Exploiting Hashtags for Adaptive Microblog Crawling

Xinyue Wang, Laurissa Tokarchuk, Félix Cuadrado and Stefan Poslad
School of Electronic Engineering and Computer Science
Queen Mary, University of London
London, UK
{xinyue.wang, laurissa.tokarchuk, felix.cuardado, stefan.poslad }@eecs.qmul.ac.uk

***Abstract*—Researchers have capitalized on microblogging services, such as Twitter, for detecting and monitoring real world events. Existing approaches have based their conclusions on data collected by monitoring a set of pre-defined keywords. In this paper, we show that this manner of data collection risks losing a significant amount of relevant information. We then propose an adaptive crawling model that detects emerging popular hashtags, and monitors them to retrieve greater amounts of highly associated data for events of interest. The proposed model analyzes the traffic patterns of the hashtags collected from the live stream to update subsequent collection queries. To evaluate this adaptive crawling model, we apply it to a dataset collected during the 2012 London Olympic Games. Our analysis shows that adaptive crawling based on the proposed Refined Keyword Adaptation algorithm collects a more comprehensive dataset than pre-defined keyword crawling, while only introducing a minimum amount of noise.**

Keywords—Social Network, Twitter, Data Crawler, Hashtag

I. INTRODUCTION

The enormous popularity of microblogs, combined with their conversational characteristic [1] has led them to become one of the most popular platforms for extracting information. Early research identified characteristics of information diffusion and users behavior on the entire microblogsphere [10][12]. Nowadays, the focus has shifted to real world event detection [9] and event summarization [5]. For instance, recent research has examined the use of such tools, primarily Twitter-based, to get knowledge about ongoing affairs [4][7], or even to dig out hints of upcoming events [2][6].

In order to identify and analyze events in the Twittersphere, a comprehensive dataset describing the event is compulsory. The majority of 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” [17]. However, the set of predefined keywords is subjective and can easily lead to incomplete data. Even given expert knowledge, keywords and specialised hashtags often arise in the midst of such events. During a football event at the London 2012 Olympics, apart from hashtags *#football* and *#olympic*, people also published tweets with *#FIFA* and *#GBRvKOR* in a game played between Britain and Korea. The prediction of keywords for situation awareness during emergencies or disasters is even harder. In these scenarios, people will communicate their observations and perceptions without the explicitly mention of the event “keywords” terms [17]. Another bias of collection is introduced by Twitter’s free API restrictions. The Twitter Streaming API provides no more than 1% of the total traffic. Most data collected during a popular event or crisis will easily hit this limit. Furthermore, the Twitter Search API only allows for retrieval of tweets within a week, making historical reconstruction difficult as well. The collection of big datasets under these restrictions has scalability issues and sometimes doesn’t provide compressive enough information about the events themselves, and therefore significantly affects the performance of the Twitter-based analysis algorithm [16].

Accordingly, the problem we address in this paper is how to automatically, i.e. no manual modification of the search terms, gather a comprehensive set of social media documents with unknown features. The proposed approach is to collect an extended set of event relevant information from the Twitter live streams by identifying extra search terms.

In designing the proposed crawling model, the challenge is to identify new hashtags that without the appearance of the original keywords in contents related to the event in question. Specifically, the novel contributions are as follows:

- We develop the recall-oriented query that exploits emerging popular hashtags and examine it in a live event by integrating it to the data crawler;
- We improve the adaptive performance by refining the keywords selection algorithm;
- We demonstrate that our method collects more relevant tweets than the most existing approaches and reducing the amount of irrelevant information.

The remainder of this paper is divided into four sections. Section II introduces the related work and distinguishes our work with those existing ones; section III details the proposed adaptive crawling models and two keyword adaption algorithms; section IV reports the evaluation of our technique, showing its performance over the 2012 Olympics dataset; and finally section V concludes our work and future directions.

II. RELATED WORK

In addition to pre-defined keywords searches for collection, attempts to use additional metrics as search criteria have also been made. Fabian et al. leverage users’ profiles, semantics meanings and metadata of tweets to generate new search

978-1-4799-1498-2/13/$31.00 ©2013 IEEE

criteria based on materials from news websites [7]. Though the accuracy is improved, the cost of calculation increases exponentially. Furthermore, these solutions focus on improving the user experience for interactive searching rather than collecting event-related tweets for ongoing affairs. On the other hand, an online crawling architecture emphasizes continuous crawling on the blogsphere [14]. However, the differences between blog and microblog make the migration difficult.

