How Events Unfold: Spatiotemporal Mining in Social Media

Ting Hua[1,*] Liang Zhao [1] Feng Chen[2] Chang-Tien Lu[1]
Naren Ramakrishnan[1]
1 Department of Computer Science, Virginia Tech
2 Department of Computer Science, University at Albany-SUNY
∗ E-mail: tingh88@vt.edu

Abstract

There has been significant recent interest in the application of social media analytics for spatiotemporal event mining. However, no structured survey exists to capture developments in this space. This paper seeks to fill this void by reviewing recent research trends. Three branches of research are summarized here—corresponding (resp.) to modeling the past, present, and future—information tracking and backward analysis, spatiotemporal event detection, and spatiotemporal event forecasting. Each branch is illustrated with examples, challenges, and accomplishments.

1 Introduction

With rapid developments in modern geo-tagged communication, especially social media, spatial computing is now engendering a revolution in the modeling and understanding of human behavior. The rise of "big data" (e.g., via channels like Twitter, Facebook, Youtube) has given a new window into studying events across the globe. It has become possible to aggregate public data to capture triggers underlying events, detect on-going trends, and forecast future happenings. Concomitantly there has been a rapid development of new computational methods for spatiotemporal mining of social media datasets.

This paper structures recent research into three directions:

- *The Past*, i.e., information tracking and backward analysis. Social media data has grown enormously over the years (in 2013, more than 400 million tweets are estimated posted by millions of users [1]). Such voluminous dynamics provides an interesting opportunity to track information on targeted topics and analyze triggers underlying them. It is also possible to capture diversity in information flows: for instance, a user can obtain information from a friend's tweets in his/her social network or obtain the necessary information externally (e.g., TV or news media). Studies of the interaction across multiple data sources provide richer contextual information into information flows.

- *The Present*, i.e., spatiotemporal event detection. It is believed that news breaks earlier in social media than in traditional media [18]. The aim of event detection is to identify ongoing events from social media data before their reporting in mainstream news outlets. With real-time data streams as input, event detection models can output spatiotemporal summaries for on-going events, including information about event occurrence time, the locations involved, and a textual event summary.

- *The Future*, i.e., spatiotemporal event forecasting. Twitter data has been shown powerful in forecasting [20] that some tweets may contain context indicating future events. For example, tweet "TRAFFIC ALERT: Rt. 20 closed due to a wreck" provided evidence of future hazard along roadways. In addition, compared to traditional documents, social media is endowed with multiple features, such as time stamps, geo-tags, and an underyling network. Utilization of these multiple features and indicating information make forecasting spatiotemporal events before their occurrence possible.

[1] https://blog.twitter.com/2013/celebrating-twitter7

2 The Past: Information Tracking and Backward Analysis

This section introduces the capture of developments involving an event, evaluating its trustworthiness [25, 27], and how we can identify the underlying event triggers [9]. Figure 1 illustrates the evolution of a civil unrest event in Mexico. This event occurred in January 2013 and involves a civil unrest event in Mexico. News media initially reported that (human) bodies were found in an suburban area on Jan 3 and Jan 4, but this event received little or no attention in social media. Next, the government captured some dogs as suspects. This event was first discussed in the news on Jan 7. A hashtag specifically denoting the event named "YoSoyCan26" was created on Jan 8, and soon spread rapidly among Twitter users. Tweets using this hashtag predominantly called for the release of the captured dogs. In the following day, news media began reporting on the chatter underyling the popular trending Twitter topic "YoSoyCan26". This online trending topic triggered a real world protest event on Jan 9. As can be seen from the development underlying the "dog protests", Twitter users could be influenced by news reports, and conversely, news media also can be influenced by information from Twitter. Topics in Twitter cover almost all themes from daily life happenings to breaking news. How to harvest the large volume and fast dynamics underyling social media to track events and analyze triggers are key questions of interest.

