Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 126 (2018) 490–497

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2018, 3-5 September 2018, Belgrade, Serbia

How to combine spatio-temporal and thematic features in online news for enhanced animal disease surveillance?

Sarah Valentin[a,b,c,*], Renaud Lancelot[a,b], Mathieu Roche[a,c]

[a] *CIRAD, Montpellier, France*
[b] *ASTRE, CIRAD, INRA, Univ Montpellier, Montpellier, France*
[c] *TETIS, APT, CIRAD, CNRS, IRSTEA, Univ Montpellier, Montpellier, France*

Abstract

Early detection of outbreaks of emerging and exotic pathogens is one of the means of preventing the introduction of infectious diseases into unaffected territories. In that context, since 2016, the French Animal Health Epidemic Intelligence team (Veille Sanitaire Internationale, VSI) monitors the online news sources through a designated Platform for Automated extraction of Disease Information from the web (PADI-web). The tool automatically detects, categorizes, and extracts information from online news reports. We focus on the combination of epidemiological features (locations, dates, diseases and hosts) extracted from free text of the news in order to automatically find similarity between different news reports. We describe an original approach based on text mining and data fusion methods and evaluate its performance on a specialized corpus.

Keywords: Animal Disease Surveillance ; Text Mining ; Fusion

1. Introduction

The international monitoring of animal health traditionally relies on analysis of infectious animal disease outbreak (hereafter referred to as disease outbreak) data from the official notifications to the World Organization for Animal Health (OIE). The official notification usually involves several administrative steps from onset to confirmation of disease outbreaks, thus limiting the awareness of unaffected countries for the potential risks of disease spread [1]. During the last two decades, several studies have highlighted the value of online news sources for global monitoring of animal (public) health threats [1,2]. While official sources share data only for a list of notifiable diseases to the OIE, unofficial sources are more flexible and inform on a variety of disease outbreak information, ranging from news on unexplained animal mortality to news on notifiable and non-notifiable diseases to the OIE and prevention and surveillance strategies against disease spread [3]. Therefore, querying the online sources allows detection of a variety of relevant news from multiple sources and languages [4]. In order to assist the French Animal Health Epidemic Intelligence team (Veille Sanitaire Internationale, VSI) a platform dedicated to automatic surveillance of online news sources, PADI-web (Platform

* Corresponding author.
E-mail address: sarah.valentin@cirad.fr

10.1016/j.procs.2018.07.283

for Automated extraction of animal Disease Information from the web), has been created[5]. This tool automatically detects, categorizes and extracts epidemiological information from online news sources (diseases, hosts, symptoms, dates, locations and number of cases)[6]. Since its launching in 2016, PADI-web has retrieved over than 25,000 news reports. Developing methodologies to find epidemiological relationships between these reports without any human intervention is of crucial importance. Such methods could be used on the daily stream of news reports collected by PADI-web in order to automatically associate related documents. Another practical application would be to analyze the huge amount of animal health data about past outbreaks which are not yet digitized (handwritten notes, newspaper articles) to produce historical reports. Finding suitable methods with regard to veterinary epidemiology is challenging, because this field involves multi-hosts diseases. A host or a disease can have different references depending on the source of the news report. Moreover, the outbreaks usually spread to many countries because of livestock trade and wild animal movements. Thus, news reports often contain a lot of spatial features, with different degrees of granularity and ambiguity.

In this paper, we use different types of epidemiological features mentioned in online news reports and apply data fusion methods to associate news that have similar epidemiological content. We studied the association of several types of features (locations, dates, diseases and hosts) on a corpus of news reports in English language, related to animal disease outbreaks[7]. Section 2 presents the related work on the use of epidemiological features and on data fusion methods in information retrieval. Section 3 presents our global process to combine the epidemiological features and our evaluation approach. Subsection 3.4 discusses the results obtained with the different fusion methods.

