Mining web data for epidemiological surveillance

Didier Breton[1], Sandra Bringay[2 3], François Marques[1],
Pascal Poncelet[2], Mathieu Roche[2]

[1] Nevantropic, France
[2] LIRMM – CNRS, Univ. Montpellier 2, France
[3] MIAp Department, AMIS Group, Univ. Montpellier 3, France

Abstract. Epidemiological surveillance is an important issue of public health policy. In this paper, we describe a method based on knowledge extraction from news and news classification to understand the epidemic evolution. Descriptive studies are useful for gathering information on the incidence and characteristics of an epidemic. New approaches, based on new modes of mass publication through the web, are developed: based on the analysis of user queries or on the echo that an epidemic may have in the media. In this study, we focus on a particular media: web news. We propose the EPIMINING approach, which allows the extraction of information from web news (based on pattern research) and a fine classification of these news into various classes (new cases, deaths...). The experiments conducted on a real corpora (AFP news) showed a precision greater than 94% and an F-measure above 85%. We also investigate the interest of tacking into account the data collected through social networks such as Twitter to trigger alarms.

1 Introduction

In the context of epidemiological surveillance, the analysis of relevant information is crucial to the decision-making process when an expert has to decide to trigger or not an alarm. The question we tackle in this article is the following: Can the flow of information exchanged on the Web be used to improve the daily monitoring of the epidemiological reality that epidemiologists themselves sometimes have difficulty to establish?

Health professionals can use news as new resources of information. However, they have to deal with the abundance of information. How to sort efficiently this pool of resources, to keep only the relevant information according to a specific issue?

The work presented in this paper is based on a collaboration between the Nevantropic company and the LIRMM laboratory. The company focuses on the development of operational monitoring of the environment at local and regional scales. In this collaboration framework, we are particularly interested in the automatic tracking of the evolution of H1N1 from heterogeneous resources of the Web. Our goal is to extract knowledge from news to provide new indicators for health authorities in order to assist

them in the decision-making process. In this paper, we present a method for automatic detection of weak signals (task of epidemiological surveillance) from a news series. This method is based on pattern research to extract information from a news corpus and on the classification of these annotated sentences of news into topics (e.g. news cases, death...). We also investigate the interest of tacking into account the data collected through social networks such as Twitter to trigger alarms.

Our contribution comes in treefold: (1) to annotate the news according to a set of concepts; (2) to classify the news into categories, (2) to identify, count and locate the cases associated with an epidemic thanks to this classification. A brief state-of-the-art is presented in Section 2. In Section 3, we present the EPIMINING approach. The conducted experiments are described in Section 4, and discussed in Section 5. In Section 6, we discuss about the information that can be obtained from social networks and mainly focus on Tweets related to disease. We thus illustrate how such an information can be useful for improving the monitoring. Finally in Section 7, we conclude with future work.

2 Background

2.1 Context

Agencies managing the traditional systems of epidemiological surveillance (e.g. Institut National de Veille Sanitaire in France, European Influenza Surveillance Schema, US CDC Center for Diease Control and Prevention[4]) generally use virologic data, clinical information from medical reports or pharmacies in order to monitor an epidemic. For example, in France, one of the objectives of the Sentinel Network[5] composed of physicians and pharmacists is to monitor, according to the medical consultations, various diseases (e.g. asthma, diarrhea, influenza-like illness...). Even if these approaches are very effective, the proposed analyzes only focus on the events of the previous weeks and only few approaches are able to monitor outbreak in real-time [1].

