OJPHI

ISDS 2015 Conference Abstracts

ISDS INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

Global Disease Monitoring and Forecasting with Wikipedia

Nicholas Generous*, Geoffrey Fairchild, Alina Deshpande, Sara Y. Del Valle and Reid Priedhorsky

Los Alamos National Laboratory, Los Alamos, NM, USA

Objective

To explore the use of Wikipedia as a data source for disease surveillance.

Introduction

Infectious disease remains costly in human and economic terms. Effective and timely disease surveillance is a critical component of prevention and mitigation strategies. The limitations of traditional disease surveillance systems have motivated new techniques based upon internet data sources such as search queries and social media. However, 4 challenges remain before internet-based disease surveillance models can be reliably integrated into an operational system: openness, breadth, transferability, and forecasting. We evaluated a new data source, Wikipedia access logs, in these 4 challenges for global disease surveillance and forecasting.

Methods

We used Wikipedia article access logs and disease incidence reports to build linear models to analyze 3 years of data for 14 disease-location contexts. Access logs for all Wikipedia articles are freely available online [1]. We used official epidemiological reports available from government health agencies and the World Health Organization.

As Wikipedia does not provide article access counts for specific countries, we used language as a proxy. We selected articles by examining the English Wikipedia article for the disease, enumerated relevant linked articles and identified corresponding articles in each language by following the inter-language wiki link.

To nowcast, we aligned the article access counts with the incidence data in order to yield time series with the same frequency. For each article we computed Pearson's correlation r against the disease time series and selected the 10 highest correlated articles. We then built a linear multiple regression model. We assessed forecasting potential by repeating the process with the article time series shifted 28 days forward and backward in 1 day increments. To evaluate whether model transferability is possible, we computed a metric r_t, the Pearson's r computed between the correlation scores r of each article found in both languages, for each pair of locations tested on the same disease.

Results

Among the 14 contexts we analyzed, 8 of the models succeeded for nowcasting and forecasting, 3 cases failed because patterns in the official data were too subtle to capture and 3 failed because the signal-to-noise ratio in the Wikipedia data was too subtle to capture. Performance fell along disease lines: all influenza and dengue models were successful, 2 of the 3 tuberculosis models were, and cholera, ebola, HIV/AIDS, and plague were unsuccessful. Table 1 summarizes the nowcasting and forecasting performance of the models. Table 2 lists the transferability scores r_t for each pair of countries tested on the same disease. In the case of influenza, both Japan/Thailand and Thailand/USA show promising preliminary results.

Conclusions

Human activity on the Internet leaves traces that contain real and useful evidence of disease dynamics. Wikipedia data are one of the few Internet data sources that can meet all 4 challenges. Wikipedia data are freely available to anyone (openness), they work in multiple locations for multiple diseases around the world with model success of r^2 up to 0.92 (breadth), Wikipedia based models can possibly be transferable to different locations with similarity of up to 0.81 (transferability), and they have forecasting value through a horizon of 28 days (forecasting).

This preliminary study has several limitations. The methods need to be tested in more contexts, a better article selection procedure is needed, and better geo-location is needed.Despite these limitations, Wikipedia access logs is a useful data source for global disease monitoring and forecasting.

			r^2 at forecast				Best forec.	
Disease	Location	Result	0	7	14	28	Days	r^2
Cholera	Haiti	Failure (SNR)	0.45	0.39	0.41	0.48	26	0.50
Dengue	Brazil	Success	0.85	0.81	0.77	0.65	−3	0.86
	Thailand	Success	0.55	0.54	0.57	0.74	28	0.74
Ebola	Uganda/DRC	Failure (SNR)	0.02	0.01	0.02	0.02	5	0.14
HIV/AIDS	China (PRC)	Failure (Official data)	0.62	0.48	0.34	0.31	−1	0.63
	Japan	Failure (Official data)	0.15	0.19	0.15	0.05	9	0.22
Influenza	Japan	Success	0.82	0.92	0.86	0.52	8	0.92
	Poland	Success	0.81	0.86	0.88	0.72	12	0.89
	Thailand	Success	0.79	0.76	0.67	0.48	−2	0.80
	United States	Success	0.89	0.90	0.85	0.66	5	0.91
Plague	United States	Failure (SNR)	0.23	0.03	0.05	0.07	0	0.23
Tuberculosis	China (PRC)	Success	0.66	0.66	0.52	0.25	−9	0.78
	Norway	Failure (Official data)	0.31	0.41	0.40	0.42	20	0.48
	Thailand	Success	0.68	0.68	0.69	0.69	9	0.69

Table 1: Summary of model performance

Disease	Location 1	Location 2	r_t
Dengue	Brazil	Thailand	0.39
HIV/AIDS	China (PRC)	Japan	−0.06
Influenza	Japan	Poland	0.45
	Japan	Thailand	0.81
	Japan	United States	0.62
	Poland	Thailand	0.48
	Poland	United States	0.44
	Thailand	United States	0.76
Tuberculosis	China (PRC)	Norway	0.19
	China (PRC)	Thailand	−0.20
	Norway	Thailand	n/a

Table 2: Summary of transferability scores

Keywords

Wikipedia; Disease Surveillance; Internet data; search queries; global

Acknowledgments

This work is supported in part by NIH/NIGMS/MIDAS under grant U01-GM097658-01 and the Defense Threat Reduction Agency (DTRA), Joint Science and Technology Office for Chemical and Biological Defense under project numbers CB3656 and CB10007. Mac Brown's participation in discussions improved this work significantly. LA-UR~14-22535

References

[1] http://dumps.wikimedia.org/other/pagecounts-raw/

***Nicholas Generous**

E-mail: generous@lanl.gov