www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 7 July 2015, Page No. 13007-13010

Event Detection in Twitter Using Mentioning Practice of the User

Chaitra M.
Dept of CSE,B.L.D.E.A CET, Vijayapur
chaitramudenur@gmail.com

Abstract*: Twitter is one of the most popular Online Social Networks (OSN's) nowadays. Twitter is valuable source of timely information. Detecting events in the twitter is difficult task; information on Twitter is overwhelmed by huge volume of uninteresting tweets. Prior works may not be appropriate, because it focuses on only textual contents. In Twitter user may also inserts non- textual contents of their interest. Event detection in Twitter Using mentioning practice of the user (EDT), a novel method that gives desired result of creation frequency of dynamic links (i.e. mentions) that users inserts in tweets to detect important events and find the impactful words of the events The experiments we conducted shows that the proposed method can efficiently detect accurate (95%) and meaningful events.*

Keywords: Twitter, Event Detection, Twitter API's, Mentions.

Introduction

Now a day's online social networking services like Twitter, Facebook, Google+, and LinkedIn play an important role in the spreading wide information in a real-time manner.Some recent observation says that events and news emerge and spread first using these media channels rather than traditional media like online news sites, blogs, television and radio breaking news. Natural disasters, celebrity news, products announcements, show that people make use of these services, discuss and exchange information; this information may fade over time. Twitter is the first medium to break important natural events such as earthquakes in a matter of seconds after they occur. It has a real time flow of tweets (text messages) coming from different source covering various kind of information in different languages and locations.

Twitter is a micro blogging service that enables user to share, discuss and forward different kind of information from personal daily life events to worldwide important events. Number of register user of the twitter around world tweeting make it as a valuable source of timely information. But information in the twitter is overloaded by unrelated topics so it is difficult to identify the events that are most interest the crowd.Event detection in twitter is difficult (challenging) at least for three reasons first tweets are posted at very fast rates and thus produces large volume data, second shortness of the tweets ideas are in brief and may not have enough information, third tweets are noisy in nature. Tweets may contain informal grammatically incorrect text with misspelling and abbreviations and also tweets include personal updates of the users, spam and self promotion processing of such tweets increases the processing time and reduces the quality of the resultTwitter produces continuous stream of tweets, from term-weighted based approach to topic-modelling based approach and including clustering based assumed as signal events. Existing mention-anomaly based event detection and tracking in Twitter [10] method focuses on already collected tweets (dataset). It can predict only those things that have happened some time back. Existing method align with external source of information this procedure is not always suitable mainly in case of controversial events or politically related events because it misrepresent the way events are arrived on twitter. Most of the existing methods deduce predefined fixed time duration.

Existing works fail to consider the social aspect of the twitter and only focuses on textual content, user inserts non-textual content of their interest in the tweets such interest is "mentioning habit" which means that referring other user screen name (using syntax @username) in tweets and hash tags (#topic name) to know what people are talking about on twitter, the topic with hash tags are called trending topics. These mentions are the dynamic links created to participate in discussion with other or specific users or in retweeting.

This method finally produces list of the events each event is described by group of related words and frequency (impaction rate) of mention creation and performs quantitative studies on English twitter of different cities in India and it is able to produce accurate events.

This paper is organized in five main sections: Introduction, Related work, Implementation, Experimental results and Conclusion and future work. Section 1 describes the introduction about twitter motivation of proposed work. Section2 covers the back ground study to understand in order to fully comprehend the system that will be implemented. Also give description about the problems or challenges in existing methods. Section3 presents proposed model that is being implemented. The flow diagram of the proposed method, algorithm, module description, flow charts, data flow diagrams are discussed. Section 4 presents graphs and results obtained from the proposed system. Data collection and parameter settings are discussed. Section 5 summarizes the research results.

1. Related work

Topic detection and tracking is the major area that tackles the problem of discovering events that most interest the crowd. [1] Shows how to extract bursty features from text streams based on modelling the text stream using an infinite state automaton, where bursts are modelled as state transitions. The brief surveys of the proposed approaches are as follows

2.1 Event detection from tweets

Term-weighting based approach [3] for each source, tweets were temporally grouped into "bag of words" style collections, or "documents", each document is corresponds to the tweets posted in a particular period of time. One of the selection criterion in steaming trend detection is [3] normalized frequency ($tf_{normi,j}$) involved utilizing only the term frequency of each element, rather than both term frequency and the inverse document frequency. Meaning the term $tf_{normi,j}$ thought in terms of frequency per million words. Each word in the document was given trending score.

