A Survey on Event Detection Methods on Various Social Media

Madichetty Sreenivasulu and M. Sridevi

Abstract Today, social media play a very important role in the world to share the real-world information, everyday existence stories, and thoughts through the virtual communities or networks. Different types of social media's such as Twitter, Blogs, News Archieve, etc., have heterogeneous information with various formats. This information is useful to the real time events such as disasters, power outage, traffic, etc. Analyzing and understanding such information on different social media are a challenging task due to the presence of noisy data, unrelated data, and data with different formats. Hence, this paper focuses on various event detection methods in different types of social media and categorizes them according to the media type. Moreover, features and data sets of various social media are also explained in this paper.

Keywords Social media · Twitter · Event detection · User generated content
YouTube · News archive

1 Introduction

Nowadays, social media play a very important role in the world. Social media can be used for sharing information's, ideas, and interests among the people throughout the world. Lot of people communicated through social media by sending and receiving messages. Social media can be different types such as Twitter, news archive, multimedia, BlogPosts, web pages, Facebook, etc. Social media generate huge amount of information it will be very much useful to the people. The information can be described as something happened outside is known as event. Social media contains

M. Sreenivasulu (✉) · M. Sridevi
Department of Computer Science & Engineering, National Institute of Technology,
Tiruchirappalli, Tiruchirappalli, Tamil Nadu, India
e-mail: sreea568@gmail.com

M. Sridevi
e-mail: msridevi@nitt.edu
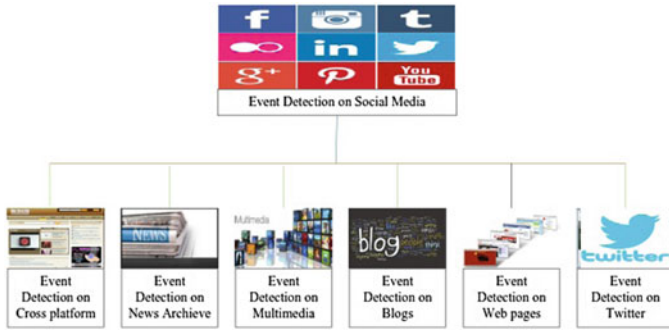
Fig. 1   Event detection on different types of social media

information about related to be event and not related to event. During a crisis, colossal information can be posted in social media. In that, determining the information related to event is challenging task on social media. Different types of social media have the different characteristics. Figure 1 describes about event detection on social media.

Whenever social event occurs, traditional media take 2 h or even days for reporting the news of a particular event, at the same time corresponding information may begin to spread promptly in the social media like Twitter. Twitter is a free social networking microblogging service which allows users for registering and post short texts (messages) on microblog called tweets. Each tweet allows 140 characters which may be text images, hyperlink, videos, etc. As per the statistics given in [1], average of 41 million users have been registered in July 2009 which was drastically increased to 317 million users, in August 2016. Whenever the user follows any another user, user is able to read the tweets and share the same tweet, called retweet. A user has the ability to read the tweet which is related to other users. In twitter, the user-generated contents are updated each and every time which will be useful to the users in the twitter [2]. During disasters, large colossal information can be generated in the microblog. At that time, it is very important to know where the disasters occur and where can be disseminated? And there is a need to know what is happening? in those locations. Extraction of the event is a challenging task. By entering pertinent word to search engine (Google, Bing, Yahoo! etc) which gives very large roll of URL which are related to some concepts of interests. At the present, researchers mainly focus on ranking of the list of documents based on the weight [3].

Users easily access the news, web, videos by leverage the search engines and video sharing websites such as Google, YouTube, Baidu and YouKu. Newswires like CNN, BBC, and CCTV also publish news videos. According to the 2014 YouTube data, for every minute, 100 h of videos are uploaded to YouTube, and each month six billion hours of videos are watched on the YouTube. These factualities describe a new challenge for the users to clench the great events available from searching video databases [4]. In general, belief of the users is initially user-generated content (UGC) and got huge popularity and then gradually dim into oblivion. Whenever

analyzing more than 350k YouTube videos, it needs at least 1 year time for uploading videos to get huge popularity. These videos are called sleeping beauties. Sleeping beauties will get huge popularity after one year which is challenging a problem. Identifying such videos will not only comfort the advertisers, but also the designers of recommendation systems who search for to maximum user satisfaction [5].

Due to ubiquitous of the We-media, information can be received and published into numerous forms at anywhere at any time with leverage the Internet.The rich cross-media information contains multimodal data in multiple media and has many audiences. The large information available on different events which have great impact compare to single media information. Hence, detection of the event in cross-media is a challenging problem in the following aspects (1) It contains the different media data which has different characteristics (2) Topics are presented among the noisy web data. Detection of the event in cross-media is very helpful for the government organization and advertising agencies [6].