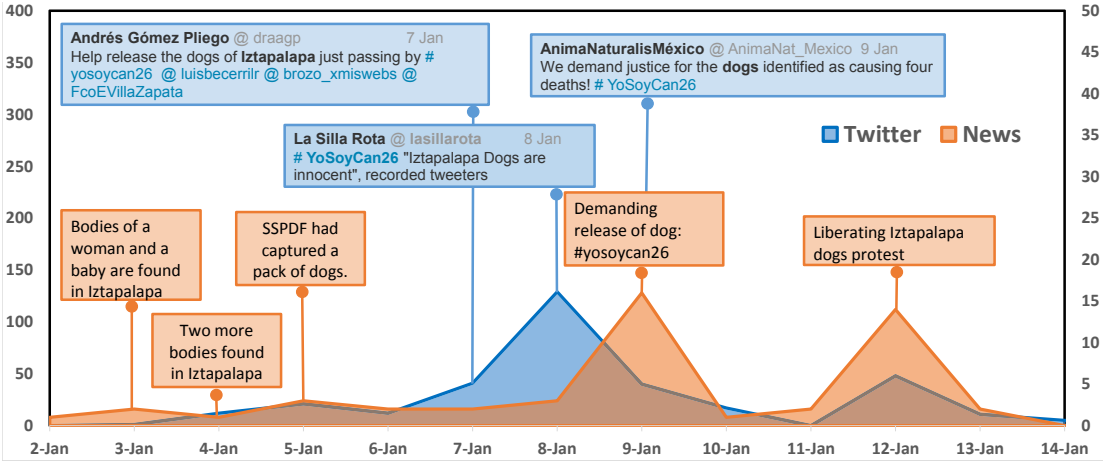

Figure 1: Evolution of "dog protest" in Mexico.

In contrast to traditional plain text, tweets often contain features such as hashtags, replying, and friendships, and harnessing such features can be crucial to dealing with the massive dynamics of social media data. There has been significant work on general theme tracking [2, 23], where themes are typically represented as mixtures of latent topics. Research on targeted theme tracking most adopt a supervised classification framework for theme extraction [16, 17]. Lin et al. use predefined keywords as query terms, and this approach requires significant human effort and can introduce bias [14, 15]. Backward analysis for identifying triggers and key players of interest is a relatively new research direction, and builds upon developments in time series modeling [12] and temporal topic mining [21].

2.1 Tracking themes in Twitter

Zhao et al. [25] suggest a framework that can track themes of targeted domain dynamically. This framework jointly models textual content and social network simultaneously. Specifically, it first builds a heterogeneous graph that includes multiple types of theme entities such as tweets, terms, and users. The connections between nodes are computed through co-occurrence, authorship, and replying relationships, and the theme snapshots are arranged on time-ordered sub-collections. The underlying parameters are estimated by minimizing the Kullback-Leibler divergence between inferred themes and ground truth data. This methodology is effective as well as efficient. By using heterogeneous relationships from Twitter, this method makes full use of Twitter textual and structural features. Meanwhile, the model also enjoys linear scalability which is ensured by conditional independence relationships among entities.

Evaluating the trustworthiness of themes is of equal importance as tracking them. Most previous work focused on evaluating the credibility of general tweets [22, 4, 6], while Zhao et al. made the first attempt at evaluating topic-focused tweets [27]. Topic-focused tweets are first identified through a text classifier that is trained through Twitter labels, and then trustworthiness of users as well as their posts are updated by an iterative propagation algorithm. Specifically, the

trustworthiness is evaluated through multiple aspects, such as trustworthiness of Twitter texts, authorship, and underlying social graph.

2.2 Identifying Triggers

Hua et al. [9] present a study that analyzes the root causes of civil unrest through tweets. Tweets related to specific protests provide insights into the root causes, i.e., who the organizers are, and how online expression reflects or contributes to such events. In addition, the causes of social events can also be viewed through the analysis of interactions between social media and traditional news streams, which support a variety of applications, including: understanding the underlying factors that drive the evolution of data sources, tracking the triggers behind events, and discovering emerging trends.

