Using Web and Social Media for Influenza Surveillance

Courtney D Corley[1], Diane J Cook[2], Armin R Mikler[3] and Karan P Singh[4]

Abstract Analysis of Google influenza-like-illness (ILI) search queries has shown a strongly correlated pattern with Centers for Disease Control (CDC) and Prevention seasonal ILI reporting data. Web and social media provide another resource to detect increases in ILI. This paper evaluates trends in blog posts that discuss influenza. Our key finding is that from 5-October 2008 to 31-January 2009 a high correlation exists between the frequency of posts, containing influenza keywords, per week and CDC influenza-like-illness surveillance data.

Keywords Health informatics, disease surveillance, public health epidemiology, information retrieval, social media analytics

1 Introduction

Influenza diagnosis based solely on the presentation of symptoms is limited as these symptoms may be associated with many other diseases. Serologic and antigen tests require that a patient with influenza-like-illness (ILI) be examined by a physician who can either conduct a rapid diagnosis test or take blood samples for laboratory testing. This suggests that many cases of influenza remain undiagnosed. While the presence of influenza in an individual can be confirmed through specific diagnostic tests, the influenza prevalence in the population at any given time is unknown and can only be estimated. In the past, such estimates have relied solely on the extrapolation of diagnosed cases, making it difficult to identify the various phases of seasonal influenza, or the identification of a more serious manifestation of a flu epidemic.

Web and social media (WSM) provide a resource to detect increases in ILI. This paper evaluates blog posts that discuss influenza, analysis show a signif-

[1] Pacific Northwest National Laboratory, Richland, WA, e-mail: court@pnl.gov

[2] School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA, e-mail: cook@eecs.wsu.edu

[3] Department of Computer Science and Engineering, University of North Texas, Denton, TX, e-mail: mikler@unt.edu

[4] Department of Biostatistics, School of Public Health, University of North Texas Health Science Center, Fort Worth, TX, e-mail: ksingh@hsc.unt.edu

icant correlation with the US 2008-2009 seasonal influenza epidemic. We briefly discuss a history of infectious disease outbreaks and recent approaches in online public health surveillance of influenza are discussed with regards to outbreak responses. Next, the data set used in our analysis is presented and the methodology for information extraction and trend analysis is outlined posting trends. Through discovery and verification of trends in influenza related blogs, we verify a correlation to Centers for Disease Control and Prevention (CDC) influenza-like-illness patient reporting at sentinel healthcare providers.

1.1 Background

Epidemics of infectious diseases have plagued humankind since historical times. There are accounts of epidemics dating back to the times of Hippocrates (459-377 B.C.) and the ancient Greeks [1]. Fourteenth century Europe lost a quarter of its 100 million people to Black Death. The fall of the Aztec empire in 1521 was due to smallpox that eradicated half of its 3 ½ million population. The pandemic influenza of 1918 caused over 20 million excess deaths in 12 months. More recently, the severe acute respiratory syndrome (SARS) outbreak of 2003 highlighted the rapid spread of an epidemic at the global level. The outbreak, emanating from a small Guangzhou province in China, spread around the world requiring a concerted response from public health administrations around the world and the World Health Organization (WHO) to curtail the epidemic [5]. The WHO and CDC [2] actively engage in worldwide surveillance of infectious diseases, and prioritize prevention and control measures at the root cause of epidemics.

The pervasiveness and ubiquity of Internet and World Wide Web resources provide individuals with access to many information sources that facilitate self-diagnosis; one can combine specific disease symptoms to form search queries. The results of such search queries often lead to sites that may help diagnose the illness and offer medical advice. Recently, Google has addressed this issue by capturing the keywords of queries and identifying specific searches that involve search terms that indicate influenza-like-illness (ILI) [4]. Published research on influenza Internet surveillance also includes search “advertisement click-through” [3] using a set of Yahoo search queries containing the words *flu* or *influenza* [9] and health website access logs [6, 7]. Other information sources, such as telephone triage services, can be useful to ILI detection. The findings in Yih et al. [11], show that telephone triage service is not a reliable measure for influenza surveillance due to service coverage; however, it may be beneficial in certain situations where other surveillance measures are inadequate.

2 Data and Methodology

Spinn3r (www.spinn3r.com) is a Web and social media indexing service that conducts real-time indexing of all blogs, with a throughput of over 100,000 new blogs indexed per hour. Blog posts are accessed through an open source Java application-programming interface (API). Metadata available with this data set includes the following (if reported): blog title, blog url, post title, post url, date posted (accurate to seconds), description, full HTML encoded content, subject tags annotated by author, and language.