Today, there is huge growth in web technology, social websites have become the flexible platform for users to access the world and exchanging the ideas (opinions). However, prodigious explosion in the extensiveness of web pages from the social media has made the strenuous problem for the users to access the hot topics and web administrators to detect the web activities properly. The content of social media is less restricted and less predictable compared to others (news articles) and also it is short, noisy and sparse [7].

2   Literature Survey

First story detection (FSD) or new event detection can be defined as identifying the event which is unseen or unknown previously. In [8], the author proposed a method which is called Inverted Index for FSD, which is efficient and it is helpful for large data streams. This approach was used by the Umass FSD system. The limitation of the proposed method is that it is not scalable to unbounded streams and processing of one document will take longer time. Hence, it will not provide good performance for Twitter due to continuous tweets, it is clear as some document takes long time for processing each document than arrival of tweets. In [9], the author proposed a method called probabilistic approaches for FSD. The main advantage of this method is that it presents the data in structured way but it is more expensive and also very less suitable for processing large amount of data.

In [10], the author proposed a method for event detection in blogs called temporal random indexing. Blogs are less structured compared to newswire and it can be updated very fast. The main limitation of this method is that the events based around keywords need to specify the keywords explicitly and require lot of volume for semantic shift. It is not suitable for a large-scale event detection.

Support Vector Machine (SVM) for event detection in the Twitter during disasters is used in [11]. It detects specific types of events and it is more attractive for small sets of events with high precision. It is unfeasible for any type of events. If large number

of events are there then it requires separate classifier which needs to be trained for each event and also requires labeled data.

In [12], the author proposed a hybrid method to detect the events from the noisy and fragmented tweet. In this paper, the author considered power outage as event and power outage detection as an event detection. Hybrid method is a combination of heterogeneous information network and supervised topic model. Heterogeneous information network contains both temporal and spatial information of tweets. Supervised topic model used heterogeneous information network for improving the accuracy of event detection compared to existing methods like SVM, Bayesian model, etc.

A method was developed for extracting the traffic events like traffic, public transport, water supply, weather, sewage, and public safety from the tweets which are posted by the observations [17]. In general, for detecting the events, there is a need to understand the tweet stream which is a challenging task in social media. In this paper, hypothesize the problem of annotating the tweet stream as a sequence labeling problem. For the sequence labeling models, there is a need for manual labeled training data. The proposed method is a novel approach for creating the automatic training label based on the instance-level domain knowledge and it indicates locations in a city and possible event occurrences. It gives comparable performance to annotated tasks which is more advantageous for training and avoids the need of the manual labeled training process.

A novel framework is proposed in [4] to better the group associated web videos to events. First, the data preprocessing is performed to select features and tag relevance learning. Next, multiple correspondence analysis is applied to explore the correlations between terms and events with the assistance of visual information. Co-occurrence and visual near duplicate feature trajectory induced from near duplicate key frames (NDKs) are combined to calculate the similarity between NDKs and events. Finally, a probabilistic model is proposed for news web video event mining, where both visual temporal information and textual distribution information are integrated.

In [7], a method is explored in the view of similarity diffusion. The method uses clustering-like pattern across similarity cascades (SCs). SCs are chain of subgraphs generated by pruning the similarity graphs with a set of thresholds and capturing the topics with the help of the maximum cliques. At last, they discovered the real topics in effective manner from huge number of candidates with the use of topic-restricted similarity diffusion method. This method is experimented on three public datasets such as MCG-WEBV, YKS and Social Event Detection 2012 (SED2012). In this experiment, they considered only social event as web topics and ignored the geographical information. It gives better performance. The summary of the methods discussed for event detection on various social media is tabulated in Table 1.

Table 1 Summaries of event detection on various social media

Sl. No.	Author and reference	Media	Method	Features	Datasets	Advantages
1	Wayne Xin Zhao et al. [13]	News articles	Burst VSM	Bursty features	68 millon web pages	• Effective for large corpus of news archieve
2	Thorsten Brants et al. [14]	News stories	IncrementalTF-IDF model	Hellinger distance document similarity normalization	TDT2, TDT3 andTDT4 corpora	• Gives good results in news stories. • Reduces the cost
3	Ting Hua et al. [15]	Twitter	Targeted-domain spatio-temporal event detection (ATSED)	Twitter location information	July 2012 to May 2013 of both twitter and GSR data	• Automatic assignment of labels to the tweets
4	Maia Zaharieva et al. [16]	MultiMedia	An unsupervised, multistaged approach	Time and location information	306, 159 Flickr photos and 969 YouTube videos	• More robust grouping of media for different platforms
5	Lingyang Chu et al. [6]	MultiMedia	Multimodality fusion framework and a topic recovery (TR)	Multimodality graph	MCG-WEBV and YouKu Sina dataset (YKS)	• Very flexible for different platforms
6	Haifeng Sun et al. [12]	Twitter	Supervised topic model with heterogeneous information	Temporal, spatial and information	10,000 labeled tweets related to power outage detection	• locate events without any additional measurements
7	Pramod Anantharam et al. [17]	Twitter	Automatic labeling to training data	Locations of city and event term	4 months tweets of San Francisco Bay area	• Uses best classification model • Suitable for traffic agents
8	Wei Wang et al. [18]	News articles	A multiple instance learning based on CNN	Labels at the document level	GSR 19795 news article related to event and 3759 articles (not related)	• Works very fast and feasible for news articles