Hua et al. also recently proposed a hierarchical Bayesian model that jointly models news and social media topics and their interactions [10]. This model jointly considers news data and tweet data in an asymmetrical frame. Such structure can significantly improve modelling performance for short texts (tweets), without loss of accuracy in long documents (news). Besides, the output of this model enables a variety of applications to understand the complex interaction between news and social media data such as: checking the topical coverage of different data sources, capturing the influencers from topic to topic, identifying key documents and key players. Some interesting conclusions can be drawn from their experimental results. For instance, news topics are generally more influential than Twitter topics; topic occurring first in one data source but growing popular in another data source might be a source of triggers.

3 The Present: Spatiotemporal Event Detection

This section introduces semi-supervised and unsupervised methods that can be used for automatical detection of targeted events. First, social media is known to be a more responsive medium than traditional news outlets [18]. Early detection with social media data is therefore of great practical use. Second, tweets not only contain plain texts, but also include spatiotemporal information such as geo-tags and time-stamps. Figure 2 shows an example spatiotemporal event detection on July 14, 2012 in Mexico. Red nodes in the figure represent event relevant tweets (a protest against president Pena Nieto) that were published on that day and can be positioned by the longitude and latitude according to their geo-tags. These tweets are mainly distributed within two clusters, corresponding to two metropolitan centers: Mexico city and Monterrey city. The aim of early event detection in social media data is to identify these spatiotemporal clusters and summarize tweets into events. Numerous previous studies have focused on detecting events from formal documents such as news articles or emails [11, 5]. However, data in Twitter streams are heavily informal, ungrammatical, and dynamic so that traditional methods cannot be applied to the mining of such noisy data.

Figure 2: Example of spatiotemporal event detection (Mexico, July, 2012).

Most existing targeted-domain event detection algorithms adopted in social media analytics are supervised methods. These methods [17, 13] usually first train a classifier to recognize tweets from the targeted domain, and then apply techniques such as Kalman filtering to detect locations from these tweets as event occurrence locations. However, the building

and maintenance of high quality labeled data requires extensive human efforts. Although most classifiers in existing work are sensitive to the features, there is no accepted methodology for feature selection. Therefore, semi-supervised and unsupervised methods are in great need.

3.1 Unsupervised event detection

Zhao et al. made an attempt to detect events of user-specified interest in an unsupervised way [24], requiring no pre-given labeled data. Starting from seed words, this model acts like a search engine that can retrieve the relevant information automatically. Specifically, given a targeted domain, the algorithm expands the seed words to domain-related terms via a tweet homogeneous graph. The expansion of seed words with co-occuring terms can be viewed as a process of knowledge acquisition, while traditional supervised methods are limited in this respect. This process is iteratively repeated so that key terms are exhaustively extracted and then weighted in each iteration. A graph induced using this expanded vocabulary is then subject to a local modularity spatial scan (LMSS) to capture both semantic similarities and geographical proximities by jointly maximizing local modularity and spatial scan statistics. This event detection framework can be the foundation to more sophisticated models, e.g., tracking the evolution of targeted themes [25].

3.2 Semi-supervised event detection

A semi-supervised method is another solution for automatic event detection. Hua et al. proposed a model that can learn pseudo-labels (from news) for Twitter data [8] and therefore save costs in human labeling without significant loss in accuracy. Specifically, this model first transfers labels from newspapers to tweets through a novel ranking algorithm, and further expands the initial label subspace by an EM inference algorithm. The noisy nature of Twitter data is a new challenge for text classification. To address these challenges, a customized text classifier for Twitter analysis is provided to combine tweets into mini-clusters by social ties. To make maximum usage of all Twitter geographical information, this model also extends spatial scan statistics with multinomial distributions, and thus factors from various location-related items (e.g., user-profile locations or geo-tags) can be considered together to improve geo-coding performance.