Topic-modeling based approaches [1][7] In the probabilistic topic modeling that is based on LDA [2], the studies have examined latent topics and their changes across time. It assumes that there is k underlying latent topics according to which documents are generated, and over the $|V|$ words in the vocabulary, each topic is represented as multinomial distribution. A document is generated by sampling a combination of these topics and then sampling words form that combination. This approach does not suit the online setting where text streams continuously arrive with time. To overcome this on-line LDA [6] method is come into place. This approach allows LDA model to work in an online fashion such that it incrementally builds an up-to-date model when new document appears.

Clustering based approaches[4][8][9]The most related topic approaches for event detection are EDCoW [5] and Twevent[8] that automatically detect generic events from tweets. For each individual word EDCoW builds signal using wavelet theory to capture the burst in the word's appearance. It removes trivial words using their corresponding auto-correlations, and clustering remaining signals based on the cross correlation between them using modularity based graph partitioning. But it lacks in scalability and computational efficiency. The cross correlation may cluster two events that are not related but happened in the same time span. Twevent uses even segments constructed using the statistical information provided by Microsoft web N-gram service and Wikipedia to present events. Thus the speed of the segmentation relies on this service. To find similarity between the detected segments, Twevent divides the time window into sub-window and, find he cosine similarity between sets of associated tweets for each sub-window This amounts to a lot of additional computations as computing frequencies of the segments in each time window requires another scan of the whole set of tweets.

2. Proposed Method

In this section first we define the event and then give overview of the proposed method. We have considered event as bursty topic

Event: If the topic is bursty, if and only if it has attracted to high level attention during some particular period. Event is characterized by bursty topic.

Overview

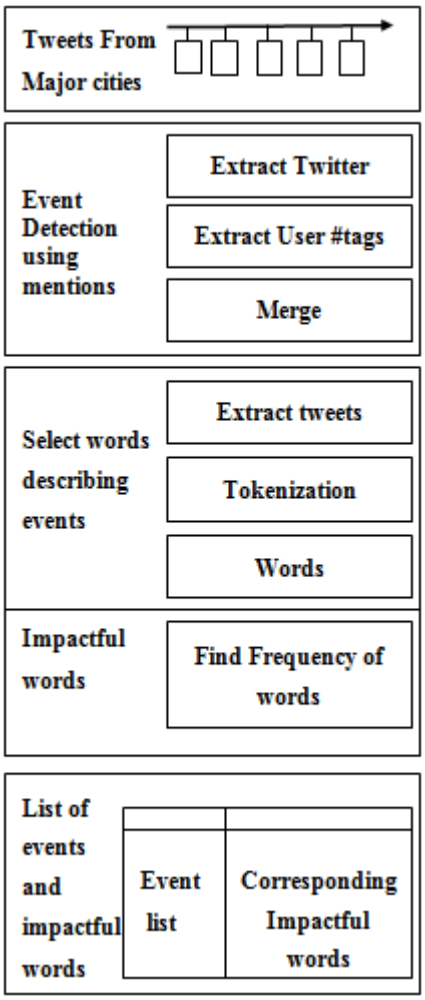

Fig. 1: Flow diagram of proposed work

This method is prepared work on Indian tweets in an English language. Here, considered the 4 major metro cities in India – Delhi, Bangalore, Chennai, and Mumbai. Locate these cities based on latitude and longitude of the cities. Even we can consider other cities and countries by changing corresponding latitude and longitude of those cities and every city in the world is identified by Twitter using id i.e. Woeid's.

The proposed method has two phases and depends on three components as follows

i. Detection of the events based on the mentions (hash tags)
ii. Selection of the word which describe the event
iii. List of the events with impactful words

First we will collect the tweets from different cities in particular country as input to the method. In the first component again there are three sub components. The first sub component extracts the Twitter hash tags (trending topics or event) from the collected tweets. The second sub component extracts the user hash tags from the collected tweets and the third sub component is used to merge both of first and second sub components. So finally we will get list of partially defined events and this list is ordered according to the impact of the events.

In the second component again there are three sub components. The first sub component extracts the tweets from each listed hash tags from first component. Tweets related to the particular event are tokenized to produce set of words from second sub component. Allthe stop words are removed during tokenization.

Finally we will get list of the words from second component. Find out the frequency of these words and

select the word which has highest frequency that word will be associated with that particular event.