3   Conclusion

The enormous amount of user-generated content (UGC) can be disseminated in social media. Analyzing and understanding the UGC are very helpful to the users at the time of disasters, power outage, traffic, etc. Event detection is one of the important tasks and targets at identifying real-world occurrences. However, event detection on social media must be efficient and accurate. In this paper, survey of different methods is explained based on event detection with various types of social media. In future, A new approach for event detection in social media will be proposed to improve the accuracy and speed.

References

1. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World wide web, pp. 591–600. ACM (2010)
2. Pervin, N., Fang, F., Datta, A., Dutta, K., Vandermeer, D.: Fast, scalable, and context-sensitive detection of trending topics in microblog post streams. ACM Trans. Manag. Inf. Syst. (TMIS) **3**(4), 19 (2013)
3. Mori, M., Miura, T., Shioya, I.: Topic detection and tracking for news web pages. In: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 338–342. IEEE Computer Society (2006)
4. Zhang, C., Xiao, W., Shyu, M.-L., Peng, Q.: Integration of visual temporal information and textual distribution information for news web video event mining. IEEE Trans. Hum.-Mach. Syst. **46**(1), 124–135 (2016)
5. Sikdar, S., Chaudhary, A., Kumar, S., Ganguly, N., Chakraborty, A., Kumar, G., Patil, A., Mukherjee, A.: Identifying and characterizing sleeping beauties on youtube. In: Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion, pp. 405–408. ACM (2016)
6. Chu, L., Zhang, Y., Li, G., Wang, S., Zhang, W., Huang, Q.: Effective multimodality fusion framework for cross-media topic detection. IEEE Trans. Circuits Syst. Video Technol. **26**(3), 556–569 (2016)
7. Pang, J., Jia, F., Zhang, C., Zhang, W., Huang, Q., Yin, B.: Unsupervised web topic detection using a ranked clustering-like pattern across similarity cascades. IEEE Trans. Multimed. **17**(6), 843–853 (2015)
8. Allan, J., Lavrenko, V., Malin, D., Swan, R.: Detections, bounds, and timelines: Umass and tdt-3. In: Proceedings of Topic Detection and Tracking Workshop, pp. 167–174 (2000)
9. Ahmed, A., Ho, Q., Eisenstein, J., Xing, E., Smola, A.J., Teo, C.H.: Unified analysis of streaming news. In: Proceedings of the 20th International Conference on World Wide Web, pp. 267–276. ACM (2011)
10. Jurgens, D., Stevens, K.: Event detection in blogs using temporal random indexing. In: Proceedings of the Workshop on Events in Emerging Text Types, pp. 9–16. Association for Computational Linguistics (2009)
11. Sakaki, T., Okazaki, M., Matsuo, Y.: Tweet analysis for real-time event detection and earthquake reporting system development. IEEE Trans. Knowl. Data Eng. **25**(4), 919–931 (2013)
12. Sun, H., Wang, Z., Wang, J., Huang, Z., Carrington, N., Liao, J.: Data-driven power outage detection by social sensors. IEEE Trans. Smart Grid **7**(5), 2516–2524 (2016)
13. Zhao, W.X., Chen, R., Fan, K., Yan, H., Li, X.: A novel burst-based text representation model for scalable event detection. In: Proceedings of the 50th Annual Meeting of the Association for

Computational Linguistics: Short Papers-Volume 2, pp. 43–47. Association for Computational Linguistics (2012)

14. Brants, T., Chen, F., Farahat, A.: A system for new event detection. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 330–337. ACM (2003)

15. Hua, T., Chen, F., Zhao, L., Chang-Tien, L., Ramakrishnan, N.: Automatic targeted-domain spatiotemporal event detection in twitter. GeoInformatica **20**(4), 765–795 (2016)

16. Zaharieva, M., Del Fabro, M., Zeppelzauer, M.: Cross-platform social event detection. IEEE MultiMed. (2015)

17. Anantharam, P., Barnaghi, P., Thirunarayan, K., Sheth, A.: Extracting city traffic events from social streams. ACM Trans. Intell. Syst. Technol. (TIST) **6**(4), 43 (2015)

18. Wang, W., Ning, Y., Rangwala, H., Ramakrishnan, N.: A multiple instance learning framework for identifying key sentences and detecting events. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 509–518. ACM (2016)