Agate : Plate-forme de suivi de catastrophes sur le web

Arnaud Saval[1,3] Yann Mombrun[2,3]

[1] GREYC, Université de Caen, France
[2] LITIS, INSA de Rouen, France
[3] EADS Defence & Security Val-de-Reuil, France
arnaud.saval@info.unicaen.fr, yann.mombrun@insa-rouen.fr

Résumé

La quantité et la qualité des informations contenus sur les blogs, forums et sites de news s'accroissent de jour en jour. Ces sources concernent des domaines variés et se diffusent en temps quasi réel partout dans le monde, malgré un fort niveau de bruit. Le domaine du risque est un bon candidat pour tester ces sources d'information. En effet, elles répondent aux besoins d'avoir une détection rapide des catastrophes et différents points de vue sur un même événement. Nous proposons une plate-forme de traitement de ces sources d'information afin d'extraire des « catastrophes » et ensuite les présenter de façon intelligible et efficace.

Mots-Clés

Traitement automatique des langues naturelles, Modélisation d'événements, Suivi des catastrophes

1 Architecture

Pour illustrer notre approche, nous avons développé une plate-forme de veille de catastrophes naturelles pour traiter des flux d'information non structurés. Cette plate-forme, AGATE[1], a été développée en partie dans le cadre du projet européen CITRINE. Elle est basée sur l'architecture d'intégration fournie par le WebLab [4]. Nous avons construit une chaîne de traitement de l'information à partir d'outils des domaines de la Recherche d'Information et du Traitement Automatique des Langues Naturelles. Cette chaîne est composée de quatre étapes : collecte, extraction d'entités, référencement géographique et indexation (Fig. 1).

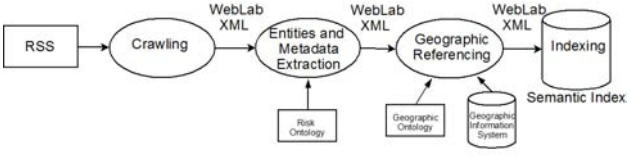

FIG. 1 – Chaîne de traitement Agate.

Nous recevons des alertes à partir des flux d'information sur Internet mis à jour en continu. Les entités extraites à partir de ces textes nous permettent de construire des entités spatiales, temporelles et sémantiques. La récupération de ces informations grâce aux flux RSS nous garantit un accès rapide aux alertes récentes. Cette méthode nous permet d'accéder de manière transparente à des sources d'information hétérogène : blogs, forums, journaux. Ces flux sont composés de métadonnées et de texte libre. Les entités d'intérêt sont extraites des flux d'information grâce à un "pipeline GATE" [2]. Quelques entités sont extraites à partir des métadonnées selon la taxonomie Dublin Core : titre, date, source… La majeure partie des entités extraites proviennent du traitement du texte libre décrivant l'alerte. Elles sont extraites suivant plusieurs domaines : catastrophes naturelles, victimes de catastrophes, information spatiales et temporelles. Une fois structurées, ces entités sont reliées aux instances de l'ontologie GeoNames[2]. Celle-ci contient des propriétés de hiérarchie entre entités géographiques (Rouen est une ville en France). Nous nous servons du modèle pivot WebLab pour représenter le texte de l'alerte, le titre, la source, les instances temporelles, les instances géographiques, les événements et leur relations. Cet ensemble d'informations est finalement indexé dans une base de données sémantique.

2 Travaux Connexes

Notre définition des événements est proche de la représentation proposée par [5] mais elle se démarque par les propriétés sémantiques associées à un événement. La modélisation des événements a déjà vu plusieurs travaux s'intéressant au domaine du risque. [3] exprime la complexité du risque à travers une méthodologie dédiée. Ces travaux traitent seulement l'information structurée et ignorent l'ambiguïté dans les alertes. [1] propose un langage d'événements étendu avec des propriétés spatiales et temporelles dédiées aux épidémies. Les travaux de [6] visent à ajouter des propriétés sémantiques à la description des blogs. La méthodologie proposée cherche à faciliter l'accès aux blogs en fonction des préférences d'un utilisateur. Certains sites tels que MediaWatch[3] et SiloBreaker[4] extraient des

[1] Accessible à http ://eads-vdr.no-ip.org :8041/agate-viewer

[2] http ://www.geonames.org
[3] http ://www.ecoresearch.net/election2008/mediawatch
[4] http ://www.silobreaker.com/

entités sémantiques, temporelles et spatiales mais n'apportent pas de corrélations entre les entités extraites.

