breaking news events. They usually consist of multiple topics that are semantically and temporally related. In order to retrieve tweets for RT event detection, our target is to capture event tweets within the Twitter stream.

A. Hashtag-based TF-IDF Vector

The conventional use of TF-IDF performs well for structured, long paragraph articles but is not that accurate for short and noisy sentences such as tweets [16]. In order to maximize the usability and accuracy of TF-IDF for tweets, we formulate the document based on the Twitter topic indicator (the hashtag # notation) rather than on single tweets [12].

A key problem for generating our improved TF-IDF vector is to define how the hashtag document is constructed. For each hashtag, its hashtag document consists of a group of multiple tweets which contain that hashtag. Those tweets are retrieved through the Twitter Search API where the hashtag is used as the query keyword. Among the tweets sent back by Twitter, only the tweets published in a certain period are used to formulate the document of that hashtag.

Tokenizing the document is the second step. Each document is analysed by an original Twitter Analyser implemented with Lucene². We adopt some basic tweet processing that is used by most existing research. Apart from stemming and stop word elimination, this analyser also removes Twitter's notations, such as @mention, #hashtag, URLs and redundant repeat characters, e.g. words such as yeeeeaaaaah will be converted into yeah. The remaining words are tokenized according to the bag of words model using Mahout³. We tokenize the document into a unigram (a single term in a sentence) as this low complexity model achieves a good accuracy.

The output tokens are used for the construction of hashtag-based TF-IDF vectors. For each hashtag document, its TF-IDF vector, i.e. the hashtag-based TF-IDF vector, is the union of all the TF-IDF values of its tokens. Unlike a conventional tweet level TF-IDF vector that suffers from the sparse vector issue, this hashtag-based TF-IDF vector combines multiple tweets to generate denser vectors. This can provide more reliable results when performing the similarity comparison between hashtag documents. Also, this approach significantly reduces the amount of vectors that need to be calculated, thus improving the computation efficiency.

B. Identifying Event-Relevant Content

For the event-relevant tweets identification, we track the stream by monitoring the volume of each emerging hashtag. The tweets collected via the Twitter Streaming API are temporally segmented into slices, defined as *time frames*. At the end of each time frame, the high frequency hashtags (more than 5 times in a time frame) were selected as potential hashtags, represented as $H(t_n)$. Then, the algorithm retrieves the Twitter Search API for tweets to construct hashtag-based

documents. Comparing the similarity distance (*dist*) of TF-IDF vectors between potential hashtags and the pre-defined initial ones, the algorithm marks the top rated ones as the event topics, and uses them as new keywords. With the additional event content introduced by the extra keywords, the algorithm is supplied with more information, and thus can capture more events topics. The following pseudocode gives a more detail explanation of the algorithm:

Algorithm CETRe Keyword Adaptation

```

1  $H_{potential} = H(t_n)$ 
2  $H_{TF}(t_n) = H_{BL}$ 
3 for  $\forall h_i \in H_{potential}$ 
4   for  $\forall h_j \in H_{TF}(t_n)$ 
5     if  $0 < dist(TFIDF(h_i), TFIDF(h_j)) < Thres$  and  $h_j \in H_{BL}$ 
6        $H_{TF}(t_n) = \{h | h \in H_{TF}(t_n) \text{ or } h = h_i\}$ 
7        $H_{TF}(t_n)$  is marked as Topic words
8     end if;
9   end for;
10 end for;

```

The threshold *Thres* is empirically set to 0.8 based on our observation of several historical events [13][14]. This value gave a good performance for almost all the test events. We employ the cosine distance measurement to quantify the differences between each pair of candidate hashtags. By only considering recent tweets, their temporal character is also considered when calculating the hashtag similarity

IV. EVALUATION WITH PLANNED AND UNPLANNED EVENTS

We evaluate CETRe against a baseline method. In the baseline method, event tweets are identified by a constant keyword list where the keywords are manually defined and remain unchanged for the entire collection period. In this section, we will first describe the datasets we collected and then illustrate how CETRe provides better feeds for event detection.

