PADI-web: un système automatique multilingue pour la veille sanitaire internationale en santé animale

Sarah Valentin[1,2], Julien Rabatel[3], Elena Arsevska[1,3], Sylvain Falala[1,3], Jocelyn de Goer[4], Alizé Mercier[1,3], Renaud Lancelot[1,3], Mathieu Roche[2,3]

[1] UMR ASTRE, Univ. Montpellier, Cirad, INRA, Montpellier, France
sarah.valentin@cirad.fr
[2] UMR TETIS, Univ. of Montpellier, AgroParisTech, Cirad, CNRS, Irstea, Montpellier, France
[3] Cirad, Montpellier, France
[4] INRA, UMR EPIA, Clermont-Ferrand, France

Mots-clés : Santé Animale, Intelligence Epidémiologique, Web, Text Mining.

Conférence visée : Ingénierie des Connaissances (IC)

Détails techniques pour la démonstration : Accès à Internet

1 Introduction

La veille en santé animale a pour objectif l'alerte précoce vis-à-vis de dangers sanitaires connus ou émergents. Elle repose sur le recueil, le suivi et l'analyse quotidienne d'informations issues de sources officielles, telles que l'Organisation mondiale de la santé animale (OIE), et de sources non-officielles telles que les médias ou les réseaux sociaux (Hartley *et al.* (2010)). Plusieurs systèmes de biosurveillance, tels que MedISys (Mantero *et al.* (2011)), GPHIN (Blench (2008)) ou HealthMap (Freifeld *et al.* (2008)), sont ainsi dédiés à l'acquisition et à la diffusion de données issues de sources informelles. Ces systèmes s'intéressent à un large éventail de risques sanitaires (maladies infectieuses humaines, animales ou végétales, risques environnementaux, etc.), mais aucun d'entre eux n'est spécifiquement dédié à la santé animale. De plus, tous reposent sur une modération humaine à une ou plusieurs étapes de leur processus. Dans ce contexte, nous présentons PADI-web [1] (Platform for Automated extraction of Disease Information from the web), un outil de biosurveillance des médias digitaux pour la détection de foyers de maladies animales (Arsevska *et al.* (2018)). PADI-web est intégré dans la thématique de Veille sanitaire internationale, au sein de la plateforme d'Epidémiosurveillance en santé animale [2] (plateforme ESA). Depuis sa première version, dédiée à la veille de sources en anglais, PADI-web a été enrichi d'un nouveau classifieur reposant sur de l'apprentissage automatique et intègre les documents multilingues.

2 PADI-web : de la collecte d'articles à l'extraction d'information

PADI-web repose sur 4 étapes successives permettant d'extraire des informations épidémiologiques à partir du contenu d'articles relatifs à des événements infectieux en santé animale.

1. https ://padi-web.cirad.fr/en/
2. https ://www.plateforme-esa.fr/

2.1 Collecte des articles

L'aspiration des articles est effectuée quotidiennement et de manière automatique via l'agrégateur Google News, grâce à des requêtes intégrées sous la forme de flux RSS. Ces flux sont des combinaisons booléennes de mots-clés développées par une approche combinant l'extraction automatique de termes et la sollicitation d'avis d'experts (Arsevska *et al.* (2016)). Deux types de requêtes sont actuellement implémentés dans PADI-web. Les requêtes spécifiques incluent le nom d'une maladie (par exemple, « avian flu OR avian influenza OR bird flu »), et visent à détecter les événements vis-à-vis de maladies d'intérêt. Les requêtes non-spécifiques consistent en une combinaison de signes cliniques et de noms d'hôtes (par exemple, « abortions AND cows »), et permettent de détecter des événements non-prédéfinis.

2.2 Nettoyage du contenu et traduction

Le contenu des articles aspirés est nettoyé afin d'en supprimer les éléments inutiles (images, hyperliens, publicité, etc.), puis est enregistré dans une base de données accompagné des métadonnées de l'article (nom de la source, date de publication et titre). PADI-web filtre les éventuels doublons en comparant l'url de chaque nouvel article à ceux déjà existants dans la base de données. Les étapes de classification et d'extraction d'information reposant sur des modèles appris en anglais, tous les articles aspirés en une autre langue que l'anglais sont préalablement traduits. La langue source est détectée grâce à la librairie *langdetect* (Python) et la traduction repose sur l'API Translator du système Microsoft Azure. Sur une période de 3 mois, l'intégration des requêtes multilingues a permis d'augmenter le nombre d'articles pertinents de 131% pour la peste porcine africaine (207 articles en anglais, 272 traduits), de 47% pour l'influenza aviaire (212 en anglais, 99 traduits) et de 67% pour la fièvre aphteuse (104 en anglais, 174 traduits).

