Computers and Electronics in Agriculture 123 (2016) 104–115

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web

Elena Arsevska [a,b,*], Mathieu Roche [c,d], Pascal Hendrikx [e], David Chavernac [a,b], Sylvain Falala [a,b], Renaud Lancelot [a,b], Barbara Dufour [f]

[a] French Agricultural Research and International Cooperation Organization (CIRAD), Unit for Control of Exotic and Emerging Diseases in Animals (UMR CMAEE), Campus international de Baillarguet, 34398 Montpellier, France
[b] French National Institute for Agricultural Research (INRA), Unit for Control of Exotic and Emerging Diseases in Animals (UMR CMAEE), Campus international de Baillarguet, 34398 Montpellier, France
[c] French Agricultural Research and International Cooperation Organization (CIRAD), Unit for Land, Environment, Remote Sensing and Spatial Information (UMR TETIS), 500 rue Jean-François Breton, 34093 Montpellier, France
[d] Laboratory of Informatics, Robotics and Microelectronics (LIRMM), UMR 5506, French National Centre for Scientific Research (CNRS), Montpellier University, 34000 Montpellier, France
[e] French Agency for Food, Environmental and Occupational Safety (ANSES), Unit for Coordination and Support to Surveillance (UCAS), 14 rue Pierre et Marie Curie, 94706 Maisons-Alfort, France
[f] Alfort Veterinary School (ENVA), 7 avenue du Général de Gaulle, 94704 Maisons-Alfort, France

ARTICLE INFO

Article history:
Received 5 March 2015
Received in revised form 17 January 2016
Accepted 18 February 2016
Available online 5 March 2016

Keywords:
Web
Disease outbreak
Text mining
Term extraction
Query
Delphi method

ABSTRACT

Timeliness and precision for detection of infectious animal disease outbreaks from the information published on the web is crucial for prevention against their spread. The work in this paper is part of the methodology for monitoring the web that we currently develop for the French epidemic intelligence team in animal health. We focus on the new and exotic infectious animal diseases that occur worldwide and that are of potential threat to the animal health in France.

In order to detect relevant information on the web, we present an innovative approach that retrieves documents using queries based on terms automatically extracted from a corpus of relevant documents and validated with a consensus of domain experts (Delphi method). As a decision support tool to domain experts we introduce a new measure for ranking of extracted terms in order to highlight the more relevant terms. To categorise documents retrieved from the web we use Naïve Bayes (NB) and Support Vector Machine (SVM) classifiers.

We evaluated our approach on documents on African swine fever (ASF) outbreaks for the period from 2011 to 2014, retrieved from the Google search engine and the PubMed database. From 2400 terms extracted from two corpora of relevant ASF documents, 135 terms were relevant to characterise ASF emergence. The domain experts identified as highly specific to characterise ASF emergence the terms which describe mortality, fever and haemorrhagic clinical signs in *Suidae*.

The new ranking measure correctly ranked the ASF relevant terms until position 161 and fairly until position 227, with areas under ROC curves (AUCs) of 0.802 and 0.709 respectively.

Both classifiers were accurate to classify a set of 545 ASF documents (NB of 0.747 and SVM of 0.725) into appropriate categories of relevant (disease outbreak) and irrelevant (economic and general) documents.

Our results show that relevant documents can serve as a source of terms to detect infectious animal disease emergence on the web.

Our method is generic and can be used both in animal and public health domain.

1. Introduction

Textual information sources on the web, such as publically available news articles, official disease reports and newsletters, have been found informative for early detection of emerging infectious disease outbreaks. Over the years, several web focused,

* Corresponding author at: French Agricultural Research and International Cooperation Organization (CIRAD), Unit for Control of Exotic and Emerging Diseases in Animals (UMR CMAEE), Campus international de Baillarguet, 34398 Montpellier, France.

E-mail addresses: elena.arsevska@cirad.fr (E. Arsevska), mathieu.roche@cirad.fr (M. Roche), Pascal.HENDRIKX@anses.fr (P. Hendrikx), david.chavernac@cirad.fr (D. Chavernac), sylvain.falala@cirad.fr (S. Falala), renaud.lancelot@cirad.fr (R. Lancelot), bdufour@vet-alfort.fr (B. Dufour).

http://dx.doi.org/10.1016/j.compag.2016.02.010
0168-1699/

event-based biosurveillance systems (further in the text web monitoring systems) have been created in order to detect infectious disease outbreak information from articles published on the web (Collier et al., 2008; Freifeld et al., 2008; Mykhalovskiy and Weir, 2006; Steinberger et al., 2008).

Despite the great potential in detection of early signals of infectious disease emergence from diverse web sources, the timeliness in detection of relevant articles is challenging due to the vast amount of ever growing publications on the web. Barboza et al. (2013) have shown that due to the access to diverse information, the web monitoring systems can detect avian influenza epizootics 12.7 days before the official notification to the World Organisation for Animal Health (OIE). In January 2014, an online post on the ProMED-mail system, referred to a local news media in Lithuania which reported complaints by hunters on increased mortality in wild boars at the border line with Belorussia (ProMED-mail, 2014). These reports are probably among the first signals of the spread of African swine fever (ASF) to a new territory well before official government reports were issued (OIE, 2014).