2. Related work

2.1. Textual representation and document ranking

Document ranking is often used to find relevant documents among a corpus with all types of documents (relevant and irrelevant). In information retrieval, document ranking is generally an assignment of computing numeric scores between queries and documents, thus allowing the retrieval of a higher number of relevant documents[8]. To calculate these scores, a common approach is to convert the text into a structured representation such as the vector space model[9]. This model allows to use analysis tools from the linear algebra area. Further on, the model encodes a document in a k-dimensional space where each component w_{ij} represents the weight of a term j in a document i. This model neglects the grammatical structure of the text, also referred to as bag-of-words representation[10]. The basic weight method is binary, with the value being 1 or 0, whether a term is present in the document or not. Another common weight method is the $tf - idf$ function, which calculates the weight w_{ij} (see formula (1)).

$$w_{ij} = tf_{ij} \times idf_j = tf_{ij} \times log(\frac{N}{df_j}) \tag{1}$$

where tf_{ij} is the frequency of the term j in the document i, N is the total number of documents from the corpus and df_j is the frequency of the term j in the whole corpus (the number of documents which contain the term).

Once the document is represented by its weighted vector, a large range of measures can be used to the compute its score of similarity with other documents[11]. Among them, a common one is the cosine similarity.

Precisely, the cosine similarity (see formula (2)) between two documents D_a and D_b is.

$$sim(D_a, D_b) = \frac{\sum_{j=1}^{E} w_{aj} \times w_{bj}}{\sqrt{\sum_{j=1}^{E} w_{aj}^2 \times \sum_{j=1}^{E} w_{bj}^2}} \tag{2}$$

where w_{aj} is the weight of the term j in document D_a, w_{bj} is the weight of the term j in document D_b, and E the total number of terms.

2.2. Epidemiological features

In the medical domain, the extraction and use of epidemiological indicators from new data sources have been increasingly studied over the last few years. In human medicine, a large range of sources can be used, such as

chief complaints[12] or more informal ones such as social media[13]. Symptoms can be extracted from unstructured textual data and gathered into syndromes, manually or through text mining methods. Syndromes can be defined as "a combination of clinical signs that repeatedly occurs in different observations, indicating a possible presence of disease"[14]. Monitoring those syndromes should allow to detect an outbreak before it has been diagnosed. However, optimal syndrome definitions adapted to each specific data sources have not yet been determined[15]. In veterinary medicine, data sources are not so numerous and can be more challenging to obtain. Moreover, due to the specificity of the social media user (i.e. the patient himself), there is not such an interest in using those sources to extract animal symptoms. As a consequence, clinical data from practitioners and laboratory data are currently the main sources used in animal syndromic surveillance[16]. An increasingly number of publication evaluate the potential of other data sources such as production data[17] or news reports[5]. The syndromic surveillance approaches are mostly cumulative: an alert is created when a deviation through a baseline level appears (for example, an increase of the mortality rate in cattle). Other initiatives focus on non-cumulative approaches, able to detect weak signals. For instance, weighted vectors were created from health-related tweets in order to gather them into clusters sharing the same content[13]. This approach aimed to detect "latent infectious disease": when a tweet, represented by its vector, could not be associated to any existing cluster, it was a candidate for potential emerging health event. To date, a few papers studied the association of several indicators to improve surveillance[18,19]. In the next section, we present methods used to combine different types of features.

2.3. Data fusion

Data fusion methods (hereafter referred to fusion) are increasingly used in the field of content analysis and information retrieval, especially for diverse and complementary data sources. For instance, fusion methods are used to combine textual and visual data features to improve multimedia retrieval[20]. Currently, there are several types and level fusion strategies. The early and the late fusion methods are the most commonly used[21]. The functions of the early and the late fusion take two inputs, which are the single modality matrices resulting from the feature extraction. The methods differ from each other in the way they integrate the results from the feature extraction.

