Monitoring Disease Outbreak Events on the Web Using Text mining Approach and Domain Expert Knowledge

Elena Arsevska[1,2], Mathieu Roche[3,4], Renaud Lancelot[1,2], Sylvain Falala[1,2], David Chavernac[1,2], Pascal Hendrikx[5], Barbara Dufour[6]

[1] CIRAD, UMR CMAEE, Montpellier, France
[2] INRA, UMR CMAEE, Montpellier, France
[3] CIRAD, UMR TETIS, Montpellier, France
[4] Montpellier University, CNRS, LIRMM, UMR 5506, Montpellier, France
[5] ANSES, UCAS, Maisons-Alfort, France
[6] EnvA, Maisons-Alfort, France
{elena.arsevska, mathieu.roche, renaud.lancelot, sylvain.falala, david.chavernac}@cirad.fr,
pascal.hendrikx@anses.fr, barbara.dufour@vet-alfort.fr

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. We propose a generic method to enrich and extend the use of different expressions as queries in order to improve the acquisition of relevant disease related pages on the web. Our method combines a text mining approach to extract terms from corpora of relevant disease outbreak documents, and domain expert elicitation (Delphi method) to propose expressions and to select relevant combinations between terms obtained with text mining. In this paper we evaluated the performance as queries of a number of expressions obtained with text mining and validated by a domain expert and expressions proposed by a panel of 21 domain experts. We used African swine fever as an infectious animal disease model. The expressions obtained with text mining outperformed as queries the expressions proposed by domain experts. However, domain experts proposed expressions not extracted automatically. Our method is simple to conduct and flexible to adapt to any other animal infectious disease and even in the public health domain.

Keywords: digital disease detection, text mining, Delphi method

1. Introduction

Emerging infectious disease outbreaks (disease outbreaks) are an incising threat to human and animal population in non-affected countries due to globalisation, movement of passengers and international trade. Traditional disease surveillance systems are based on a structured multilevel health infrastructure, which can lead to delays in reporting of disease outbreaks (Chan et al., 2010). As a complement to the traditional disease surveillance systems, recently created biosurveillance systems focus on timely detection of disease outbreaks by monitoring the information published on the web. In order to acquire information about disease outbreaks, the common biosurveillance system use terms issued from medical dictionaries or terms proposed by experts (Mantero et al., 2011; Freifeld et al., 2008; Collier et al., 2007).

Currently, no precise work presents how the biosurveillance systems identify the terms for monitoring the web. The work presented in this paper fills this missing gap.

We investigate the use of domain expert knowledge and a text mining approach to propose expressions in order to improve the acquisition of web pages with information about disease outbreaks. We evaluate our approach in the animal health domain and one disease in particular, African swine fever (ASF).

2. Methods

The work presented in this paper is part of the methodology that we currently develop for the French epidemic intelligence team in animal health (VSI[1]). The VSI team focuses in monitoring animal disease outbreaks which occur outside France. Our contribution is to detect early signals of animal disease emergence from the information published on the web. Figure 1 illustrates the framework of our methodology.

To acquire web pages (documents), we use web mining techniques based on expressions, such as names of diseases and associations between terms, such as clinical signs and hosts (step 1, Figure 1). For automatic classification of retrieved documents we use supervised approaches in machine-learning, i.e. Naive Bayes (NB) and Support Vector Machine (SVM) classifiers, previously trained on a corpus of documents labelled according to the content by veterinary epidemiologists - user specialists (step 2, Figure 1). The learnt models serve as a basis to classify new documents into relevant - disease outbreak documents, and irrelevant - economy and general documents. In a previous work we showed that the classifiers correctly categorized 545 ASF documents into appropriate categories (disease outbreak, economy and general) with an accuracy of 0.75 for NB and 0.73 for SVM (Arsevska et al., 2016b). Lastly, using text mining techniques, we extract information from the relevant (disease outbreak) documents, such as name of the disease, date, and place of the event, affected hosts and clinical signs (step 3, Figure 1).