Recently, Yahoo and Google have proposed systems which take advantage of the mass of information now available online for epidemiological surveillance. In 2008, [2] have examined the relationship between searches for influenza and actual influenza occurrences, using search queries from the Yahoo! search engine. The principle is based on the assumption that when a person has disease symptoms, he tends to query the Web like: "What are the symptoms of this disease?", "Which web sites deal with this disease?". Using the keywords chosen by the web users and their location, it is possible to define what are the trends of the users and

[4] http://www.invs.sante.fr/,
http://www.ecdc.europa.eu/en/activities/surveillance/EISN/Pages/home.aspx,
http://www.cdc.gov/

[5] http://websenti.b3e.jussieu.fr/sentiweb/

consequently to predict potential outbreaks. [3] made a similar proposal by using the Google search engine to predict in advance the H1N1 epidemic peaks. The results of these two experiments showed that these approaches predicted an increase of the epidemic up to 5 weeks in advance from the US CDC. Even if these approaches are very effective, they require to access to the content of the user's requests and also to have a sufficient number of users to define a prediction model.

2.2 State-of-the-art

Different approaches are based on the extraction of information available in Web documents (news, reports, and so forth) in order to predict knowledge [4–6].

The principle generally used is the following one: From a large volume of Web documents, they extract features such as numbers and location. The collected numbers are often used to display with different colors (more or less dark) information that may be located on a map. For example, systems such as MedISys, Argus, EpiSpider, HealthMap, or BioCaster[6] support the global and real-time monitoring of a disease for a country. These systems are not intended to replace the traditional collaborative systems based on the exchange of official data, but allow to trigger a pandemic alert by integrating data from regions or countries for which official sources are limited or unavailable. However, these approaches suffer from some drawbacks. Because of the aggregate view, it is difficult to monitor an epidemic with a low granularity (time and space). For example, it could be interesting for the epidemiologist to identify which city or village develop new cases instead of having the information for a country. Moreover, most of the systems rarely support a fine result classification (e.g. difference between new cases or deaths). Knowing that in a country, there are occurrences of the H1N1 virus is relevant but, classifying the information retrieved in the news into new cases or new deaths is also informative. Finally, many methods return documents but not relevant segments in these documents. The epidemiologists have to read all the documents to find a section of interest.

In order to predict relevant information, the first stage consists in extracting relevant features in texts. For this extraction process, a lot of methods use patterns [7, 4]. These ones match entity classes by using regular expressions and lists of terms from the studied domain. For instance, the lists include verbs of infection, named entity and so on [4]. To extract information and build knowledge bases of epidemiologic studies, other methods use machine learning approaches [8]. This kind of supervised method has an important limit: a lot of labeled data are necessary

[6] http://medusa.jrc.it,
http://biodefense.georgetown.edu/projects/argus.aspx,
http://www.epispider.org,
http://www.healthmap.org,
http://www.biocaster.org

in order to learn a model.

Our objective in this paper is to address the limitations of the previous approaches. We are interested in the echo that may have an epidemic in the media through news that we classify automatically according to their content into very specific categories (i.e. new cases, new deaths). For this, we first use an extraction method to annotate the news based on pattern recognition and a classification algorithm that takes into account the number of patterns retrieved in the news.
The classification based on an unsupervised approach is not done at the level of the document but at the level of the segments in the documents. Finally, in order to assist the decision-maker, the epidemiologist, we provide different visualizations of the results either as graphical statistics (histogram, pie chart), or as geographical representations of events using GoogleMap.

3 The EPIMINING approach

In this section, we present the overall EPIMINING approach detailed in the Figure 1.

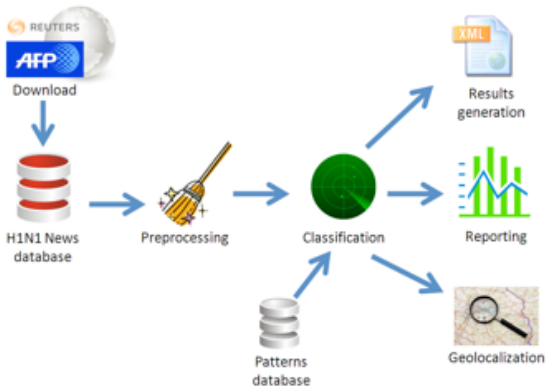

Fig. 1. EPIMINING approach

3.1 Acquisition and pre-processing of the corpus

To feed the News database, we queried sites such as Reuters or the French equivalent AFP. We used keywords associated with the disease (e.g. swine flu, H1N1, influenza...). We tokenize and tag words that appear in the retrieved news with the TreeTagger tool [9]. For example, let us consider the subpart of the second sentence of the news presented in figure 2:

Fig. 2. An example of H1N1 news.

