Adaptive Identification of Hashtags for Real-time Event Data Collection

Xinyue Wang, Laurissa Tokarchuk, Felix Cuadrado, and Stefan Poslad

Abstract the widespread use of Microblogging services, such as Twitter, makes them a valuable tool to correlate people's personal opinions about popular public events. Researchers have capitalized on such tools to detect and monitor real world events based upon this public, social, perspective. Most Twitter event analysis approaches rely on events tweets collected through a set of pre-defined keywords. In this paper, we show that the existing data collection approaches risks losing a significant amount of event-relevant information. We propose a refined adaptive crawling model, to detect emerging popular topics, using hashtags, and monitor them to retrieve greater amounts of highly associated data for the events of interest. The proposed adaptive crawling model expands the queries periodically by analysing the traffic pattern of hashtags collected from a live Twitter stream. We evaluated this adaptive crawling model with a real-world event. Based on the theoretical analysis, we tuned the parameters and ran three crawlers, including one baseline and two adaptive crawlers, during 2013 Glastonbury music festival. Our analysis shows that adaptive crawling based upon a Refined Keyword Adaptation algorithm outperforms the others. It collects the most comprehensive set of keywords, and with the minimal introduction of noise.

1 Introduction

The enormous popularity of Microblogs, combined with their conversational characteristic [1] (leading to multiple short updates and used as a medium to express opinions) has led them to become one of the most popular platforms for researchers to extract public information. Early attempts were conducted to identify character-

School of Electronic Engineering and Computer Science
Queen Mary University of London, London, UK.
E-mail: {xinyue.wang,laurissa.tokarchuk,felix.cuadrado,stefan.poslad}@qmul.ac.uk

Goal! Aaron Ramsey. Penalty. GB 1-1 Korea. **#football #olympic**

And just like that **#FIFA** awards **#GBR** a penalty. **#GBRvKOR**

Fig. 1: Tweets about the 2012 Olympic Games

istics of information diffusion and users' behavior on the entire Microblogsphere [7, 12, 15]. Nowadays, the research focus has shifted to more specific problems, such as real world event detection [11] and event summarization [5].

As one of the most popular Microblogging services, Twitter[1] provides people with a platform to share their observations and opinions online. This simple version of a blog service allows users to post short messages (tweets) up to 140 characters. Users can not only update their thoughts through the website, but also post tweets using their mobile devices through either a cellular network or Short Message Service (SMS). This easy access to Twitter facilitated the dramatic growth of the number of Twitter users. With thousands of posts published every second[2], Twitter also becomes a precious resources pool for researchers to analyse public reaction and behavior under event scenario.

For instance, recent research has examined the use of such tools, primarily Twitter-based, to get knowledge about ongoing affairs [4, 6, 9], or even to dig out hints of upcoming events [2, 8]. Becker et al. use Twitter, along with other social media sites, to retrieve content associated with a planned event [4]. Sakaki et al. use Twitter to detect the occurrence and location of earthquakes even before the disaster hits [2].

In order to identify and analyze events among the entire Twittersphere (also called Twitterverse), a comprehensive dataset describing the event is compulsory. The majority of collection techniques collect tweets from the live Twitter stream by matching a few search keywords or hashtags. For example, Starbird and Palen collected information about the 2011 Egyptian uprising by using the keywords"*egypt, #egypt, #jan25*" [3], Nichols et al. collected sport related tweets using keywords "*worldcup*" and "*wc2010*" [20]. However, the set of predefined keywords is subjective and can easily lead to incomplete data. Moreover, even given expert knowledge, keywords and specialised hashtags often arise in the midst of such events. For example, Fig. 1 shows two tweets relating to the London 2012 Olympics (the football event). It is straightforward to determine that the first one is related to the 2012 Olympics football event, whereas the second one, which refers to the same event, is much harder to distinguish. Fig. 2 illustrates how this will result in the loss of

[1] Twitter Home page, https://twitter.com/

[2] New Tweets per second record, and how: https://blog.twitter.com/2013/new-tweets-per-second-record-and-how

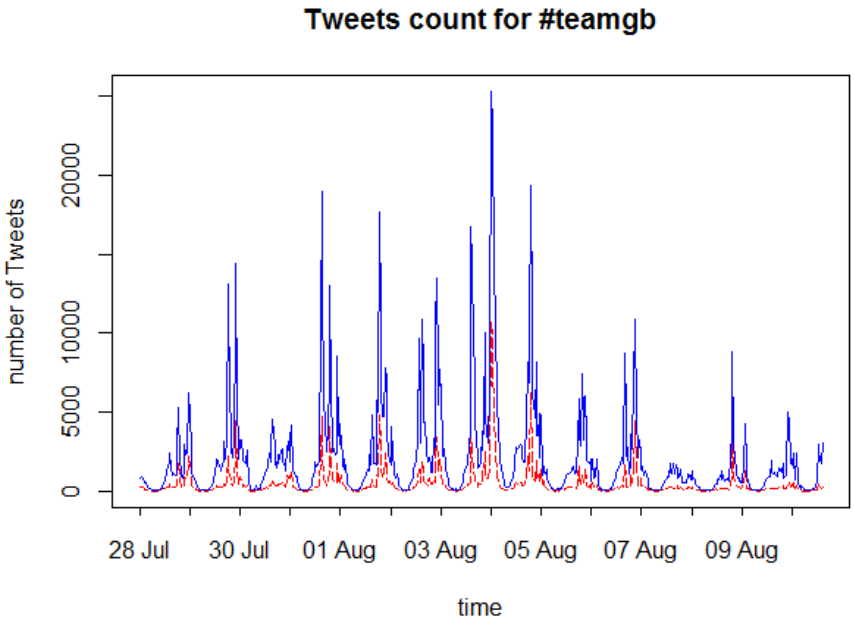


Fig. 2: Comparison of tweet volume crawled by keyword *Olympic* (lower, red dashed line) versus *Olympic* & *#teamgb* (higher, blue solid line) during the 2012 London Olympics

a significant amount of event related information. The blue solid line is the traffic generated by using both *Olympic* or *#teamgb* as keywords, while the red dashed line represents the volume of tweets solely retrieved using a keyword *Olympic*. It is clear that the trend for both lines is the same, but the volume varies. A larger amount of event information can be fetched if other keywords are introduced. This issue is even more severe when using Microblogs for situation awareness during emergencies or disasters. People will communicate their observation and perception about events, even without explicitly mentioning the title of the event [20].

Moreover, Twitter's APIs data access restrictions[3][4][5], greatly complicate collecting all the social media documents corresponding to one event. Lanagan et al. mentioned that incomplete tweet datasets significantly affects the performance of their

[3] Search API only returns tweets within 7 days, and the rate limit of Search API is not specified in the official documentation.Version 1.0

[4] Streaming API provides real-time services but only returns 1% of total number of tweets.Version 1.0

[5] At time of publication, access to the full Firehose stream of tweets is allowed only if a large amount of money is paid, e.g. PowerTrack costs $2,000 per month plus $0.10 per 1,000 tweets delivered. Retrieved from: http://gnip.com/pr_announcing_power_track

event detection algorithm [19]. In fact, the Twitter API restrictions not only introduce difficulties on live tweets retrieval, but they also make it harder to recapture data once the events of interest are finished.