In this paper, we focus on the first step of the methodology. More precisely, we propose a method to enrich and extend

[1] Veille Sanitaire Internationale

the use of the different expressions as queries. We investigate a text mining approach (Section 2.1) and domain expert knowledge (Section 2.2) as a source of expressions for monitoring the web; and we evaluate the performance of the different expressions to acquire relevant web pages. Figure 2 illustrates the method for identification of terms for monitoring the web.

2.1. Expressions Obtained with Text Mining

Terms are automatically extracted from a corpus of relevant documents for a certain disease (e.g., ASF), using BioTex[2], a tool that combines linguistic and statistic information adapted to biomedical area (Lossio-Ventura et al., 2016). More precisely, using BioTex, the terms are selected based on two principles:
i) use of relevant combination of information retrieval techniques and statistical methods (e.g., TF-IDF, OKAPI and C-value measures), and
ii) use of list of syntactic structures learnt with relevant thesaurus in the biomedical area, such as the Medical Subject Headings (MeSH).
The terms extracted with BioTex can be either simple, one term (e.g., "pig", "fever") or composed, multi-term (e.g. "dead wild boar", "devastating haemorrhagic fever"). We tested our approach on multi-word terms.
Next, a user specialist - veterinary epidemiologist, does a preliminary selection of relevant terms to the disease of interest, i.e. terms that best describe the "clinical sign" and the "host" and constructs associations thereof (e.g., "high mortality" AND "wild boars").

[2]BioTex, BIOmedical Term EXtraction, available from: http://tubo.lirmm.fr/biotex/.

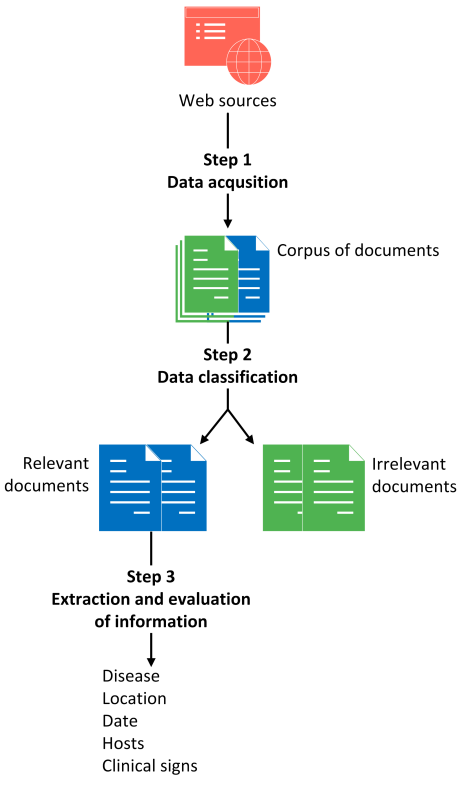


Figure 1: Methodology of the system for monitoring the web

2.2. Expressions Proposed by Domain Experts

Using a Delphi method expert elicitation, domain experts on a certain disease, propose expressions that characterize the disease of interest and can be used as queries for early detection of signals of its emergence on the web. Figure 3 illustrates an example of a Delphi method on - line questionnaire that we used for our experiment (Section 3.2). Delphi method expert elicitations have been successfully applied to answer public health problems, such as to prioritize exotic infectious diseases of importance to the European Union (Economopoulou et al., 2014) or to evaluate the quality of text mining approaches for construction of terminology for syndromic surveillance in animal health (Furrer et al., 2015).
The domain experts evaluate the preselected list of relevant terms and the association thereof (Section 2.1) and validate the expressions which characterize the disease of interest and the queries suitable to build queries.

