# Event Detection in Social Media Data

Wenwen Dou, Xiaoyu Wang, William Ribarsky, Michelle Zhou*
University of North Carolina at Charlotte, IBM Almaden Research Center*

**ABSTRACT**

Recent research has shown that a considerable fraction of social media streams are about "events". Collectively, events serve as a succinct summary of social media streams. Individually, event and its sub-events, reveal the evolution of certain social phenomena over time. In addition, analyzing relationships between events and people's responses to the events provides rich information on the masses' opinions towards an event; this may further shed light on the impact of public policies dealing with the events.

Event detection, therefore, is an important and practical task to identify and make sense of the overwhelming amounts of social media data. In this paper, we explore and summarize popular tasks in the domain of event detection. More specifically, we present four tasks: New Event Detection, Event Tracking, Event Summarization, and Event Association. We use these four tasks to illustrate main purposes for performing event detection in social media space, and further present their application domains.

**Keywords**: event detection, social media, visualization.

**Index Terms**: H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

## 1 INTRODUCTION

Social media provides a great source of information. The online conversations have undergone tremendous growth over the past few years. Some of the conversations are personal status updates, usually more relevant to a user's social circle. While a large portion of the conversations in the social media space are instead responses triggered by events. Such events include natural disasters (e.g. hurricanes, earthquakes), political events (e.g. presidential elections), protests and marches, etc. Take Occupy Wall Street as an example, the OWS movement is a widely participated event known to use social media to advertise and spread nationwide. The movement is long lasting and widespread without central leadership, which creates challenges in understanding and responding to the movement. Given the critical role of social media in the OWS movement, it is a great source to understand and analyze such events.

An event is commonly considered as an occurrence at a specific time and place [1, 2, 3, 4, 5, 6, 7]. However, in the social media space, certain social campaign/movements do not necessarily happen in a physical location. We revise the definition of an event in the context of social media as [20]:

"*An occurrence causing change in the volume of text data that discusses the associated topic at a specific time. This occurrence is characterized by topic and time, and often associated with entities such as people and location.*"

Recent research has demonstrated that one of the common uses of social media is reporting and discussing events users are experiencing: Sakaki et al. [11] showed that mining of relevant tweets can be used to detect earthquake events and predict the earthquake center in real time. Becker et al. [12] proposed to identify real-world events through exploring a variety of techniques for learning multi-feature similarity metrics for social media documents. Their evaluation results showed that events could be effectively detected from large-scale images provided by Flickr.

Although numerous research papers have focused on presenting methods and systems for extracting event-related information from social media and newswire, few has reviewed the proposed systems from a task-specific perspective.

In this paper, we focus on the task of identifying major events and sub-events when analyzing social media data. The timeline of events serves as a succinct summary of the massive social media space. Through further analysis of the responses to each individual event over time, one can gain knowledge regarding the event itself, people's opinions towards the event, as well as inferring causal relationships between events (sub-events) and responses.

## 2 TASKS

We surveyed related work from several research communities including Information Retrieval, Human Computer Interaction, and Information Visualization. We categorize four tasks that are highly relevant to event detection. In the following subsections, we describe each task and provide examples drawn from previous work.

### 2.1 New Event Detection (NED) (What's new)

"*New Event Detection refers to* ***identifying the first story*** *on topics of interest through constantly monitor news streams.*"
In the TDT (Topic Detection and Tracking) community, NED is characterized as "queryless information retrieval" for the times when analysts do not know what to look for (i.e. detecting the unexpected [14]). NED is an automated process that makes binary event identification on whether a document discusses a new topic that has not been reported before [3]. Therefore, such NED systems are very powerful in situations where novel information needs to be ferreted out from a mass of rapidly growing data, such as in the domain of financial marketing, news analyses, intelligence gathering, natural disaster etc. For example, a good NED system would be one that correctly identifies the article that reports the earthquake's occurrence as the first story [1]. And subsequently, it will help identify the additional coverage on death toll and rescue efforts.

Specifically, NED consists of two subtasks: Retrospective NED and Online NED [8]. The former subtask discovers previously unidentified events in an accumulated collection, while the latter aims at identifying new events from live text streams in real time.

#### 2.1.1 Retrospective NED

Retrospective NED utilizes comparison metrics to detect new events. This process picks documents reporting new events by comparing them with all the stories that have arrived in the past. Metrics, such as cosine similarity, Hellinger similarity, KL divergence etc. are widely used in this NED process to determining how closely related two documents are [1].