In this paper, we aim to present an automatic event content collection method that gathers a set of tweets, without preliminary knowledge of the events, by just relying on initial search terms for live events. We introduce an adaptive Microblogging crawling model that allows comprehensive information about an event to be retrieved. By embedding the Keyword Adaptation (KwA) algorithm, this adaptive crawling model can collect an extended set of specific instances of an event. This is achieved by monitoring the Twitter live stream with only the initial keywords, without manual modification of the search terms. In designing the adaptive crawling model, the challenge is to identify extra search terms, beyond the original keywords, appearing in content related to the event in question. Specifically, compared with the previous work [31], which only evaluates the KwA algorithm theoretically with existing datasets, the novel contributions of the paper are as follows:

- We investigate the use of trend detection method in our proposed adaptive crawling model and prove that it is insufficient to identify event relevant topics.
- We examine the proposed adaptive crawling model for real-time events by retrieving multiple datasets with an exemplar type of real-time event.
- We demonstrate that the adaptive crawling based upon a Refined Keyword Adaptation algorithm identifies event topics in real time. Also, its collects additional relevant tweets, while greatly reducing the amount of irrelevant information.

The remainder of this paper is divided into four sections. Section 2 introduces the related work and distinguishes our work from exiting work; section 3 introduces the functions and restrictions of Twitter service; section 4 details the proposed adaptive crawling models; Section 5 reports the evaluation of our technique, showing its performance over the 2013 Glastonbury Festival; and finally section 6 concludes our work and discusses some future directions.

2 Related Work

Online content collection and analysis has been a popular research issue for years. Much work exists pertaining to structured articles collection from online platforms [17, 23]. Later, researchers tried to improve the traditional content collection and analysis approaches by taking advantage of additional information [32]. Some researches detected latent features (e.g. topics) to obtain a better understanding of the event in question [26]. However, the differences between traditional websites (i.e. news portals and blogs) and Microblogs with respect to resource deployment and contents structure make the transplant of Web-based to Microblog methods difficult. This section will review the existing work relating to online crawling, topic detection and similarity measurement of text, specifically under the Twitterverse.

Crawling a set of online documents, relating to an event of interest, can be achieved by simple keywords searching. This approach has been adopted by some early attempts on tweet collection and analysis [2, 3], but crawling on a pre-defined keywords set didn't provide satisfactory results. In addition to these kind of aforementioned approaches, attempts to use more than keywords as search criteria have also been made [4, 9, 13]. For example, Becker et al. examine the use of precision and recall-oriented strategies to automatically identify event features, then generating queries to retrieve content from diverse social media sites for planned events. Unlike our proposed model, which solely relies on initial terms, they use event announcements from sites such as Last.fm[6] to aid query formulation [4]. Rather than use other websites, Fabian et al. leverage several metrics from Twitter, such as users' profiles, semantics meanings and metadata of tweets, to generate new search criteria from news websites [9]. Although this additional material facilitates a higher precision and recall rate on search results, the processing cost of these exponentially increases. While planned events can draw on extra material from announcements and news sites, such material for unplanned events is almost impossible to obtain. Furthermore, these solutions were designed to improve the user experience of interactive searching rather than collect additional event-related tweets for real world ongoing affairs. In addition, event tweets can also be fetched from a particular group of target users[33, 34]. This kind of approach chooses the users that are involved in or related to the event as the initial seed for collection. For example, 11 car-related companies are selected as seeds when collecting tweets for 2012 Super Bowl [33]. It is similar to the pre-defined keyword crawler as the initial seeds are fixed. Although our crawling target is different, it is possible to apply our idea in their scenario to support seed adaptation.

Recently, Twitter has attracted unprecedented attention with the research efforts on the detection of trending topics under different circumstances. Some researchers report some success with the detection of event topics and content in large Twitter datasets [2, 8, 19]. However, these types of techniques analyse tweets and track the inherent topic on a large datasets which only represents the state of the Twitterverse at a particular point in time. Namely, these researchers concentrated on building an accurate model in an offline fashion. On the other hand, some researchers explored the traffic characterization of text streams [24, 25] for real time identification of the emerging topics. This kind of approach tracks the evolution of topics by identifying frequent terms in a specific time interval. Rather than identifying general trending topics for multiple events, our objective is to identify an extended set of topic terms for a single event. In addition, the proposed model utilizes the relations between topic terms rather than measuring them separately. The other kind of online topic detection approach builds a model for each topic by capitalizing on the statistical relations between vocabularies [26, 27]. Their conclusion is based on the observation that some particular words appear in the documents belonging to the same topic more frequently, while others less so. However, the main drawback is that they rely on the prior training to construct an accurate topic model. Explicitly, they require

[6] Last.fm website: http://www.last.fm/

the use of human annotated tweets during training stage, i.e. background knowledge about the event need to be known in advance, which is not feasible enough for real-time topic detection. Moreover, statistic based approaches for short text modeling in microblogging environments remain an open research issue since the effectiveness of a trained topic model can be highly affected by the length of the documents [28].

In order to identify as many event-related documents as possible, a measurement to evaluate their relevance to the events of interest is necessary. The majority of existing research relies on the traditional TF-IDF text vector and distance measurements to assess the similarity [5, 10, 29]. Though TF-IDF, is widely used in Natural Language Processing as a measurement of words' importance and offers great performance for long paragraph text-mining, its accuracy for shorter tweets-alike document is still unsure [22]. In fact, Microblog posts are naturally unstructured with many colloquial expressions and often do not comply with the normal syntax used in the Web. Only spare TF-IDF vectors can be formulated from tweets, which this is not a qualified input for traditional distance measurements. Recently, an attempt to associate tweet-level features with other metadata was conducted [30], however it still measured the event in a static way without considering the temporal evolution of the topics. In this paper, we propose to use a similarity measurement for a time series to overcome the above problem.

3 Social Microblogging Service: Twitter

Topic Indicator Conversational Hashtags: Twitter has sometimes been described as "the SMS of the Internet[7]" due to its conversational characteristic. This is supported by its well-known @ mention, RT retweet and # hashtag annotation. In this work, the hashtag annotation is of special interest as it allows users to indicate what the message is about when they publish a tweet [14]. By adding a # mark before the topic words, users can generate their own topic indicator at any moment. Twitter's user interface automatically associates a hyperlink for each hashtag to allow people to retrieve all tweets with the same hashtag in just a click. As the adaptive crawling framework is designed to collect data on a specific topic, this characteristic is adopted and explored.