Therefore, automated identification of relevant articles on the web is the first step towards an effective event-based biosurveillance. In order to increase specificity in detection of relevant articles on the web, the current web monitoring systems widely use disease related search terms and Boolean queries (using the operators AND, OR, AND NOT, e.g., "african swine fever" OR "swine fever" AND NOT "classical swine fever") and proposed by domain experts (Mantero et al., 2011), trained analysts (Mykhalovskiy and Weir, 2006) or based on a medical ontology (Collier et al., 2010). However, up to this point, no detailed work exists on how the current web monitoring systems identify the terms to detect signals of infectious disease emergence, especially in animal health. Moreover, identification of disease related vocabulary in animal health, faces additional challenges, such as multiple clinical signs and multiple hosts (Santamaria and Zimmerman, 2011; Smith-Akin et al., 2007).

Limited number of studies exploited the text mining approaches in order to construct terminology for infectious animal diseases (Anholt et al., 2014; Arsevska et al., 2014; Furrer et al., 2015) and therefore we propose an innovative methodology for identification of terms to build queries for monitoring the web for new and exotic infectious animal disease outbreaks. The method is based on automatic extraction of terms from relevant corpora of disease outbreak articles and identification of relevant terms using domain expert knowledge. The method is based among other on machine learning techniques and on a new function for ranking of the automatically extracted terms.

The methodology that we propose is generic and can be applied both to animal and public health domain.

For our experiments we use data on ASF. We choose this disease, because it is highly contagious and mortal in porcine animals; it has neither vaccine nor treatment and due to trade barriers the affected countries suffer great economic losses. This disease, endemic in sub-Saharan Africa and Sardinia (island in Italy) is an emerging threat to the European countries after its introduction for the first time in 2007 in the Caucasus region of Europe (Sánchez-Vizcaíno et al., 2013).

This rest of the work is organised as follows: Section 2 presents the related work, Section 3 presents our methodology, Section 4 presents our experiments and the results, Section 5 discusses the results, and Section 6 concludes the paper.

2. Related work

The earliest automatic web monitoring system, the Global Public Health Intelligence Network (GPHIN) founded by the Public Health Agency of Canada in 1997, in order to detect disease outbreak articles of potential relevance uses mainly two news aggregator feeds, Al Bawaba which covers information from the Middle East and North Africa and Factiva which covers information from more than 32,000 web sources worldwide. Once detected, articles are selected using a scanning tool based on a custom-built taxonomy of search terms and Boolean queries, updated regularly by GPHIN's human analysts. The search terms have a relevance score automatically attributed to the retrieved article; and an article is rejected if it has a relevance score below an established threshold (Keller et al., 2009; Mykhalovskiy and Weir, 2006).

The Medical Information System (MedISys), founded by the Joint Research Centre (JRC) of the European Commission (EC) in 2004, retrieves articles from diverse web sources, such as more than 150 medical web sites and 1000 news portals worldwide in 40 languages. The retrieval is based on a list of predefined multilingual terms for each disease included in the system (alert definitions) and a combination of search terms, as proposed by domain experts. The MedISys covers more than 200 alert definitions for different public health-related subjects. A dedicated algorithm developed by the JRC team scans in real time the incoming articles for the alert definitions and keeps the articles that satisfy those criteria. An article is kept by the system and displayed in the disease category for dengue if the term "dengue" appears at least three times in the text of the article. However, if the text of the article includes an irrelevant term (proposed by experts), such as the term "concert", the article will not be selected. An article is also selected if two combinations of relevant search terms appear in the text of the article, such as "dengue" and "outbreak" and regardless of how many times they appear in the news article (Mantero et al., 2011; Steinberger et al., 2008).

The Argus system, hosted at the Georgetown University Medical Centre, United States, since 2004 detects events on the web that might threaten the human, plant, and animal health globally, except United States. It collects, in an automated process, media news pages, including blogs and official sources, such as World Health Organisation (WHO) and OIE, and interprets their relevance according to a specific set of concepts, search terms and Boolean queries relevant to each infectious disease covered by the system. Argus does not use scientific journals as a primary source to identify emerging events. Regional experts, collectively fluent in more than 40 languages, review manually the acquired articles before they are posted in the system (Nelson et al., 2010).

HealthMap, founded by the Boston Children's Hospital in 2006, draws terms from a continually expanding dictionary of pathogens (human, plant, and animal diseases) and geographic names (country, province, state, and city). Using a Bayesian classifier, articles are categorised by disease and location, automatically tagged according to their relevance and then overlaid on an interactive geographic map (Freifeld et al., 2008). The web search criteria used by HealthMap include disease names (scientific and common), symptoms, keywords, and phrases in seven languages (Brownstein et al., 2008). The system integrates outbreak data from multiple electronic sources, including online news aggregators (e.g., Google News), Really Simple Syndication (RSS) feeds, expert curated accounts (e.g., ProMED-mail), multinational surveillance reports (e.g., Eurosurveillance), and validated official alerts (e.g., from WHO and OIE) (Keller et al., 2009).

The conceptual framework for the BioCaster project founded in 2006 is a multilingual ontology – a structured public health vocabulary and terms, such as names of diseases, agents, clinical signs, syndromes and hosts, as well their relations, such as clinical signs or pathogen agents which affect a particular host. Terms are identified for eight Asia–Pacific languages by domain experts in biology, epidemiology, genetics and computational linguistic and linked with sources such as ICD10, MeSH, SNOMED CT and