Early fusion consists of preliminary combination of features into a unique multimodal representation (for instance, textual and visual features[22]). This representation is used as an input for the "decision step". This step, also called "learning phase", can be as simple as the calculus of a similarity matrix[21]. It can also involve more sophisticated approaches such as the manual attribution of scores by experts[20] or the use of machine learning methods[22]. The main advantage of early fusion is that one unique matrix goes through the learning phase, which reduces the computing time and leverages the correlation between the concatenated features. The main disadvantages are the increase of the representation space and the difficulty to combine features into a common representation[21,22].

In late fusion, the decision step is first performed on unimodal features. The outputs of the decision step are then combined into a single final dataset. The advantage is that the features are combined at the same level of representation (for instance, similarity matrices). The main drawbacks are the increase of computing time and the potential loss of correlation[21]. For both early and late fusion, a weight can be applied at the combination step to control the influence of each modality. A visual representation of both methods applied to our process is proposed in section 3.2.

3. Our global process

3.1. Corpus representation

Our corpus is a set of 442 news reports written in English language, all related to animal disease outbreaks. Depending on the news, the strength of the link with a specific disease outbreak is variable: some reports contain very precise information about a new disease outbreak, whereas other describe preventive and control measures or economic consequences against disease outbreaks. The corpus contains information about the news report itself (publication date, title, content, url, etc.) as well as the feature candidates (locations, diseases, hosts and dates). These candidates are automatically identified by the information extraction module of PADI-web, which is based on data mining and rule-based approaches. In this process, each candidate has been manually labelled as correct or incorrect by a veterinary epidemiologist and a computer scientist[5]. The list of unique candidates classified as correct included

59 hosts, 81 diseases, 370 temporal features and 794 spatial features. We can gather these types of features in two main categories: (i) spatio-temporal features (location, date), (ii) thematic features (host and disease features).

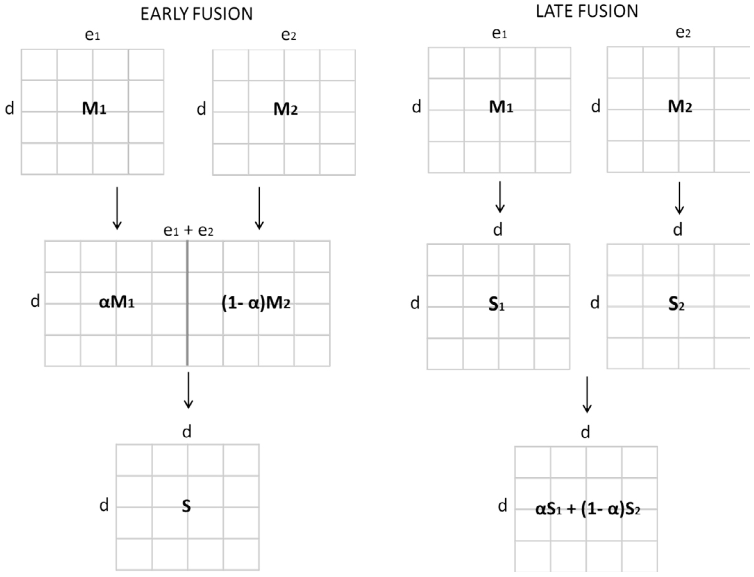
For each type of features, we convert the corpus into a document-term matrix in which the rows represent the document (i.e. the news report) and the columns represent the distinct values of each type of features. The weights used are either boolean (0 or 1), or the $tf - idf$ values as expressed in formula (1).

3.2. Fusion process

3.2.1. Fist step of the fusion process

As the fusion functions take two input matrices, we fuse both the spatio-temporal features and the thematic features (i.e. host and disease features). Depending on the weight used (boolean or $tf - idf$), the matrices have the same range of values. Therefore, with the feature combination, we simply concatenate the matrices. For the "decision step" described in section 2.3 we calculate the cosine similarity. Thus, the outputs of this step are called "similarity matrices". The Figure 1 illustrates the fusion process. In our study, M_1 is the disease feature matrix (resp. the spatial feature matrix) and M_2 is the host feature matrix (resp. the temporal feature matrix). We vary the weight used for the linear combination (α) from 0 to 1.