Twitter API and Rate Limits: Twitter provides three public APIs to the developers and researchers for designing and implementation their desired tools: the Search API, the Streaming API and the Representational State Transfer (REST) API[8]. Of these, the Streaming APIs is used in our proposed model. This is because the Streaming API is the only interface that offers real-time access to the public tweets timeline. This API sends back 1% of the whole tweets volume in its core

[7] The SMS of the Internet: http://www.wisitech.com/blog/the-sms-of-the-internet

[8] Twitter API Documentations: https://dev.twitter.com/

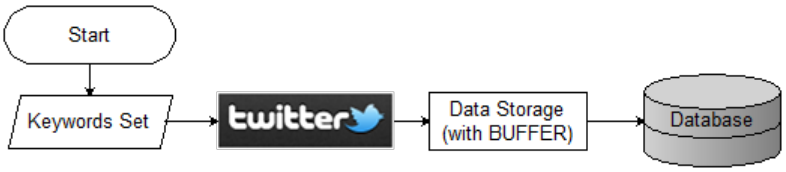


Fig. 3: System Flow of Simple Twitter Crawling Model (Baseline)

database by using sample() function for each normal OAuth[9] enabled user. This 1% limitation also applies to the filter method of the Streaming API. It is possible to use the method to generate a query to extract all tweets with specific criteria, e.g. keywords. However, the full amount of content is available only when the retrieved volume is less than 1% of the total traffic of Twitter. Otherwise, that 1% will be spread out across keywords, that is, only a subset of tweets will be retrieved for each individual keyword.

In the proposed adaptive crawling model, the filter method in the Streaming API is used to collect event relevant tweets. Twitter allows a maximum of 400 keywords for a single query and thus our search query was similarly limited.

4 Twitter Crawling Model Design

A Twitter crawler is a program that collects tweets or users' information through Twitter API matching a set of search criteria. In this section, a novel adaptive crawling model will be introduced. This adaptive crawling model is based upon the simple keyword crawler but embedded with a keyword adaptation algorithm running in real time.

4.1 Twitter Crawling Model

In this work, we are interested in keyword-based crawling, where every matching tweet will contain at least one of the defined search keywords. Compared with the simple (baseline) Twitter crawling model, the adaptive Twitter crawling model enables the adaptive crawling algorithm to leverage keyword adaptation in real-time.

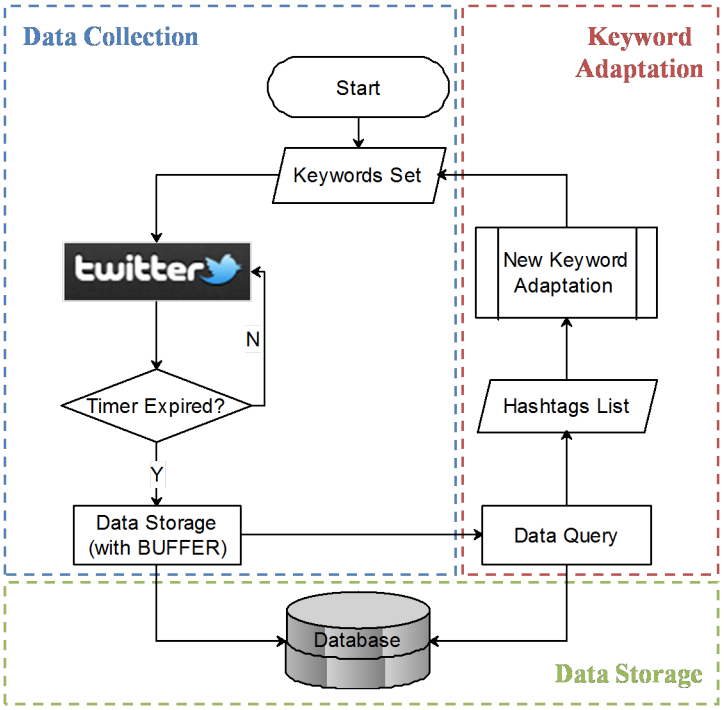


Fig. 4: System Flow of the Adaptive Twitter Crawling Model

4.1.1 Baseline Crawling

The baseline crawling model defines and uses a constant keywords set. In this model, a keywords set is used for focused crawling of a specific event. The keywords are manually defined according to the event of interest and remain unchanged for the entire collection period. The system flow of this crawling model is illustrated in Fig.3 After sending the keywords with a query to the Twitter Streaming API, the qualified tweets will be returned as a stream. These tweets are stored in a database system. We use the dataset collected by this model as a ground truth in the evaluation section as this crawling model is used by most of the existing research.

4.1.2 Adaptive Crawling

The system structure of the adaptive crawling model is similar to the baseline crawling model for the Data Collection and Data Storage Components. The difference is the additional *Keyword Adaptation* component, as illustrated by Fig. 4. This component enables the application of the Simple Keyword Adaptation algorithm

[9] OAuth: http://oauth.net/

and Refined Keyword Adaptation algorithm described in the next section when crawling data in real-time events.

In this model, the data collection process is triggered by the same set of predefined keywords as the baseline's. The keyword adaptation feature enables the identification of popular event-related hashtags by using the Keyword Adaptation Algorithm (KwA). At the end of every time frame, the KwA is run over the previous time frame to generate a new keyword set. Finally, a query that encodes all the words in the keywords set is sent to the Twitter API and the time frame timer restarted.

We exploit the traffic characteristics of hashtags gathered via Twitter Streaming API to realize keyword adaptation. Research shows that hashtags, a kind of user-defined index term that start with #, have been used as topical markers to link relevant topics and events when people express their interests [14]. Exploiting hashtags for keyword searching not only reduces the complexity in getting the semantic meaning from tweets but also increases the efficiency of data analysis.

4.2 Keyword Adaptation (KwA) Algorithm

The goal of a keyword adaptation algorithm is to automatically find the list of hashtags, beyond the initial set of keywords, appearing in tweets related to the event of interest. By using "automatic", we mean that keywords should be classified without manual intervention. Therefore, the essential problem is to figure out what kind of hashtags help with extra event-related information retrieval.

In our first attempt we apply the idea of trend detection in the Simple Keyword Adaptation algorithm. We assume the hashtags that appear frequently in tweets with initial keywords are related to the event. However, when evaluating Simple Keyword adaptation in the adaptive crawler we found that the Simple Keyword Adaptation algorithm introduces a lot of noise. Furthermore, due to the rate limit restriction from Twitter, the volume of the event-related tweets retrieved by this approach was far less than the volume collected by simply using the baseline crawler. This approach collected a large amounts of noise.

In order to balance the efficiency and performance of crawling content under Twitter API restrictions, we designed the Refined Keyword Adaptation algorithm. In this section, full details about both versions of Keyword Adaptation algorithm are presented.

4.2.1 Simple Keyword Adaptation Algorithm (SKwA)

In this SKwA, the collection of hashtags within that fixed time frame is represented as $H_{tf}(t_n) = \{h_1, h_2, ...\}$. The keywords set, sent to Twitter API in Fig. 3, at any time frame n, can be represented as $H(t_n) = \{h_1, h_2, ...\}$, where $h_k(k = 1, 2,)$ is an individual hashtag. Here, we use *keywords* to indicate hashtags that were eventu-

ally sent to Twitter for data collection. At the same time, the model also keeps two hashtags frequency lists, one for the whole collection period and the other one for the current time frame. At the moment when any time frame n is passed, the hashtags frequency list for whole collection period is represented as $freq(t_n)$, while hashtags frequency list for the n_th time frame is written as $freq_{tf}(t_n)$. The frequency list for the whole collection period updates every time frame, while the other list updates within the time frame when a new tweet arrives. The hashtag list and the frequency list have a one-to-one correspondence, i.e. the frequency count of a hashtag h_k at n_{th} time frame is $freq_{tf}^{h_k}(t_n)$. The Frequency List Update algorithm is defined in Algorithm 1.

