Cartographie de dépêches pour analyser le développement d'épidémies

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 enjeu très important notamment dans le cas de politique publique de la santé. Parmi les différentes études menées, les études descriptives permettent de recueillir des informations sur le nombre de cas et les caractéristiques d'une pathologie. Les approches actuelles sont principalement basées sur des recueils d'information, sur l'analyse de requêtes d'utilisateurs ou sur l'écho que peut avoir une épidémie dans les médias notamment via les documents disponibles sur le Web. Dans ce dernier cas, l'une des difficultés essentielle est d'arriver à extraire, à partir de très gros volumes de données, les informations utiles pour cartographier le développement d'une épidémie. Dans cet article, nous proposons une approche, appelée EPIMINING, permettant, pour une maladie, de classer des dépêches de presses selon différents critères (e.g., nouveaux cas, décès, maladie, bilan). Les expérimentations menées sur des corpus issus de Reuter et de l'AFP montrent une précision supérieure à 94% et une F-mesure supérieure à 85%.

Mots-clés : Epidémiologie, 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. Par exemple l'objectif du réseau Sentinelles[1] composé entre autre de médecins et de pharmaciens est de surveiller, en fonction des consultations médicales, différentes maladies (e.g., crises d'asthme, diarrhée aiguë, maladie de Lyme, Syndrome Grippaux, ...). Même si ces approches sont très efficaces, les analyses proposées ne

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

font que reporter des événements passés la semaine précédente ou les quinze derniers jours et seules quelques approches permettent de suivre en temps réel l'épidémie (Tsui *et al.*, 2003). Récemment une nouvelle approche (Ginsberg *et al.*, 2009) est apparue et utilise les requêtes effectuées sur un moteur de recherche pour prédire à l'avance les pics d'épidémie. Le principe est basé sur l'hypothèse que lorsqu'une personne commence à ressentir les symptômes d'une maladie, elle a tendance à effectuer des requêtes du type : "*quels sont les symptômes de la grippe H1N1 ?*", "*quels sont les sites qui parlent de H1N1 ?*". A l'aide de ce type d'informations et de la localisation du lieu de la requête, il est alors possible de définir quelles sont les tendances des usagers et donc prédire les épidémies potentielles. Les résultats des expérimentations ont montré qu'avec une telle analyse il est possible d'anticiper un pic d'épidémie quinze jours à l'avance. Même si cette approche est très efficace, elle nécessite d'avoir accès aux contenus des différentes requêtes des utilisateurs et également d'avoir un nombre suffisant d'utilisateur afin de déterminer le modèle de prédiction. Enfin, d'autres approches (e.g., (Collier *et al.*, 2008; Zant *et al.*, 2008; Zhanga *et al.*, 2009)) utilisent les informations contenues dans les documents disponibles sur le Web (dépêches, reporting). Ce type de système n'a pas pour objectif de remplacer les systèmes collaboratifs basés sur l'échange de données officielles, mais de permettre une vigilance pandémique intégrant des données issues de régions ou pays pour lesquelles les sources officielles ne sont peu ou pas disponibles. Par exemple, les systèmes tels que MedISys[2], Argus [3], EpiSpider [4], HealthMap [5] ou BioCaster [6] offrent à l'utilisateur une vision globale et en temps réel de la présence d'une maladie dans un pays. Le principe généralement utilisé est le suivant: à partir d'un grand volume de documents du Web, ils extraient des critères concernant les nombres et la localisation. Le nombre reccueilli est principalement utilisé pour afficher avec différentes couleurs (plus ou moins foncées) les informations qui peuvent être situées sur une carte grâce à la localisation. Ces approches sont très pertinentes pour obtenir une vision globale de la 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 des résultats (e.g., nouveaux cas, décès, 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 pouvoir classer les documents en nouveaux cas ou 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 objectif dans cet article est de répondre aux limitations des approches précédentes. Nous nous intéressons également à l'écho que peut avoir une épidémie dans les médias et nous offrons de classer les documents en fonction de leur contenu en se