* wdou1@uncc.edu

While majority Retrospective NED follows a set of common models and algorithms, many variations of the document representation, similarity metrics and clustering algorithms are also proposed in the literature [4, 5, 6]. Specifically to social media streams, Sayyadi et al. [7] have developed a Retrospective NED algorithm that creates a keyword graph and used community detection methods to discover and describe events.

### 2.1.2 Online NED

When a document comes in, the Online NED system compares it with all previous events and computes a pair-wise similarity score in real time. During this process, single-pass clustering is widely used to process incoming news stories one-by-one to determine whether a new event has occurred [8]. Such clustering methods will generate a similarity score that will be further used to classify new events from the story sources. If the score exceeds a certain threshold, the Online NED system will mark the document as a new event; otherwise the document is labeled as old and merged into the corresponding topics.

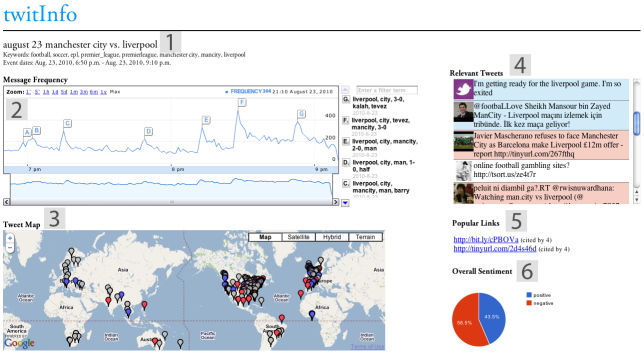
Online NED is shown to be powerful in detecting events in newswire (less update frequency), when applied to microblogs (e.g. Tweets), such process faces additional challenges from the frequent update of the massive amount of fragmented documents. Specifically, such problems include a much higher volume of data as well as its noise. To detect new events from a stream of Twitter posts, Petrovic et al. [13] presented an algorithm based on locality-sensitive hashing to deal with the large number of tweets generated every second. Their results showed that the proposed algorithm achieved significant speedup in processing time while maintaining competitive performance comparing to traditional NED approaches.

## 2.2 Event Tracking

*"Event tracking task refers to the study of how events unfold."*

Once a new event (e.g., an emergency or humanitarian crisis) has been detected, the logical next step is for people to track the development of such event. Tracking how an event and responses to the event evolve over time is related to the problem of information diffusion, especially on how information regarding a specific event spread in social networks. Studying how hashtags (represents ideas and sometimes events) spread within a Twitter user network, Romero et al. [22] found significant variation in the ways that widely used hashtags on different topics spread. In particular, hashtags on politically controversial topics are particularly persistent while natural analogues of Twitter idioms and neologisms are particularly non-persistent. To model contagions, Centola et al. [23] designed a model to assess the impact of an isolated event on the social networks. Complex contagions depend primarily on the width (the number of ties) of the bridges across a network, not just their length (the geodesic a bridge spans). Leskovec et al. [27] proposed a framework for tracking "memes", which act as signatures of topics and events propagating and diffusing over the web. Their meme tracking approach provided a coherent representation of the news cycle – the daily rythms in the news media. In particular, through analysing 1.6 million mainstream media sites and blogs, a typical lag of 2.5 hours was observed between the peaks of attention to a phrase in the news media and in blogs respectively.

With interests in specific events, recent research has demonstrated how to track an event over time and extract important information to support situational awareness during crisis or inform public policies [17, 18]. Vieweg et al. [17] studied two natural hazards events – the Red River Floods and the Oklahoma Grassfires – through analyzing tweets to identify information that may contribute to situational awareness. Through closely tracking the two events especially on situation updates, the authors demonstrated that relevant information can be extracted




during emergency situations. Tumasjan et al. [18] closely followed the event of German national election and demonstrated that tweets' political sentiment are in close correspondence to the parties' political positions, which in turn indicate that the content of Twitter messages can be used to predict elections. Recent studies have shown that computer meditated communication

Figure 1: The TwitInfo user interface. Area1 denotes the input query on the event of interest. Area2 shows the volume of tweets regarding the event over time. Area3 depict the geolocations of the tweets. Area4 shows a sample of the tweets. Area5 highlights popular links mentioned in the tweets. Area6 presents the aggregated setiment.

involves self-organizing behavior produces accurate results often in advance of official communications [15, 16].

## 2.3 Event Summarization (What happened)

"*Event Summarization is about creating a summary of events based on bursty features identified from a text corpus*."