Algorithm 1 Frequency List Update

Require: H_{tf}, $freq_{tf}^{h_k}$
1: **for** $\forall h$ in the incoming tweets **do**
2: **if** $\exists h_k = h : h_k \in H_{tf}(t_n)$ **then**
3: $freq_{tf}^{h_k}(t_n) = freq_{tf}^{h_k}(t_n) + 1$;
4: **else**
5: $H_{tf}(t_n) = H_{tf}(t_n) + 1$;
6: $freq_{tf}^{h_k}(t_n) = 1$
7: **end if**
8: **end for**

Apart from these two frequency lists, a minimum frequency ($freq_{min}$), as a threshold for being a keyword, and an array of blacklist hashtags (H_{black}) are also used in the simple adaptive crawler to help with adaptation. The pseudocode in Algorithm 2 details this version of the Keyword Adaptation algorithm.

Algorithm 2 Simple Keyword Adaptation (SKwA)

Require: H_{tf}, $freq_{tf}^{h_k}$
1: **for** $\forall h \in H_{tf}(t_n)$ **do**
2: **if** $h \in H_{blacklist}$ **or** $freq_{tf}^{h_k}(t_n) < freq_{min}$ **then**
3: $H_{tf}(t_n) = \{h_k | h \in H_{blacklist}, h_k \neq h\}$;
4: $freq_{tf}(t_n) = \{freq_{tf}^{h_k}(t_n) | freq_{tf}^{h_k}(t_n) \in freq_{tf}(t_n), h_k \neq h\}$
5: **else**
6: $H(t_n) = H(t_{n-1}) \cup \{h_k | freq_{tf}^{h_k}(t_n) \in Topn(freq_{tf}^{h_k}(t_n))\}$;
7: $freq(t_n) = freq(t_{n-1}) \cup \{freq_{tf}^{h_k}(t_n) \in Topn(freq_{tf}^{h_k}(t_n))\}, where n = N - num[H(t_{n-1})]$
8: **end if**
9: **end for**

This algorithm keeps at most $N = 400$ keywords for querying Twitter every 10 minutes, where N is the maximum number of hashtags in keywords set. When a new hashtag appears, the algorithm will check whether or not it already exists in the keywords set $H(t_n)$. If it is a query keyword, its whole period frequency list is

incremented by 1. Otherwise, the hashtag is stored in the time frame hashtags list temporarily. When the timer expires, hashtags in the time frame hashtags list are sorted according to their frequency. Top ones will be added to the keywords set. In other words, hashtags with a low frequency within time frame n don't become a keyword.

This SKwA employs three noise reduction steps to avoid overwhelming the new keyword set with non-related keywords. First, the threshold for being a keyword, $freq_{min}$, helps to filter out the unusual hashtags. While those hashtags can be relevant to the event of interest, they are not worthy of collection because they only generate a tiny amount of traffic. In addition, the introduction of these low frequency hashtags will significantly increase the calculation cost, both in space and time. As a result, we set the $freq_{min}$ with an empirical value to be once per minute. Second, by discarding the long term, low frequency, items, the crawler can improve the utility of N keywords. This mechanism functions as follows: for hashtag h has low values for a long period ($freq_{tf}(t_n)$,$freq_{tf}(t_{n-1})$, , $freq_{tf}(t_{n-m})$), it will be removed from the keywords set. Last, the introduction of a keyword blacklist allows noisy keyword to be manually filtered. The blacklist is empty when the crawler is started. Users can identify and add non-related words to the blacklist during the collection period. The algorithm will check this list every time when it identifies new search terms so it can discard the words that are in the blacklist. For the experiments in this paper, the blacklist words belong to either general association with news channels (e.g. BBC and CNN) or as hashtags used by follow up and follow back activities (e.g. teamfollow and followback).

4.2.2 Refined Keyword Adaptation Algorithm (RKwA)

Our initial attempts show that extra traffic can be produced when using the proposed SKwA when running with the adaptive crawler. However, we found the dataset collected through SKwA also contains a large amount of non-related tweets: the longer the crawler runs, the larger the proportion of noisy tweets. The noise, i.e. non-related tweets, eventually overwhelm the event-related data, which results in a chaotic dataset. This issue is caused by the fact that the algorithm relies on the collected content: a clean dataset will help the crawler to better adapt; a noisy dataset always becomes even noisier.

In order to reduce the impact of noisy information on the adaptive dataset, the traffic pattern of hashtags is exploited to classify those potential keywords according to their relevance to the events. The problem is how to modify the SKwA so the adaptive crawler collects a greater amount of highly event associated data without significantly increasing the dataset noise.

The refined version first automatically gets a hashtags list based on the SKwA. The list is then passed to an extended part of the keyword adaptation algorithm for assessing the elements' relevance to the event. Here, we introduce the *correlation coefficient* to evaluate the relevance. In order to calculate the correlation between two hashtags, we subdivide the time frame into several time slots. The frequency

counts of each time slot is represented by $freq_{tf}^{h_k}(t_n)$. This array indicates the frequency counts of a hashtag h_k in all the time slots within the n_{th} time frame. The collection of initial keywords is represented as $H_{seed} = \{h_1, h_2, ...\}$. Instead of using $H(t_n)$, the keywords set will be sent to Twitter API at the end of each time frame and is written in the form $H_{fin}(t_n)$, $H(t_n)$. It is a temporal list which holds the same result as that used by SKwA. The pseudocode is updated as the Algorithm 3.

Algorithm 3 Refined Keyword Adaptation (RKwA)

Require: $H_{seed} = H(t_n) \cup H_{fin}(t_{n-1}), H_{fin}(t_n) = H_{BL}$
1: **Execute Algorithm 2 SKwA**
2: **for** $\forall h_x \in H_{seed}$ **do**
3: **for** $\forall h_y \in H_{fin}(t_n)$ **do**
4: **if** $h_y \in H_{BL}$ **and** $cor(freq_{tf}^{h_x}(t_n), freq_{tf}^{h_y}(t_n)) > Thres_1$ **then**
5: $H_{fin}(t_n) = \{h | h \in H_{fin}(t_n)$ **or** $h = h_x\}$;
6: **else if** $h_y \neq H_{BL}$ **and** $cor(freq_{tf}^{h_x}(t_n), freq_{tf}^{h_y}(t_n)) > Thres_2$ **then**
7: $H_{fin}(t_n) = \{h | h \in H_{fin}(t_n)$ **or** $h = h_x\}$;
8: **end if**
9: **end for**
10: **end for**

