A Nonparametric Method for Early Detection of Trending Topics

Stanislav Nikolov*,† and Devavrat Shah†
{snikolov,devavrat}@mit.edu
†Department of EECS, Massachusetts Institute of Technology
*Twitter, Inc.

Abstract

Online social networks can be used as networks of human sensors to detect important events [3] — from a global breaking news story to an incident down the street. It is important to be able to detect such events as early as possible. To do so, we propose a nonparametric method that predicts *trending topics* on Twitter by comparing a recent activity signal for a topic to a large collection of historical activity signals for trending and non-trending topics. We posit that the signals observed for each class of topics were generated by an unknown set of *latent source* signals for that class according to a stochastic model depending on the *distance* between the observation and its latent source, and propose a class estimator based on this model. Using our method, we are able to detect trending topics in advance of Twitter 79% of the time, with a mean early advantage of 1 hour and 26 minutes, while maintaining a true positive rate of 95% and a false positive rate of 4%. In addition, our method allows for tradeoffs between error types and relative detection time, scales to large amounts of data, and provides a broadly applicable framework for nonparametric classification.

Empirical Observations

On Twitter, users can post short, public messages known as *Tweets*. There are over 400 million Tweets written every day, many of which can be considered *about* one or more topics. For example, this tweet by one of the authors (Twitter handle @snikolov) *"Stuyvesant High School Taps 'Stuy Mafia' at Google, Foursquare to Enhance Computer Science Program via @Betabeat http://betabeat.com..."* is about "Stuyvesant High School", "Computer Science", and so on. Because of the public nature of Twitter, topics can spread and gain popularity. Topics that gain sudden widespread popularity start *trending* i.e. they are featured on a list of top ten *trending topics* on Twitter.

Trending topics can typically be detected by a sudden high-magnitude spike in activity over some baseline of activity [2][1]. However, this sudden spike is often preceded by lower magnitude activity that is indicative of the topic's imminent popularity. This suggests that we can detect trending topics earlier by observing this early activity. Thus, we propose to predict whether a topic will become *trending* by comparing recent time series of activity for the topic to historical time series of activity leading up to other topics becoming trending, and historical time series of activity for topics that did not become trending.

We define the *activity signal* for a topic in terms of the rate $\rho[n]$ of Tweets about that topic over time, at time bins $n = 1, \ldots, N_{obs}$. We observe that activity is typically characterized by spikes above a baseline rate, so we further transform the rate to normalize away the baseline ($\rho_b[n] = (\rho[n]/b)^\beta$, $b = \sum_n \rho[n]/N_{obs}$) and emphasize spikes ($\rho_{b,s}[n] = |\rho_b[n] - \rho_b[n-1]|^\alpha$), according to parameters $\alpha \geq 1, \beta \geq 1$ (we used $\alpha = 1.2$, $\beta = 1$). In addition, we convolve the result with a smoothing window to eliminate noise and effectively measure the volume of Tweets in a sliding window ($\rho_{b,s,c}[n] = \sum_{m=n-N_{smooth}+1}^{n} \rho_{b,s}[m]$). Finally, because the spread of topics can reasonably be thought of as a branching process, and branching processes exhibit exponential growth, we measure the volume $\rho_{b,s,c}$ at a logarithmic scale ($\rho_{b,s,c,l}[n] = \log \rho_{b,s,c}[n]$).

Data Model

Activity signals, even within a single class, are incredibly diverse. Rather than training a model to distinguish between the activity signals of trending and non-trending topics, we assume no model structure at all and instead propose the following nonparametric model relating observed activity signals (*observations*) to their class labels. Suppose there are two classes: $+$ (topics that were trending at some point during the period of interest) and $-$ (topics that were never trending during the period of interest). We posit that there are a number of distinct *latent source* signals in each class that account for all observations in that class. Let us call them $\mathbf{t}_1, \ldots, \mathbf{t}_n$ for $+$ and $\mathbf{q}_1, \ldots, \mathbf{q}_\ell$ for $-$. Each observation labeled $+$ is assumed to be a noisy version of one of the latent sources $\mathbf{t}_1, \ldots, \mathbf{t}_n$. Similarly, each observation labeled $-$ is assumed to be a noisy version of one of the latent sources $\mathbf{q}_1, \ldots \mathbf{q}_\ell$. We do not know what the latent source signals are or even how many there are. We only know a stochastic model that relates an observation to its latent source.