A. Dataset Overview

We run the CETRe algorithm under planned and unplanned event scenarios as defined by Becker [17]. CETRe is expected to perform well for both types of events.

The planned event is the 2014 Sochi Winter Olympic Games, while the unplanned event is about the 2014 Malaysia Airlines missing Flight 370. The Sochi 2014 is a real-world event which happens during 2014-02-10, 16:00, GMT to 2014-02-24, 16:00, GMT. Compared with the baseline approach which identifies 6,146,399 tweets, the content-based approach provided another 3 million tweets, resulting in a dataset with 9,140,180 tweets. It also distinguishes 2495 extra keywords, apart from the baseline keywords *sochi* and *#olympic2014*. The overall situation is similar for the unplanned MH370 missing plane event. The content-based approach identified extra keywords and tweets compared to the baseline approach. We selected a sub-period, from 2014-03-09, 22:30, GMT to 2014-03-15, 09:00, GMT to evaluate the algorithm. By using the initial keyword *MH370* and *Malaysia Airlines*, the baseline crawler collected 3,019,001 event tweets. Within the same period, the content-based

² Apache Lucene: <https://lucene.apache.org/>

³ Apache Mahout: <https://mahout.apache.org/>

TABLE I. ACCURACY OF KEYWORD IDENTIFICATION (SOCHI 2014)

Sample period \ Accuracy	Precision	Recall	F-Measure
02-12 06:11	90.00%	77.14% [▲]	83.08%
02-18 02:23	86.36% [▼]	45.24%	59.38%
02-19 16:13	88.89%	23.88% [▼]	37.65%
02-23 03:24	97.50% [▲]	69.64%	81.25%
Average	91.30%	55.63%	67.36%

[▲]([▼])indicate the highest (lowest) value among all the sample periods

crawler identified 574 more keywords and resulted in a 3,742,616 tweet dataset.

B. Accuracy of Keyword Identification

The quality of keywords identification is essential for event content retrieval. We examine the accuracy of this procedure based upon the recall-precision rate.

For the whole collection period, the CETRe algorithm adapted keywords 1384 and 647 times. In order to examine the accuracy, we randomly select 50 periods. For each period, the identified keywords are manually categorized into event-related (*TP*) keywords and event-nonrelated keywords (*TN*). For the recall rate calculation, we retrieved the tweets in the previous collection period, and manually tagged each potential keyword as an event-related potential keyword (*TP+FN*) or event-nonrelated keyword (*TN+FP*). Our categorisation strategy classifies a keyword as event-related one only when it explicitly refers to the event of interest. Hashtags such as *canada* and *lost* won't be classified as event-related keywords since they also introduce new noisy tweets for the events. Two labellers participated in the manually labelling task. A third person was introduced only when the two disagree with each other.

For example, the CETRe algorithm identified 30 new keywords at 6:11 12th Feb, comprising 27 event related keywords (e.g. *#erinhamlin* and *#speedskating*) and 3 event-nonrelated keywords (e.g. *#everywhere* and *#russia*). Therefore, the precision rate for this period is $27/30 = 90\%$. In the previous time frame, i.e. 05:59 to 6:10, 103 high frequency hashtags were identified, among which 35, are related to the event of interest. As a result, the recall rate is $27/35 = 77.14\%$.

Table I illustrates the accuracy measurements of parts of the random selected sample periods for the Sochi 2014. The last row gives the average values of all the 50 periods. The precision of our algorithm is high - about 90% for identifying event topics accurately. Although the lowest recall rate 24% happened when rate limited⁴, the overall recall is acceptable, at more than 55%. This indicates that more than half of the potential words are picked for the event-tweets collections. Considering that one general event word (e.g. *figureskating*) can be used to retrieve tweets with more specific topics (e.g. *freedance*), the recall rate of event tweets retrieval should be higher than this. For the other event, the overall F-measure remains at the same level, but the precision rate falls, as shown in the Table II. This decrease results from: 1) some keywords

TABLE II. ACCURACY OF KEYWORD IDENTIFICATION (MH370 MISSING PLANE)