The initial keys H_{seed} and correlation measurements *cor* are defined based on the following hypotheses:

Hypothesis 1 *(H1): the initial keywords used for both baseline crawler and adaptive crawler are the most representative words that describe the event of interest.*

Hypothesis 2 *(H2): trending keywords for an event during one particular or several sequential time frames are likely to exhibit similar traffic patterns.*

Hypothesis 2.1 *(H2.1): the frequency of occurrence of two trending keywords shows a linear relationship. Namely, when keyword A appears more, the frequency of keyword B will also increase, and vice versa.*

Consequently, the initial keywords used by the baseline crawler and adaptive crawler with SKwA are also selected as initial keys in RKwA. A popular linear correlation measurement, i.e. Pearson correlation, which is defined by the following equation, is chosen as the measurement of similarity between related keywords.

$$cor = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{1}$$

Here, sequence X represents the $freq_{tf}^{h_x}(t_n)$ and Y for the $freq_{tf}^{h_y}(t_n)$ in the algorithm. That is to say, hashtag $h \in H_{tf}(t_n)$, as calculated by SKwA, is only retained in RKwA if it has a high correlation with one of the seed keywords. For example, #100aday is a trending hashtag but irrelevant to the event. It was detected as a keyword by SKwA, but it was successfully excluded by RKwA because of its low correlation to the initial hashtag.

The threshold values for $Thres_1$ and $Thres_2$ also need to be set for executing RKwA. We use a single variable approach to choose their values: one of the thresholds was fixed while the other one changing gradually. We found that changing of $Thres_1$ didn't bring too much impact on the result, but the differences introduced by changing of $Thres_2$ is notable there is always a range of threshold values that can make the signal to noise ratio higher than others. Therefore, the final value we choose is $Thres_1 = 0.5$ and $Thres_2 = 0.8$.

5 Evaluation of Adaptive Crawling Model

The purpose of the evaluation is to test if the proposed adaptive crawling model helps to collect additional data without introducing too much noise.

In our previous work [31], the experiments were conducted on a historical dataset of 2012 London Olympic Games and with only a theoretical analysis. Whereas, here we apply both SKwA and RKwA to our adaptive crawling model and test them with a large public event in real-time. Our aim is to demonstrate that the information gain is at the same level as we showed, and the signal to noise ratio (i.e. ratio between the event related information and event irrelevant information) is much more significant than our previous estimation.

In more detail, an overview of the collected datasets, including one for baseline crawler and one for adaptive crawler with SKwA and RKwA respectively, is described first. Then, we will analyse and evaluate the proposed adaptive crawling model by classifying the retrieval keywords and tweets. Accordingly, the relevance of keywords and that of tweets from all the three datasets will be assessed with a quantitative method.

5.1 Dataset Overview

The datasets were collected during the 2013 Glastonbury festival [10] period. Three crawlers were run for the tweets collection, first the baseline crawler, and then the two instances of the adaptive crawler, with the SKwA and RKwA respectively. Only "Glastonbury" is used as the initial keyword for all the three crawlers.

Table 1 and Fig. 5 illustrate the tweet volume collected for "Glastonbury" from 2013-6-28, 19:00:00, BST to 2013-7-1, 07:00:00, BST. The collection period lasted 60 hours, with more than half a million tweets collected from the baseline crawler alone. The number of tweets collected by SKwA is almost twenty times the number collected by the baseline crawler. In Table 1, the column "unique" is the number of tweets that appear only in that dataset. Providing that all the crawlers start with the same initial keyword "Glastonbury", SKwA and RKwA dataset should contain

[10] What is Glastonbury: http://www.glastonburyfestivals.co.uk/information/what-is-glastonbury

Table 1: Tweet volume generated by different crawling approaches

	Baseline	SKwA	RKwA
Tweet Count	550,417	10,433,355	2,472,953
Unique Tweet	10,275 (1.8%)	9,534,735 (91.4%)	1,252,577 (50.6%)

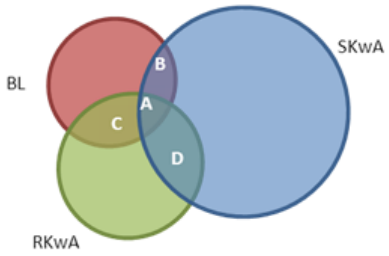


Fig. 5: Number of common tweets in baseline, SKwA and RKwA datasets: A = 214218, B = 2084, C = 323840, D = 682318.

all the tweets in the baseline dataset, i.e. the cell indicated the unique number of tweets in the baseline dataset should be zero. However, some of the tweets, even if they contain the initial keywords, cannot be retrieved by the SKwA due to the 1% rate limitation. When the number of keywords increases, the volume of tweets containing those keywords also increases and are more likely to exceed the rate limit. Fig. 6 shows the traffic volume, every 5 minutes for each of the three datasets, and provides a graphical view for examining the period when any of the crawlers hit the Twitter rate limits. According to the figure, it is obvious that the number of tweets approached 2900/min in the SKwA dataset. Based on an empirical test, this value is close to the upper tweet volume limit when accessing the Streaming API free of charge. Due to the reason that SKwA crawler is always rate limited, tweets showing up in the baseline dataset can be lost in the SKwA dataset, and therefore results in the unique tweets in baseline dataset. Compared with the SKwA dataset, the RKwA dataset contains many more tweets than the baseline dataset, i.e. almost all the tweets in baseline dataset also showed in RKwA dataset. It also included some tweets from SKwA and 50% unique tweets. Though the RKwA crawling also hit the rate limit at some points according to the Fig. 6, it still achieved an acceptable performance most of the time.

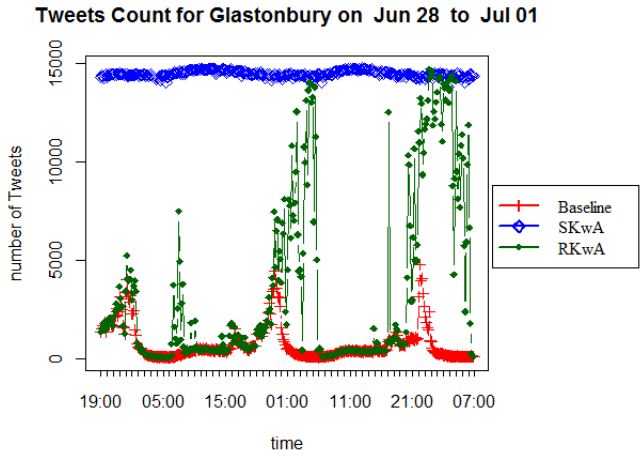


Fig. 6: Tweet Volume for Glastonbury Festival (10 mins interval).

5.2 Evaluating the Keyword Adaptation Algorithm

The previous section qualitatively illustrates the overall statistics of the three datasets. The next step is to quantitatively analyse to show that the adaptive crawling model helps to extract extra relevant event information. This section details the evaluation procedures for revealing whether or not the extra tweets are all related to the event in question.