A precision and recall-oriented search query generation technique is presented by Becker et al. in [4]. In this work, the authors present a strategy to automatically identify event features, generating queries to retrieve content from diverse social media sites for planned events. While their query identifies content across different social media sites, our work focus on maximizing the utility of content from a single platform. The algorithm used in [8] is closest to our initial approach (SKwA) in the sense that it also identifies new event-related search terms. However, their work provides no quantitative analysis of the performance or dataset characteristics. Instead of using the TF-IDF measure for whole tweets, we only focus on the hashtags because for three reasons: first, the accuracy of TF-IDF on Tweets-alike documents is still uncertain [19]; second, hashtags are used as topical markers to link relevant topics and events [11]; third, exploiting hashtags for keyword searching reduces the complexity in getting semantic meaning and increases the efficiency of data analysis.

III. TWITTER CRAWLING MODEL DESIGN

A. *System Flow of Twitter Crawling Model*

A Twitter crawler collects tweets through the Twitter API that matches a set of search criteria. In this work, we are interested in keyword-based crawling, where every matching tweet will contain at least one of the defined search keywords.

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. All the retrieved tweets are stored in a database. As this approach is the one adopted by most existing research, we use a dataset collected by this model as the ground truth in our evaluation.

2) *Adaptive Crawling*

The system structure of the adaptive crawling model is similar to the baseline, except an additional *Keyword Adaptation* feature. This feature enables the application of either the Simple Keyword Adaptation algorithm (SKwA) or the Refined Keyword Adaptation (RKwA) algorithm described in the next section.

In this model, the data collection process is started by the same set of predefined keywords as the baseline. The keyword adaptation feature enables the identifying of popular event-related hashtags by using the Keyword Adaptation Algorithm. Then, those hashtags are added to the keywords retrieval set at the end of every time frame. Finally, a query that encodes all the words in the keywords set is sent to the Twitter API when the timer restart and another iteration of adaptation begin.

B. *Keyword Adaptation Algorithm*

The goal is to automatically find a list of hashtags, beyond the initial set of keywords, appearing in tweets related to the event of interest that gathers extra event-related information retrieval. In our first attempt, we assume that the hashtags that frequently appear in the baseline tweets help to collect additional relevant information. However, the Simple Keyword Adaptation algorithm enclosed this assumption introduces lot of noise. We then propose the Refined Keyword Adaptation algorithm to improve the performance. In this section, full details of both Keyword Adaptation algorithms are presented.

1) *Simple Keyword Adaptation Algorithm (SKwA)*

In this algorithm, the collection of hashtags within a fixed time frame is represented as $H_{tf}(t_n) = \{h_1, h_2, ...\}$, while the keywords set, sent to Twitter API at any time frame n, is $H(t_n) = \{h_1, h_2, ...\}$ where h_k is an individual hashtag. Two frequency lists are maintained, one for the whole collection period $freq(t_n)$, and the other for the current time frame $freq_{tf}(t_n)$. $freq(t_n)$ updates every time frame, while $freq_{tf}(t_n)$ updates when a new tweet arrives. The hashtag the frequency lists map in pair, i.e. the frequency of a hashtag h_k for the whole collection period is $freq(t_n)[k]$, is $freq_{tf}(t_n)[k]$ at time frame n. A minimum threshold frequency $(freq_{min})$ for a hashtag to be considered as keyword, and an array of blacklist hashtags (H_{black}) are also used. The pseudocode below describes the proposed algorithm:

```
Algorithm Simple Keyword Adaptation (SKwA)
  for $\forall h \in H_{tf}(t_n)$
    if $h \in H_{blacklist}$ or $freq_{tf}(t_n)[k] < freq_{min}$
      $H_{tf}(t_n) = \{ h_k \mid h_k \in H_{tf}(t_n), h_k \neq h\}$
      $freq_{tf}(t_n)=\{freq_{tf}(t_n)[k] \mid freq_{tf}(t_n)[k \in freq_{tf}(t_n), h_k \neq h\}$
    else
      $H(t_n) = H(t_{n-1}) \cup$
                $\{h_k \mid freq_{tf}(t_n)[k] \in Top\ n\ (freq_{tf}(t_n)[k]);$
      $freq(t_n) = freq(t_{n-1}) \cup$
            $\{freq_{tf}(t_n)[k] \mid freq_{tf}(t_n)[k] \in Top\ n\ (freq_{tf}(t_n)[k]);$
                              where $n=N-num(H_{t-1})$.
    end if;
  end for;
```

This algorithm keeps at most N[1] keywords for querying Twitter every 5 minutes. When a new hashtag appears, the algorithm will check whether it is in $H(t_n)$. If it already is a query keyword, its $freq(t_n)$ is incremented by 1. Otherwise, the hashtag is stored to $H_{tf}(t_n)$ temporarily. When the timer expires, hashtags in $H_{tf}(t_n)$ are sorted according to their frequency. Top ones will be added to the keywords set. In this step, hashtags with low $freq_{tf}(t_n)$ don't become keywords. Besides, hashtag will be removed from the keywords set if it got low frequency for a long period of time.