"10 deaths had occurred in adults all under the age of 65 in England"

The associated lemmatized sentence, composed of the original form of each word (i.e. first element), the grammatical category (i.e. second element) and the lemma (i.e. third element) is:

"10/CD/Card deaths/NNS/death had/VHD/have occurred/VVN/occur in/IN/in adults/NNS/adult all/RB/all under/IN/under the/DT/the age/NN/age of/IN/of 65/CD/Card

3.2 Annotation of the news

Pre-treated news are automatically annotated thanks to a Pattern Database which enables to identify the relevant concepts. We apply an approach similar to the one described in [10] who details different Information Extraction (IE) tools in order to find specific information in free texts. Like our method, the developed tools used patterns associated to

part-of-speech knowledge. Note that the EPIMINING system described in this paper is more specific to the epidemiology domain. To recognize patterns in documents, we rely on their linguistic characteristics and other syntactic rules of their arrangement. Specifically, the tagged documents are parsed in order to detect patterns. The analysis started by applying a set of syntactic rules to locate all the patterns present in different sections of the document. A filter is then applied to favour the longest pattern among several patterns sharing the same lemmatised words. For example, in the sentence of the Figure 2, we identify the concept PERSON thanks to the presence of the lemmatized word "adult". Similarly, the concept YEARS_OLD is retrieved via the pattern series: <PERSON> followed by the expression "under the age of" followed by the number 65.

The Pattern Database is composed of patterns specified by an expert. These patterns were identified after a textual analysis of the news content. They take into account the specificities of the news regarding the other types of text documents. Applying this method, we identify in the previous example the concepts: NUMBER, DEATH, PERSON, YEARS_OLD, CITY. To refine the information about the location in the news, we use a database of geographic information (Geolocalisation database). After this step, documents are labelled for an easier classification: When this was possible, each sentence is associated with a number of sick and dead people, a location, a date... Finally, we obtain the following annotations:

"<NUMBER>10</NUMBER><PERSON><DEAD>death</DEAD>
<AGE> under 65</AGE></PERSON> <CITY> London </CITY>".

3.3 Classification

A news can contain information, which can be classified into various categories. For example, we can find in the same new information about sick and dead patients. Consequently, the news classification at the document level is often not relevant. To obtain a fine classification, we decide not to classify the news but sentences of these news. The classification is performed as follows: Each class is associated with a set of patterns. If patterns of a defined class are retrieved in a news, the one is associated to this class. For instance, the news of the Figure 2 is associated to the class Death because we have found the news with the concept DEAD. For each association between a sentence and a class, we calculate the EPIMINING score according to the following heuristic. The score equals 1 if all the elements that are expected are found in the sentence (e.g. for the association between a sequence containing a date, a number of death, a geographical location and class Death). The score is based on the reliability of extracted information. For example, if the location is not in the sentence, the search is expanded in the nearest sentence to find the missing information and the score is decreased.

4 Experiments

In order to evaluate the performance of our approach, two data sets in French were used for experiments: a database of 510 AFP news over the period of September 2009 to February 2010 and a database of 353 Reuters news over the period of January 2009 to February 2010. To analyze the quality of the returned results, 477 AFP news, and 7147 sentences have been manually annotated. The objective was to evaluate the news classification into four categories. The first two ones depends on when the cases mentioned in the news are listed: "New cases" corresponds to the description of information about new patient at a given time and "Report" corresponds to older cases. The last two categories correspond to the categorization of the patient: "Dead patient" and "Sick patient". Two types of evaluations have been conducted (1) by considering the documents as objects to be classified and (2) by considering the sentences. To evaluate the results of these two classifications, we measure the precision (ratio of relevant documents found on the total number of selected documents), recall (ratio of relevant documents found in the total number of relevant documents), and the F-measure (harmonic average between precision and recall).

