OJPHI

ISDS 2015 Conference Abstracts

ISDS INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

Eliciting Disease Data from Wikipedia Articles

Geoffrey Fairchild*[1, 3], Lalindra De Silva[2], Sara Y. Del Valle[1] and Alberto M. Segre[3]

[1]Analytics, Intelligence, and Technology Division, Los Alamos National Laboratory, Los Alamos, NM, USA; [2]University of Utah, Salt Lake City, UT, USA; [3]University of Iowa, Iowa City, IA, USA

Objective

To improve traditional outbreak surveillance systems by utilizing the content of Wikipedia articles.

Introduction

Traditional disease surveillance systems suffer from several disadvantages, including reporting lags and antiquated technology, that have caused a movement towards internet-based disease surveillance systems. Recently, Wikipedia access logs (e.g., McIver 2014[1], Generous 2014[2]) have been shown to be effective in this arena. Much richer Wikipedia data are available, though, including the entire Wikipedia article content and edit histories.

We study two different aspects of Wikipedia content as it relates to unfolding disease events: 1) we demonstrate how to capture case, death, and hospitalization counts from the article text, and 2) we show there are valuable time series data present in the tables found in certain articles.

We argue that Wikipedia data cannot only be used for disease surveillance but also as a centralized repository system for collecting disease-related data in near real-time.

Methods

Most outbreak articles we surveyed contained a variety of useful information in the text (e.g., dates, locations, case and death counts, demographics). These data are generally swiftly updated as new information become available, and sources are often provided so that external review can occur. In order to recognize certain key phrases in the Wikipedia article narrative, we trained a *named-entity recognizer* (NER). NERs are *sequence labelers* (they label sequences of words). We trained Stanford's NER to automatically identify three entity types: 1) DEATHS, 2) INFECTIONS, and 3) HOSPITALIZATIONS.

We demonstrated the viability of tabular data using the Ebola virus epidemic in West Africa article. We elicited 39 unique tables from the 5,137 revisions made to the article from March 29, 2014 to October 14, 2014. For each affected country, each table contained case and death count time series.

Results

To test the NER's performance, we averaged precision, recall, and F1 score results from 10-fold cross-validation. Our 14-article training set achieved precision of 0.812 and recall of 0.710, giving us an F1 score of 0.753. The classifier's performance is respectable and will likely improve given a larger, more expansive training set.

To determine the accuracy and timeliness of the Wikipedia West African Ebola epidemic time series, we used Caitlin Rivers' crowd-sourced Ebola data as ground truth. We compared the 39 Wikipedia epidemic time series to the ground truth data by computing the root-mean-square error (RMSE). The average RMSE values for each country's time series are listed in Table 1. The RMSE values are low, indicating that the time series found on the Wikipedia article are both timely and accurate.

Conclusions

Internet data are becoming increasingly important for disease surveillance because they address some of the existing challenges, such as the reporting lags inherent in traditional disease surveillance data, and they can also be used to detect and monitor emerging diseases. Additionally, internet data can simplify global disease data collection. We envision this work being incorporated into a community-driven open-source emerging disease detection and monitoring system. A community-driven effort to improve global disease surveillance data is imminent, and Wikipedia can play a crucial role in realizing this need.

Average cases and deaths RMSE across all table revisions.

Country	Mean Cases RMSE	Mean Deaths RMSE
Guinea	3.790	2.701
Liberia	18.168	11.983
Nigeria	0.310	0.189
Senegal	0.403	0.008
Sierra Leone	18.847	12.015
Spain	18.243	0.050
United States	0.174	0.000

Keywords

natural language processing; named-entity recognition; Ebola; Wikipedia; disease surveillance

Acknowledgments

This work is supported in part by NIH/NIGMS/MIDAS under grant U01-GM097658-01 and the DTRA Joint Science and Technology Office for Chemical and Biological Defense under project numbers CB3656 and CB10007. LANL is operated by Los Alamos National Security, LLC for the Department of Energy under contract DE-AC52-06NA25396.

References

1. McIver DJ, Brownstein JS. Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time. PLOS Computational Biology. 2014 Apr 17;10(4):e1003581.
2. Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R. Global Disease Monitoring and Forecasting with Wikipedia. PLOS Computational Biology. 2014 Nov 13;10(11):e1003892.

***Geoffrey Fairchild**
E-mail: gfairchild@lanl.gov