2) *Refined Keyword Adaptation Algorithm (RKwA)*

Our initial attempts show that extra traffic, both event-related and noise, is generated when using the proposed SKwA. Furthermore, the longer the crawler runs, the larger the proportion of noise. Eventually, the noise will overwhelm the

[1] N: maximum number of keyword (400 in Twitter Streaming API V1.0).

event-related data, resulting in a meaningless dataset. This is because the algorithm highly relies on the collected contents.

In order to reduce the impact of noises in the adaptive dataset, the traffic pattern of hashtags is exploited to classify potential keywords according to their relevance to the events. In the Refined Keyword Adaptation algorithm, SKwA is modified to enable the collection of a greater amount of highly event-associated data without significantly increasing the noise.

RKwA first automatically gets a keywords list based on the algorithm in SKwA. The collection of initial hashtags seed is represented as $H_{ini} = \{h_1, h_2, ...\}$. The keywords set at the end of each time frame is written as $H_{Fin}(t_n)$. This keywords set contains hashtags from SKwA's $H(t_n)$ which have high correlation with the initial seeds. The pseudocode is as follows:

```
Algorithm Refined Keyword Adaptation (RKwA)
Execute SKwA
    $H_{Fin}(t_n) = H_{ini}$
    for $\forall h_i \in H(t_n)$
        for $\forall h_j \in H_{Fin}(t_n)$
            if $H_{Fin}(t_n) = H_{ini}$ and $cor(h_i, h_i) > Thres_1$
                $H_{Fin}(t_n) = \{h|h \in H_{Fin}(t_n)$ or $h = h_i\}$
            else if $H_{Fin}(t_n) \neq H_{ini}$ and $cor(h_i, h_i) > Thres_2$
                $H_{Fin}(t_n) = \{h|h \in H_{Fin}(t_n)$ or $h = h_i\}$
            end if;
        end for;
    end for;
```

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

Hypothesis 1 *(H1)*: *the initial keywords used for both baseline crawler and SKwA 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 pattern to each others.*

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 are selected as starting seeds in RKwA. The Pearson correlation is chosen as the measurement of similarity between related keywords.

In this algorithm, Hashtag $h \in H_{tf}(t_n)$, as calculated by SKwA, is only retained in RKwA if it has high correlation with one of the seed keywords. In order to calculate the *correlation* between two hashtags, we subdivide the time frame into several time slots, and the sequence is the frequency counts of each time slot. We use a single variable approach to set the value of $Thres_1$ and $Thres_2$ as 0.5 and 0.8 respectively.

IV. Evaluation of Keyword Adaptation Algorithm

A. Dataset Overview

A dataset collected during the 2012 Olympic is used in this experiment. Two data crawlers were run: baseline crawling model and adaptive crawling with SKwA. Both crawlers employed "Olympic" and "London2012" as initial keywords. Consequently, two separated datasets were collected.

TABLE I. Tweet Volume of Collected Datasets

Tweets count *Dataset*	**Total**	**Unique**
Baseline	14,916,105	5,323,011(77%)
SKwA	58,759,453	49,166,359(36%)

As shown in TABLE I. , millions of Olympic related tweets were generated during 27th of July to 11th of August. The column "unique" is the number of tweets that appeared only in that dataset. Both crawlers collected on the same keywords, so ideally, the SKwA adaptive crawler should collect all tweets in the baseline dataset, i.e. Baseline Unique should equal zero. However, some of the tweets, even with the initial keywords, are not retrieved by the SKwA adaptive crawler due to the 1% rate limitation. As the number of keywords increases, the volume of tweets containing those keywords also increases with a potential to exceed the rate limit. According to the collected data, Twitter returns up to 3000 tweets every minute when rate limited. Since the 1% tweet volume is spread out across all keywords, the available volume for tweets carrying the initial keywords is reduced in the SKwA dataset, which results in the unique tweets in the baseline dataset.