5.2.1 Evaluation Setup

The aim of this experiment is to verify that RKwA performs better than SKwA in retrieving a greater amount of event-related information while retaining the noise (non-related tweets) to signal (event-related tweets) ratio at an low level. The following hypothesis act as a condition for evaluating the performance of SKwA and RKwA:

Hypothesis 3 *(H3): a tweet is likely to only be about one topic which is described by a hashtags, and therefore its correlation to an event of interest is determined by its hashtags.*

H3 determines whether or not the tweet's hashtags affect the tweet's relevance to an event of interest. Based on this hypothesis, we design the procedures for evaluating the performance of the adaptive crawling model as follows:

Table 2: The Hashtag Category and Grading Strategy

Hashtag Category	Specification	score
Related (C1)	hashtags contain the terms Glastonbury, band names or song names that appear during the festival	+2
Possibly-related (C2)	hashtags stand for media which broadcast the event, as well as emotional hashtags and those emerged with ongoing affairs (nextyear, best-seats, etc)	+1
Non-related(C3)	hashtags showing no particular relationship with the event	-2
Not known (C4)	non-English hashtags that the manual taggers didn't identify	0
Non-keyword hashtags	Hashtags that have not been selected as key-words	-1

5.2.2 Labeling Keywords Manually

In order to filter out noisy tweets, the first step is to distinguish between the related and non-related keywords by manually labeling: hashtags shown in the keywords set are manually classified into corresponding categories. Three independent participants are involved in this labeling process. The final result is based on the average produced by two independent participants. A third labeller was introduced in the case of a disagreement.

Hashtags in different time periods were labelled according to how closely they are related to Glastonbury Festival. For example, "#glasto2013" is definitely related, while "#6hobbs" is more complicated to classify. It could be related since it represents a program for BBC Radio Music which always broadcasts information about music. However, it may also include information other than Glastonbury Festival. In our grading strategy, it was classified as possibly-related. All the hashtags were labelled into five categories based on the criteria in Table 2.

5.2.3 Classifying tweets According to the Manual Labeling

In this step we classify whether or not a tweet is related to the event based on the hashtags it contains using the grading system in Table 2. Each hashtag is assigned a score and the final grade of a tweet is the sum of all the hashtags' scores.

By using this strategy, tweets with a grade more than 0 are classified as related tweets, and those less than or equal to 0, as non-related tweets. The grading system can identify non-related tweets even if it carries related hashtags. For example, "*Friday night! Meet new people - FREE! onclique #meetingpeople #Bristol #instagram #Glastonbury #Manchester #onclique*" are classified as non-related tweet. The final

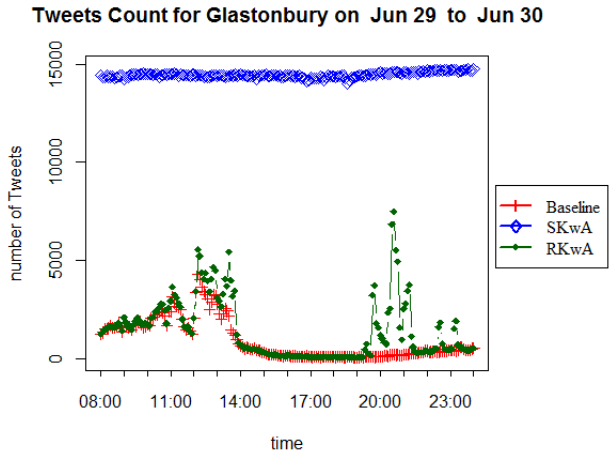


Fig. 7: Tweet Volume (Evaluation Period) for Glastonbury Festival

grade is -2 because the score introduced by #Glastonbury is cancelled out by other negative instances.

Therefore, the baseline, SKwA and RKwA datasets were all classified into two sub datasets, related and non-related tweets datasets. Finally, we compare the proportion of related and non-related tweets in all datasets to check the levels of noise introduced and the proportion of event-related information retained.

5.3 Experiment Results

A subset of the Glastonbury data was selected for the evaluation. The test set is during 8:00 29_{th} June to 00:00 30_{th} June as this is the period where both the SKwA and RKwA worked properly with normal behavior (e.g. not suddenly polluted by noise).

The tweets volume of all the crawlers during the selected period is shown in Fig. 7. According to this figure, the first fluctuation appeared at 16:00, while the highest traffic period started at night from about 20:00, and reached the peak at about 23:00. This is because the famous music performers started to show up in the afternoon and the performances finished in midnight. It is clear that the adaptive crawler with the SKwA was always rated limited, while the other adaptive crawler identified extra tweets during the peak period, when compared with the baseline crawler.

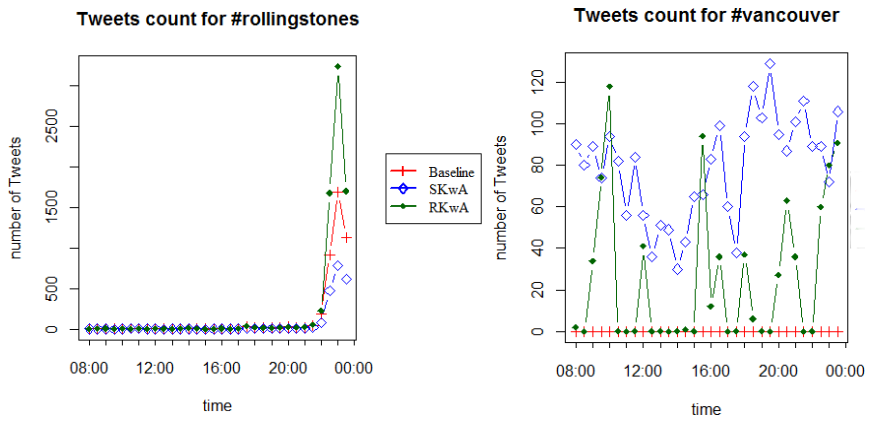
Table 3: Hashtag Count of Manually Labeling Categories

Keywords count Category	BL	SKwA	RKwA	RKwA to SKwA ratio
Related (C1)	1	15	66	440.00%
Possibly-related (C2)	0	16	22	137.50%
Non-related(C3)	0	1360	96	7.06%
Not known (C4)	0	500	30	6.00%
Total	*1*	*1891*	*214*	*11.32%*

5.3.1 Relevance of Identified Topical keywords

In Table 3, each row is the number of keywords in the corresponding category. The first column describes the keyword composition for the baseline crawler (BL). The value 1 shows in the first category and indicates it only maintains single keywords during the whole crawling period. Namely, the baseline crawler doesn't adapt the keyword set. According to the figures here, the SKwA did provide an extra 30 $(15+16-1)$ event keywords. But clearly, its retrieval keyword set is very noisy as C3 keywords dominates most of the SKwA keywords set. The statistics in the third column shows that the RKwA performs much better than the SKwA. The last column makes this clear: it is the RKwA to SKwA (RS) ratio between the number of $C_x(x = 1,2,3,4)$ keywords from RKwA crawler to that from the SKwA crawler. It is clear that the RKwA reduced the proportion of C3 keywords in SKwA dataset by more than a thousand, i.e. the noisy keywords are dropped to only 7.06%. Meanwhile, the RKwA identified more C1 and C2 keywords compared with the SKwA. The RS ratio for C1 and C2 keywords reached 440% and 137.5% respectively. This indicates that by using the proposed RKwA, the event-related terms are more likely to be identified, while the introduction of noisy keywords is controlled. This provides preliminary evidence that the RKwA performs much better than the SKwA.