4 The Future: Spatiotemporal Event Forecasting

The development of spatiotemporal events usually contain several different stages. Figure 3 shows an example of influenza outbreak in November 2014. Taking Louisiana state as an example, this state was "healthy" in week 45 as its inuenza-like illness (ILI) activity level was minimal (green), became "lightly infected" in week 46 (low ILI activity level), and ended with becoming "seriously infected" in week 47 (high ILI activity level). It is clear that the evolution of a spatiotemporal event is not only impacted by its current stage but also influenced by its geographical and temporal neighborhood.

Figure 3: Influenza outbreak on Week 47 ending Nov 22,2014 in southern region.

Unlike traditional plain text, social media data is multi-dimensional, including spatial feature "geo-tags", temporal feature "publish date", textual feature "content", and influence feature "friend relationships". How to utilize social media data for forecasting is an active research topic of current interest. However, dynamic patterns of features (keywords) and the geographic heterogeneity of social media data bring critical challenges. Some studies utilize regression or SVM models to predict the occurrence of future events. Their difference lies in the features they used, where some adopted tweet volume or sentiment scores [1, 3, 7] while others may use more informative features such as semantic topics. Most of them ignored the geographical information which is in fact one of the most important features of an event. Beyond these simple solutions, here we introduce two novel methods that can forecast spatial events through social media data.

4.1 Forecasting through HMM

With tweet streams as input, the model introduced by Zhao et al. [26] can forecast spatiotemporal events (e.g., the one shown in Figure 3) involving multiple stages. This model is built by modeling the evolution of events, which can therefore predict the events at multiple stages. Some Twitter forecasting models only focus on prediction of temporal pattern [19], while this work jointly modeled the structural contexts, geo-locations, and time in one frame. The spatial information is harnessed through assignments of geographical priors, the accurate sequence likelihood is estimated through dynamic programming, and historical geographical information is used here for the prediction of new event location. It also enables the understanding for the relationship of inside and outside event venue under the tweet observations as it is built to model the evolutionary development of events.

4.2 Forecasting with multi-task learning

Supervised methods (e.g., LASSO) are of great use in identifying static features such as keywords, and unsupervised models (e.g., DQE) are suitable to handle dynamical features. Tweet streams are known to contain both sets of features. Zhao et al. proposed a novel multi-task learning framework that can concurrently address both the statical and dynamical features [28]. Specifically, given locations (e.g., cities) as input, the proposed model is able to forecast events for all locations simultaneously. One secret of its success is that this model can extract and utilize shared information among locations and therefore effectively increases the sample size for each individual location. The other advantage is that the model considers both the static features from a predefined vocabulary (made by domain experts) and dynamic features from DQE [24] in one multi-task feature learning framework. Different strategies are used to control the common set of features and thus balance the homogeneity and diversity between static and dynamic terms.

5 Conclusion

Rapid developments in social media bring new opportunities for the spatial computing community. This paper reviews the most popular research branches of social media event mining, including theme tracking and backward analysis, on-going event detection, and future event forecasting. Tracking and backward analysis is a relatively new research branch and will continue to attract more attention. The key issues in this area may include: identifying information of users' interests, correlating multiple data sources, and evaluating the influences between users/topics/events. Compared to forecasting and backward analysis, on-going event detection of targeted domain is relatively well developed. Unsupervised and semi-supervised methodologies that require less cost in human effort should be widely used in practical applications soon. Prediction via social media is the most popular research direction but far from well studied. The most challenging problems in this topic are how to model event evolution and how to extend existing forecasting models to social media data.

6 Acknowledgement

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract D12PC00337. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US government.

References

[1] Marta Arias, Argimiro Arratia, and Ramon Xuriguera. Forecasting with twitter data. In *ACM Transactions on Intelligent Systems and Technology (TIST)*, volume 5, page 8. ACM, 2013.

[2] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[3] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. In *Booktitle of Computational Science*, volume 2, pages 1–8. Elsevier, 2011.

[4] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.

