TMEDS: Twitter based Minor Event Detecting System

(Invited Paper)

R. Uday Kiran*†, Masaru Kitsuregawa*‡
*The University of Tokyo, Tokyo, Japan
Email: {uday_rage, kitsure}@tkl.iis.u-tokyo.ac.jp
†National Institute of Information and Communications Technology, Tokyo, Japan
‡National Institute of Informatics, Tokyo, Japan

Extended Abstract

Finding events from Twitter data is an important research issue in data mining. The current event detection algorithms focus on finding all major events happening within a particular time frame. One cannot ignore the knowledge pertaining to minor events. The reason is that minor events, such as land slides and nuclear leaks, can provide useful and valuable information to the users in real-world applications. Discovering minor events are challenging due to the noisy nature of Twitter. To handle noisy data, researchers have focused on finding minor events for a specific topic, say land slides [1]. As a result, these approaches lack the generalization of discovering several events happening together at a particular time frame.

We have been investigating the behavior of major and minor events in Twitter for the past two years. As hashtags provide useful information about an event, we have analyzed hashtags with respect to their frequency and temporal appearances. In our investigation, we have observed that when an event happens, hashtags of the corresponding event exhibit periodic behavior within the event's time frame. Based on this observation, we introduce a generic Twitter based Minor Event Detection System (TMEDS) that tries to discover several events happening together at a particular time frame.

A. Twitter based Minor Event Detection System

Our system has three main components: (i) finding top-k hashtags (ii) discovering recurring patterns by modeling Twitter data as timeseries and (iii) finding events. Figure 1 shows the architecture of our system. The three think dotted rectangles represent each component of our system. We briefly explain each of these components.

Our TMEDS collects raw tweets using Twitter's application program interface (APIs). The set of raw tweets are fragmented into distinct subsets with respect to time, say minute, hour, and day. The hashtags appearing in each subset of Tweets are extracted and modeled as a document. Next, top-k hashtags are identified using **term frequency and inverse document frequency** (TF-IDF). The TF-IDF facilitates our system to preserve the knowledge pertaining to rarely appearing hashtags by pruning out commonly appearing hashtags. These steps are performed in our first component of our system.

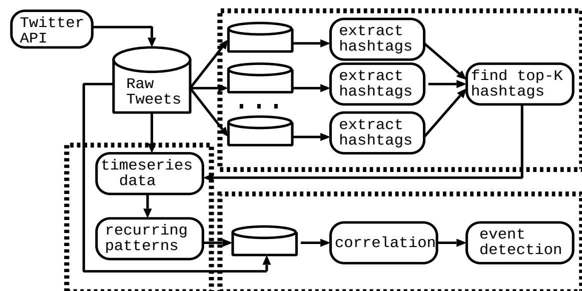

Fig. 1. System architecture.

The second component plays a crucial role in our system. Our observation (discussed in Section 1) is implemented in this component to find useful information pertaining to minor events and their durations of occurrences. This information (i.e., periodic durations of the events) play a key role in our final component, which segments the data and discovers minor events. In the second component, the top-k hashtags appearing in raw tweets are modeled as time series, and recurring pattern mining is performed to discover recurring pattern. Each recurring pattern represents a set of hashtags that have exhibited periodic behavior for particular time intervals within the data. Briefly, our recurring pattern model is as follows.

Let $H = \{h_1, h_2, \cdots, h_n\}$ be the set of hashtags. Let $e = (ts, X)$ be an event, where $X \in H$ is a pattern and $ts \in R$ is the timestamp at which the corresponding event has occurred in the data. Let $TSD = \{e_1, e_2, \cdots, e_m\}$ denote the timeseries data. For an event $e = (ts, Y)$, such that $X \subseteq Y$, it is said that X occurs in e and such a timestamp is denoted as ts^X. Let $TS^X = \{ts_k^X, \cdots, ts_l^X\}$, where $1 \leq k \leq l \leq m$, denote an **ordered set of timestamps** at which X has occurred in TSD. Given the user-defined $period$, a recurring pattern Z is represented as follows:

$$Z \quad [support = x\%, \{[ts_a^X, ts_b^X], [ts_c^X, ts_d^X], \cdots\}].$$

The above pattern says that Z has appeared in $x\%$ of the events, and its cyclic repetitions have been observed from ts_a^X to ts_b^X, from ts_c^X to ts_d^X and so on. The intervals $[ts_a^X, ts_b^X]$ and $[ts_c^X, ts_d^X]$ are called the **periodic-intervals** of Z and they

TABLE I
SOME OF THE INTERESTING RECURRING PATTERNS AND THEIR EVENTS DISCOVERED IN TWITTER.