In addition, the extra event-related keywords can pull extra event content, especially for the RKwA dataset, as shown in Fig.8.a. The volume of the band name keyword "*rollingstones*" has its peak at the same time for all the three crawlers, though the volume varies. The difference in information gain between the baseline and the RKwA crawling illustrates that the adaptive crawling has the potential to fetch additional event-related information. More specifically, the RKwA dataset contains more tweets with *#rollingstones* than either the SKwA dataset or the baseline dataset. Surprisingly, the SKwA dataset maintains the lowest volume of tweets containing "*rollingstones*". Considering that the SKwA adaptive crawler was rate limited all the time and collected tweets with more keywords, this phenomenon is caused by the spread of the space of other non-related traffics. Apart from this, the apparent differences for the volume of tweets containing "*#vancouver*" in the SKwA dataset and the RKwA dataset in Fig.8.b proved that the RKwA is also able to reduce the impact of irrelevant keywords. Random spikes of "*#vancouver*" for RKwA dataset shown in Fig.8.b are introduced by tweets that carry both the event related



(a) Trending event keyword *#rollingstones*

(b) Event irrelevant keyword *#vancouver*

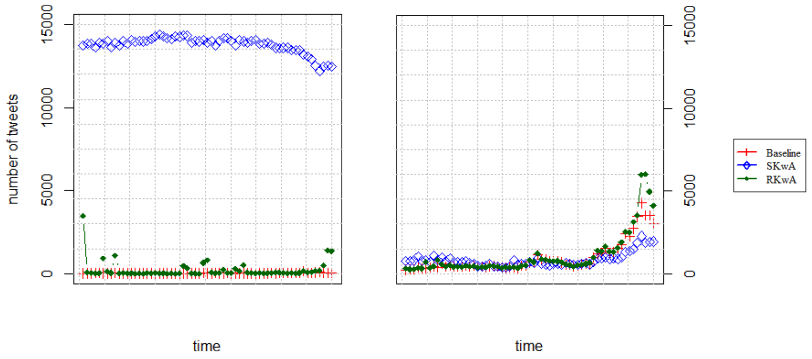
Fig. 8: Comparison of event-related vs. event-irrelevant keywords in the three datasets

keywords and the "*vancouver*". However, the event related keyword in such tweets is not "*glastonbury*", so it can't be retrieved by the baseline crawler. For example, when the RKwA adaptive crawler retrieved tweets for the keyword "*bbcglasto*", tweet like "Trending *#rollingstones #bbcglasto #vancouver #chic* #followme" was also collected.

5.3.2 Relevance of Collected Tweets

The original intention of proposing this adaptive crawler is to fetch extra event-related tweets. We examine this over all datasets. Specifically, tweets from the SKwA and RKwA datasets were classified according to the final grade. In the baseline tweets classification task, the final grade is calculated by referring to the manually labelled results of the other two adaptive datasets. If the baseline tweets contain any of the labelled keywords, the same value will be added to or deducted from the final grade.

Fig. 9 shows the traffic volume of irrelevant event tweets and the event-related tweets in these three different datasets. Fig. 9 a) illustrates that SKwA introduces a great amount of noise - most of the traffic (about 94%) from SKwA dataset is irrelevant to the event. The green line for RKwA clearly illustrates that RKwA crawler performs well in reducing the amount of noise. Compared with the SKwA dataset, the irrelevant event tweets in the RKwA dataset are relatively few. According to Fig. 9 b), the SKwA only introduced extra event content at the beginning of the evaluation period. At the time when the event content is increasing, i.e. after 20:00, the SKwA loses a large proportion of event content even compared with the baseline



(a) Irrelevant event tweets (b) Event-related tweets

Fig. 9: Proportion of Tweets and Their Relevances to the Event in the Three Datasets

one. Because of the rated limited condition, the amount of tweets that can be fetched is fixed. When most of the collection channel was occupied by the event-irrelevant tweets, the amount of event related tweets is significantly reduced. On the other hand, the RKwA performance well on collecting extra event-related information: compared with baseline crawling, the RKwA crawler fetched more than 70% extra event-related information at the event peak. On average, the RKwA can identify about 100 more event tweets every minute compared to baseline crawling. These extra event-related tweets give additional event information. The observation here is that RKwA performs well. It supplies extra event-related tweets while reducing the noise in the SKwA.

One interesting phenomenon is that there's also a small amount of noise in the baseline dataset. Even though the word "*glastonbury*" is highly specific to the festival, it also introduces noise because there are tweets that contained Glastonbury but 1) did not talk about the event itself; 2) were spam tweets. The proposed grading strategy is not always successful in tackling the first problem but is designed to deal with the second one. It works well for spam tweets that are published to spread trending topic and hashtags. These kinds of tweets contain many hashtags without any plain text or content. The grading strategy can identify them by reducing the final score when non-keyword hashtags appear. This kind of spam is one of the most prevalent sources of the irrelevant tweets in the baseline dataset.

6 Conclusion

In this paper, we focus on finding out a solution for crawling Microblog feeds in real-time. By exploiting the hashtags from Twitter feeds, we proposed an adaptive crawling model that reviews the retrieved content to identify new keywords for automatic live event tweets collection. In order to improve the reliability and robustness,

we further refined the KwA to support higher precision. Based on the evaluation results, we have shown that:

- The trend detection based SKwA is not efficient enough to identify event keywords for adaptive crawling, as it introduces too much noise;
- The RKwA performs well in reducing non-related keywords, and distinguishes an extra amount of the event-related keywords from the noisy hashtags;
- The adaptive crawler based on RKwA is able to collect extra event-related tweets (70%) compared to the baseline crawling approach, while maintaining a noise level below 35 tweets per minute.

Future work for this adaptive crawling model includes an improvement of the new keyword selection schema and the use of an auto initial seed setup. Currently, the threshold value for the correlation is set to be a fixed value. If the system can automatically update the thresholds without losing the real-time efficiency, the performance will be more stable. Also, this can reduce the chance of hitting the rate limit. Another improvement regarding to the keyword selection is the automatic selection of baseline keywords, i.e. initial seeds. Furthermore, research towards identifying and validating additional metrics for accessing the adaptation is also a goal of our future research. The aim is to combine other additional metrics with the RKwA to improve the performance of our adaptive crawler.