B. Experiment Setup

The aim of this experiment is to verify that the noise (non-related tweets) to signal (event-related tweets) ratio is reduced by using of the RKwA. The following hypothesis act as a condition for evaluating the performance of RKwA:

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

H3 determines whether the tweet's hashtags affect the tweet's relationship to an event of interest. Based on this hypothesis, we design a procedure for performance evaluation of the adaptive crawling model with RKwA as follows:

1) *Using new keyword to filter dataset*

The dataset will change according to the new list H_{Fin}. Although the true volume produced by RKwA adaptive crawler is unknown (rate limiting would be applied proportionately to the new, reduced keyword list), we can still conclude that the RKwA is better than SKwA if RKwA can retain most of the event-related tweets and reduce noisy tweets. In this case, the RKwA dataset is composed of tweets with keywords identified by RKwA in SKwA datasets.

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.

Hashtags in different time periods were labeled according to how closely they are related to Olympic events. For example, "#2012olympic"is definitely related, while "#harrypotter" is more complicated: it could be related since the opening ceremony features Harry Potter, but isn't related to the Olympic in the longer term. Accordingly, hashtags were labeled into four categories as shown in TABLE II. The final list was based on the average results of 5 independent taggers.

TABLE II. THE HASHTAG CATEGORY AND GRADING STRATEGY

Hashtag Category	Specification	score
Related (C1)	Contains baseline criteria, team name or event name	+3
Possibly-related (C2)	Country name, reference of specific temporal meaning	+1
Non-related(C3)	Media companies, generic emotions	-3
Not known (C4)	Non-English hashtags that the manual taggers didn't identify	0
Non-keyword hashtags	Hashtags that not been selected as keywords	-1

3) *Classify tweets according to the manually labeling*

In this step we classify whether a tweet is related to the event based on the hashtags it contains using the grading system in TABLE II. 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 less than or equal to 0, as non-related tweets. Therefore, the baseline, the SKwA and the 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 the RKwA dataset to check the levels of noise reduction achieved and the proportion of event-related information retained.

C. Results

We evaluate the RKwA adaptive crawler by applying RKwA to timeslot 20:00 to 21:00 on 4th Aug 2012, when the Men's 4x100m Medley Relay final was taking place, and to the time slot 21:30 to 22:30 on 5th Aug 2012, when the Men's 100m final happened. Both experiments produced similar results. Due to the space constraints, only the results for the first experiment are presented here.

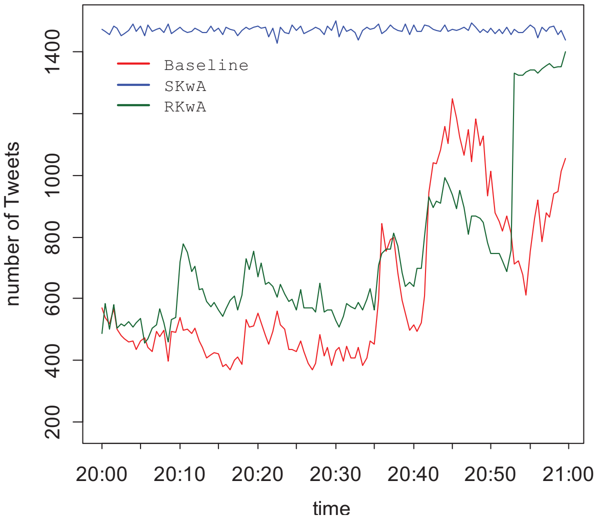


Fig. 1. Tweet volume over the three datasets for Men's 4x100m Medley

We first compare the traffic volume generated by the three different approaches. Fig. 1 gives the tweets count during our test period. The traffic in red and blue are the real traffic collected by the corresponding crawlers, while the green one is filtered traffic from the blue one. Overall, the volume of data filtered by RKwA is approximately half of the amount collected by SKwA and generally higher than the baseline. One exception is from around 20:35 till 20:55, where the traffic volume from RKwA is lower than the baseline. During this period, many people started to tweet about the Olympic Men's 4x100m Medley Relay. The baseline crawler collected nearly 18000 event-tweets missed by RKwA due to both the number of keywords used for this algorithm and rate limiting. The RKwA dataset is a subset of the SKwA dataset, and thus those 18000 tweets from the baseline do not appear in the RKwA dataset. While this reduces the volume of information extracted by RKwA in this scenario, in a live crawling scenario, given the same rate limit, the volume of tweets collected by RKwA would be greater than that collected by the baseline crawler.