Fig. 1. Early fusion and late fusion of two matrices, M_1 and M_2, representing d documents (i.e. news reports) with resp. e_1 and e_2 features. S, S_1 and S_2 are the similarity matrices, α is the weight.



3.2.2. Second step of the fusion process

In order to get the final combination, and therefore to evaluate it, we use the output matrices from which we obtained the best results at step 1 and perform a basic weighted linear combination. M_{DH} being the output matrix from the disease-host feature fusion, M_{ST} being the output matrix from the spatio-temporal feature fusion and β being the weight, the final matrix M_F is then calculated as follows (formula (3)):

$$M_F = \beta \times M_{DH} + (1 - \beta) \times M_{ST} \tag{3}$$

In this study, we vary β from 0 to 1.

3.3. Experiments

3.3.1. Evaluation protocol

To evaluate the different fusion models, we consider them as ranking functions. We randomly selected 20 news reports from our corpus described in subsection 3.1. For each of these news, a veterinary epidemiologist manually retrieved all the other relevant news reports from the corpus. Determining whether a report is relevant or not is not trivial, since two documents can be linked at different levels (mentioning the same disease or containing the same geographical references for instance). Moreover, the theoretical spatio-temporal scale to link two events varies depending on the disease. In this study, two news reports are related if they describe exactly the same event (i.e. same disease, same host, same place and same date) or if they describe two events of the same ongoing outbreak (i.e. same disease, same host, but not necessarily same location or date). In the latter, the spatial window can vary, but has to refer to the same country as the original event. For example, with Report 1 being one of the randomly selected news, Report 2 and Report 3 are considered as relevant (Figure 2). On the contrary, Report 4 is considered as irrelevant, since the described event is in a different country.

Fig. 2. Examples of news reports.

Report 1: "Last Updated Friday, February 13, 2015: The Canadian Food Inspection Agency says it has confirmed mad cow disease, or bovine spongiform encephalopathy (BSE), in a beef cow from Alberta."

Report 2: "The Canadian Food Inspection Agency (CFIA) confirmed February 13, 2015 that a case of mad cow disease has been found in Alberta, the first case in Canada since 2011."

Report 3: "South Korean quarantine authorities are examining Canadian beef imports, following a recent notification from the Canadian government of an outbreak of bovine spongiform encephalopathy, or mad cow disease."

Report 4: "A single case of "mad cow disease" has been identified in the Republic of Ireland, from a dairy located in County Louth.The suspect cow was identified early in June 2015, following its sudden death."

For each document from the set of 20 news reports, the values of the output matrices from the second step of the fusion (detailed in subsection 3.2) are used to rank all the other documents from the corpus (in decreasing order of similarity score). Our approach to quantitatively evaluate the ranking is described in subsection 3.3.2.

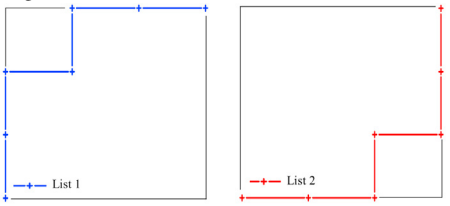
3.3.2. Evaluation criteria to evaluate the fusion process

We evaluate the ranking performance with an approach based on the ROC (Receiver Operating Characteristics) curve. The ROC-curve depicts the trade-off between the maximization of the true positive rate and the minimization of the false positive rate[23]. It indicates the ability of a ranking method to give higher scores to relevant elements (in our case, couples of related articles) than to irrelevant ones (in our case, couples of unrelated articles). The area under the ROC curve (AUC – Area Under Curve), which is an equivalent to the Wilcoxon rank statistics, is then viewed as a global measure of the ranking quality[24]. We illustrate how ROC curves work regarding ranking evaluation with an example. Let L_1 and L_2 be two lists of elements ranked by two different functions. Each element is represented by "+" (i.e. relevant element) or a "-" (i.e. irrelevant element):

- L_1 = {(+),(+),(-),(+),(-),(-)}
- L_2 = {(-),(-),(+),(-),(+),(+)}

As represented in Figure 3, for each "+", the curve increases one unit in the Y-axis direction. For each "-", the curve increases one unit in the X-axis direction. At a consequence, the AUC of the best ranking function (here, corresponding to L_1) is greater than that of a function giving a poorer ranking (here, corresponding to L_2).

