# Analyse de dépêches pour l'épidémiologie

Didier Breton, Mathieu Roche, Pascal Poncelet, François Marques

# Analyse de dépêches pour l'épidémiologie

Didier Breton[1], Mathieu Roche[2], Pascal Poncelet[2], and François Marques[1]

[1] NEVANTROPIC, 16 Bis Avenue du 14 Juillet, 97300 Cayenne, France
db.nev@ntropic.fr, fm.nev@ntropic.fr
[2] LIRMM – CNRS, 161 rue Ada, 34095 Montpellier Cedex 5, France
mroche@lirmm.fr,poncele@lirmm.fr

**Abstract** : L'analyse épidémiologique est à l'heure actuelle un enjeux très important notamment dans le cas de politique publique de la santé. Dans cette démonstration, nous présentons l'approche EPIMINING permettant, à partir de dépêches de presses, de suivre et d'illustrer via l'application GoogleMap des informations sur le développement de nouveaux cas ou sur l'évolution des bilans.

**Mots-clés** : Suivi d'épidémiologie, classification de dépêches, virus H1N1.

## 1 Introduction

Les systèmes traditionnels de surveillances d'épidémie (e.g, *Institut de Veille Sanitaire*, *European Influenza Surveillance Schema*, *US Center for Diease Control and Prevention*) utilisent généralement des données virologiques, cliniques ou des informations issues des rapports médicaux ou des pharmacies afin de pouvoir suivre le développement d'une épidémie. Même si ces approches sont très efficaces, les analyses proposées ne font que reporter des événements passés la semaine précédente ou les quinze derniers jours. Récemment de nouvelles approches se sont intéressées à l'utilisation de données disponibles sur le Web afin de pouvoir extraire des connaissances liées au suivi de maladies. Ainsi de nouveaux systèmes comme MedISys[1], Argus [2], EpiSpider [3], HealthMap [4] ou BioCaster [5] ont fait leur apparition et offre à l'utilisateur une vision globale et en temps réel de la présence d'une maladie dans un pays. Ces approches sont très pertinentes pour obtenir une vision globale de présence de maladie mais souffrent des défauts suivants : la vision agrégée proposée ne permet pas de suivre le déroulement d'une épidémie à une faible granularité ; elles proposent rarement une classification

[1] http://medusa.jrc.it

[2] http://biodefense.georgetown.edu/projects/argus.aspx

[3] http://www.epispider.org

[4] http://www.healthmap.org

[5] http://www.biocaster.org

des résultats (e.g., nouveaux cas, décés, bilan) ; elles ne retournent que les documents utilisés sans préciser les segments pertinents. En effet, dans le premier cas il est, par exemple, important pour suivre le développement d'une épidémie dans un pays d'analyser dans quelle ville ou village se développe de nouveaux cas. Savoir que dans un pays il y a apparition du virus H1N1 est pertinent mais savoir qu'il s'agit de nouveaux cas ou des nouveaux décès offre plus d'informations pour assurer le suivi de l'épidémie. Enfin, ne retenir que les segments pertinents du document offre la possibilité d'avoir un résumé du contenu du document sans avoir à parcourir ce dernier.

Notre contribution dans cette démonstration est la suivante. D'une part, nous classons automatiquement les dépêches d'articles de Reuters et de l'AFP pour ne retenir que celles qui concernent véritablement l'épidémie suivie. Cette approche est basée sur un algorithme de classification qui tient compte d'un certain nombre de motifs extraits dans les documents. Ces motifs ont été définis après une analyse textuelle fine du contenu de dépêche et tient également compte du fait que les documents décrits sous la forme de dépêches possèdent des particularités par rapport à d'autres types de documents textuels. D'autre part, nous proposons pour une maladie de pouvoir classer les documents selon différents critères (e.g., nouveaux cas, décès, bilan). Enfin, de manière à aider le décideur, nous proposons différentes visualisations des résultats sous la forme de statistiques ou sous la forme d'un positionnement géographique de la dépêche à l'aide de GoogleMap.