S.No	Recurring Pattern	Periodic duration	Cause for the events
1	{yyc, uttarakhand}	[2013-06-21 01:08, 2013-07-01 04:27]	On June 20, Uttarakhand, a state in India, and Alberta, a province in Canada have witnessed heavy floods.
2	{nuclear, hibaku} (In Japanese, hibaku means radiation.)	[2013-05-06 22:33, 2013-05-24 22:13], [2013-07-01 06:17, 2013-07-14 06:21]	(i) A Japanese minister has visited Chernobyl, Ukraine to learn from the recovery from the severe nuclear accident. (ii) People were tweeting about detection of Plutonium at a point 12 KM from Fukoshima nuclear reactor.

provide key information pertaining to the periodic duration of hashtags. The model and a pattern-growth algorithm to discover recurring patterns is presented in [2].

In our final component, we use periodic-intervals of a recurring pattern to discover events. As a result, this step is performed for every discovered recurring pattern. The following steps are performed in this component. (i) We select a recurring pattern and perform data segmentation by collecting only a portion of raw tweets that have appeared within the periodic-intervals of corresponding pattern. Let D denote the collected data. (ii) A set of hashtags in a pattern can represent either a single event or multiple events happening together at a particular time frame. To handle this dilemma, we find correlation between the hashtags in D using statistical measures, such as all-confidence and normalized google distances. In our system, we have employed all-confidence as it satisfies the null-invariance property. This property facilitates to disclose genuine correlation relationships without being influenced by the object co-absence in a database. (iii) In the final step, we apply an event detection algorithm that uses correlation between the hashtags to discover events [3].

B. Experimental Results

We have been crawling the Twitter dataset since 11-3-2011. Our actual data collection consists of user profiles, timestamp and tweet contents including retweets. The dataset contains 260 million users as active users and 783 million tweets of multiple languages. More details regarding the collection of this data are presented in [4]. For our experiment, we have considered the tweets containing English characters for the period 1-May-2013 to 28-August-2013. The raw data contains 44 million tweets. We have split this data into hourly basis, extracted hashtags, and identified top-1000 hashtags using TF-IDF. Next, we have modeled raw tweets containing top-1000 hashtags as timeseries, applied recurring pattern mining to discover the sets of hashtags and their periodic durations within the data. Next, we discovered events from the recurring patterns.

Table I lists some of the recurring patterns and the events related to those patterns. Figures 2 (a) and (b) show the frequencies of the terms present in patterns {yyc, uttarakhand} and {nuclear, hibaku} on a daily basis. It can be observed from the daily frequencies of these hashtags that they are appearing less frequently in our data, however, they are providing useful information about the events relating to floods and nuclear leaks. Thus, our system is able to discover minor events of heterogenous nature.

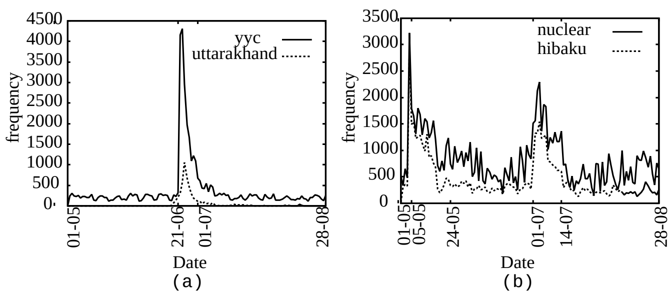

Fig. 2. Frequency of hashtags at different days in database. Date is of form 'dd-mm'. Year of this date is 2013

REFERENCES

[1] A. Musaev, D. Wang, C. Cho, and C. Pu, "Landslide detection service based on composition of physical and social information services," in *ICWS*, 2014, pp. 97–104.

[2] R. U. Kiran, H. Shang, M. Toyoda, and M. Kitsuregawa, "Discovering recurring patterns in timeseries." To be appeared in EDBT 2015.

[3] O. Ozdikis, P. Senkul, and H. Oguztüzün, "Semantic expansion of tweet contents for enhanced event detection in twitter," in *ASONAM*, 2012, pp. 20–24.

[4] G. Rattanaritnont, M. Toyoda, and M. Kitsuregawa, "Analyzing patterns of information cascades based on users' influence and posting behaviors," in *TempWeb '12*, 2012, pp. 1–8.