Silence or Outbreak - a Real-Time Emergent Topic Identification System (RealTIS) for Social Media

Ning Lu[1] **, Zhen Yang**[1]*, **Jian Huang**[2], **Yaxi Wu**[1], **Hesong Wang**[1]

[1]College of Computer Science, Beijing University of Technology, Beijing, China 100124
[2] Central University of Finance and Economics, Beijing, China 100081
yangzhen@bjut.edu.cn

Abstract

This paper presents **RealTIS**, a **Real**-time emergent **T**opic **I**dentification **S**ystem for user-generated content on the web via social networking services such as Twitter, Weibo, and Facebook. Without user intervention, our proposed RealTIS system can efficiently collect necessary social media posts, construct a quality topic summarization from the vast sea of data, and then automatically identify whether the emerging topics will be out-breaking or just fading into silence. RealTIS uses a time-sliding window to compute the statistics about the basic structure (motifs) variation of the propagation network for a specific topic. These statistics are then used to predict unusual shifts in correlations, make early warning and detect outbreak. Besides, this work also illustrates the mechanism by which our proposed system makes early warning happen.

Introduction

It is critical to identify an emergent topic outbreak on social network early and rapidly, both for timely alerting government agencies and the general public and taking effective public intervention measures. For one thing, an emergent topic outbreak on social network tends to reflect the occurrence of major incident in real world. For example, when the tsunami struck the coasts of Asia and East Africa, massive information came forth immediately, and early warning and outbreak detection would provide a great help for government in deploying rescue resources, racing against time to rescue victims, and thus assuring the safety of life and property(Yang, Yao, and Tu 2020) . For another, information explosion in cyber space has great influence on real society. Web popular brand, as a typical example, is becoming increasingly known and hailed with more and more consumers post, comment, and forward the related content.

Being critical, an early and rapid identification of emergent topic outbreak is also a challenging task. There is in fact a consensus among researchers that emergent hot topic event usually experiences a development cycle of three phases: latency, diffusion and recession, which is also confirmed by Xie et al. (Xie et al. 2010) who further divides the period of diffusion into three sub-phases: gemination, acceleration and maturation. In his study on how to extract the sequence of key events in a news story from the raw numbers of tweets and retweets that take place during these events, Chierichetti et al. (Chierichetti et al. 2014) proposed a heart pulse pattern that is confirmed again in Meladianos et al. (Meladianos et al. 2015). While these empirical studies improve the development of topic identification system, it remains an demanding and outstanding task to predict, prior to accelerating phrase, whether a topic is about to experience outbreak or just about to fade into silence. In relation to this task, we've made two contributions in this study as below:

- We find that the critical factor in determining whether a topic is about to outbreak or fade into silence is the formation of key network basic structure in the process of circulation, i.e, motifs structure(Milo et al. 2002) of specific category.
- We develop RealTIS, a real-time emergent topic identification system. RealTIS system can efficiently collect necessary social media posts, construct a quality

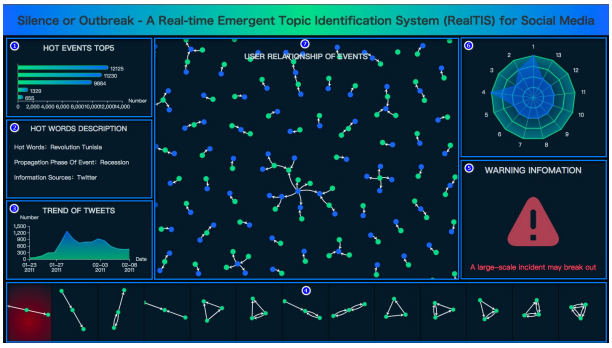
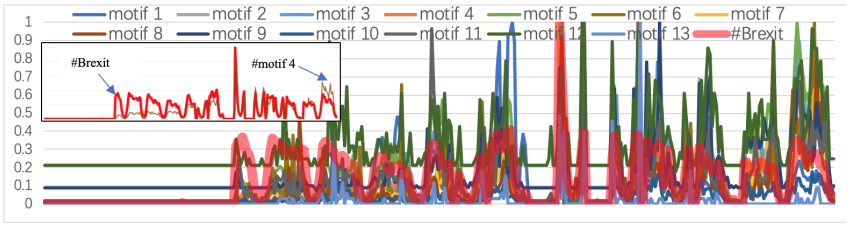
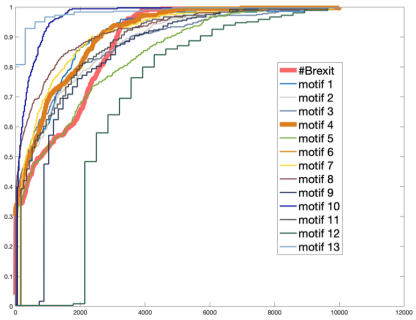


Figure 1: The user interface of our RealTIS system. (1) Identification of hot topic event; (2) Description of key words about the event; (3) Development tendency of event; (4) Motifs;(5) Warning on outbreak of emerging event; (6) Radar statistics graph of motif sub-graph real-time quantity; (7) Liaison network of current event users.