In the next step of our evaluation, we want to check whether the keywords identified by RKwA are more informative and useful for tweet collection than those identified by SKwA. The number of keywords (excluding Olympic and London2012) retained by the SKwA in comparison to the RKwA during 21:20 to 21:30 in 4th Aug 2012 is 15.45%, broken down by category: C1: 21.06%, C2: 11.61%, C3: 8.45%, C4 40.00%. RKwA only retains less than 10% of the noisy (C3) keywords while retaining an acceptable ratio of related (C1 and C2) keywords. This indicates the important event-related keywords are more likely to be retained by using RKwA. In addition, RKwA keeps most data for trending events, as shown in Fig. 2. The tweets retained for the crucial event period is equal in both algorithms (green and blue lines merge).

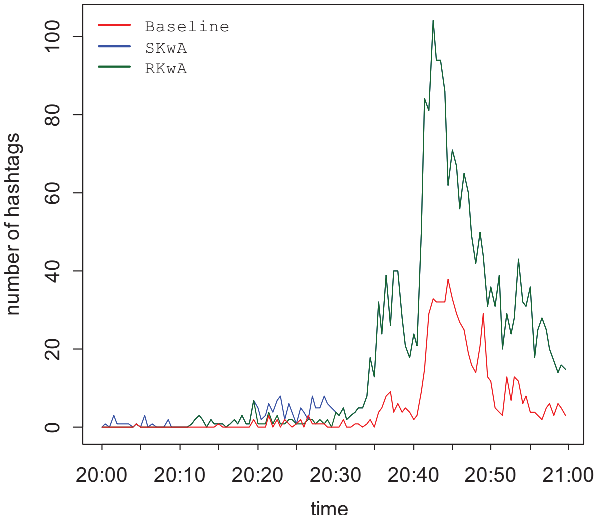
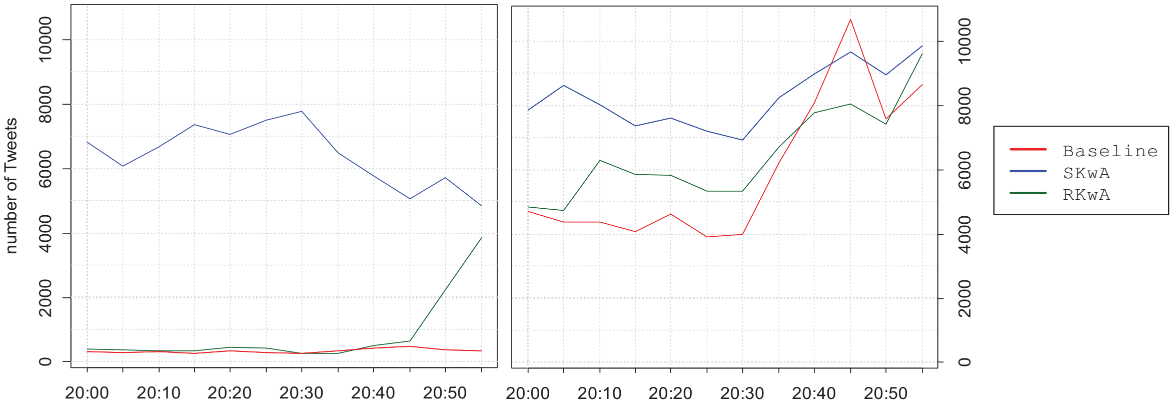


Fig. 2. Traffic pattern of #phelps in three datasets for Men's 4x100m Medley

In general, the trends in traffic volume are the same for all the three lines. The difference of information gain between baseline and adaptive illustrates that the adaptive crawling fetches additional event-related information. More specifically, since Fig. 1 and Fig. 2 illustrate the same period of time, it is obvious that even if the traffic volume retained by the RKwA dataset at 20:45 to 20:55 is low, it still maintains the event-relevant information gain shown in the SKwA dataset. Furthermore, while the total volume of tweets has gone down and the information gain has increased, it is clear that RKwA has achieved significant reduction in noise.