Data is selected from an arbitrary time period of 20 weeks, beginning at 5-October 2008 and ending at 31-January 2009. (total 97,955,349; weblogs 69,627,831; forums 1,986,656; mainstream media 21,543,027; other 4,797,835). Weblog, micro-blog and mainstream media items containing characteristic keywords are extracted from Web and online social media published in the same time frame. Characteristic keywords include flu, influenza, H5N1, H3N1, and other keywords relevant to this task. This paper defines the *blog-world* to be English language, non-spam, blog posts. We also consider the following terms to be equivalent: blog post & blog item and blogger & blog site. Indexing, parsing and link extraction code was written in Python, parallelized using pyMPI and executed on a cluster at the Center for Computational Epidemiology and Response Analysis at the University of North Texas. This compute resource has eight nodes (2.66GHz Quad Core Xeon processors), 64 core, 256 GB memory, 30 TB of network storage [8, 10].

In our analysis, we extract English language items from the blog-world index when a lexical match exists to the terms *influenza* and *flu* anywhere in its content (misspellings and synonyms are not included). The blog items are grouped by month, week (Sunday to Saturday) and by day of week. The extracted blog items containing influenza keywords are termed flu-content posts or ***FC-posts***. FC-post trends can be monitored using the social media mining methodology presented in this paper. This methodology facilitates identification of outbreaks or increases of influenza infection in the population. This paper's most significant finding is a strong correlation between the frequency of FC-posts per week and Centers for Disease Control and Prevention influenza-like-illness surveillance data.

3 Results

We hypothesize that the frequency of blog-world flu-posts correlates with a patient reporting of influenza-like-illness during the US flu-season. To verify this statement we compare our data to Centers for Disease Control and Prevention surveillance reports from sentinel healthcare providers. The CDC website states the Outpatient Influenza-like-illness Surveillance Network (ILINet) consists of about 2,400 healthcare providers in 50 states reporting approximately 16 million patient

visits each year. Each provider reports data to CDC on the total number of patients seen, and the number of those patients with influenza-like-illness (ILI) by age group. For this system, ILI is defined as fever (temperature of 100F [37.8C] or greater) and a cough and/or a sore throat in the absence of a known cause other than influenza [2].

The CDC ILINet surveillance and FC-post per week data are plotted in Figure 1. CDC influenza-like-illness symptoms per visit at sentinel US healthcare providers labels the primary Y-axis. The secondary Y-axis marks the FC-post per week frequency normalized by the corresponding blog-world week post count. Correlation between the two data series measured with a Pearson correlation coefficient, *r*. Evidence to support our hypothesis (a correlation exists between CDC ILINet reports and Web and social media mined FC-post frequency) is the Pearson's correlation coefficient evaluated between the two data series. The Pearson correlation evaluates to unity if the two data series are exactly matching, *r=1*. If no correlation exists between the data series, the Pearson correlation evaluates to zero, *r=0*.

In our analysis, the 20 ILI and FC-post data points correlate strongly with a high Pearson correlation, *r=0.626*, and the correlation is significant with 95% confidence.

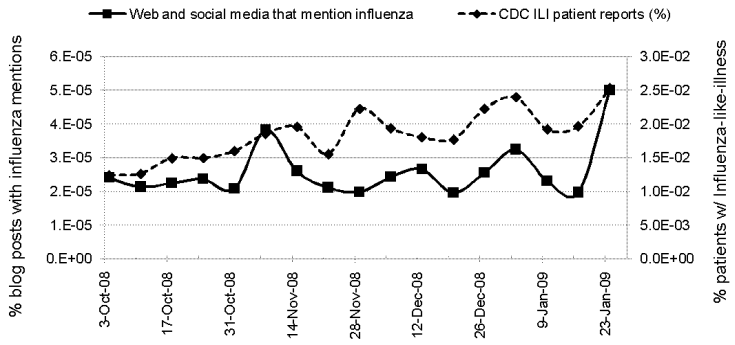

Fig. 1: CDC ILINet vs Normalized FC-post Frequency per Week. Each FC-posts per week data point is normalized by the corresponding blog-world posts per week count. Pearson correlation = 0.626, 95% confidence.