2.3 Classification

L'étape de classification est une étape cruciale dans le processus de PADI-web, car elle permet de filtrer la quantité d'articles qui seront présentés à l'utilisateur en rejetant les articles non-pertinents (non liés à un danger sanitaire). Le classifieur de PADI-web est issu d'un apprentissage automatique supervisé. Une sélection de différents modèles est entrainée une fois par jour sur un corpus d'apprentissage. Le modèle qui obtient les meilleures performances est sélectionné pour la classification des nouveaux articles (actuellement, il s'agit d'un classifieur de type Random Forest, qui obtient une exactitude (*accuracy*) moyenne de 0.97 en validation croisée). Le corpus d'apprentissage est un corpus annoté de 600 articles (200 articles pertinents et 400 articles non pertinents), pouvant être directement enrichi par l'utilisateur. À partir de l'interface, l'utilisateur peut en effet attribuer une classe à chaque nouvel article, indépendamment de la classe attribuée par le classifieur. Cette fonctionnalité permet de corriger les éventuelles erreurs de classification et d'augmenter facilement le jeu d'apprentissage. De plus, le module est générique : l'utilisateur peut créer autant de nouvelles tâches de classification que nécessaire (sous condition d'inclure un jeu de données annotées pour l'apprentissage). Les classes correspondant à chaque tâche de classification sont attribuées indépendamment les unes des autres par le classifieur.
Depuis sa mise en fonctionnement en février 2016, PADI-web a aspiré plus de 66 000 articles [3], dont 15 000 articles classés comme pertinents. Un échantillon de 100 articles aléatoirement sélectionnés dans la base de donnée de PADI-web a été manuellement évalué par deux épidémiologistes. L'exactitude (*accuracy*) sur cet échantillon est de 0.92.

2.4 Extraction d'information

La dernière étape de PADI-web consiste en l'extraction des indicateurs épidémiologiques dans le contenu des articles pertinents. Ce module est issu d'un apprentissage supervisé dé-

3. Les requêtes multilingues ayant été intégrées récemment, elles ne sont pas comptabilisées.

taillé et évalué par Arsevska *et al.* (2018). Brièvement, les noms de maladie, les hôtes et les symptômes sont détectés grâce à un dictionnaire créé manuellement et régulièrement enrichi, prenant en compte les synonymes pour chaque type d'hôte ou de maladie. Les localisations et les dates sont extraites respectivement grâce au gazetier GeoNames (Ahlers (2013)) et à HeidelTime, un système d'étiquetage d'expressions temporelles à base de règles (Strotgen & Gertz (2010)). Le principe mis en oeuvre est détaillé par Arsevska *et al.* (2018).

3 Interface de PADI-web

3.1 Recherche d'information

Les articles stockés dans la base de données de PADI-web sont consultables via une interface dédiée. Par défaut, les 10 derniers articles aspirés et classés comme pertinents sont affichés. Un large choix de filtres permet à l'utilisateur d'effectuer des recherches plus détaillées. Les articles peuvent être filtrés en fonction de différents attributs tels que la date de leur publication, leur classe (pertinent ou non pertinent) ou encore le nom de leur source. L'utilisateur peut également effectuer sa recherche sur la base du contenu des articles en utilisant les entités épidémiologique extraites (maladie, hôte, etc.) ou en recherchant un mot ou expression de son choix dans le titre ou le corps de l'article.

3.2 Visualisation et annotation

L'utilisateur peut accéder aux métadonnées, aux informations extraites et au contenu de chaque article des résultats d'une requête (Figure 1). Les entités extraites sont listées dans un encart et identifiées dans le texte avec une icône spécifique de chaque type afin de faciliter la visualisation des informations essentielles. Pour chaque entité, une fenêtre contenant des informations complémentaires peut être affichée. Un lien vers Google Maps est associé à chaque entité géographique. A partir de cette interface, l'utilisateur peut manuellement annoter la pertinence de l'article et des entités extraites. Les annotations sont automatiquement enregistrées et prises en compte lors des requêtes ultérieures.

3.3 Exports

Les résultats issus des requêtes peuvent être exportés sous différents formats. Le nombre d'articles correspondant à la requête en fonction du temps peut être visualisé par un histogramme, en utilisant plusieurs niveaux d'agrégation temporelle (par jour, mois ou année). L'utilisateur peut également exporter le jeu de données contenant les entités épidémiologiques extraites, en choisissant parmi différents formats (csv, json ou xls).

4 Conclusion

Nous proposons un outil de biosurveillance dédié à la veille en santé animale et adapté à une utilisation quotidienne par les épidémiologistes. Outre sa spécificité vis-à-vis du domaine vétérinaire, PADI-web repose sur des approches issues d'apprentissage automatique et de fouille de texte permettant de produire des données structurées et directement exploitables par les experts. L'interface permet à l'utilisateur de personnaliser ses requêtes et d'accèder rapidement aux informations pertinentes. Nous envisageons d'enrichir PADI-web d'un module d'extraction de signaux faibles afin d'identifier des informations épidémiologiques fines, telles que les mesures de lutte et de prévention ou les états d'alerte.

