

Conception d'une solution de détection d'événements basée sur Twitter

Christophe Servan, Catherine Kobus, Yongchao Deng, Cyril Touffet, Jungi Kim, Inès Kapp, Djamel Mostefa, Josep Crego et Jean Senellart

SYSTRAN, 5 rue Feydeau, 75002 Paris, France
firstname.familyname@systrangroup.fr

RÉSUMÉ

Cet article présente un système d'alertes fondé sur la masse de données issues de Tweeter. L'objectif de l'outil est de surveiller l'actualité, autour de différents domaines témoin incluant les événements sportifs ou les catastrophes naturelles. Cette surveillance est transmise à l'utilisateur sous forme d'une interface web contenant la liste d'événements localisés sur une carte.

ABSTRACT

Design of a solution for event detection from Tweeter

This article presents a detection system based on Big Data from Tweeter. The goal of the tool is to supervise news from various domains such as sport events or natural disasters. This monitoring is transmitted to the user as a website, which contains a list of events located on a map.

MOTS-CLÉS : Détection d'événements, Masse de Données, Tweeter.

KEYWORDS: Event Detection, