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 [1] — from a global breaking news story to a fire down the street. It is important to be able to detect such events as early as possible. 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 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. Using our stochastic model, we derive a class estimator based on the ratio of conditional class probabilities. We are able to detect trends 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%. Our method allows for tradeoffs between TPR, FPR, 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 messages known as *Tweets*. There are over 400 million Tweets written per 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. Some topics gain sufficient popularity and start *trending* i.e. they are featured on a list of top ten trending topics on Twitter. We observe that trending topics are distinguished from nontrending topics by their pattern of activity leading up to the time the topic is detected by Twitter (the true onset). In particular, the activity of a soon to be trending topic is characterized by frequent sharp jumps of high magnitude over some baseline activity. Let $\rho[n]$ be the discrete derivative at time step n of the volume $v(t)$ of Tweets about a topic over time, for $n = 1, \ldots, N$. We shall call this the rate of Tweets. We construct a baseline-normalized rate $\rho_b[n] = (\rho[n]/b)^\beta$ by removing the baseline rate $b = (\sum_n \rho[n])/N$

Timeseries activity is incredibly diverse. Rather than training a model to distinguish between trending and non trending topics, which would require us to assume a model structure, we propose a nonparametric method that relies directly on large amounts of data as a proxy for the latent model structure

Data Model

We posit that all observed signals were generated by an unknown set of *latent source* signals. A
Stochastic model
Class Probability propto ...etc.
All we have to do is compute distances to examples.
We can do it in parallel.

Results

Figure with example early detection (¿ 2 hrs early) Figure with Distribution of early and late in best case Figure with Tradeoffs (ROC curve/envelope and distribution of early/late at two or three points)

Conclusion

Broadly applicable

References

[1] Siqi Zhao, Lin Zhong, Jehan Wickramasuriya, and Venu Vasudevan. Human as real-time sensors of social and physical events: A case study of twitter and sports games. *CoRR*, abs/1106.4300, 2011.