Final output produces list events and corresponding impactful words.

Frequency of the word (F) is calculated as follows

$$F = \frac{No.of words repated in a token set}{Total no.of words in token set} \quad (1)$$

Algorithm 1: Event Detection in Twitter (EDT)

Step 1: Check for the Twitter connection

If connection is successful

Get an instance of twitter object

else

null.

Step 2: Select city (or cities)

Step 3: Read top N tweets from cities

Step 4: Extract Twitter #tag trends

Step 5: Extract user #tags

Step 6: Make a list of top M #tags

Step 7: Merge #tags

Step 8: Extract top Z tweets

Step 9: Find out the frequency of the words

Step 10: Make a list of events

---

3. Experimental results

4.1 Data collection

In prior works they collecting data (tweets) from Twitter and keep it in corpus (a large and structured set of text). They collect past data that ranges from one period of time to another (may collect one year data). Apply their method on the collected data and find events in that corpus. Now Twitter made policy restrictions on collection of data from twitter to collect millions of tweets we need to pay for that so it is not possible to collect millions tweets for free. To overcome this restriction we have collected tweets directly from the twitter by connecting using access tokens. There is rate limitation on number of calls we make to twitter.

4.2 Parameter Setting

Input parameters are as follows

- **Select the cities to collect tweets from:**
  These are the cities from which the tweets of normal user will be collected. The user generated #tags are discovered from these tweets. Set this parameter from 100 to 400
- **Number of tweets to fetch from each city:**
  This tells how many number of tweets to fetch from each of the cities. Set this parameter from 100 to 400
- **Select cities to collect trends from:**
  This shows from where all we want to collect the twitter generated trends from. Select any or all of four metro cities (Bangalore, Chennai, Delhi, and Mumbai)
- **Number of tweets to fetch from the final #tags:**
  This tells us how many tweets to select from each of the merged hash tag topics. The tokens and impactful words are all selected from these tweets. Set this parameter from 100 to 400
- **Number of items to display:**
  This tells us how many number of news items to be displayed in the final output. Set this parameter from 5 to 25 with 5 differences
- **Number of impactful words to display:**
  This tells us how many of impactful words to be displayed for each of the news items which were discovered in the above step. Set this parameter from 5 to 25 with 5 difference

4.3 Stop words

After pre-processing, stop words are considered and used as filter. Stop word is defined as word which has no meaning and not relevant. In this method totally 512 words I have considered as stop words. 75% of these words are most frequent in collected tweets (or data). If the word is identified as stop word in the token set then it is immediately removed from the consideration as probable trending topic. And also the word is considered as stop word if it matches the following criteria.

- If the word appeared in over 250 of the 512
- If the word has total frequency of at least all through 512
- If the word is preposition or conjunction
- If the word is derivative of the 500 or more.

The experiment is evaluated using precision, F-measure (Frequency measure), Duplication event rate (DE-Rate) and also considered running time of our method over two runs. Precision is state or quality being precise, F-Measure is the harmonic mean of the precision and recall. Here we considering the events that are identified by the experiment in comparison with events identified by the Twitter. All these measures requirecalculation of true positives- This means that items identified as events by both experimental and twitter. Precision is measured using false positive- the items identified as events by experimental method but not by Twitter. Recall is measured using false negative- the items identified asevents by Twitter but not by the experimental method.Recall is only considered to calculate the F-measure. DE- Rate indicates the percentage of the duplicate events among the detected events. In this method DE-Rate is zero because it take cares of duplicated event while putting each detected events in event list at final will check whether the event is exist in list or not if the event is not exist in the list, add that event into list if it is existed in the list, discard that event.

We used Precision, F-measure and DE-rate to measure the performance of the system

