Twitter Event Detection

David Kimmey

Indiana University - Purdue University Fort Wayne

Follow this and additional works at: http://opus.ipfw.edu/stu_symp2015

Part of the Computer Sciences Commons

Recommended Citation
http://opus.ipfw.edu/stu_symp2015/43

This is brought to you for free and open access by the IPFW Student Research and Creative Endeavor Symposium at Opus: Research & Creativity at IPFW. It has been accepted for inclusion in 2015 IPFW Student Research and Creative Endeavor Symposium by an authorized administrator of Opus: Research & Creativity at IPFW. For more information, please contact admin@lib.ipfw.edu.
1. Introduction

What methods are available for Twitter event detection? Predicting the present (i.e. "contemporaneous forecasting" or "nowcasting") is a hot topic of interest to corporations, central banks, and other government agencies. Furthermore, Twitter data is becoming a popular source of social trajectory data for "nowcasting" (e.g. geo-economic events, the discovery of unusual social events, geographic disease/influenza trends (Figure 1), and social questions). However, the defining and detecting of events (i.e. a thing that happens, especially one of importance) is a non-trivial task.

2. Goal

Given: a set of raw (i.e. unprocessed) Twitter data
Output: find a flu event, and the list of Tweets that occur during the event for use in future data mining work (e.g. descriptive and predictive models of the flu event).

3. Methods

Twitter Event Detection

David L. Kimmey
Department of Computer Science - Purdue University Fort Wayne * Fort Wayne, Indiana 46805

Faculty Sponsor: Jin Soung Yoo

2. Goal

Given: a set of raw (i.e. unprocessed) Twitter data
Output: find a flu event, and the list of Tweets that occur during the event for use in future data mining work (e.g. descriptive and predictive models of the flu event).

4. Results

Twitter Event Detection Method output

Tweet Frequency Bursts Method:

* flu event detected in Tweets from March 9th to March 13th 2015

Tweet Deviation / Anomaly Detection Method:

no flu event detected in Tweets (Tweet freq. is not above top blue line)

Tweet Probabilistic Soft Logic Method:

* flu event detected in Tweets by PSL Rules from March 8th to March 16th 2015

4. Results

Twitter Event Detection Method output

Tweet Frequency Bursts Method:

* flu event detected in Tweets from March 9th to March 13th 2015

Tweet Deviation / Anomaly Detection Method:

no flu event detected in Tweets (Tweet freq. is not above top blue line)

Tweet Probabilistic Soft Logic Method:

* flu event detected in Tweets by PSL Rules from March 8th to March 16th 2015

4. Results

Twitter Event Detection Method output

Tweet Frequency Bursts Method:

* flu event detected in Tweets from March 9th to March 13th 2015

Tweet Deviation / Anomaly Detection Method:

no flu event detected in Tweets (Tweet freq. is not above top blue line)

Tweet Probabilistic Soft Logic Method:

* flu event detected in Tweets by PSL Rules from March 8th to March 16th 2015

References


5. Summary

- The dark gray and blue density areas of the PSL density map align closer to the red, yellow, and green density areas of the U.S. Department of Health’s Heatmap (Figure 1). Therefore, the PSL method provides better results than the Frequency Bursts and the Deviation/Anomaly detection methods.

- The PSL method produced the largest corpus of tweets with more words and higher frequency counts per word. The larger corpus will be an advantage in future data mining work.

- The Deviation/Anomaly Detection method was not able to find this event because of the event’s low frequency of tweets.

5. Summary

- The dark gray and blue density areas of the PSL density map align closer to the red, yellow, and green density areas of the U.S. Department of Health’s Heatmap (Figure 1). Therefore, the PSL method provides better results than the Frequency Bursts and the Deviation/Anomaly detection methods.

- The PSL method produced the largest corpus of tweets with more words and higher frequency counts per word. The larger corpus will be an advantage in future data mining work.

- The Deviation/Anomaly Detection method was not able to find this event because of the event’s low frequency of tweets.

6. Conclusion

Overall, the Probabilistic Soft Logic (PSL) method identified more of the event area than the Frequency Bursts and Deviation/Anomaly detection methods. In addition, PSL seems to have the advantage of finding events in low Tweet frequency environments, and PSL is not dependent on the selection of the time series window (i.e. tweets per hour vs tweets per day). Future work will be needed with larger data sets, and to see if the tweets from the PSL method will produce an effective model after data mining.