In Table 1, results of the tests conducted on the news classification are reported. The best results are obtained for classes "Report" and "Dead patient". This is justified by the fact that the distinction between illness and death is not always present in the news and by the fact that the concept of novelty is more difficult to detect. Even when the analysis is conducted by an expert, the difference between the two classes is not necessarily obvious to capture.

Table 2 presents the experiments conducted on the classification of the sentences. We worked with different EPIMINING score values corresponding to the search for patterns in different sentences close to the evaluated segment. With a high confidence score (i.e. [50..100]), we obtain the best precision (83.6%).

Finally, Tables 1 and 2 show that the EPIMINING approach focuses on precision. Indeed, the patterns are often quite restrictive to return relevant elements. To increase the recall, we can consider the sentences with a large EPIMINING confidence as shown in Table 2 .

5 Discussion

A prototype dedicated to healthcare professionals was set up by the Nevantropic company. Figure 3 shows an interface of this tool that presents various indicators that can be used for decision-making.
On the left, the evolution of the number of cases of sick and died people identified through the news dealing with H1N1 are presented over several months or years. On the right, the cases are located on a GoogleMaps at a given time. Of course, only the indicators derived from the news

Classes	Retrieved and relevant	Retrieved	Relevant	Precision	Recall	F-Measure
Dead	100	106	128	94.3%	78.1%	85.5%
Ill	43	55	65	78.2%	66.2%	71.7%
Report	88	103	114	85.4%	77.2%	81.1%
New	48	59	78	81.4%	61.5%	70.1%

Table 1. News classification

Confidence	Retrieved and relevant	Retrieved	Relevant	Precision	Recall	F-Measure
[0..25[	20	46	280	43,5%	7,1%	12,3%
[25..50[	58	97	280	59,8%	20,7%	30,8%
[50..100]	112	134	280	83,6%	40,0%	54,1%
[0..100]	190	277	280	68,6%	67,9%	68,2%

Table 2. Sentence classification

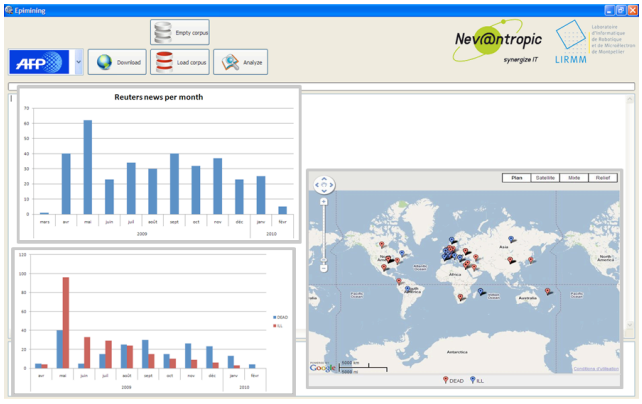

Fig. 3. EPIMINING tool

classification are presented on this image but of course, it is the combination of several indicators that make sense for healthcare professionals who must take a decision. For example, the tool can be used by epidemiologists who should or should not trigger an alarm or by physicians to guide their diagnosis during the visit of a patient in suspected cases of epidemic in the country where he travel back. The proposed architecture for the monitoring of H1N1 is of course adaptable to other types of epidemics.

The type of approach presented in this paper, based on the exploitation of massive data published on the Web, like the approaches proposed by Yahoo and Google, are relevant because they help to early alert health authorities. The results of these methods must be considered as new and indispensable sources of information that have to be crossed with more traditional sources of information provided by the agencies managing the traditional systems of epidemiological surveillance, either to confirm, disprove or in most cases to clarify. These methods are especially useful in geographic areas that do not have a conventional surveillance infrastructure but where the deployment of the Internet is already well advanced.