Event Summarization (ES) is a popular and challenging task that several research communities have attempted to address. Research has been conducted to achieve this task through both content-based (e.g., unstructured text analysis) and structured methods (e.g. meta-information).

Specifically, for content-based modeling efforts, Kleinberg [25] extracted bursty features from text streams through modeling the text streams using an infinite-state automaton through the use of machine learning and statistical methods. To group bursty features into events, Fung et al. [9] further proposed a feature-pivot clustering approach that also identifies hot periods for bursty events. To summarize public health related events through analyzing tweets, Paul and Dredze [10] used the Ailment Topic Aspect Model to detect events such as flu outbreaks. Their results showed a faster detection of flu outbreak using tweets as opposed to CDC's measure (percentage of specimens test positive for influenza).

In addition to summarizing events based on modeling the content of social media data, recent work has also taken advantage of meta-information such as geolocations, named entities, and sentiment. Marcus et al. [21] proposed a system – TwitInfo – for visualizing and summarizing events on Twitter. TwitInfo allows users to browse a large collection of tweets using a timeline-based display that highlights peaks of high tweet activity. Users can further drill down to subevents, and explore via geolocation, sentiment, and popular URLs. One limitation of TwitInfo is that it requires users to specify what event they want to explore, such as the soccer event seen in figure 1. The peaks of Twitter activity,

which denote subevents, are detected purely based on volume of tweets regardless of the content.

To address this limitation, Dou et al. [20] developed LeadLine, which summarizes and visualizes events based on the 4Ws (who, what, when, where) commonly used investigative journalism.

Assuming events has been detected from social media data, to summarize highly-structured and recurring events such as sports, Chakrabarti et al. further extracted a few tweets that best describe the chain of interesting occurrences in each event. They formalized the problem of summarizing events and provided a solution based on learning the underlying hidden state representation of each event via Hidden Markov Models.

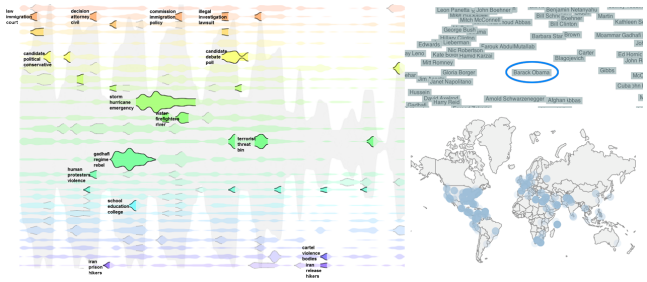


Figure 2: The LeadLine user interface. Left view presents information regarding what and when for each event. Top right presents who and their relationships. Bottom right displays information regarding where events occurred.

### 2.4 Event Associations

"*Event association focuses on analyzing the relationship between events. Such analysis could reflect the impact of a certain event on other events.*"

The majority body of work in TDT has been focusing on how to detect topics and novel events [24]. News stories are usually organized into a flat hierarchical structure in which news stories discussing the same topics (or events) are grouped into clusters. But over the recent years, research on event association has been booming given the need for ***exploring complex evolution relationships between events***. To generalize event episodes across different events of the same nature, Wei et al [2] proposed an event evolution pattern discovery technique that identifies event episodes together with their temporal relationships that occur frequently in a collection of events of the same type. Yang et al. [3] constructed an event evolution graph to present the structure of events for efficient browsing and extracting of information. As opposed to previous approaches that organize events in a hierarchical fashion by topics and time, Yang et al constructed event evolution graphs based on event timestamp, event content similarity, temporal proximity, and document distribution proximity.

To investigate dependencies - generally causal - among events within a news topic, Nallapati et al. [19] proposed the concept of *event threading*. It refers to the process of recognizing events and identifying dependencies among them. Modeling dependencies among events could help users explore and navigator through events faster. More recently, Shahaf and Guestrin [26] proposed a novel way to navigate within a collection of news articles and discover hidden connections among stories. More specifically, the proposed system could automatically identify a coherent chain linking two given news articles. For example, the system can recover the chain of events starting with the decline of home prices (Jan 2007), and ending with the more recent health-care debate.

## 3 Conclusion

In this paper, we summarized four tasks related to analyzing events from social media data. The four tasks include new event detection, event tracking, event summarization, and event association. The four tasks enable organization of existing methods and systems based on a task-specific perspective. Such organization could further help collectively evaluate the methods for addressing each task and potentially identify the missing pieces to practically tackle individual tasks when analyzing social media data.