3. Experiments

As mentioned previously, for this paper we experimented with data on African swine fever (ASF). ASF is a highly contagious and mortal disease in domestic and wild porcine animals; it has no vaccine nor treatment and the main measure for control and eradication is stamping-out. Besides negative implications on the health of the porcine population, affected countries suffer great economic losses due to trade barriers.
ASF is endemic in sub-Saharan Africa and Sardinia (island

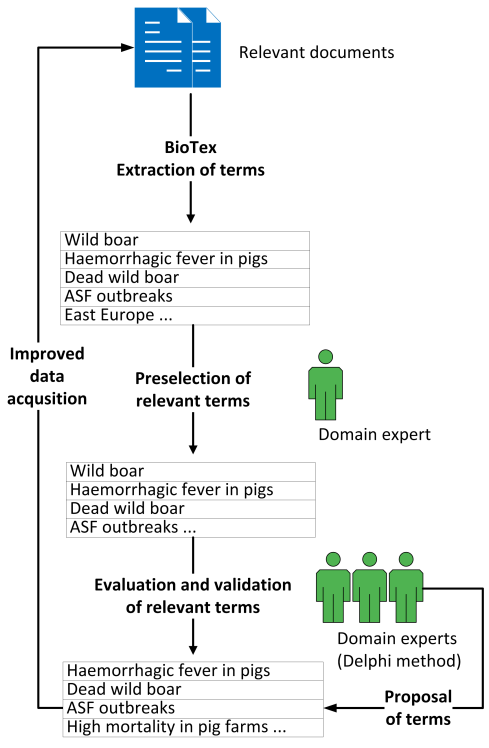
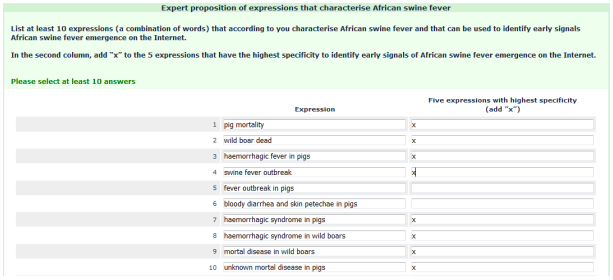


Figure 2: Method for identification of terms for monitoring the web

in Italy); however, in 2007 ASF appeared for the first time in the Caucasus region in Europe and in the following years infected several eastern European countries, including in 2014 several Baltic countries. Therefore, there is a risk of spread of ASF towards other western and southern European countries (Sánchez-Vizcaíno et al., 2013).

3.1. Corpus

To obtain a list of expressions with text mining, we used two corpora of ASF relevant documents (181 Google news articles and 45 PubMed abstracts) - published between 2011 and 2014, which we retrieved in August and September 2014. Using BioTex we extracted 1,200 multi-terms from each corpus. Twenty five terms described an ASF "clinical sign" i.e., terms describing fever, mortality and haemorrhagic clinical signs. Twenty five terms described an ASF "host" (including term synonyms), such as pigs, swine, wild pigs and wild boars. A total of 506 expressions - associations between "clinical sign" and "host" were included in the experiment (e.g., "devastating haemorrhagic fever" AND "domestic pigs", "devastating haemorrhagic fever" AND "wild boars", etc.).
To obtain a list of expressions from domain experts, we elicited 21 experts for ASF using a Delphi method on-line questionnaire (Figure 3, Section 2.2). The experts proposed 186 expressions related to ASF. Fifty three expressions contained an association between "clinical sign" and "host". As written in natural language and where necessary, the expressions were simplified to fit the criteria "clinical sign" and "host", e.g., "haemorrhagic fever in pigs" was simplified as "haemorrhagic fever" AND "pigs". After removing the duplicates, 32 expressions were included in the experiment.


Expert proposition of expressions that characterise African swine fever

List at least 10 expressions (a combination of words) that according to you characterise African swine fever and that can be used to identify early signals African swine fever emergence on the Internet.

In the second column, add "x" to the 5 expressions that have the highest specificity to identify early signals of African swine fever emergence on the Internet.

