IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

Online Identification of Real World Events from Twitter

Prof. R. M. Shete

Assistant Professor, Computer Science & Engineering Department, DMIETR, Wardha, Maharashtra, India

Abstract: Recently the size and popularity of social media is rising. System which extracts the useful information is growingly needed. Social media is using for sharing reviews on real time events by people. So, Public using twitter for posting the massages which are related to the events takes place in real time. Descriptive and collective keywords are used of characterizing the stories and events. Here, the purpose is to distinguish between the real-world events and non event posts. This approach is based on aggregate statistics of topical similar post cluster.

Keywords: Identification of events, Twitter, Features extraction, Clustering, Real world event.

I. INTRODUCTION

For reviling innermost secrets and breaking news, internet has become a major source and communication channel. Therefore, web is nothing but a mirror of and part of the real world. Recently some social networking sites are gaining much following and attention. One of the famous of them is “Twitter”. In the recent years, twitter the micro blogging service has become very popular for expressing broadcasting news, opinions and simply connecting with the friends. More than 6 million people are using the twitter for sharing their reviews and also for stay connected to their friends, family and colleague. Maximum size of one tweet is 140 characters. A short message on twitter posted by people can reflect on real time events. Therefore, this user generated data on online media is useful identifying the real world events.

In proposed work, the focus is on the detection of current topic on social media content. Events and its related posts will be identified by clustering techniques where the similar tweets will group together. Then the feature extraction will be applied for classifying clusters according to events. Then these features will be used to identify whether it is an event cluster and non-event cluster. The effectiveness of the techniques used will be validating over millions of messages.

When some broad major event happen, three factors are the main contributors to the rapidly spread of information materialized in exchanged messages between users of a social service network. 1) the ubiquity nature of today’s social network services, they are available nowadays by any internet connected device like a personal computer or a smart-phone; 2) the ease of use and agility of entering or forward information is also a key factor that lead some messages to be spread very fast on the network and go viral and 3) the life span of the messages is also an interesting feature of those social network services. The assumption made is that each online social user is regarded as a sensor and each message as sensory information. It is pointed that two major substantial differences of the social text stream data over general text stream data: 1) social text stream data contains rich social connections (between the information senders/authors and recipients/reviewers) and temporal attributes of each text piece; and 2) the content of text piece in the social text stream data is more context sensitive.

Access of structured information which is seen in databases and unstructured information which is seen in documents or unstructured fields of texts both benefits to information processing applications. So, access of such texts information, also having the benefits of linguistic analysis of text, as contrary of shallower “word basis” analysis. As there are lots of methods, techniques that can be tested on natural language texts, its effect in the amount of search in the natural language processing fields.

II. RELATED WORK

M. Cataldi, C. Schifanella and L. Di Caro [4] proposed two measures, term frequency to calculate nutrition for each word and a page rank measure. After that Bursty keywords are obtained using nutrition trend. Then by using graph based approach for bursty keywords generates the topic boundary. Sayyadi, Maykov and Hurst used graph approach in which clustering of keywords is done by matching pairs. They used community detection algorithm in which made a graph whose nodes are clustered. Also the topic extraction is carried out by identifying document with similar term. Lehmann, Kleinberg and Backstorm [5] have used the graph for short phrases. Phrases are connected by edges.

One of the method modelled called Latent Dirichlet Allocation (LDA), the idea of knowing the most breaking news by calculating the bursty terms in document [6].

This avoids the other topics by capturing the high peak. So first find bursty term then cluster them for event detection. In some graph based approach, the first step is to tag the terms, then group it and then find the interest in social media [7].

DOI 10.17148/IJARCCE.2016.5527

III. SEPARATION OF EVENT AND NON-EVENT CONTENT

Here, the purpose is to address event detection problem using an online filtering and clustering framework and then explained types of features which are extract for clusters.

1. Classification and Clustering Framework
The incremental is elected to use. To cluster effectively, online clustering algorithm is used which cluster a stream of twitter messages effectively in real time. For that there is a need to choose a scalable clustering algorithm and there is not required the number of cluster knowledge as over time the user generated data and new real time events get added to the stream. On the basis of these observations, we can propose use threshold parameter incremental clustering algorithm. This kind of algorithm will consider each message in turn and identify the proper clustering assignment using similarity of twitter messages to given clusters.