References

1. Dejin Zhao and Mary B.Rosson. “How and why people Twitter: the role that micro-blogging plays in informal communication at work”. In *Proceedings of the ACM 2009 international conference on Supporting Group work* (GROUP ’09). pp. 243-252.
2. Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. “Earthquake shakes Twitter users: real-time event detection by social sensors”. In *Proceedings of the 19th international conference on World Wide Web* (WWW ’10). pp. 851-860.
3. Kate Starbird and LeysiaPalen. “(How) will the revolution be retweeted?: information diffusion and the 2011 Egyptian uprising”. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (CSCW ’12). pp. 7-16.
4. Hila Becker, Dan Iter, Mor Naaman, and Luis Gravano. “Identifying content for planned events across social media sites”. In *Proceedings of the fifth ACM international conference on Web search and data mining* (WSDM ’12). pp. 533-542.
5. Deepayan Chakrabarti and Kunal Punera, “Event summarization using tweets”. In Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM ’11). pp. 66-73.
6. Sophia B. Liu and LeysiaPalen,“Spatiotemporal mashups: A survey of current tools to inform next generation crisis support”. In *Proceedings of the 6th International Conference on Information Systems for Crisis Response and Management*(ISCRAM ’09)
7. Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. “A few chirps about twitter”. In *Proceedings of the first workshop on Online social networks* (WOSN ’08). pp.19-24.
8. Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. “Predicting elections with twitter: What 140 characters reveal about political sentiment”. In *Proceedings 4th International AAAI Conference on Weblogs and Social Media* (ICWSM ’10). pp178-185.

9. Fabian Abel, IlknurCelik, Geert-Jan Houben, and Patrick Siehndel. “Leveraging the semantics of tweets for adaptive faceted search on twitter”. In *Proceedings of the 10th international conference on the Semantic Web* (ISWC’11). pp1-17.
10. Albert Bifet, Geoffrey Holmes, and Bernhard Pfahringer.“MOA-TweetReader: real-time analysis in twitter streaming data”. In *Proceedings of the 14th International Conf on Discovery Science* (DS?11).pp. 46?60
11. Sasa Petrovi, Miles Osborne, and Victor Lavrenko. “Streaming first story detection with application to Twitter”. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (HLT ’10). pp. 181-189.
12. Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. “Social networks that matter: Twitter under the microscope”. Dec 2008.
13. Feng Liang, RunweiQiang, and Jianwu Yang. 2012. “Exploiting real-time information retrieval in the microblogosphere”. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries* (JCDL ’12). pp. 267-276.
14. Oren Tsur and Ari Rappoport. “What’s in a hashtag?: content based prediction of the spread of ideas in microblogging communities”. In *Proceedings of the Fifth ACM international conference on Web search and data mining* (WSDM ’12). pp. 643-652.
15. HaewoonKwak, Changhyun Lee, Hosung Park, and Sue Moon. “What is Twitter, a social network or a news media”. In *Proceedings of the 19th international conference on World Wide Web* (WWW ’10). pp. 591-600.
16. Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. “Twitinfo: aggregating and visualizing microblogs for event exploration”. In *Proceedings of the 2011 annual conference on Human factors in computing systems* (CHI ’11). pp. 227-236.
17. Mehdi Naghavi and Mohsen Sharifi. “A Proposed Architecture for Continuous Web Monitoring Through Online Crawling of Blogs”. *International Journal of UbiComp*, 3(1), pp. 11-20.
18. Siqi Zhao, Lin Zhong, JehanWickramasuriya and VenuVasudevan, “Human as real-time sensors of social and physical events: a case study of Twitter and sports games”, *Technical Report TR0620-2011*, Rice University and Motorola Mobility, June 2011
19. Lanagan, James and Smeaton, Alan. “Using Twitter to Detect and Tag Important Events in Sports Media”, In *Proceedings of the Fifth ACM international conference International AAAI Conference on Weblogs and Social Media* (ICWSM ’11)
20. Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. “Summarizing sporting events using twitter”. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces(*IUI ’12). pp. 189-198.
21. Jie Yin, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. “Using Social Media to Enhance Emergency Situation Awareness”, *Intelligent Systems*, IEEE , vol.27, no.6, pp.52-59, Nov.-Dec. 2012
22. Fernando Perez-Tellez, David Pinto, John Cardiff.And Paolo Rosso. “On the difficulty of clustering company tweets”. In *Proceedings of the 2nd International Workshop on Search and Mining UserGenerated Contents* (SMUC ?10), pp. 95?102.
23. April Kontostathis, Leon Galitsky, William M. Pottenger, Soma Roy, and Daniel J. Phelps. “A survey of emerging trend detection in textual data mining”. *Survey of Text Mining*, pp. 185–224, 2003.
24. Michael Mathioudakis and Nick Koudas.“TwitterMonitor: trend detection over the twitter stream”, In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, June 06-10, 2010, Indianapolis, Indiana, USA
25. Mario Cataldi, Luigi D.Caro, and Claudio Schifanella.“Emerging topic detection on Twitter based on temporal and social terms evaluation”.In *Proceedings of the 10th International Workshop on Multimedia Data Mining*, p.1-10, July 25-25, 2010, Washington, D.C.
26. Loulwah AlSumait, Daniel Barbar, and Carlotta Domeniconi.“On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking”. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining* (ICDM’08), p.3-12, December 15-19, 2008

27. Jey Han Lau, Nigel Collier, and Timothy Baldwin. “On-line trend analysis with topic models: #twittertrends detection topic model online”. In *Proceedings of the 24th International Conference of on Computational Linguistics*, pp. 1519-1534.
28. Liangjie Hong and Brian D. Davison. “Empirical study of topic modeling in Twitter”. In *Proceedings of the First Workshop on Social Media Analytics* (SOMA ’10). ACM, New York, NY, USA, 80-88.
29. Andrea Varga, Amparo E. Cano, and Fabio Ciravegna. “Exploring the similarity between social knowledge sources and twitter for cross-domain topic classification”. In *Proceedings 11th International Semantic Web Conference on Knowledge Extraction and Consolidation from Social Media*, (ISWC2012).
30. Dhekar Abhik and Durga Toshniwal, “Sub-event Detection of Natural Hazards Using Features of Social Media Data”, In *International World Wide Web Workshop on Social Web for Disaster Management* (SWDM’13)2013: Rio de Janeiro, Brazil.
31. Xinyue Wang, Laurissa Tokarchuk, Felix Cuadrado and Stefan Poslad. “Exploiting Hashtags for Adaptive Microblog Crawling”, In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (ASONAM 2013)
32. Mehmet B. Baray, Hakan Kurt. “On-line new event detection and tracking in a multi-resource environment”. *Unpublished master’s thesis*, Bilkent University, Computer Engineering Department.
33. Changhyun Byun, Yanggon Kim, Hyeoncheol Lee, and Kwangmi Ko Kim. 2012. “Automated Twitter data collecting tool and case study with rule-based analysis”. In *Proceedings of the 14th International Conference on Information Integration and Web-based Applications and Services* (IIWAS ’12). pp. 196-204.
34. Matko Boanjak, Eduardo Oliveira, Jose Martins, Eduarda Mendes Rodrigues, and Luos Sarmento. “TwitterEcho: a distributed focused crawler to support open research with twitter data”. In *Proceedings of the 21st international conference companion on World Wide Web* (WWW ’12 Companion). pp. 1233-1240.