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