Emerging Health Threats Journal 2011, **4**:s119. **doi:** 10.3134/ehtj.10.0119
© 2011 M Torii *et al.*; licensee Emerging Health Threats Journal.
www.eht-journal.org

ABSTRACT

Document classification toward efficient event-based biosurveillance

M Torii, N Nelson[1], and D Hartley[1]

Division of Integrated Biodefense, ISIS Center, Georgetown University Medical Center, Washington, DC, USA
E-mail: torii@isis.georgetown.edu

Objective

This paper describes ongoing efforts in enhancing automated document classification toward efficient event-based biosurveillance.

Introduction

Event-based biosurveillance is a practice of monitoring diverse information sources for the detection of events pertaining to human health.[1–3] Online documents, such as news articles on the Internet, have commonly been the primary information sources in event-based biosurveillance.[4–8] With the large number of online publications as well as with the language diversity, thorough monitoring of online documents is challenging. Automated document classification is an important step toward efficient event-based biosurveillance. In Project Argus, a biosurveillance program hosted at Georgetown University Medical Center, supervised and unsupervised approaches to document classification are considered for event-based biosurveillance.

Methods

In Argus operations, analysts are requested to label online documents that they read in their regular surveillance work. Currently, two document classes, relevant and irrelevant, are assumed. With such labeled articles, a customized classifier is trained for target geographic regions/languages using a machine-learning algorithm. Documents retrieved from a Boolean keyword search can be classified (filtered) or ranked according to the relevancy scores assigned. In addition, we considered dynamic grouping of documents, in contrast to classification into predefined classes. To reflect analysts' perspective in clustering documents, we try to weight features (for example, keywords) based on information extracted from class-labeled documents and/or past event reports.

Results

We have tested the proposed framework to facilitate supervised machine-learning classifiers on past data. The framework has been implemented in the Argus surveillance workflow. We are in the process of evaluating the performance of trained classifiers in operational settings. Meanwhile, we became aware of inherent challenges in the framework that could affect performance of classifiers, which include class-imbalance in training data sets, that is, few labeled irrelevant (or relevant) articles may be available because of labeling bias (or to the inherent class-imbalance), and dominant topics in labeled examples, for example, articles on seasonal influenza.

Using past data, we observed that informative subsets could be derived using document clustering, for example, k-means clustering.

Conclusions

We observed promising results on automated document classification in our preliminary experiments. Previously good results have been reported by other related studies in this domain.[4–8] Meanwhile, Boolean queries created and maintained by expert analysts have also been found effective in Project Argus. The utility of automated document classifiers in contrast to the Boolean keyword search should be evaluated in real-life surveillance settings in the future.

Acknowledgements

This paper was presented as an oral presentation at the 2010 International Society for Disease Surveillance Conference, held in Park City, UT, USA, on 1–2 December 2010.

References

1 Hartley DM, Nelson NP, Walters R, Arthur R, Yangarber R, Madoff L, *et al.* The landscape of international event-based biosurveillance. *Emerg Health Threats J* 2010;**3**:e3.

[1] These authors contributed equally to this paper.

2 Keller M, Blench M, Tolentino H, Freifeld CC, Mandl KD, Mawudeku A, *et al.* Use of unstructured event-based reports for global infectious disease surveillance. *Emerg Infect Dis* 2009;**15:**689–95.

3 Walters R, Harlan P, Nelson NP, Hartley DM. Data sources for biosurveillance. In: Voeller JG (ed) *Wiley Handbook of Science and Technology for Homeland Security: Risk Analysis*. Hoboken, NJ, USA, 2009.

4 Nelson NP, Brownstein JS, Hartley DM. Event-based biosurveillance of respiratory disease in Mexico, 2007–2009: connection to the 2009 inuenza A(H1N1) pandemic? *Euro Surveill* 2010;**15**, pii = 19626. Available online http://www.eurosurveillance.org/ViewArticle.aspx? ArticleId = 19626.

5 Brownstein JS, Freifeld CC. HealthMap: the development of automated real-time internet surveillance for epidemic intelligence. *Euro Surveill* 2007;**12:**E071129–071125.

6 Mawudeku A. Mining the Internet: GPHIN. In *The International Meeting on Emerging Diseases and Surveillance (IMED 2007)*. Vienna, Austria, 2007.

7 Collier N, Doan S, Kawazoe A, Goodwin RM, Conway M, Tateno Y, *et al.* BioCaster: detecting public health rumors with a Web-based text mining system. *Bioinformatics* 2008;**24:** 2940–1.

8 Steinberger R, Fuart F, Groot Evd, Best C, Etter Pv, Yangarber R. *Text Mining from the Web for Medical Intelligence*. OIS Press, The Netherlands, 2008.