Fig. 3. ROC curves obtained from List 1 and List 2.



One of the main advantage of this method is to be resistant to imbalanced data (imbalanced number of relevant and irrelevant elements for instance)[25]. Whatever the proportion of relevant elements, if they all are at the top of the ranked list, the ROC curves will be strictly similar with AUC = 1.

In our context, for each news report A_k, we evaluate the similarity with the other news report A_i where $i \in [1, 20], i \neq k$ (see subsection 3.3.1). So a relevant (resp. irrelevant) element is a relevant (resp. irrelevant) couple A_k–A_i. The average of the AUC for the 20 reports is calculated. Note that with our evaluation dataset, 441 couples are taken into account for analyzing the results presented in the following section.

3.4. Results

3.4.1. First step of the fusion process

In Table 1, the case $\alpha = 0$ corresponds to the use of the host features matrix alone, and the case $\alpha = 1$ corresponds to the use of the disease features matrix alone. In Table 2, the case $\alpha = 0$ corresponds to the use of the temporal features matrix alone, and the case $\alpha = 1$ corresponds to the use of the spatial features matrix alone. In those particular cases, early and late fusion results are strictly similar since the output is the similarity matrix of each feature matrix.

If using only those unimodal similarity matrices for ranking, temporal features obtain the lowest AUC for both boolean (AUC = 0.562) and $tf - idf$ (AUC = 0.564) matrices (Table 2). The best AUC is obtained with spatial features alone (AUC = 0.933). This is consistent with the fact that spatial features have a higher degree of granularity, whereas vagueness is more often present in temporal features (*'last week'*, for example).

Both early and late fusion give very good results, for both boolean and $tf - idf$ matrices (AUC ranging from 0.851 to 0.938). Although all the differences are not necessarily significant, the fusion of disease features with host features slightly increases the AUC. The highest AUC is obtained with the late fusion with $\alpha = 0.6$ (AUC = 0.903) (Table 1). Regarding spatio-temporal features, several combinations of fusion and α values give the best AUC value (AUC = 0.938).

Table 1. AUC comparison of different fusion methods to combine disease and host features.

	Early fusion		Late fusion	
α	AUC(boolean)	AUC($tf - idf$)	AUC(boolean)	AUC($tf - idf$)
0	0.866	0.857	0.866	0.857
0.1	0.883	0.871	0.886	0.875
0.2	0.883	0.879	0.889	0.879
0.3	0.885	0.882	0.893	0.882
0.4	0.888	0.883	0.898	0.883
0.5	0.893	0.884	0.902	0.883
0.6	0.897	0.886	**0.903**	0.882
0.7	0.895	0.888	0.901	0.883
0.8	0.894	0.890	0.899	0.883
0.9	0.894	0.891	0.896	0.887
1	0.896	0.896	0.896	0.896

Table 2. AUC comparison of different fusion methods to combine spatial and temporal features.

	Early fusion		Late fusion	
α	AUC(boolean)	AUC($tf-idf$)	AUC(boolean)	AUC($tf-idf$)
0	0.562	0.564	0.562	0.564
0.1	0.851	0.862	0.856	0.882
0.2	0.853	0.896	0.869	0.906
0.3	0.861	0.924	0.892	0.918
0.4	0.883	0.933	0.911	0.926
0.5	0.918	0.936	0.924	0.93
0.6	0.935	0.937	0.932	0.933
0.7	**0.938**	**0.938**	0.936	0.935
0.8	**0.938**	0.935	0.937	0.936
0.9	**0.938**	0.932	**0.938**	0.937
1	0.932	0.933	0.932	0.933

3.4.2. Second step of the fusion process

We evaluate the combination of the best fused matrix for disease and hosts features (obtained by the late fusion of boolean matrices with $\alpha = 0.6$), with each of the five best matrices obtained by the spatio-temporal fusion : early fusion of boolean matrices with $\alpha = 0.7$ (Mod1), $\alpha = 0.8$ (Mod2) and $\alpha = 0.9$ (Mod3), early fusion of $tf-idf$ matrices with $\alpha = 0.7$ (Mod4) and late fusion of boolean matrices, $\alpha = 0.9$ (Mod5).