## References


[1] Giridhar Kumaran , James Allan , Andrew Mccallum. Classification Models for New Event Detection. *Applied Optics*, 15:2513–2519, August 1980.

[2] Chih-Ping Wei; Yu-Hsiu Chang. "Discovering Event Evolution Patterns From Document Sequences," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* , vol.37, no.2, pp.273-283, March 2007.

[3] Christopher C. Yang, Xiaodong Shi, and Chih-Ping Wei. 2009. Discovering event evolution graphs from news corpora. *Trans. Sys. Man Cyber. Part A* 39, 4 (July 2009), 850-863.

[4] Yiming Yang, Jian Zhang, Jaime Carbonell, and Chun Jin. 2002. Topic-conditioned novelty detection. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '02). ACM, New York, NY, USA, 688-693.

[5] Giridhar Kumaran and James Allan. 2004. Text classification and named entities for new event detection. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '04). ACM, New York, NY, USA, 297-304.

[6] Thorsten Brants, Francine Chen, and Ayman Farahat. 2003. A System for new event detection. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (SIGIR '03). ACM, New York, NY, USA, 330-337.

[7] Hassan Sayyadi, Matthew Hurst, and Alexey Maykov. Event detection and tracking in social streams. In Proceedings of International Conference on Weblogs and Social Media (ICWSM), 2009.

[8] Rui-Feng Xu; Wei-Hua Peng; Jun Xu; Xiao Long; , "On-line new event detection using time window strategy," *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on* , vol.4, no., pp.1932-1937, 10-13 July 2011.

[9] Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Philip S. Yu, and Hongjun Lu. 2005. Parameter free bursty events detection in text streams. In *Proceedings of the 31st international conference on Very large data bases* (VLDB '05).

[10] Michael J. Paul and Mark Dredze. You are what you tweet : Analyzing Twitter for public health. In Proceedings of International Conference on Weblogs and Social Media (ICWSM), 2011.

[11] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (WWW '10). ACM, New York, NY, USA, 851-860.

[12] Hila Becker, Mor Naaman, and Luis Gravano. 2010. Learning similarity metrics for event identification in social media. In *Proceedings of the third ACM international conference on Web search and data mining* (WSDM '10). ACM, New York, NY, USA, 291-300.

[13] Sasa Petrovic, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to Twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*

(HLT '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 181-189.

[14] Juha Makkonen, Helena Ahonen-Myka, and Marko Salmenkivi. 2003. Topic detection and tracking with spatio-temporal evidence. In *Proceedings of the 25th European conference on IR research* (ECIR'03), Fabrizio Sebastiani (Ed.). Springer-Verlag, Berlin, Heidelberg, 251-265.

[15] Leysia Palen, Sarah Vieweg, Sophia B. Liu, and Amanda Lee Hughes. 2009. Crisis in a Networked World. *Soc. Sci. Comput. Rev.* 27, 4 (November 2009), 467-480.

[16] Sutton, J, Palen, L, and Shklovski, I. (2008). Backchannels on the front lines: emergent uses of social media in the 2007 Southern California Wildfires. In *Proceedings of the 5th International ISCRAM Conference* (Wasington, DC, May 2008). 1-9.

[17] Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the 28th international conference on Human factors in computing systems* (CHI '10). ACM, New York, NY, USA, 1079-1088.

[18] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pages 178–185, 2010.

[19] Ramesh Nallapati, Ao Feng, Fuchun Peng, and James Allan. 2004. Event threading within news topics. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management* (CIKM '04). ACM, New York, NY, USA, 446-453.

[20] Wenwen Dou, Xiaoyu Wang, Drew Skau, William Ribarsky and Michelle X. Zhou. LeadLine: Interactive Visual Analysis of Text Data through Event Identification and Exploration. To appear. IEEE Conference on Visual Analytics Science and Technology, 2012.

[21] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. 2011. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 annual conference on Human factors in computing systems* (CHI '11). ACM, New York, NY, USA, 227-236.

[22] Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web* (WWW '11). ACM, New York, NY, USA, 695-704.

[23] Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. American Journal of Sociology, 113(3):702–734, November 2007.

[24] James Allan, editor. *Topic detection and tracking: event-based information organization*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[25] Jon Kleinberg. 2002. Bursty and hierarchical structure in streams. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '02). ACM, New York, NY, USA, 91-101.

[26] Dafna Shahaf and Carlos Guestrin. 2010. Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '10). ACM, New York, NY, USA, 623-632.

[27] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '09). ACM, New York, NY, USA, 497-506.