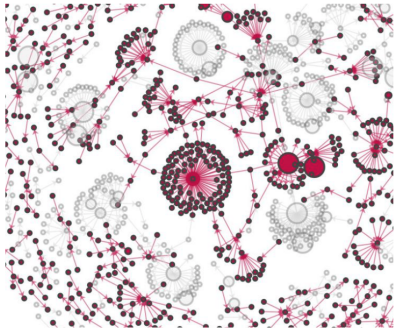
*Corresponding author: Zhen Yang, yangzhen@bjut.edu.cn, Beijing University of Technology, http: //www.dmslab.net.



(a) Normalized number of tweets with hashtag #Brexit vs. Normalized number of motifs in the social networks made up of users who at lease post one tweet with hashtag #Brexit, from January 20, 2019 to August 30, 2019.



(b) Cumulative Distribution of normalized number of tweets with hashtag #Brexit vs. Cumulative Distribution of normalized number of motifs.



(c) The users belong to motif 4 highlighted in red in a sample social networks made up of users who at lease post one tweet with hashtag #Brexit.

Figure 2: The RealTIS system consists of two parts: (a) Back end: sub-system of detecting emergent event; (b) Front end: sub-system of warning about event outbreak.

topic summarization from the vast sea of data, and then automatically identify whether the newly emerging topics will be breaking out or fading into silence based on the motifs structure of network, degree information and quantities of tweets. The user interface of our RealTIS system is shown in Figure 1.

System Design and Results Analysis

RealTIS is made up of two major components:

- **Detection technology of emergent event.** Identifying events in real time on Twitter is a challenge, due to the heterogeneity and immense scale of the data. To address this challenge, we've developed an AP incremental clustering algorithm to detect and track topic development in an incremental way while incorporating new tweets regarding the topic at the same time(Frey and Dueck 2007).
- **Warning technology of event outbreak.** We find that the outbreak of an event is always related to the formation of several specific types of basic structure (motifs). We thus use a time-sliding window to compute the statistics about the basic structure (motifs) variation of the propagation network for a specific topic. These statistics are then used to identify unusual shifts in correlations, make early warning and detect outbreak.

To illustrate and evaluate our methods, we built a tweet collection corpora which contains about 4 millions of tweets from January 20, 2019 to August 30, 2019. As an example, our proposed RealTIS, first of all initializes the topic of $\#Brexit$ and traces the development of the topic using hashtag labels in an incremental way each and every new cycle (usually 1 day). Figure 2(a) illustrates the normalized number of tweets with hashtag $\#Brexit$ vs. normalized number of motifs in the social networks made up of users who at least posted one tweet with hashtag $\#Brexit$. To explore which motifs best predict the development of topic $\#Brexit$, as shown in Figure 2(b), we compared the cumulative distribution of normalized number of tweets with hashtag $\#Brexit$ vs. cumulative distribution of normalized number of motifs. Using KL and Euclidean distance to measure the similarity of cumulative distribution, it is found that the variation of motif 4 is a good indicator to tracking the topic $\#Brexit$. In Figure 2(c), the users belong to motif 4 highlighted in red in a sample social networks made up of users who at lease post one tweet with hashtag $\#Brexit$. Similarly, we find that the variation of other network topics is also closely correlated to variation of one or several basic network structures (motifs).

Impact and Future Work

RealTIS works well in blocking the well-targeted rumor attack, for example against a particular celebrity or company. It can also assist governments in observing the impact of hot topic event on twitter and preventing the outbreak of those events which would otherwise negatively affect social stability. In the future, we will build a bigger corpora of twitter events with a complete development process with the view to promoting the performance of our system. Relevant demonstration videos can be found on https://youtu.be/Lr0zqMDKprM.

Acknowledgments

This work was supported by the Industrial Internet Innovation Development Project, the Industrial Technology Basic Public Service Platform Project, and the Great Wall Scholar Program.

References

Chierichetti, F.; Kleinberg, J.; Kumar, R.; Mahdian, M.; and Pandey, S. 2014. Event detection via communication pattern analysis. In *ICWSM 2014*, 51–60.

Frey, B. J.; and Dueck, D. 2007. Clustering by passing messages between data points. *Science*, 315(5814): 972–976.

Meladianos, P.; Nikolentzos, G.; Rousseau, F.; Stavrakas, Y.; and Vazirgiannis, M. 2015. Degeneracy-Based Real-Time Sub-Event Detection in Twitter Stream. In *ICWSM 2015*, 248–257.

Milo, R.; Shen-Orr, S.; Itzkovitz, S.; Kashtan, N.; Chklovskii, D.; and Alon, U. 2002. Network motifs: simple building blocks of complex networks. *Science*, 298(5594): 824–827.

Xie, K.; Zhao, S.; Gang, C.; and Wenjing, C. 2010. Research on lifecycle principle and group decision making of network public sentiment emergency. *Wuhan University of Technology - Social Science Edition*, 23(4): 482–486.

Yang, Z.; Yao, Y.; and Tu, S. 2020. Exploiting Sparse Topics Mining for Temporal Event Summarization. In *2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC)*, 322–331.