6 What's about a more real time information?

In this section, we consider another kind of information that can be very useful for helping to evaluate the propagation of epidemics. In the previous sections, we focused on information available on news. That means that this information is basically evaluated by a journalist grouping and aggregating together different data or information. In an other way, the development of social and collaborative Web 2.0 underlines the central and active role of users in collaborative networks. Easy to create and manage these tools are used by Internet users to communicate about themselves. Thus, this data represents an important source of information that can be used for helping epidemiological surveillance. For instance, Twitter is a platform for microblogging, i.e. a system for sharing information where users can either follow other users who post short messages (140 characters) or can be followed. Furthermore, Tweets are associated with meta-informations such as date or location. For instance, from Tweets we can extract the following messages "*I have a huge headache...*" expressed in New York in November or "*... gasrointestinal problems are not good. go 2 a doc!*" from Los Angeles in December.

We have investigated this new kind of media. Basically by using the MeSH (*Medical Subject Headings*) National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a twelve-level hierarchy that permits the search to be carried out at various levels of specificity. At the most general level of the hierarchical structure are very broad headings such as "Anatomy" or "Mental Disorders". More specific headings are found at more narrow levels, such as "Ankle" and "Conduct Disorder". In 2011, 26,142 descriptors were available in MeSH. We conducted some experiments by focusing on the "Disease" part of the hierarchy and we queried Twitter by using all the terms of the corresponding hierarchy. We thus collected 1,801,310 tweets in English from January 2011 to February 2011.

For instance, Figure 4 reports the results of the number of occurrences of terms "Pneumonia", "Leukemia" and "Hepatitis" over the period. It is interesting to notice that, for the decision maker, two peaks are important for the "Hepatitis" (i.e. end of january 2010 and beginning of january 2011). By using the same tools as in EPIMINING, we can easily locate the origins of these tweets as illustrated in Figure 5.

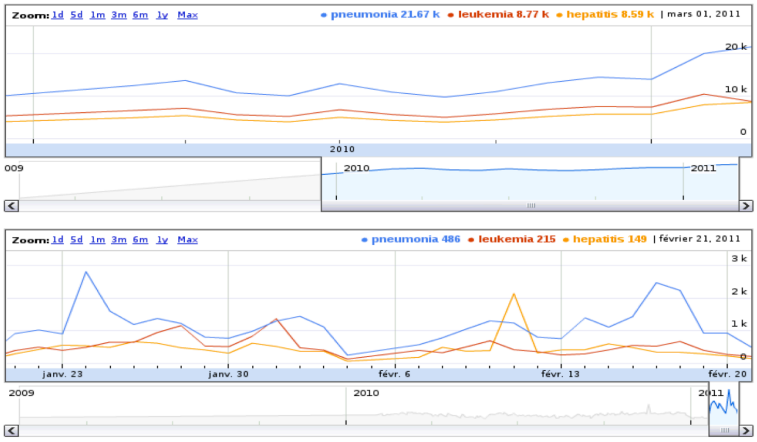

Fig. 4. Occurrences of diseases "Pneumonia", "Leukemia", 'Hepatitis" from January 2011 to February 2011

