Exploiting Textual Source Information for Epidemiosurveillance

Elena Arsevska, Mathieu Roche, Renaud Lancelot, Pascal Hendrikx, Barbara Dufour

Exploiting Textual Source Information for Epidemiosurveillance

Elena Arsevska[1], Mathieu Roche[1], Renaud Lancelot[1],
Pascal Hendrikx[2], and Barbara Dufour[3]

[1] Cirad, Montpellier, France
[2] Anses, Paris, France
[3] EnvA, Maisons-Alfort, France

Abstract. In recent years as a complement to the traditional surveillance reporting systems there is a great interest in developing methodologies for early detection of potential health threats from unstructured text present on the Internet. In this context, we examined the relevance of the combination of expert knowledge and automatic term extraction in the creation of appropriate Internet search queries for the acquisition of disease outbreak news. We propose a measure that is the number of relevant disease outbreak news detected in function of the terms automatically extracted from a set of example Google and PubMED corpora. Due to the recent emergence we have used the African swine fever as a disease example.

Keywords: terminology extraction, internet disease surveillance.

1 Introduction

The new and exotic infectious diseases are an incising threat to countries due to globalization, movement of passengers, and international trade. With the traditional reporting schemes, often there are miss, delays or underreporting of disease outbreaks; leading to unawareness of countries about potential disease threats. As the Internet is a source of numerous and dynamic information, services need tools that could refine the search and detect the information of interest. Two important systems of the state-of-the-art, MediSys (Mantero *et al.* 2011) and Healthmap (Collier 2012) are based on a series of automatic steps to detect and acquire disease related news. The algorithms rely upon predefined templates, such keywords or patterns. Internet search queries have been proposed as inexpensive method to detect signals of diseases (ex. avian influenza) (Polgreen *et al.* 2008). In the face of many diseases and even more symptoms, the analysts face another challenge: How to identify appropriate queries for Internet disease surveillance? One option is to use the terms from existing thesaurus (e.g., MeSH). In this paper we present a new combined approach of selection of terms automatically extracted from relevant scientific and non-scientific corpora in order to identify most appropriate search queries for the detection of disease outbreak news on the Internet. As it is a recently emerging disease we use African swine fever (ASF) as a disease example.

S. Closs et al. (Eds.): MTSR 2014, CCIS 478, pp. 359–361, 2014.

2 How to Extract Relevant Information?

The methodology we propose consists of four stages: data acquisition, information retrieval, information extraction and information evaluation. Here we focus on the automatic term extraction and evaluation by domain experts in order to improve the relevance of the search queries for the detection of disease outbreak news on the Internet. For automatic extraction of terms from documents, we have used the BioTex tool which combines linguistic and statistic information adapted to biomedical domain (Lossio *et al.* 2014). More precisely, with Biotex (i) the list of syntactic structures of terms are learnt with relevant sources for our study (e.g., MeSH), and (ii) the relevant combination of information retrieval techniques (e.g. TF-IDF, OKAPI, and C-value measures). The aim of our work consists of weighting the terms extracted according to different sources of information. Therefore we propose a measure (see formula (1)) that privileges the terms extracted from the relevant sources and the high ranking obtained with Biotex.

$$w(t) = \sum \alpha_i \times \frac{1}{rank_{Si}(t)} \quad \text{with } \alpha_i \in [0,1] \text{ and } \sum \alpha_i = 1 \qquad (1)$$

where t is the term, Si is the information from the Internet source, $rank_{Si}$ is the automatic Biotex rank of the term t from a source Si and where α_i is the weight attributed by experts to Si.

3 Experiments

Two principal sources of information were used in this work: Google and PubMED. The search queries were applied for the period from 01/01/2011 to 10/06/2014 on the 10th of June 2014. The Google corpus was acquired with the search query: “african swine fever outbreak” that resulted in 497 news. Only 123 HTML pages, reporting an ASF outbreak (place, time, animals affected, symptoms etc.) were considered as relevant to this work. The PubMED corpus was consisted of 232 abstracts that contained the term “african swine fever” in the title. Only 66 abstracts were selected as relevant to the epidemiology of ASF. 1200 terms were extracted and ranked from the Google and PubMED corpora. Domain experts identified 67 (5,6 %) terms from Google and 85 (7,1 %) from PubMED as relevant to describe an ASF case or outbreak, including acronyms and synonyms. According to this evaluation, the attributed weight (α_i) for Google was 0,4 and 0,6 for PubMED. For example the weight given to the term “asf outbreaks” based on the formula (1) was (1/5)*0,6+(1/1034)*0,4 = 0,12. This term used as a query enabled to identify 67 disease outbreak news not identified previously.

4 Conclusion and Future Work

Our work shows that both Google and PubMed could serve as sources of terms for Internet search queries (with PubMED giving 20% more relevant terms). We believe that search-term surveillance may represent an inexpensive way of performing supplemental disease surveillance. The use of search queries is not limited to ASF; it could also be used to monitor other infectious diseases or even symptoms (e.g., abortion, mortality). For this preliminary study, we limited our experiments to a small set of examples. In future we intend to test the relevance of a more precise set of terms or combinations thereof as Internet search queries.

Acknowledgements. This work was supported by the French Ministry of Agriculture, Food and Forestry and the French Agricultural Research Centre for International Development. We thank Sylvain Falala and David Chavernac for their technical support for this work.

References

1. Mantero, J., Belyaeva, E.E., Linge, J.P.: How to maximize event-based surveillance web-systems: the example of ECDC/JRC collaboration to improve the performance of MedISys. Publications Office of the European Union (2011)
2. Lossio Ventura, J.-A., Jonquet, C., Roche, M.: Teisseire, Towards a Mixed Approach to Extract Biomedical Terms from Text Corpus. Int. J. Knowl. Disc. Bioinfo. 4(1), 1–15 (2014)
3. Collier, N.: Uncovering text mining: A survey of current work on web-based epidemic intelligence. Glob. Public Health 7(7), 731–749 (2012)
4. Polgreen, P.M., Chen, Y., Pennock, D.M., Nelson, F.D.: Using Internet Searches for Influenza Surveillance. Clin. Infect. Dis. 47(11), 1443–1448 (2008)