Let the observation $\mathbf{s}$ be the most recent N_{obs} samples of an infinite stream $\mathbf{s}_\infty$ of activity. An observation $\mathbf{s}$ is *generated* by a latent source signal $\mathbf{q}$ according to the stochastic model

$$\mathbb{P}(\mathbf{s} \text{ generated by } \mathbf{q}) \propto \exp\left(-\gamma d(\mathbf{s}, \mathbf{q})\right) \tag{1}$$

where d is a symmetric, positive definite, and convex distance function (we used the Euclidean norm) and γ is a scale parameter. To determine whether the observation belongs to $+$ or $-$, we make use of a set of example, or *reference* activity signals $\mathcal{R}_+$ from $+$ and $\mathcal{R}_-$ from $-$. Under our model the observation must belong to $+$ if it has the same latent source as one of the reference signals in

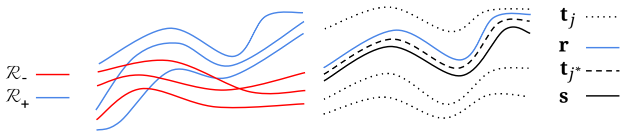

Figure 1: Left: Reference signals from each class. **Right**: Finding the latent source signal $\mathbf{t}_{j^*}$ that minimizes $d(\mathbf{s}, \mathbf{t}_j) + d(\mathbf{r}, \mathbf{t}_j)$, i.e. the latent source signal most likely to have generated both $\mathbf{s}$ and $\mathbf{r}$.

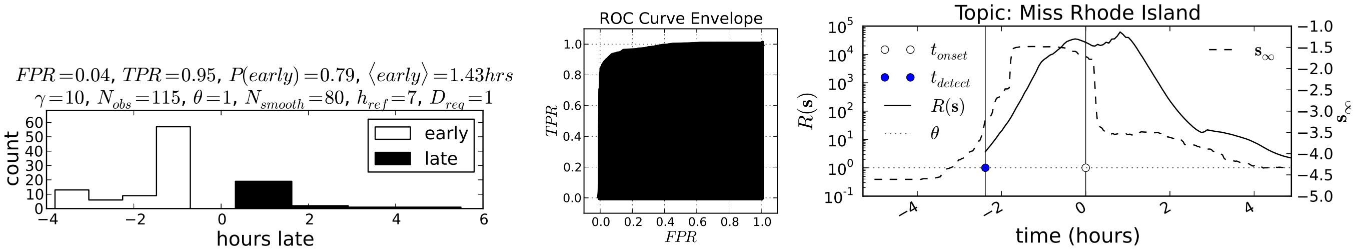

Figure 2: Our method is capable of early detection of trending topics while maintaining a low rate of error and provides the flexibility for tradeoffs between error types.

$\mathcal{R}_+$. Similarly, the observation must belong to $-$ if it has the same latent source as one of the reference signals in $\mathcal{R}_-$. Hence, the probability that the observation belongs to $+$ is

$$\mathbb{P}(+ \mid \mathbf{s}) = \sum_{\mathbf{r}\in\mathcal{R}_+} \mathbb{P}(\mathbf{s} \text{ belongs to } +, \mathbf{s} \text{ shares a latent source with } \mathbf{r}) = \sum_{\mathbf{r}\in\mathcal{R}_+} \sum_{j=1}^{n} \mathbb{P}(\mathbf{s} \text{ generated by } \mathbf{t}_j, \mathbf{r} \text{ generated by } \mathbf{t}_j)$$