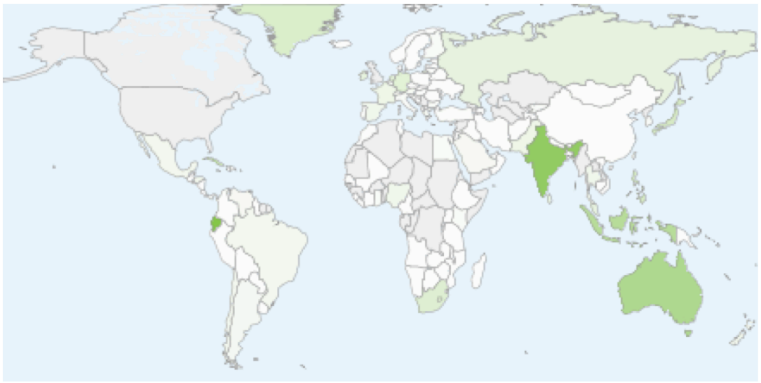

Fig. 5. Localization of Tweets for Pneumonia

Interestingly, we can notice that lots of tweets are exchanged in Ecuador or in Russia. A closer analysis highlights that one alert have been triggered by FluTrackers[7] in Ecuador and that, following the exchanges of tweets, a same kind of alert has been triggered in Russia.

[7] http://www.flutrackers.com/forum/showthread.php?t=158136.

7 Conclusion and Future Work

In this paper, we have proposed a new approach, called EPIMINING, to monitor epidemics, based on automatic knowledge extraction and news classification. EPIMINING have been illustrated by a prototype for monitoring indicators on the H1N1 epidemic. The advantage of our approach, by measuring the echo of an epidemic in the media, is to be complementary to traditional surveillance networks and user's queries analysis proposed by Yahoo and Google systems for instance. The perspectives associated to our proposal are numerous. We can easily improve the classification with learning methods in order to automatically extract the representative patterns of a class. In addition, we plan to extend our approach to other types of textual datasets (e.g. weblogs). We also plan to combine this method with the ones based on other types of datasets (air transport, meteorological, entomological data...). Finally, to answer to our initial question, we can say that the data issued from the web seem to be relevant variables, which can be included into the models of epidemics to better anticipate and predict their dynamics. Furthermore, as illustrated in the last section of the paper, it is more and more important to consider social network to improve the anticipation of epidemics. Knowing, for instance, that some people have fever, headache, gastrointestinal problems, muscle pain at the same time and in the same location is clearly important to better anticipate the propagation of an epidemy.

References

1. Tsui, F.C., Espino, J., Dato, V.M., Gesteland, P.H., Hutman, J., Wagner, M.: Technical description of rods: A real-time public health surveillance system. The Journal of the American Medical Informatics Association **10** (2003) 399–408
2. Polgreen, P., Chen, Y., Pennock, D., Forrest, D.: Healthcare epidemiology: Using internet searches for influenza surveillance. Invited article in Clinical Infectious Diseases – Infectious Diseases Society of America **47** (2008) 14431448
3. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature (2009) 1012–1015
4. Collier, N., Doan, S., Kawazoe, A., Goodwin, R., Conway, M., Tateno, Y., Ngo, Q., Dien, D., Kawtrakul, A., Takeuchi, K., Shigematsu, M., Taniguchi, K.: Biocaster: detecting public health rumors with a web-based text mining system. Bioinformatics **24**(24) (2008) 2940–2941
5. Zant, M.E., Royauté, J., Roux, M.: Représentation événementielle des déplacements dans des dépêches épidémiologiques. In: TALN 2008, Avignon (2008)
6. Zhanga, Y., Danga, Y., Chena, H., Thurmondb, M., Larsona, C.: Automatic online news monitoring and classification for syndromic surveillance. Decision Support Systems **47**(4) (2009) 508–517

7. Turchin, A., Kolatkar, N.S., Grant, R.W., Makhni, E.C., Pendergrass, M.L., Einbinder, J.S.: Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. Journal of the American Medical Informatics Association : JAMIA **13**(6) (2006) 691–695
8. Lu, Y., Xu, H., Peterson, N.B., Dai, Q., Jiang, M., Denny, J., Liu, M.: Extracting epidemiologic exposure and outcome terms from literature using machine learning approaches. Int. J. Data Min. Bioinformatics **6**(4) (2012) 447–459
9. Schmid, H.: Probabilistic Part-of-Speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing. (1994) 44–49
10. Muslea, I.: Extraction patterns for information extraction tasks: A survey. In: AAAI-99 Workshop on Machine Learning for Information Extraction. (1999) 1–6