A la recherche des tweets porteurs d'informations journalistiques

Benjamin Rosoor*, Laurent Sebag*, Sandra Bringay**,***, Mathieu Roche ***

* Web Report – contact@webreport.fr
** Dépt. MIAP, Université Montpellier 3
*** LIRMM, CNRS, Université Montpellier 2 – {bringay,mroche}@lirmm.fr

1 Introduction

Le succès des réseaux sociaux ne fait plus aucun doute et leurs taux d'activité ont atteint des niveaux sans précédent. Twitter qui est l'un de ces réseaux, permet aux internautes de « microblogguer », c'est-à-dire d'envoyer des messages courts, des « tweets », de moins de 140 caractères et de lire les messages des autres utilisateurs. En 2010, plus de 6 millions de tweets sont produits chaque jour. Une des applications associées à ces données consiste à détecter automatiquement et à analyser en temps réel des sujets émergents et/ou des histoires qui font le "buzz" sur le réseau. Pour les journalistes et autres analystes, détecter ces tendances le plus tôt possible puis suivre leur évolution sont des tâches cruciales. Par exemple, Kostkova *et al.* (2010) montrent l'intérêt de suivre les messages concernant la grippe pour un système d'alerte efficace de la maladie et une meilleure compréhension de son évolution. Récemment, Boyd *et al.* (2010) ont travaillé sur l'activité appelée « retwit » qui consiste à faire suivre les messages d'autres utilisateurs signifiant qu'ils sont appréciés, qu'ils apportent une information récente, inédite ou encore insolite.

Le système LANGMA développé par la société « Web Report » en collaboration avec le LIRMM est dans la lignée de ces méthodes automatiques. Il vise à fournir un support pour produire puis vérifier des informations (tweets) sur les catastrophes naturelles qui, si elles sont publiées par un site public, seront qualifiées de « scoop ». Cet outil se rapproche de la méthode proposée par Sakaki *et al.* (2010) qui détecte les tremblements de terre au Japon via les tweets et dont est issu le site Toretter (http://toretter.com/). Notre approche fondée sur des méthodes de fouille de textes est décrite dans la section suivante. Les résultats expérimentaux obtenus à partir de données réelles sont synthétisés en section 3.

2 Méthode de Fouille de Textes pour filtrer les tweets

Les documentalistes du projet LANGMA disposent d'une interface graphique en mode Web (voir Figure 1) pour gérer les principaux éléments :

- Les sources (flux RSS, tweets, status, Facebook, sites web) sont aspirées à fréquence régulière paramétrable (*toutes les minutes à toutes les heures*).
- Les informations sont issues des sources et filtrées par les méthodes d'analyse et de classification (*informations non vérifiées*), ces sources sont ensuite sélectionnées et vérifiées par les journalistes (*informations en cours de vérification*) avant d'être

classées et diffusées aux agences de presse (*informations vérifiées*). Notre approche décrite ci-dessous permet de proposer à l'utilisateur les tweets les plus proches du thème des catastrophes naturelles.

La méthode développée dans le cadre du projet LANGMA s'appuie sur trois étapes :

(1) Acquisition et représentation d'un corpus d'apprentissage. La première étape est une phase d'acquisition de corpus afin d'obtenir les données d'apprentissage utiles à notre système. Une fois le corpus acquis, ce dernier est représenté de manière vectorielle par une approche dite *sac de mots*. Un traitement préalable consiste à éliminer les mots fonctionnels (préposition, articles, etc) puis à radicaliser (Porter, 1980) les autres mots du corpus. Une interface graphique sécurisée est dédiée à ce processus.

(2) Représentation vectorielle des thèmes. Chaque document est étiqueté par un expert comme catastrophe naturelle. Un vecteur moyen par thème général (catastrophe naturelle ou non) et/ou par sous-thème (types de catastrophes naturelles) peut alors être construit.