Sample period \ Accuracy	Precision	Recall	F-Measure
03-10 16:20	79.17%	82.61% [▲]	80.85%
03-12 04:08	64.29%	42.86% [▼]	51.43%
03-14 09:40	90.00% [▲]	45.00%	60.00%
03-15 05:52	69.57% [▼]	57.14	62.75%
Average	78.35%	60.69%	67.27%

in Malaysian are ignored; 2) this event is less likely to generate event specific keywords since most of the tweets mentioned *MH370* explicitly. Reason 1) is also responsible for a higher recall rate for this unplanned event.

C. Real-time Event Detection

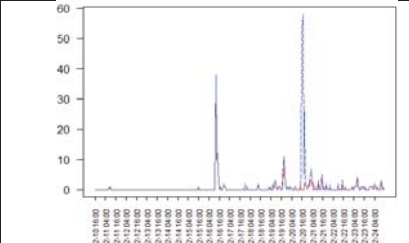
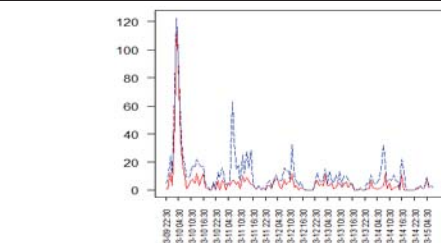
In fact, the extra amount of keywords not only enriches the datasets with extra information, but also helps with RT event detection. To examine whether or not the extra keywords introduce sufficient additional material for finer event detection, we employ a parameter free and computational efficient method that detects the bursty behaviour by tracking the abnormal changes of a specific term [18]. The authors model the probability of the number of tweets that contain a specific term (N_i) in a time window as a binomial distribution. When monitoring the stream in real-time, an event is detected whenever the real count N_i is much larger than the mean value of the distribution. We adopt the idea to detect the key moments for keywords identified by the content-based algorithm. Table III illustrates the results of burst detection for example keyword *maoasada* and *pray4mh370*. The number of bursts that can be detected from the content-based datasets is larger than those from the baseline dataset. Additionally, the CETRe algorithm also helps to keep track of some important moments of the event. For example, the final of the figure skating was on 20th Feb, which is detected from the content-based dataset, but is not detected using the baseline dataset. Similarly, a tweet burst at 6:30 11th Mar related to the announcement about the identity of fake passport carriers. This was missed when detecting the corresponding burst on the baseline dataset.

D. Discussion

Overall, the evaluation results show that the CETRe algorithm identifies a larger amount of event keywords with a high precision rate (more than 80% on average) and with a good recall rate. The benefit of these extra event keywords is the extra amount of event content that can be used for finer event detection. We also demonstrate that we are able to detect more bursts with the datasets generated by the CETRe algorithm than the one generated by the baseline approach. Some of the bursts prove to be important moments during the event. Namely, the CETRe algorithm provides better temporal characteristics for the RT burst detection. When monitoring an event, these extra bursts may represent important topic changes that convey significant event information. Consequently, our CETRe approach based on Twitter hashtag

⁴ Twitter Streaming API provides no more than 1% of the total traffic

TABLE I. BURST DETECTION FOR KEYWORDS OF CONTENT-BASED ADAPTIVE CRAWLER

Keywords	<i>maoasada</i>			<i>pray4mh370</i>		
Event						
Hashtag count						
Bursts detection	Real events	Baseline	Content	Real events	Baseline	Content
	Intensive report on Asada's final show	02-16 12:00 02-16 13:00 02-16 14:00	02-16 12:00 02-16 13:00 02-16 14:00 02-17 22:00	Families of missing passangers arrived at Malaysia airport	03-10 23:30	03-10 23:30
	Ladies Short Program	02-19 08:00 02-19 09:00	02-19 06:00 02-19 09:00	11 th annocement from Malaysia Airlines		03-11 01:30
	Ladies Free Skating Final	-	02-20 16:00 02-20 17:00	Confirmation of the identity for the fake passport holder		03-11 06:30
	Reports on her Final&Mistakes	02-23 08:00 02-24 12:00	02-23 08:00 02-24 12:00	Final conversation from the MH370 was revealed		03-12 11:30
				-	03-12 23:30	03-12 23:30
				Reports indicateds MH370 fled 4 more hours after disapear	03-13 03:30	03-13 03:30
				An American officer said MH370 try to ping the satellite after disapear		03-14 01:30
				Reports reveal that MH370 fled to Andaman island	03-14 08:30	03-14 06:30 03-14 07:30
				A press by Malaysia government	03-14 16:30	03-14 16:30 03-14 17:30

