A Bayesian Graphical Model to Discover Latent Events from Twitter

Wei Wei weiwei@cs.cmu.edu
Kenneth Joseph kjoseph@cs.cmu.edu
Wei Lo spencer w lo@zju.edu.cn
Kathleen M. Carley kathleen.carley@cs.cmu.edu

Introductions. Online social networks like Twitter and Facebook produce an overwhelming amount of information every day. However, research suggests that much of this content focuses on a reasonably sized set of ongoing events or topics that are both temporally and geographically situated. These patterns are especially observable when the data that is generated contains geospatial information, usually generated by a location-enabled device such as a smartphone. In this paper, we consider a data set of 1.4 million geo-tagged tweets from a country during a large social movement, where social events and demonstrations occurred frequently. We use a probabilistic graphical model to discover these events within the data in a way that informs us of their spatial, temporal and topical focus. Quantitative analysis suggests that the streaming algorithm proposed in the paper uncovers both well-known events and lesser-known but important events that occurred within the timeframe of the dataset. In addition, the model can be used to predict the location and time of texts that do not have these pieces of information, which accounts for the much of the data on the web.

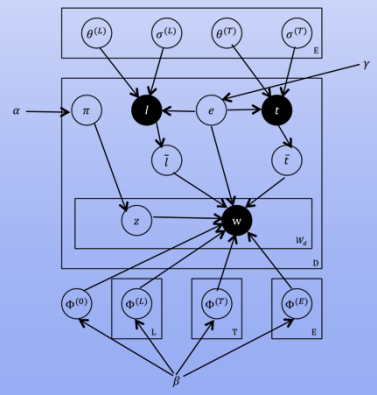

Figure 1 Graphical representation of the Probabilistic model

Methodology. We define event to be a combination of latent distributions over time, location and text. Tweets associated with a specific event are drawn from the corresponding distributions that belong to this event. We use graphical model illustrated in Figure 1 to characterize the relationships between tweet and events. Here t, l and w are the time of the tweet, location of the tweet and the actual words of the tweet. Time and location are drawn from the corresponding Gaussian distributions with parameters θT, σT, θT and σT while the text w are drawn from a Multinomial distribution characterized by parameter ϕ. Each such distributions have a different parameter setting for a different event. The goal of the learning is to discover the event related parameters based on twitter data.

E1	jan25	arrested	Egypt	Ghonim
	burn	injustice	Libya	tortured
E2	guilt	minimum	death	hurts
	Arif	home	pulse	lord of
E3	scar	pharmacist	disease	immediately
	eye	urticaria	evil	transplantation
E4	live	promise	tireless	condensed
	need	granulate	thanks	traipse
E5	end	voice	winter	lord, thou
	god	I want	lord	to god

Figure 4 Top 8 words associated with each of the 5 events

Case Study. We experimented our method on a twitter data set collected over the country of Egypt from Oct, 2011 to Nov, 2013. To illustrate the model we picked representative 5 events discovered among a larger set of 100 events. Figure 2 is a spatial visualization over the country of Egypt with the contour plot illustrating the density of the location distribution of each event. Figure 3 is the corresponding temporal distributions and Figure 4 being the lexical distributions. Based on Figure 4, we recognized that the first event is about the initial Egypt revolution demonstration that happened in Carlo, Egypt on Jan 15 of 2011. We see that both spatial, temporal and lexical distributions matched what we found about the event on Wikipedia.

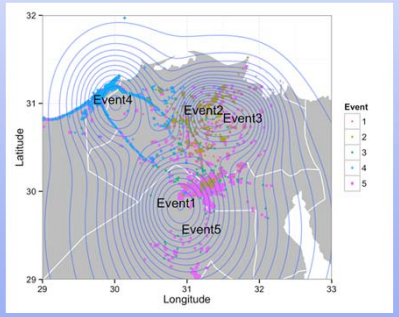

Figure 2 Spatial distribution of 5 events

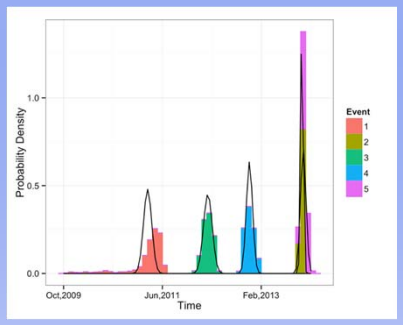

Figure 3 Temporal distribution of 5 events

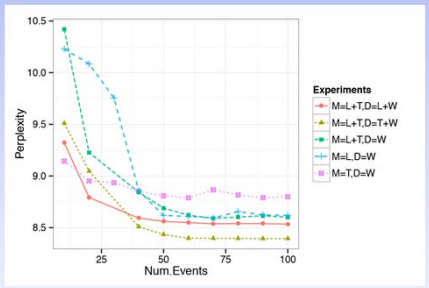

Figure 5 Perplexity over the number of events

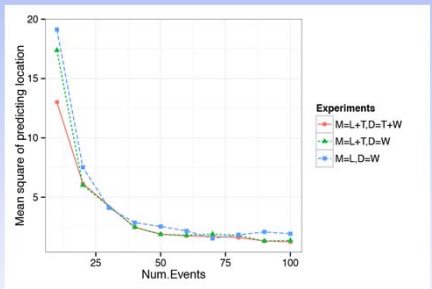

Figure 6 Prediction error of time over the number of events

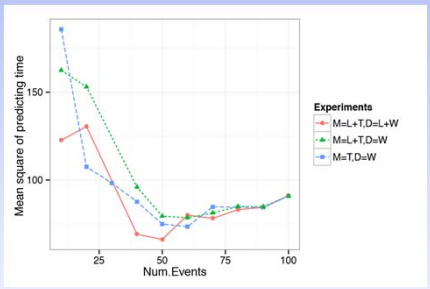

Figure 7 Prediction error of location over the number of events

Prediction Results. We also conducted large scale experiments to predict some information in the tweets. For example, we can predict the text of a tweets based on time and location, which might give us some hint of the events. We can also predict the time/location of the tweets based on other information based on our event model. Figure 5,6 and 7 are the mean square error of prediction of text, time and location over the number of events on different models. Here we see that with the increase of number of events, the prediction error decreases dramatically. We can also see that the full Bayesian model with al the components always perform better than other alternative models.

This work was supported in part by the Office of Naval Research (ONR) through a MURI N000140811186 on adversarial reasoning, MINERVA N000141310835 and by the Center for Computational Analysis of Social and Organization Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research or the U.S. government.