Table 3. AUC comparison of different weights to combine spatio-temporal and thematic (i.e. diseases and hosts) features.

β	AUC(Mod1)	AUC(Mod2)	AUC(Mod3)	AUC(Mod4)	AUC(Mod5)
0	0.938	0.938	0.938	0.938	0.938
0.1	0.985	0.984	0.983	0.983	0.984
0.2	**0.986**	**0.986**	0.985	0.984	0.985
0.3	**0.986**	**0.986**	**0.986**	0.984	**0.986**
0.4	0.982	0.984	0.985	0.981	0.984
0.5	0.976	0.978	0.979	0.975	0.978
0.6	0.967	0.971	0.972	0.967	0.97
0.7	0.955	0.958	0.961	0.952	0.957
0.8	0.939	0.941	0.944	0.942	0.941
0.9	0.932	0.933	0.933	0.933	0.933
1	0.903	0.903	0.903	0.903	0.903

The five models give comparable results. We obtain a clear improvement of the AUC values, reaching 0.986 for several β values (Table3). The best AUC values are obtained with low β values, which corresponds to give more weight to the spatio-temporal matrix. This finding is relevant with the high degree of granularity of the spatial features, and with the fact that animal diseases and geographical areas can be highly correlated (e.g., a disease outbreak reported in an Asian country is more likely to be due to avian influenza).

4. Conclusion and future work

In this study, we introduce a method to combine spatio-temporal and thematic features extracted from online news reports, based on document vectorization and data fusion methods. To our knowledge, this is the first time that such combined approach is applied to epidemiological features. The evaluation demonstrates that the proposed method is promising and can have excellent results to retrieve documents sharing the same epidemiological content. In future work, we aim to evaluate other types of fusion methods and similarity measures. We intend to compare our results to current models of word embedding, which allow to capture different semantic similarities between words. We will also continue the manual annotation of the corpus in order to increase the number of labelled relationships between the online news reports.

Acknowledgements

This work was funded by the French General Directorate for Food (DGAL), the French Agricultural Research Centre for International Development (CIRAD) and the SONGES Project FEDER and Occitanie. This work was also supported by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004.