a) non-related tweets b) related tweets

Fig. 3. Tweets volume in three different datasets for Men's 4x100m Medley

In fact, the retained ratio of event-related tweets is more important. In order to examine whether the RKwA also performs well for retaining event-related tweets, we classify tweets by applying the grading strategy. Fig. 3 shows the traffic volume of event-related tweets and noisy tweets in the three datasets. Note that the volume shows in green is a set filtered from the SKwA dataset. Fig. 3a) illustrates that the proposed RKwA performs well on reducing the amount of noise. Overall, 86.90% of noise is eliminated, rising 94.56% during 20:00 to 20:45. The sudden increase of non-related tweets in RKwA dataset from 20:40 indicates that proper $Thres_1$ and $Thres_2$ change slightly in different time frames. For example, settings of $Thres_1$=0.75 and $Thres_2$=0.8 at 20:45 to 21:00 will guarantee a flat RKwA line during the entire test period in Fig. 3a). Fig. 3b) shows that RKwA also does an acceptable job for retaining event-related information: 78.26% of event-related information in the SKwA dataset is preserved in the RKwA dataset. Given that the rate limit doesn't change, the RKwA will collect more related information than the SKwA or the baseline.

V. Conclusion and Future Work

In this paper, we focus on finding a solution for crawling microblog feeds in real-time. By exploiting hashtags from Twitter feeds, we proposed a recall-oriented adaptive crawling model (SKwA) that identifies new keywords for automatic live event tweets collection. In order to improve the reliability, we further refined the adaptive model (RKwA) to support higher precision. Based on the evaluation results, we have shown that RKwA performs well in reducing non-related keywords, while retaining more significant event-related keywords. Furthermore, it maintains 78.26% of event-related tweets and removes 86.90% non-related tweets from the SKwA dataset.

Future work includes the improvement of the new keyword selection schema and exploratory study of hashtags' traffic patterns. Currently, the threshold values, $Thres_1$ and $Thres_2$ are fixed values. If the system itself can automatically choose these thresholds without losing real-time efficiency, the performance of the RKwA will be more stable. Automatic initial seeds can also improve the stability and increase the RKwA/SKwA ratio of related keywords. Furthermore, research on exploring hashtags' traffic patterns, e.g. correlation analysis of hashtags' traffic and co-occurrence of hashtags, is our next target.

VI. Acknowledgment

The authors acknowledge support from the EU FP7 funded project TRIDEC (Grant No. 258723) and from the China Scholarship Council (for Xinyue Wang).

VII. References

[1] Zhao D., Rosson M. B. "How and why people Twitter: the role that microblogging plays in informal communication at work". In GROUP'09

[2] Sakaki T., Okazaki M., Matsuo Y., "Earthquake shakes Twitter users: real-time event detection by social sensors". In WWW '10.

[3] Starbird K., Palen L., "(How) will the revolution be retweeted?: information diffusion and the 2011 Egyptian uprising". In CSCW '12.

[4] Becker H., Iter D., Naaman M., Gravano L. "Identifying content for planned events across social media sites". In WSDM '12.

[5] Chakrabarti D., Punera K., "Event summarization using tweets". In ICWSM '11

[6] Tumasjan, A., Sprenger, T. O., Sandner, P. G., Welpe, I. M. "Predicting elections with twitter: What 140 characters reveal about political sentiment". In ICWSM '10.

[7] Abel F., Celik I., Houben G., Siehndel P.. "Leveraging the semantics of tweets for adaptive faceted search on twitter". In ISWC'11

[8] Bifet, A., Holmes, G., Pfahringer, B. "MOA-TweetReader: real-time analysis in twitter streaming data". In DS'11

[9] Petrović S., Osborne M., Lavrenko V., "Streaming first story detection with application to Twitter". In HLT '10.

[10] Huberman B. A., Romero D. M., Wu F.. "Social networks that matter: Twitter under the microscope". Dec 2008.

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

[12] Kwak H., Lee C., Park H., Moon M., "What is Twitter, a social network or a news media?". In WWW '10

[13] Marcus A., Bernstein M. S., Badar O., Karger D. R., Madden S., Miller R. C.. "Twitinfo: aggregating and visualizing microblogs for event exploration". In CHI '11

[14] Naghavi, M., Sharifi, M. "A Proposed Architecture for Continuous Web Monitoring Through Online Crawling of Blogs". IJU, 3(1), 2012

[15] Zhao S., Zhong L., Wickramasuriya J., Vasudevan V., "Human as real-time sensors of social and physical events" June 2011

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

[17] Nichols J., Mahmud J., Drews C.. "Summarizing sporting events using twitter". In IUI '12

[18] Yin J., Lampert A., Cameron M., Robinson B., Power R., "Using Social Media to Enhance Emergency Situation Awareness," Intelligent Systems, IEEE , 27(6), 2012

[19] Perez-Tellez F., Pinto D., Cardiff J., Rosso P. "On the difficulty of clustering company tweets". In SMUC '10