Please select at least 10 answers

	Expression	Five expressions with highest specificity (add "x")
1	pig mortality	x
2	wild boar dead	x
3	haemorrhagic fever in pigs	x
4	swine fever outbreak	x
5	fever outbreak in pigs	
6	bloody diarrhea and skin petechae in pigs	
7	haemorrhagic syndrome in pigs	x
8	haemorrhagic syndrome in wild boars	x
9	mortal disease in wild boars	x
10	unknown mortal disease in pigs	x

Figure 3: Print screen of the on-line questionnaire, along with one response from a domain expert that proposes expressions which characterize African swine fever

In total, we tested as queries 538 ASF - related expressions obtained with text mining and validated by a domain expert, and expressions proposed by domain experts.

3.2. Evaluation

We evaluated the performance of the expressions as queries on the Google news site in September 2015. We limited our search to news articles published from 2011 until 2014, with all the terms of the expression in the title or/ and the body of the news article.
The performance was analysed on the first 100 URL's retrieved by each expression.

The content of each web page was evaluated as follows:
i) Highly relevant pages (HRP) covered as a main content a suspicion or confirmation of ASF outbreaks;
ii) Closely relevant pages (CRP) covered as a main content a suspicion or confirmation of other disease outbreaks;
iii) Almost relevant pages (ARP) covered as a main content general health issues; and
iv) Irrelevant pages (IP) covered content unrelated to health and disease outbreaks.
This set of elements served as a baseline for evaluation of different parameters of the expressions. However, in this paper we limit our evaluation to the performance as queries of the expressions obtained with text mining and validated by a domain expert, and the expressions proposed by a panel of 21 domain experts (i.e. Delphi method).
For each expression, we calculated precision, recall, and F-score according to threshold of *n* retrieved web pages. Precision was the number of retrieved relevant web pages, divided by the number of retrieved web pages. Recall was the number of retrieved relevant pages, divided by the number of relevant web pages. F-score was the harmonic mean of precision and recall.

3.3. Results and Discussion

From 538 expressions tested, 29 expressions returned HRP web pages, 42 expressions returned HRP + CRP web pages and 58 expressions returned HRP + CRP + ARP web pages. Overall, the average performance results for precision, recall, and F-score in all three categories of *n* retrieved relevant web pages (HRP, HRP + CRP and HRP + CRP + ARP) were higher for the expressions based on text mining approach, compared to the expressions proposed by domain experts. Figure 4 illustrates the results of the performance as queries of the expressions which retrieved relevant web pages.

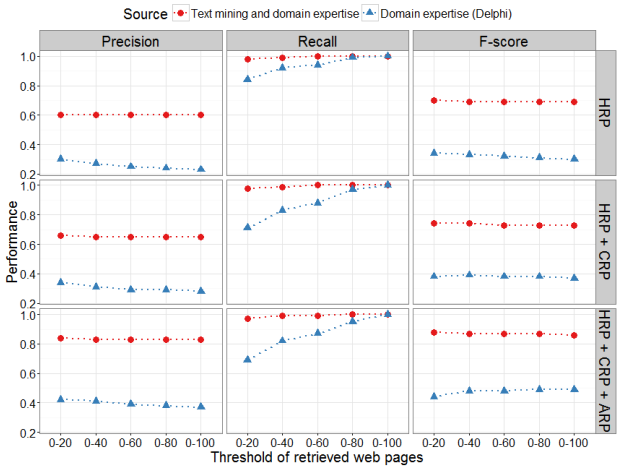


Figure 4: Average performance results (precision, recall and F-score) of the expressions which retrieved relevant web pages. Results are grouped by source of the expressions (text mining and domain expertise or domain expertise), threshold of *n* retrieved web pages (from top 20 to top 100), and content of web pages, i.e., highly relevant (HRP), closely relevant (CRP), and almost relevant (ARP).