L'article est organisé de la manière suivante. Dans la section 2 nous présentons quelques copies d'écran de l'application EPIMINING. Nous concluons dans la section 3 en proposant quelques perspectives.

# 2 Visualisation




Figure 1: EPIMINING en cours d'analyse de documents

La figure 1 illustre le fonctionnement de EPIMINING lors de l'analyse de documents

issus des dépêches. L'objectif de cette étape est d'extraire au sein du document les parties associées aux critères de la classe d'appartenance (e.g., décès, bilan), la localisation (e.g., ville ou pays) et les nombres associés. La figure 2 décrit quelques visualisations

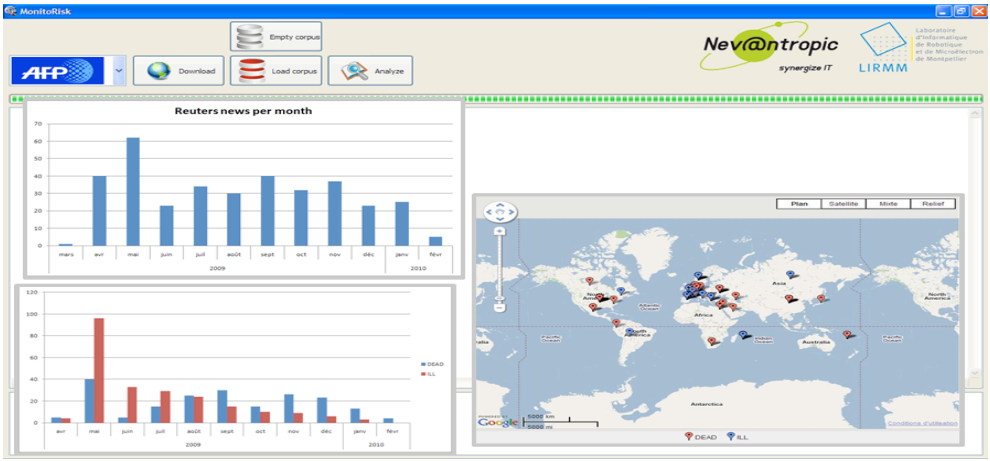


Figure 2: Quelques résultats obtenus

proposées. L'histogramme en haut à gauche décrit le nombre de documents abordant la thématique du virus H1N1 dans les dépêches Reuteurs au cours de l'année 2009. Etant donné qu'une dépêche peut aborder à la fois des nouveaux cas ou des bilans, l'histogramme en bas à gauche montre le nombre de phrases des documents pour chacune des catégories. Enfin, à droite, les différents documents sont positionnés sur la carte et les couleurs correspondent aux classes malades (rouge) ou mort (bleu).

# 3 Conclusion

Dans cet article nous avons présenté l'application EPIMINING de suivi d'épidémies que nous avons appliqué sur des données issues de Reuters et de l'AFP et concernant le développement de l'épidémie H1N1 au cours de l'année 2009. Outre le fait que via notre approche nous pouvons analyser l'écho d'une épidémie dans les médias, nous offrons à une analyse précise en fonction des différentes classes d'analyses (e.g., maladie, décès, bilan) et nous situons finement la référence géographique de la requête. Ainsi notre approche est complémentaire des travaux menés sur l'analyse des données disponibles sur le Web. Les travaux que nous menons actuellement consistent à affiner la classification des différentes dépêches et à permettre d'extraire des motifs représentatifs des épidémies. En outre, nous souhaitons étendre notre approche à d'autres types de jeux de données (e.g., blogs, données officielles) afin de pouvoir extraire de nouvelles informations et coupler ces dernières à d'autres bases de données (e.g., transport aériens, réseaux routiers) afin de mesurer différents canaux de transmission d'une épidémie.