Using Twitter to Estimate H1N1 Influenza Activity

Alessio Signorini[†], Philip M. Polgreen[‡], Alberto Maria Segre[†]
Departments of Computer Science[†] and Internal Medicine[‡], The University of Iowa

Objective

This paper describes a system that uses Twitter to estimate influenza-like illness levels by geographic region.

Background

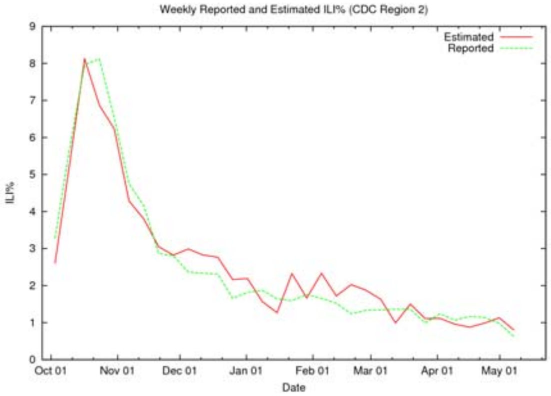
Twitter is a free social networking and micro-blogging service that enables its millions of users to send and read each other's "tweets, " or short messages limited to 140 characters. The service has more than 190 million registered users and processes about 55 million tweets per day [1]. Despite a high level of chatter, the Twitter stream does contain useful information, and, because tweets are often sent from handheld platforms on location, they convey more immediacy than other social networking systems.

Methods

We collected and stored all public tweets beginning October 1, 2009 that matched a set of pre-specified search terms (*e.g.,* flu, swine, influenza, tamiflu, oseltamivir, h1n1, etc.*).* After culling, *e.g.,* non-U.S. and non-English tweets, and applying appropriate stemming algorithms, the remaining tweets were used to produce a dictionary of English word equivalents. We compiled daily and weekly usage statistics for each dictionary term, both nationally and at the CDC's influenza reporting-region level [2]. These weekly term-usage statistics were then used to estimate weekly levels of influenza-like illness (ILI). Using Support Vector Regression [3], a supervised machine-learning method generally applied to solve classification problems [4], we trained our system using weekly term-frequency statistics from tweets issued outside of CDC Region 2 (New York and New Jersey) and ILI values reported by the CDC for the weeks October 4-10, 2009 through May 16-22, 2010. We then used the resulting system to estimate ILI in CDC Region 2, thus performing an out-of-sample validation.

Results

The figure shows estimated weighted ILI values for CDC Region 2 (New Jersey and New York) produced by our system when trained on Twitter traffic exclusive of CDC Region 2. The predicted weekly ILI values are shown in red, with ILI values reported later by the CDC in green. Our regional model approximates the epidemic curve reported by ILI data with an average error of 0.37% (min=0.01%, max=1.25%) and a standard deviation of 0.26%. Similar results were obtained when estimating ILI at a national level.



Conclusions

Our results demonstrate that Twitter traffic can be used to provide real-time estimates of disease activity. Our ability to quickly detect trends which are then confirmed by observations from traditional surveillance approaches make this new form of surveillance a promising area of research at the interface between computer science, epidemiology, and medicine.

References

1. Lorica, B. Twitter by the numbers. O'Reilly Radar, http://radar.oreilly.com/2010/04/twitter-by-the-numbers.html, April 14, 2010 (Accessed 15 August 2010).

2. Centers for Disease Control and Prevention. Overview of Influenza Surveillance in the United States,http://www.cdc.gov/flu/weekly/fluactivity.htm, 2010 (Accessed 15 August 2010).

3. Harris Drucker, Chris J.C. Burges, Linda Kaufman, Alex Smola and Vladimir Vapnik (1997). "Support Vector Regression Machines". Advances in Neural Information Processing Systems 9, NIPS 1996, 155-161, MIT Press.

4. Cristianini N and Shawe-Taylor J. *An Introduction to Support Vector Machines and Other Kernel-based*

Learning Methods. Cambridge University Press, 2000.