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

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

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

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

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

focalisant sur les caractéristiques suivantes : bilans, malades, nouveaux cas ou décès. Ainsi pour classer automatiquement les dépêches nous utilisons un algorithme de classification qui tient compte de l'apparition d'un certain nombre de motifs extraits dans les documents. Ces motifs ont été défini après une analyse textuelle fine du contenu des dépêche et tiennent 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, une analyse fine des dépêches permet de proposer une classification non plus au niveau du document mais de la phrase. Enfin, de manière à aider le décideur, nous proposons différentes visualisations des résultats soit sous la forme de statistiques, soit sous la forme d'une représentation géographique des événements à l'aide de GoogleMap.

L'article est organisé de la manière suivante. Dans la section 2 nous présentons l'approche EPIMINING. Les expérimentations menées sont décrites et discutées dans la section 3. Enfin dans la section 4, nous concluons en proposant quelques perspectives.

2 L'approche EPIMINING

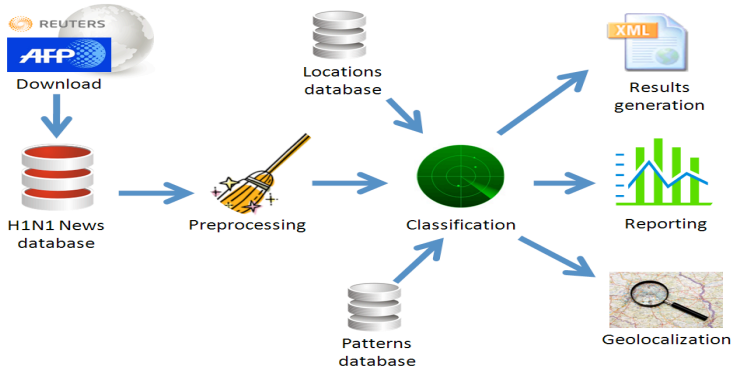

Figure 1: Processus d'EPIMINING

Dans cette section nous présentons le processus global de l'approche EPIMINING qui est décrit Figure 1. Dans un premier temps les dépêches sont extraites via des requêtes contenant des mots clés associés à la maladie (e.g., "Virus H1N1", "grippe porcine", "grippe mexicaine", "grippe A", etc.). Une analyse morphosyntaxique et une lemmatisation sont alors réalisées sur les documents retenus à l'aide de TREETAGGER[7]. L'objectif de cette phase est de repérer et d' identifier parmi les documents les parties utiles pour la comparaison de concepts. Par exemple, considérons le texte suivant : "*Un homme de 28 ans est mort hier à Lille*". La phrase lemmatisée associée est la suivante : "Un *DET:ART* un, homme *NOM* homme, de *PRP* de, 28 *NUM* @card@, ans *NOM* an, est *VER:pres* être, mort *NOM* mort, à *PRP* à, Lille *NAM* Lille". Dans

[7]http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

un premier temps, la base de motifs (*Patterns Database*) est utilisée pour identifier au sein des documents des concepts significatifs. Par exemple, nous pouvons extraire le concept PERSON grâce à la présence du mot lemmatisé "homme". De la même manière le concept YEARS OLD est extrait via l'enchaînement du motif <PERSON> "de" suivi par "an". En appliquant le même principe nous identifions dans le document les concepts : <PERSON>, <YEARS OLD> et $< DEAD >$. La localisation d'une ville est traitée différemment. En effet, il faut d'une part pouvoir retrouver une correspondance entre les motifs correspondant à une ville (e.g., "<PRP = "à"> <NOM="nom commençant par une majuscule">) et d'autre part récupérer les informations complètes de localisation. Ces informations complémentaires sont obtenues en utilisant les données stockées dans la base *Locations Database*. A l'issue de cette étape le document est taggé pour en faciliter la classification. Ainsi la phrase initiale devient "*<PERSON>homme</PERSON> <AGE>28</AGE> <DEAD> mort </DEAD> <TOWN> Lille</TOWN> <REGION> Nord Pas de Calais </REGION> <COUNTRY> France </COUNTRY>*".