In the group HRP, which interested us most, the expressions obtained with text mining had an average precision of 0.6, recall of 0.99 (from 0.95 to 1.0) and F-score of 0.69 (from 0.69 to 0.7); compared to the expressions proposed by the domain experts which had an average precision of 0.27 (from 0.23 to 0.36), recall of 0.92 (from 0.81 to 1.0) and F-score of 0.32 (from 0.3 to 0.38).
In the other groups of retrieved web pages, such as the HRP + CRP, the expressions obtained with text mining had an average precision of 0.65, recall of 0.99 (from 0.95 to 1.0) and F-score of 0.75 (from 0.73 to 0.74); compared to the expressions proposed by the domain experts which had an average precision of 0.31 (from 0.28 to 0.39), recall of 0.86 (from 0.63 to 1.0) and F-score of 0.38 (from 0.37 to 0.39).
In the group HRP + CRP + ARP retrieved web pages, the average precision of the expressions obtained with text mining was 0.83 (from 0.83 to 0.85), recall of 0.99 (from 0.95 to 1.0) and F-score of 0.87 (from 0.86 to 0.88); compared to the expressions proposed by the domain experts which had an average precision of 0.4 (from 0.37 to 0.47), recall of 0.84 (from 0.52 to 1.0) and F-score of 0.47 (from 0.4 to 0.5).
The 22 expressions obtained with text mining which retrieved HRP web pages described fever (n=21) and mortality (n=1) clinical signs in pigs, wild pigs and wild boars. The seven expressions proposed by the experts which retrieved HRP web pages described fever (n=2), mortality (n=2), haemorrhagic (n=2) and skin/ mucous (n=1) clinical signs in pigs and wild boars. Figure 5 illustrates the expressions which retrieved HRP pages.

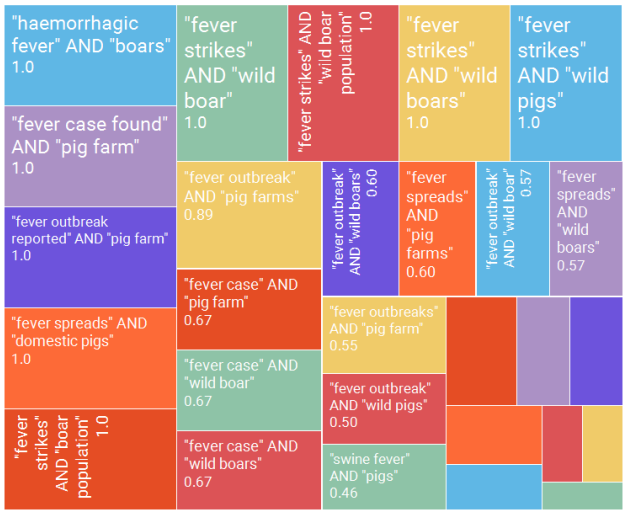


Figure 5: Expressions which retrieved HRP web pages. Each square size corresponds to the F-score for the top ten retrieved HRP web pages.

The high performance results as queries of the expressions obtained with text mining and especially the expressions which retrieved HRP web pages are convenient to us, as for an efficient detection of signals of disease emergence we are looking for news articles with a content about suspected or confirmed disease outbreaks.
Further on, the text mining approach was easy to apply and it was not time consuming compared to the on-line questionnaire of the Delphi method expert elicitation which took more than one month from formulation of the questions, answering of the participants and reaching a consensus. In future we intend to eventually modify the Delphi method by conducting personal interviews with a smaller number of domain experts (Arsevska et al., 2016a).
Finally, we believe our method gives a valuable source of terminology for syndromic surveillance in animal health - monitoring of different clinical signs in animal hosts for the purposes of early detection of signals of disease outbreaks. As noted by other authors (Santamaria and Zimmerman, 2011; Smith-Akin et al., 2007; Furrer et al., 2015), and our previous work (Arsevska et al., 2016a; Arsevska et al., 2016b), text mining approaches in animal health face challenges such as multiple hosts and less formal vocabulary. That is why we currently evaluate our approach on four more animal infectious diseases, Schmallenberg, Foot-and-mouth disease, Bluetongue and Avian influenza, and in two more languages (besides English), that is Spanish, and French.