3 Expérimentations

Pour valider notre modèle, la chaîne de traitement précédente a été mise en œuvre afin de récupérer des alertes de trois sites d'informations (Reuters, GDACS et RSOE)[5]. Nous avons construit et alimenté une base de données de plus de 20000 alertes depuis Avril 2008. Nous avons sélectionné, au hasard, un ensemble de 1500 alertes et pour chacune nous nous sommes assurés que l'extraction des entités était correcte. Nous avons défini deux requêtes à tester sur cet ensemble d'alertes validées : « Tainted milk China end of 2008 » (a) et « Floods Myanmar May 2008 » (b). Nous comparons les résultats retournés par notre modèle implémenté sans (Modèle I) et avec (Modèle II) les propriétés sémantiques. Ces résultats sont présentés dans un tableau 1 triés par date d'apparition.

Modèle *I*	Modèle *II*
Babies killed by tainted milk in China Two deaths sentenced over tainted milk	Babies killed by tainted milk in China Arsenic contamination in Guangxi Two deaths sentenced over tainted milk

Tainted milk China end of 2008 (a)

Modèle *I*	Modèle *II*
Red flood alert in Myanmar Devastation on Myanmar aid mission	Tropical Storm - Asia - Myanmar Red flood alert in Myanmar Devastation on Myanmar aid mission Reports of malaria outbreaks

Floods Myanmar May 2008 (b)

TAB. 1 – Titres des phénomènes pour les requêtes (a) et (b)

Ces résultats montrent que le Modèle I retrouve des phénomènes intéressants. Le Modèle II retrouve au moins les mêmes et met en avant des phénomènes négligés par le premier. Avec le Modèle II, les résultats donnent plus de détails pour la requête (a) et une meilleure description des causes et conséquences pour la requête (b). En conclusion, le modèle avec les propriétés sémantiques produit une meilleure description de la catastrophe et de son contexte.

4 Gestion d'alertes

FIG. 2 – Interface utilisateur de Agate

Cette application dispose d'une interface graphique (Fig. 2) destinée à la gestion et la recherche d'alertes. Elle permet donc à la fois de se tenir facilement au courant des dernières catastrophes naturelles et technologiques et de rechercher parmi les informations capitalisées à des fins d'archivage.
Après traitement, les flux d'informations pertinents sont sélectionnés pour l'affichage selon le profil de l'utilisateur (les ouragans dans les îles Caraïbes en 2008 par exemple). Lorsqu'un utilisateur clique sur une alerte, sa description complète est présentée avec une mise en relief des entités d'intérêt (géographiques, catastrophes naturelles et nombre de personnes touchées). Celui ci peut alors demander quels sont les événements qui partagent des relations selon leur proximité spatiale, temporelle et sémantique.
En plus de la recherche textuelle standard par mots clés, les utilisateurs peuvent restreindre les alertes reportées par le système grâce à une taxonomie des catastrophes et une liste de pays d'intérêt. Les requêtes géographiques sont dessinées à l'aide de polygones sur la carte.

[5] http ://www.reuters.com, http ://www.gdacs.org, http ://hisz.rsoe.hu

5 Conclusion

Nous avons vu que l'exploitation de propriétés sémantiques en plus des propriétés spatiales et temporelles d'entités dans un texte permet de découvrir des relations insoupçonnées entre les événements décris. Ceci est possible grâce à la représentation des phénomènes à travers une ontologie contenant des propriétés d'occurrences temporelles et de transformations spatiales. Ce modèle s'applique particulièrement bien aux phénomènes physiques tels que les catastrophes naturelles. Nos futurs travaux s'appliqueront à mieux formaliser la découverte de contexte et la détection de phénomène dans le but d'améliorer la précision de l'agrégation d'événement.

Références

[1] H. Chaudet. Steel : A spatio-temporal extended event language for tracking epidemic spread from outbreak reports. In *Proceedings of KR-MED*, 2004.

[2] H. Cunningham. Gate : A framework and graphical development environment for robust nlp tools and applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002.

[3] A. Daude, J. Provitolo, E. Dubos-Paillard, J. Gaillard, E. Eliot, P. Langlois, and E. Propeck. Spatial risks and complex systems : methodological perspectives. 2007.

[4] P. Giroux and al. Weblab : An integration infrastructure to ease the development of multimedia processing applications. *ICSSEA*, 2008.

[5] K. Hornsby and M. Egenhofer. Modeling moving objects over multiple granularities. *Annals of Mathematics and Artificial Intelligence*, 36(1) :177–194, 2002.

[6] S. Rajbhandari, F. Andres, M. Naito, and V. Wuwongse. Semantic-augmented support in spatial-temporal multimedia blog management. *LNCS*, 2007.