To detect each events cluster in the streams, here the variety of different revealing features using the statistics of twitter cluster messages can be calculated.

2. Event Features by Cluster-Level
Generally the computation of twitter message clusters takes place to know characteristics that may help identify clusters which are associated with events. Here explained different broad categories of features that explore several aspects of the clusters. Below, specifically considered topical, temporal, Twitter-centric and social features.

A. Temporal Features: The volume of posts or twitter messages for a particular event at the particular time describes unique characteristics. It means relation between twitter messages and time is important to consider in calculations. Temporal behavior is the key to effectively detect events in our framework. Here the set of temporal features will be collect to characterize the volume of frequent cluster terms as per the time.

B. Social Features: To get the communication message and interaction, social features is to be design. This interaction behavior might be different like events, non-event messages and other twitter centric activities. User interactions on twitter can be ‘tweets’, ‘retweets’, ‘replies’, and mentions.

C. Topical Features: In topical features the topical coherence of cluster is described. It takes place on the basis of hypothesis that clusters of event tend to revolve around a central topic and not the non-event clusters. As non-event clusters usually centers less terms that don’t reflect a one theme. Posts in event clusters generally share more terms. It means possibility is more in event clusters to appearance of more terms.

D. Twitter-centric Features: Here, the goal of proposed work is to identify event and non-event data. Here to be highlighting the differences between non-event clusters that relates to twitter- centric activities, that to be are a specific class of non-event messages. To address this task, a set of features will be designed that will target occurring of patterns in non-event clusters with twitter-centric behavior which includes presence of multi word hashtags and tag usage.

3. Classification of Event
Then in this, the machine learning concept will be use. As using above features, then train an event classifier by applying machine learning techniques. First identify event clusters. Compute the features of each cluster. Then use the model of classification to predict the probability. Because of large number of data on twitter, there is a possibility that the classifier may label many clusters as events. In searching event scenario, where user find current event information, it is necessary to display the identified event cluster’ selected subset. In that case, the classifier will select the top events as per the probability of belonging to the event class.

IV. CONCLUSION

Here, the end to end approach is presented for detecting the real world event data on twitter. The main aspect is to separate real world event and non-event from user generated tweets on twitter.

REFERENCES

[1] H. Becker, M. Naaman and L. Gravano, “Beyond trending topics: Real-world event identification on Twitter,” in Proc. ICWSM: 5th Int. AAAI Conf. Weblogs and Social Media, 2011.

[2] Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Göker and Ioannis Kompatsiaris, “Sensing Trending Topics in Twitter”, IEEE Transactions on Multimedia, Vol.15, No.6, October 2013.

[3] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer, “Detecting and tracking political abuse in social media,” in Proc. ICSWM: 5th Int. AAAI Conf. Weblogs and Social Media, 2011.

[4] M. Cataldi, L. Di Caro and C. Schifanella, “Emerging topic detection on Twitter based on temporal and social terms evaluation,” in Proc. MDMKDD: 10th Int. Workshop Multimedia Data Mining, New York, NY, USA, 2010, pp. 4:1–4:10, ACM.

[5] J. Leskovec, L. Backstrom, and J. Kleinberg, “Meme-tracking and the dynamics of the news cycle,” in Proc. KDD: 15th ACM Int. Conf. Knowledge Discovery and Data Mining, New York, NY, USA, 2009, pp. 497–506.

[6] D. A. Shamma, L. Kennedy, and E. F. Churchill, “Peaks and persistence: Modeling the shape of microblog conversations,” in Proc. CSCW: ACM Conf. Computer Supported Cooperative Work, New York, NY, USA, 2011, pp. 355–358

[7] S. Papadopoulos, Y. Kompatsiaris, and A. Vakali, “A graph-based clustering scheme for identifying related tags in folksonomies,” in Proc. DaWaK: 12th Int. Conf. Data Warehousing and Knowledge Discovery. Berlin, Germany: Springer-Verlag, 2010, pp. 65–76.

DOI 10.17148/IJARCCE.2016.5527