Traditionnellement les dépêches utilisées comportent différentes informations qui peuvent être associées à différentes classes. Aussi, notre objectif n'est pas non seulement de classer les différentes dépêches mais également de repérer dans ces dernières les segments correspondants aux différentes classes. En d'autres termes nous extrayons dans les dépêches les segments associés par exemple aux catégories "nouveaux cas", "malade", "décès" ou "bilan". Pour permettre cette opération, nous utilisons à nouveau la base de motif. Cette dernière est composée de motifs spécifiés soit par un expert, soit obtenus par apprentissage. Pour repérer les différents segments nous considérons tout d'abord une phrase afin d'extraire lorsque cela est possible un nombre de malade, un nombre de mort, un lieu, une date etc. et nous associons à chaque phrase un indice de confiance représentatif de la classe cherchée. Par défaut cet indice de confiance est positionné à une valeur maximale en début d'analyse de chaque phrase. Ensuite ce dernier diminue en fonction de la fiabilité de l'information extraite. Par exemple si le lieu ne se trouve pas dans la phrase en cours, la recherche est étendue dans le voisinage de la phrase afin de rechercher les informations manquantes et l'indice de confiance est diminué pour prendre en compte le fait que l'information n'est pas dans le segment traité. Ce choix est motivé par le fait que nous recherchons véritablement des phrases qui correspondent aux motifs décrits dans la base.

3 Expérimentations

De manière à évaluer les performances de notre approche, deux jeux de données en français on été utilisés pour les expérimentations. Une base de dépêches AFP de 510 articles sur la période septembre 2009 à février 2010. Une base de dépêches REUTERS de 353 articles sur la période janvier 2009 à février 2010. De manière à pouvoir analyser la qualité des résultats retournés, 477 dépêches AFP et 7147 phrases ont été annotées manuellement.

Les premières expérimentations ont porté sur la classification des dépêches. Pour cela, nous évaluons la précision qui correspond au rapport du nombre de documents pertinents trouvés sur le nombre total de documents sélectionnés, le rappel qui correspond

au rapport du nombre de documents pertinents trouvés au nombre total de documents pertinents et enfin la F-mesure qui est la moyenne harmonique entre la précision et le rappel. La classification est réalisée de la manière suivante : si un article contient une information relative aux motifs de la classe, celle-ci est considérée comme appartenant à la classe. Ainsi, par exemple, pour la classe "Dead", si l'article contient au moins un motif de la classe, celui-ci est considéré comme retourné pertinent. Dans la table 1 nous reportons le résultat des expérimentations menées. Comme nous pouvons le constater les meilleurs résultats sont obtenus pour les classes "Dead" et "Report". Ceci est justifié par le fait que les dans ces deux cas, les informations contenues dans les dépêches sont souvent exprimées à l'aide d'une information sur le pays (et non plus la ville) et avec des motifs facilement repérables. Le cas des classes "Ill" et "New" est en effet plus problématique car même lors d'une analyse réalisée par un expert la différence entre les deux classes n'est pas forcément évidente.

	Retournés pertinents	Retournés	Pertinents	Précision	Rappel	F-Mesure
Dead (Mort)	100	106	128	94,3%	78,1%	85,5%
Ill (Malade)	43	55	65	78,2%	66,2%	71,7%
Report (Bilan)	88	103	114	85,4%	77,2%	81,1%
New (Nouveau Cas)	48	59	78	81,4%	61,5%	70,1%