[5] Gabriel Pui Cheong Fung, Jeffrey Xu Yu, Philip S Yu, and Hongjun Lu. Parameter free bursty events detection in text streams. In *Proceedings of the 31st international conference on Very large data bases*, pages 181–192. VLDB Endowment, 2005.

[6] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. Tweetcred: Real-time credibility assessment of content on twitter. In *Social Informatics*, pages 228–243. Springer, 2014.

[7] Jingrui He, Wei Shen, Phani Divakaruni, Laura Wynter, and Rick Lawrence. Improving traffic prediction with tweet semantics. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1387–1393. AAAI Press, 2013.

[8] Ting Hua, Feng Chen, Liang Zhao, Chang-Tien Lu, and Naren Ramakrishnan. Sted: semi-supervised targeted-interest event detectionin in twitter. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1466–1469. ACM, 2013.

[9] Ting Hua, Chang-Tien Lu, Naren Ramakrishnan, Feng Chen, Jaime Arredondo, David Mares, and Kristen Summers. Analyzing civil unrest through social media. In *Computer*, number 12, pages 80–84. IEEE, 2013.

[10] Ting Hua, Ning Yue, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. Topical analysis of interactions between news and social media. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2016.

[11] Giridhar Kumaran and James Allan. 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*, pages 297–304. ACM, 2004.

[12] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM, 2009.

[13] Rui Li, Kin Hou Lei, Ravi Khadiwala, and Kevin Chen-Chuan Chang. Tedas: A twitter-based event detection and analysis system. In *Proceedings of the 28th International Conference on*, pages 1273–1276. IEEE, 2012.

[14] Cindy Xide Lin, Qiaozhu Mei, Jiawei Han, Yunliang Jiang, and Marina Danilevsky. The joint inference of topic diffusion and evolution in social communities. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 378–387. IEEE, 2011.

[15] Cindy Xide Lin, Bo Zhao, Qiaozhu Mei, and Jiawei Han. Pet: a statistical model for popular events tracking in social communities. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 929–938. ACM, 2010.

[16] Jimmy Lin, Rion Snow, and William Morgan. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 422–429. ACM, 2011.

[17] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, pages 851–860. ACM, 2010.

[18] Zeynep Tufekci and Christopher Wilson. Social media and the decision to participate in political protest: Observations from tahrir square. In *Booktitle of Communication*, volume 62, pages 363–379. Wiley Online Library, 2012.

[19] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *ICWSM*, volume 10, pages 178–185, 2010.

[20] Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. Automatic crime prediction using events extracted from twitter posts. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 231–238. Springer, 2012.

[21] Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM, 2006.

[22] Wouter Weerkamp and Maarten de Rijke. Credibility-inspired ranking for blog post retrieval. In *Information Retrieval*, volume 15, pages 243–277. Springer, 2012.

[23] Xintian Yang, Amol Ghoting, Yiye Ruan, and Srinivasan Parthasarathy. A framework for summarizing and analyzing twitter feeds. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 370–378. ACM, 2012.

[24] Liang Zhao, Feng Chen, Jing Dai, Ting Hua, Chang-Tien Lu, and Naren Ramakrishnan. Unsupervised spatial event detection in targeted domains with applications to civil unrest modeling. In *PLOS ONE*, volume 9, 2014.

[25] Liang Zhao, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. Dynamic theme tracking in twitter. In *Proceedings of the 3rd IEEE International Conference on Big Data*, pages 561–570, 2015.

[26] Liang Zhao, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. Spatiotemporal event forecasting in social media. In *Proceedings of the 15th SIAM International Conference on Data Mining*, pages 963–971, 2015.

[27] Liang Zhao, Ting Hua, Chang-Tien Lu, and Ray Chen. A topic-focused trust model for twitter. In *Computer Communications*. Elsevier, 2015.

[28] Liang Zhao, Qian Sun, Jieping Ye, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. Multi-task learning for spatio-temporal event forecasting. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1503–1512, 2015.