use, tries to minimise the calculation cost, whilst maintaining an acceptable level of accuracy. Compared to previous work [13], our approach demonstrates significant improvements for new keywords identification and more relevant event content retrieval.

Future work focuses on two aspects: the refinement of the RT keywords adaptation algorithm and the characterisation of different events. Currently, the time when a hashtag is identified as a keyword, lags behind the moment that topic becomes a trend. If the term can be picked and discarded at exactly the time when it is trending, the algorithm becomes more robust for RT event topic detection. Characterising different types of events is also of interest. We observed that the algorithm tends to pick event-specific keywords for the planned Sochi 2014, while catching more general keywords for the unplanned MH370 missing plane event.

REFERENCES

[1] H. Kwak, C. Lee, H. Park, and S. Moon, What is Twitter, a social network or a news media?. In WWW '10, pp. 591-600

[2] S. Petrović, M. Osborne, and V.r Lavrenko. Streaming first story detection with application to Twitter. In ACL-HLT' 11, pp. 181-189.

[3] amianto.com, Social Media: How To Get The Most Out Of Twitter To Make Your Event A Success, Report, September 2010

[4] P.S.Earle,D.C.Bowden,M.Guy. Twitter earthquake detection: earthquake monitoring in asocial world. Annals Of Geophysics, 2012, 54(6).

[5] E. Benson, A. Haghighi, and R. Barzilay. Event discovery in social media feeds. In ACL-HLT '11, pp. 389-398.

[6] F. Abel, I. Celik, G. Houben, and P. Siehdnel. Leveraging the semantics of tweets for adaptive faceted search on twitter. In ISWC'11, Part I. pp.1-17..

[7] F. Liang, R. Qiang, and J. Yang. Exploiting real-time information retrieval in the microblogosphere. In JCDL '12, pp. 267-276.

[8] H. Becker, F. Chen, D. Iter, M. Naaman and L. Gravano: Automatic Identification and Presentation of Twitter Content for Planned Events. ICWSM '11

[9] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In WWW '10, pp. 851-860.

[10] J. Lanagan, A. F. Smeaton. Using Twitter to Detect and Tag Important Events in Sports Media. In ICWSM '11

[11] J. Nichols, J. Mahmud, and C. Drews. Summarizing sporting events using twitter. In IUI '12, pp. 189-198.

[12] O. Tsur, A. Rappoport. What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In WSDM '12, pp. 643-652.

[13] X. Wang, L. Tokarchuk, F. Cuadrado, and S. Poslad. Exploiting hashtags for adaptive microblog crawling. In ASONAM '13, pp. 311-315

[14] X. Wang, L. Tokarchuk, F. Cuadrado, and S. Poslad. Adaptive Identification of Hashtags for Real-time Event Data Collection. To be appeared in LNSN.

[15] C.C. Yang and X. Shi. Discovering event evolution graphs from newswires. In WWW '06.

[16] F. Perez-Tellez, D. Pinto, J. Cardiff, and P. Rosso. On the difficulty of clustering company tweets. In SMUC '10, pp.95-102..

[17] H. Becker. Identification and Characterization of Events in Social Media. PhD thesis, Columbia University, 2011.

[18] J. Yin, A. Lampert, M. Cameron, B. Robinson and R. Power. Using Social Media to Enhance Emergency Situation Awareness, *Intelligent Systems*, IEEE 2012 Vol. 27(6) pp. 52-59