Table 1: Classification des dépêches

Le tableau 2 présente les expérimentations menées sur la classification de phrases. Nous avons choisi la classification la plus restrictive pour les phrases. C'est-à-dire que pour qu'une phrase soit classée comme complètement pertinente, il faut pouvoir extraire de celle-ci la date, les motifs associés à la catégorie (e.g., motifs contenant un nombre de cas pour une maladie ou un bilan), ainsi que la localisation géographique. Les valeurs associées à l'indice de confiance indiquent les différentes valeurs testées. Par exemple une confiance [0..25[indique que nous acceptons la classe comme pertinente si la confiance renvoyée est comprise dans l'intervalle 0 à 25. Bien entendu, plus les valeurs sont proches de 100 plus la phrase considérées correspond à un motif d'une classe, i.e. la valeur de précision est de 83,6% pour des valeurs comprises entre 50 et 100 donc dans un voisinage très proche de la phrase. De manière à illustrer les raisons des valeurs de rappel et de précision obtenues, considérons l'exemple suivante : "*La France entière compte toujours 32 décès de malades porteurs du H1N1 pandémique depuis le début de l'épidémie.*" Lors de l'analyse des motifs associés à la classe EPIMINING peut extraire dans une seule phrase tous les concepts importants (e.g., lieux, nombre, catégorie décès) et donc l'indice de confiance est de 100. Dans le cas de la phrase : "*Huit décès supplémentaires ont ainsi été signalés depuis le dernier point de la ministre de la Santé, jeudi, portant à 76 le nombre de personnes décédées en métropole depuis le début de l'épidémie.*" il n'est pas possible d'extraire tous les concepts et donc l'indice de confiance est de 45. Par exemple, il faut analyser les phrases du voisinage pour rechercher que les informations concernent bien le virus H1N1.

	Retournés pertinents	Retournés	Pertinents	Précision	Rappel	F-Mesure
Confiance [0..25[	20	46	280	43,5%	7,1%	12,3%
Confiance [25..50[	58	97	280	59,8%	20,7%	30,8%
Confiance [50..100]	112	134	280	83,6%	40,0%	54,1%
Confiance [0..100]	190	277	280	68,6%	67,9%	68,2%

Table 2: Classification des phrases

4 Conclusion

Dans cet article nous avons proposé une nouvelle approche de suivi d'épidémie et avons illustré les résultats de notre prototype sur l'épidémie du virus H1N1. L'avantage de notre approche est, à partir de l'analyse de dépêches, d'analyser l'écho d'une épidémie dans les médias. Notre approche est complémentaire des travaux menés actuellement via les réseaux de suivi d'épidémie (e.g., Sentinelles) et des travaux basés sur l'analyse des requêtes d'utilisateurs. L'architecture proposée actuellement est tout à fait adaptable à d'autres types d'épidémies. 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) afin de pouvoir extraire des informations complémentaires et les coupler à d'autres bases de données (e.g., transport aériens, réseaux routiers) pour mesurer la manière dont les épidémies peuvent être transmises.

References

COLLIER N., DOAN S., KAWAZOE A., GOODWIN R., CONWAY M., TATENO Y., NGO Q., DIEN D., KAWTRAKUL A., TAKEUCHI K., SHIGEMATSU M. & TANIGUCHI K. (2008). Biocaster: detecting public health rumors with a web-based text mining system. *Bioinformatics*, **24**(24), 2940–2941.

GINSBERG J., MOHEBBI M. H., PATEL R. S., BRAMMER L., SMOLINSKI M. S. & BRILLIANT L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, p. 1012–1015.

TSUI F.-C., ESPINO J., DATO V. M., GESTELAND P. H., HUTMAN J. & WAGNER M. (2003). Technical description of rods: A real-time public health surveillance system. *The Journal of the American Medical Informatics Association*, **10**, 399–408.

ZANT M. E., ROYAUTÉ J. & ROUX M. (2008). Représentation événementielle des déplacements dans des dépêches épidémiologiques. In *TALN 2008*, Avignon.

ZHANGA Y., DANGA Y., CHENA H., THURMONDB M. & LARSONA C. (2009). Automatic online news monitoring and classification for syndromic surveillance. *Decision Support Systems*, **47**(4), 508–517.