(3) Classification d'un nouveau tweet. Dans la dernière phase, nous calculons la similarité d'un nouveau tweet par rapport aux vecteurs moyens. Pour calculer cette similarité, nous appliquons la mesure cosinus bien connue en Fouille de Textes. Lorsqu'un message partage souvent les mêmes descripteurs (c'est-à-dire les mêmes mots radicalisés) avec un thème (et/ou sous-thème), ce message est automatiquement associé à celui-ci.

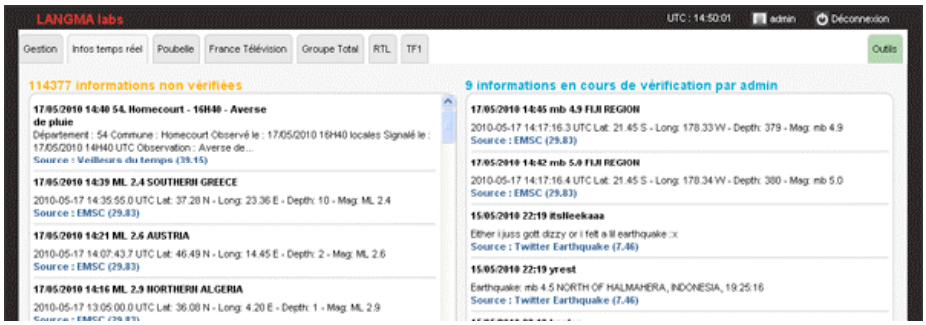

FIG. 1 – *Interface graphique de gestion des tweets*

(*informations non vérifiées et en cours de vérification*)

3 Expérimentations

Pour distinguer des textes évoquant des catastrophes parmi un ensemble de tweets, nous nous sommes appuyés sur un corpus de validation de 135 tweets en anglais dont 74 traitent du thème des catastrophes naturelles. L'évaluation s'appuie sur les mesures de précision, rappel et F-mesure. Dans ces expérimentations, lorsqu'un document possède une valeur de similarité supérieure à SCOS (Seuil de COSinus), le document est associé à la thématique des catastrophes naturelles. Dans nos expérimentations, le meilleur compromis semble être une valeur de SCOS égale à 0.2 qui fournit un excellent rappel et une précision de bonne qualité également. Le rappel élevé obtenu avec un tel seuil sur plusieurs jeux de données permet de confirmer les résultats présentés dans cet article. Notons que des expérimentations

selon les différents types de catastrophes naturelles (inondation, tremblement de terre, marée noire, tempête et tornade) sont présentées dans (Rosoor *et al.*, 2010).

SCOS	0.2	0.3	0.5
Rappel	**0.986**	0.892	0.432
Précision	0.924	**0.971**	**1.000**
F-score	0.954	0.930	0.603

TAB. 1 – *Caractérisation des tweets.*

4 Conclusions et perspectives

Dans cet article, nous avons décrit une approche destinée à des journalistes/documentalistes qui exploite les nouvelles publications massives comme les tweets. Notre méthode permet d'identifier automatiquement des messages liés à une thématique donnée. Nous avons implanté une interface et réalisé des expérimentations sur des jeux de données réelles. Dans nos futurs travaux, nous souhaitons prendre en compte les spécificités lexicales, syntaxiques et graphiques des tweets.

Références

Boyd D., S. Golder, G. Lotan (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. Proc. of HICSS.

Kostkova P., E. Quincey, G. Jawaheer (2010). The potential of Twitter for early warning and outbreak detection, Proc. of ECCMID, 2010

Porter M.F. (1980) An algorithm for suffix stripping, Program. 14(3) p130−137

Rosoor B., L. Sebag, S. Bringay, P. Poncelet, M. Roche (2010). Quand un tweet détecte une catastrophe naturelle... Proc. of VSST.

Sakaki T., M. Okazaki, Y. Matsuon (2010). Earthquake shakes Twitter users: real-time event detection by social sensors, Proc. of WWW, p.851–860

Summary

With over 15 million of users of Twitter in 2010, millions of messages (less than 140 characters) are available. Based on the LANGMA project (collaboration Web Report / LIRMM), the developed software aims to process large amount of data. Our software identifies the tweets of a given topic in order to facilitate the work of the journalist.