4.1 Precisionis defined as:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

4.2 Recall is defined as:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

4.3 F-Measureis defined as

$$\text{F-Measure} = 2.\frac{P.R}{P+R} \quad (4)$$

Where, TP= True positive

FP= False positive
FN= False negative

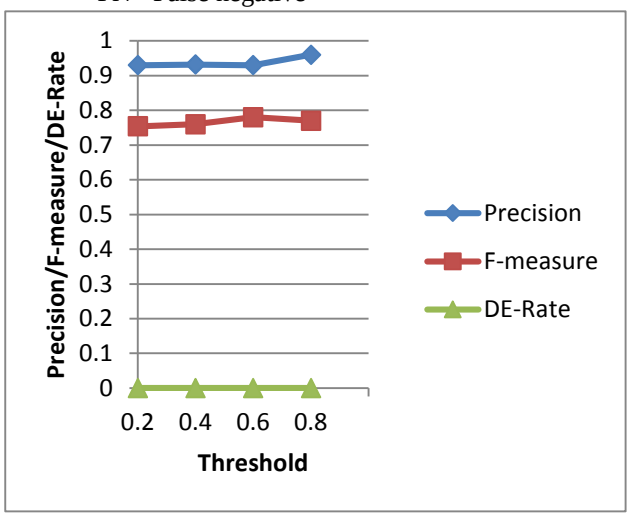

Fig 4.1: Precision, F-Measure, DE-Rate for different values of merging threshold

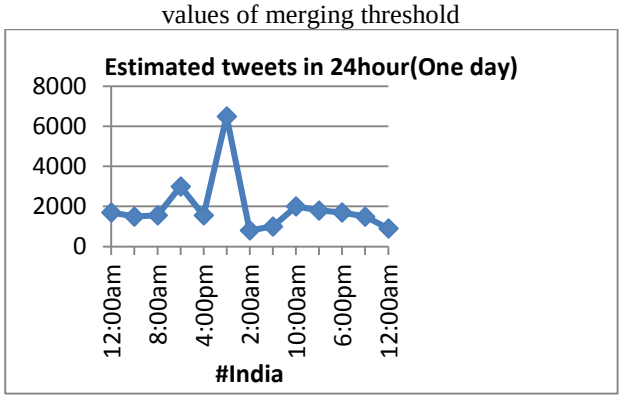

Fig 4.2: Tweets estimated for #India in 24-hour

Fig 4.2 shows estimated tweets from trending topic India in duration of24-hour i.e. one day

Table 4.1: Event detection results

Parameters	Values
Precision	95 %
F-Measure	73.7%
DE-Rate	0%
Running Time (over two runs)	22sec

4. Conclusion

Twitter is valuable source of information but the Twitter is overloaded by lot of information so it is difficult to identify the relevant information related to the particular interest. We have outlined an efficient method EDT- Event detection using mentioning practice of the user to detect the events in Twitter. We have considered hash tag mentions used by the user and considered both user generated hash tags and Twitter generated hash tags. Prior methods are work on already collected data i.e. some time back. In contrast to prior works EDT focuses on current trends and predicts what happening live. It gathers data from both Twitter own analytics (Trending topics) and also by collecting user tweets. Calculate our own trends, thus getting best of both and also it focuses on the way the topics are created in Twitter and then we will select word which best describes the event based on the frequency of mention i.e. number of occurrences in the token set. The word which has high frequency is selected as associated word for the detected events. Hash tags are from core of the Twitter thus we get better results. **Acknowledgements:** The author would like to thank Pushpa Patil for her helpful suggestions in preparing this paper.

References

[1] J. Kleinberg, "Bursty and hierarchical structure in streams," in KDD, 2002, pp. 91-101

[2] L. AlSumait, D. Barbar´a, and C. Domeniconi, "On line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking," in ICDM '08, 2008, pp. 3–12.

[3] J. Benhardus and J. Kalita, "streaming trend detection in twitter," IJWBC, vol. 9 no. 1, 2010, pp. 122-139.

[4] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in WSDM, 2011, pp. 177– 186.

[5] J. Weng and B. S. Lee, "Event detection in Twitter," in ICWSM, 2011, pp. 401-408.

[6] J. H. Lau, N. Collier, and T. Baldwin, "On-line trend analysis with topic models: #twitter trends detection topic model online," in COLING, 2012, pp. 1519–1534.

[7] H. Yuheng, J. Ajita, D. S. Dor´ee, and W. Fei, "What were the tweets about? topical associations between public events and twitter feeds," in ICWSM, 2012, pp. 154–161.

[8] C. Li, A. Sun, and A. Datta, "Twevent: Segment based event detection from Tweets," in CIKM, 2012, pp. 155-164.

[9] R. Parik and K. Karlapalem, "Et: events form tweets," in companinon WWW, 2013, pp. 613-62

[10] A. Guille and C. Favre "Mention-anomaly-based Event Detection and Tracking in Twitter", in ASONAM, 2014. pp.1-8.

Author Profile

Chaitra M. received B.E degree in ISE from Tontadarya College of Engineering, Gadag. Pursuing M.Tech in CSE in BLDEA College of Engineering and Technology, Vijaypur