4. Conclusion

Monitoring the web based on expressions which describe clinical signs and hosts may help detect early signals of disease emergence ahead of traditional surveillance methods.
We have presented a combined approach (text mining and domain expert knowledge) for identification of expressions to build queries for improved monitoring of disease emergence on the web. Our approach is generic and applicable to other animal infectious diseases and even in public health domain.

5. Acknowledgements

We thank the experts that participated in the Delphi method. This work was supported by a grant from the French Ministry of Agriculture, Food and Forestry (DGAL), the French Agricultural Research Centre for International Development (Cirad), and SONGES Project[3] (FEDER and Languedoc-Roussillon).

6. Bibliographical References

Arsevska, E., Roche, M., Hendrikx, P., Chavernac, D., Falala, S., Lancelot, R., and Dufour, B. (2016a). Identification of associations between clinical signs and hosts to monitor the web for detection of animal disease outbreaks. *International Journal of Agricultural and Environmental Information Systems*, In press, March.

Arsevska, E., Roche, M., Hendrikx, P., Chavernac, D., Falala, S., Lancelot, R., and Dufour, B. (2016b). Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web. *Computers and Electronics in Agriculture*, (123):104–115, April.

Chan, E. H., Brewer, T. F., Madoff, L. C., Pollack, M. P., Sonricker, A. L., Keller, M., Freifeld, C., Blench, M., Mawudeku, A., and Brownstein, J. S. (2010). Global capacity for emerging infectious disease detection. *Proc Natl Acad Sci U S A*, 107(50):21701–21706, December.

[3]http://textmining.biz/Projects/Songes

Collier, N., Kawazoe, A., Jin, L., Shigematsu, M., Dien, D., Barrero, R. A., Takeuchi, K., and Kawtrakul, A. (2007). A multilingual ontology for infectious disease surveillance: rationale, design and challenges. *Language Resources and Evaluation*, 40(3-4):405–413, October.

Economopoulou, A., Kinross, P., Domanovic, D., and Coulombier, D. (2014). Infectious diseases prioritisation for event-based surveillance at the European Union level for the 2012 Olympic and Paralympic Games. *Euro Surveill.*, 19(15).

Freifeld, C. C., Mandl, K. D., Reis, B. Y., and Brownstein, J. S. (2008). HealthMap: global infectious disease monitoring through automated classification and visualization of Internet media reports. *J Am Med Inform Assoc*, 15(2):150–157, April.

Furrer, L., Küker, S., Berezowski, J., Posthaus, H., Vial, F., and Rinaldi, F. (2015). Constructing a Syndromic Terminology Resource for Veterinary Text Mining. *Proceedings of the conference Terminology and Artificial Intelligence 2015*, pages 61–70.

Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2016). Biomedical term extraction: overview and a new methodology. *Information Retrieval Journal*.

Mantero, J., Belyaeva, J., and Linge, J. (2011). How to maximise event-based surveillance web-systems the example of ECDC/JRC collaboration to improve the performance of MedISys. Technical report, Publications Office, Luxembourg.

Sánchez-Vizcaíno, J. M., Mur, L., and Martínez-López, B. (2013). African swine fever (ASF): five years around Europe. *Vet. Microbiol.*, 165(1-2):45–50, July.

Santamaria, S. L. and Zimmerman, K. L. (2011). Uses of Informatics to Solve Real World Problems in Veterinary Medicine. *Journal of Veterinary Medical Education*, 38(2):103–109, June.

Smith-Akin, K. A., Bearden, C. F., Pittenger, S. T., and Bernstam, E. V. (2007). Toward a veterinary informatics research agenda: an analysis of the PubMed-indexed literature. *Int J Med Inform*, 76(4):306–312, April.