4 Future Work

Once FC-posts have been extracted, one can further monitor influenza outbreaks by evaluating the perspective of blog authors. Bloggers having a direct knowledge of influenza infection are more valuable to disease surveillance than those who author objective or opinion items. Bloggers who persistently author FC-posts are less

likely to be infected with influenza and are more likely to be writing about avian influenza (bird flu). The following post excerpts demonstrate influenza-content author perspective.

> **Self Identified**: ``What began as an irritating cold became what I think might be the flu last night. I woke up in bed around three this morning with sore muscles, congested lungs/nose and chills running throughout my body."
>
> **Secondhand**: ``According to ESPN.com, Ravens quarterback Troy Smith has lost 'a considerable amount of weight' while being hospitalized with tonsillitis and flu-like symptoms. Smith and veteran Kyle Boller likely won't play in Sunday's season opener, leaving the workload to rookie Joe Flacco and Joey Harrington, who was signed Monday."
>
> **Objective (or opinion)**: ``Domesticated birds may become infected with avian influenza virus through direct contact with infected waterfowl or other infected poultry, or through contact with surfaces or materials like that of water or feed that have been contaminated with the virus."

Identifying the perspective of influenza keyword posts facilitates determining its contribution to disease surveillance, three author perspectives are identified. A FC-post is either a self-identification of having ILI symptoms, secondhand (or by proxy) of another individual having ILI or the post is an opinion or objective article containing ILI keywords. Secondhand knowledge can be writing about a friend, schoolmate, family-member or co-worker but a blogger could also post details on famous individual such as a sports player. The season opening of American football coincides with the data and many FC-posts identify athletes who are unable to play because of an ILI. Automatic classification of the influenza post author's perspective is ongoing research.

5 Conclusion

Web and social media provide a novel disease surveillance resource. We presented a method which evaluates blog posts containing keywords influenza or flu and the results from analysis show strong co-occurrence with the US 2008-2009 flu season. This paper's key finding is that from 5-October 2008 to 31-January 2009 a high correlation exists between the frequency of posts, containing influenza keywords, per week and Centers for Disease Control and Prevention influenza-like-illness surveillance data.

Acknowledgments

We would like to thank the National Science Foundation (NSF) for partial support under grant NSF IIS-0505819 and the Technosocial Predictive Analytics Initiative, part of the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory (PNNL). PNNL is operated by Battelle for DOE under contract DE-ACO5-76RLO 1830. The contents of this publication are the responsibility of the authors and do not necessarily represent the official views of the NSF.

References

1. Bailey, N.: The Mathematical Theory of Epidemics. Griffin (1957)
2. CDC-Website: Influenza surveillance reports. Website (accessed 25 July 2009). http://www.cdc.gov/flu/weekly/fluactivity.htm
3. Eysenbach, G.: Infodemiology: tracking flu-related searches on the web for syndromic surveillance. AMIA Annual Symposium proceedings pp. 244–8 (2006)
4. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature 457(7232), 1012–4 (2009). DOI 10.1038/nature07634
5. Heymann, D., Rodier, G.: Global surveillance, national surveillance, and sars. Emerging Infectious Diseases 10(2) (2004)
6. Hulth, A., Rydevik, G., Linde, A., Montgomery, J.: Web queries as a source for syndromic surveillance. PLoS ONE 4(2), e4378 (2009)
7. Johnson, H.A.,Wagner, M.M., Hogan,W.R., Chapman,W., Olszewski, R.T., Dowling, J., Barnas, Analysis of web access logs for surveillance of influenza. Studies in health technology and informatics 107(Pt 2), 1202–6 (2004)
8. Miller, P.: pympi–an introduction to parallel python using mpi. Livermore National Laboratories (2002). URL https://computing.llnl.gov/code/pdf/pyMPI.pdf
9. Polgreen, P.M., Chen, Y., Pennock, D.M., Nelson, F.D.: Using internet searches for influenza surveillance. Clin Infect Dis 47(11), 1443–8 (2008). DOI 10.1086/593098
10. Rossum, G.V., Drake, F.: Python language reference. Network Theory Ltd (2003). URL http://www.altaway.com/resources/python/reference.pdf
11. Yih, W., Teates, K., Abrams, A., Kleinman, K., Kulldorff, M., Pinner, R., Harmon, R., Wang, S., Platt, R., Montgomery, J.: Telephone triage service data for detection of influenza-like illness. PLoS ONE 4(4), e5260 (2009)