$$\propto \sum_{\mathbf{r}\in\mathcal{R}_+} \sum_{j=1}^{n} \exp\left(-\gamma\left(d(\mathbf{s},\mathbf{t}_j) + d(\mathbf{r},\mathbf{t}_j)\right)\right) \approx \sum_{\mathbf{r}\in\mathcal{R}_+} \exp\left(-\gamma \min_j \left(d(\mathbf{s},\mathbf{t}_j) + d(\mathbf{r},\mathbf{t}_j)\right)\right) \approx \sum_{\mathbf{r}\in\mathcal{R}_+} \exp\left(-C\gamma d(\mathbf{s},\mathbf{r})\right). \tag{2}$$

The second to last approximation relies on the fact that for large enough γ, the term with the smallest exponent will dominate the sum over the latent sources. For the last approximation, we observe that the global minimum of $d(\mathbf{s},\mathbf{t}) + d(\mathbf{r},\mathbf{t})$ over all signals $\mathbf{t}$ is $Cd(\mathbf{r},\mathbf{s})$ for some $C > 0$ and is achieved at $\mathbf{t}^* = (\mathbf{s}+\mathbf{r})/2$. The approximation is valid when the minimizing latent source $\mathbf{t}_{j^*}$ is sufficiently close to the global minimizer $\mathbf{t}^*$. We approximate $\mathbb{P}(- \mid \mathbf{s})$ in a similar fashion. In essence, because the latent sources are unknown, we cannot directly compare the observation with them. Instead we compare the observation to the reference signals, using them as a proxy for the latent sources. Figure 1 illustrates this. We classify the observation according to $f(\mathbf{s}) = \operatorname{sgn}(R(\mathbf{s}) - \theta)$, where $R(\mathbf{s}) = \mathbb{P}(+|\mathbf{s})/\mathbb{P}(-|\mathbf{s})$ and θ is a threshold. We can optionally require D_{req} consecutive "detections" to declare something a trending topic. This approach has an appealing interpretation: to classify an observation, one simply computes the distance from the observation to all examples from each class. This can be done in parallel on enormous data sets.

Results and Conclusion

We collected 500 topics that were trending at some point during June 2012, and 500 that were not. We then collected 10% of the Tweets from June 2012 containing those topics. We used a 50/50 split between reference signals and observations and performed detection in a window of size $2h_{ref}$ hours, centered around the true onset of the trending topic if the topic was trending or chosen randomly otherwise. We performed detection for a range of parameters $\gamma, N_{obs}, h_{ref}, N_{smooth}, \theta$ and D_{req} to evaluate tradeoffs between detection errors and relative detection time. Using this small sample of Tweets, our method is capable of detecting trends in advance of Twitter 79% of the time, with a mean early advantage of 1 hour and 26 minutes, while maintaining a 95% true positive rate and a 4% false positive rate (Figure 2 (left)). Figure 2 (right) shows an example early detection in which our method detected the trending topic "Miss Rhode Island" over 2 hours in advance of Twitter. In Figure 2 (center) the envelope of ROC curves for all combinations of parameters shows the ability of our method to perform well under a variety of tradeoffs between types of error. It should be noted that Twitter's trending topic detection method may need to be more conservative to avoid low-quality trending topics, and that on Twitter, trending topics compete for the top ten spots, whereas our method is based on a score threshold alone. Nevertheless, our results demonstrate the effectiveness of our method for trending topic detection, as well as its potential as a broadly applicable framework for scalable nonparametric classification in the presence of large amounts of data.

References

[1] Mathioudakis, M., and Koudas, N. Twittermonitor: Trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data* (New York, NY, USA, 2010), SIGMOD '10, ACM, pp. 1155–1158.

[2] @twitter. To trend or not to trend. `http://blog.twitter.com/2010/12/to-trend-or-not-to-trend.html` (accessed 9 September, 2012).

[3] Zhao, S., Zhong, L., Wickramasuriya, J., and Vasudevan, V. Human as real-time sensors of social and physical events: A case study of twitter and sports games. *CoRR abs/1106.4300* (2011).