References

1. C. Y. Bahk, D. A. Scales, S. R. Mekaru, J. S. Brownstein, C. C. Freifeld, Comparing timeliness, content, and disease severity of formal and informal source outbreak reporting, BMC Infectious Diseases 15 (1) (2015) 135.
2. F.-J. Tsai, E. Tseng, C.-C. Chan, H. Tamashiro, S. Motamed, A. C. Rougemont, Is the reporting timeliness gap for avian flu and H1N1 outbreaks in global health surveillance systems associated with country transparency?, Globalization and Health 9 (1) (2013) 14.
3. ProMED-mail, Undiagnosed deaths, swine - Lithuania: wild boar, RFI (2014).
URL `http://www.promedmail.org/post/2175896`
4. G. Lejeune, R. Brixtel, A. Doucet, N. Lucas, Multilingual event extraction for epidemic detection, Artificial Intelligence in Medicine 65 (2) (2015) 131–143.
5. E. Arsevska, S. Falala, J. De Goer, R. Lancelot, J. Rabatel, M. Roche, PADI-web: platform for automated extraction of animal disease information from the web, in: Proceedings of LTC - Language and Technology Conference, 2017, pp. 241–245.
6. E. Arsevska, M. Roche, P. Hendrikx, D. Chavernac, S. Falala, R. Lancelot, B. Dufour, Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web, Computers and Electronics in Agriculture 123 (2016) 104–115.
7. J. Rabatel, E. Arsevska, J. De Goer, S. Falala, R. Lancelot, M. Roche, PADI-web corpus: news manually labeled, CIRAD Dataverse, 2017.
8. S. T. Strat, A. Benoit, P. Lambert, H. Bredin, G. Qunot, Hierarchical late fusion for concept detection in videos, in: Fusion in Computer Vision, Springer, 2014, pp. 53–77.
9. M. J. Pazzani, D. Billsus, Content-based recommendation systems, in: The adaptive web, Springer, 2007, pp. 325–341.
10. M. W. Berry, M. Castellanos, Survey of text mining II, Vol. 6, Springer, 2008.
11. W. H. Gomaa, A. A. Fahmy, A survey of text similarity approaches, International Journal of Computer Applications 68 (13).
12. M. Conway, J. N. Dowling, W. W. Chapman, Using chief complaints for syndromic surveillance: a review of chief complaint based classifiers in North America, Journal of Biomedical Informatics 46 (4) (2013) 734–743.
13. S. Lim, C. S. Tucker, S. Kumara, An unsupervised machine learning model for discovering latent infectious diseases using social media data, Journal of Biomedical Informatics 66 (2017) 82–94.
14. R. D. Fricker, B. L. Hegler, D. A. Dunfee, Comparing syndromic surveillance detection methods: EARS versus a CUSUMbased methodology, Statistics in Medicine 27 (17) (2008) 3407–3429.
15. K. Hennings, "What is syndromic surveillance?", Syndromic surveillance: reports from a national conference, Morbidity and mortality weekly report 53 (supplemental) (2003) 7–11.
16. F. C. Dorea, F. Vial, Animal health syndromic surveillance: a systematic literature review of the progress in the last 5 years (2011-2016), Veterinary Medicine: Research and Reports 7 (2016) 157–170.
17. A. Madouasse, A. Marceau, A. Lehébel, H. Brouwer-Middelesch, G. van Schaik, Y. Van der Stede, C. Fourichon, Use of monthly collected milk yields for the detection of the emergence of the 2007 French BTV epizootic, Preventive Veterinary Medicine 113 (4) (2014) 484–491.
18. C. Faverjon, M. G. Andersson, A. Decors, J. Tapprest, P. Tritz, A. Sandoz, O. Kutasi, C. Sala, A. Leblond, Evaluation of a multivariate syndromic surveillance system for West Nile Virus, Vector-Borne and Zoonotic Diseases 16 (6) (2016) 382–390.
19. F. Vial, W. Wei, L. Held, Methodological challenges to multivariate syndromic surveillance: a case study using Swiss animal health data, BMC Veterinary Research 12 (1).
20. S. Clinchant, J. Ah-Pine, G. Csurka, Semantic combination of textual and visual information in multimedia retrieval, in: Proceedings of the 1st ACM international conference on multimedia retrieval, ACM, 2011, p. 44.
21. E.-P. Soriano-Morales, Hypergraphs and information fusion for term representation enrichment. Applications to named entity recognition and word sense disambiguation., Ph.D. thesis, Université de Lyon - Lumière Lyon 2 (2018).
22. C. G. M. Snoek, Early versus late fusion in semantic video analysis, in: In ACM Multimedia, 2005, pp. 399–402.
23. C. Ferri, P. Flach, J. Hernndez-Orallo, Learning decision trees using the area under the ROC curve, in: ICML, Vol. 2, 2002, pp. 139–146.
24. H. Saneifar, S. Bonniol, P. Poncelet, M. Roche, From terminology extraction to terminology validation: an approach adapted to log files, Journal of Universal Computer Science 21 (4) (2015) 604–636.
25. M. Roche, J. Azé, Y. Kodratoff, M. Sebag, Learning interestingness measures in terminology extraction. A ROC-based approach., in: ROCAI, 2004, pp. 81–88.