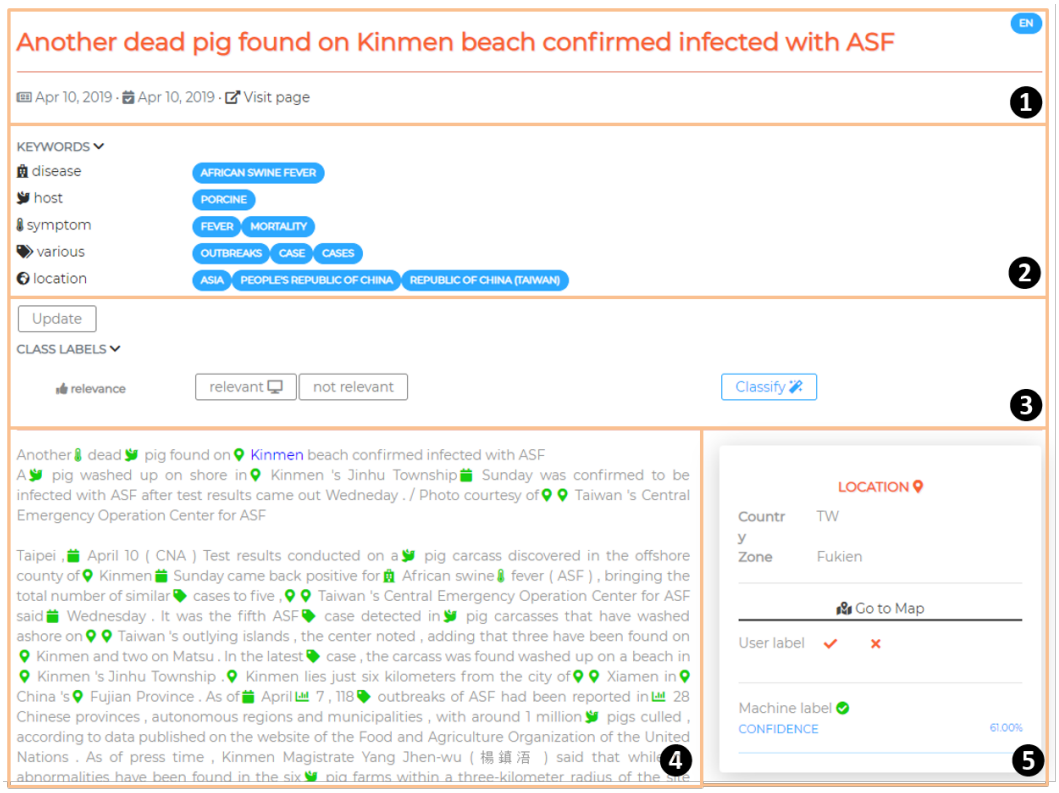

FIGURE *1 – Visualisation d'un article traité par PADI-web, contenant 1. les métadonnées de l'article (titre, date de publication, lien url vers l'article source), 2. la liste des mots-clés tagués, 3. la classe prédite par le classifieur, 4. le texte nettoyé avec les entités épidémiologiques extraites et 5. les informations liées à l'entité géographique sélectionnée 'Kinmen'.*

Références

AHLERS D. (2013). Assessment of the Accuracy of GeoNames Gazetteer Data. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*, p. 74–81, New York, NY, USA : ACM.

ARSEVSKA E., ROCHE M., HENDRIKX P., CHAVERNAC D., FALALA S., LANCELOT R. & DUFOUR B. (2016). Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web. *Computers and Electronics in Agriculture*, **123**, 104–115.

ARSEVSKA E., VALENTIN S., RABATEL J., DE GOËR DE HERVÉ J., FALALA S., LANCELOT R. & ROCHE M. (2018). Web monitoring of emerging animal infectious diseases integrated in the French Animal Health Epidemic Intelligence System. *PLOS ONE*, **13**(8), e0199960.

BLENCH M. (2008). Global public health intelligence network (GPHIN). In *8th Conference of the Association for Machine Translation in the Americas*, p. 8–12.

FREIFELD C. C., MANDL K. D., REIS B. Y. & BROWNSTEIN J. S. (2008). HealthMap : Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports. *Journal of the American Medical Informatics Association*, **15**(2), 150–157.

HARTLEY D., NELSON N., WALTERS R., ARTHUR R., YANGARBER R., MADOFF L., LINGE J., MAWUDEKU A., COLLIER N., BROWNSTEIN J., THINUS G. & LIGHTFOOT N. (2010). The landscape of international event-based biosurveillance. *Emerging Health Threats Journal*, **3**(0).

MANTERO J., BELYAEVA J., LINGE J., EUROPEAN COMMISSION, JOINT RESEARCH CENTRE & INSTITUTE FOR THE PROTECTION AND THE SECURITY OF THE CITIZEN (2011). *How to maximise event-based surveillance web-systems : the example of ECDC/JRC collaboration to improve the performance of MedISys.* Luxembourg : Publications Office. OCLC : 870614547.

STROTGEN J. & GERTZ M. (2010). HeidelTime : High Quality Rule-Based Extraction and Normalization of Temporal Expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, p. 321–324.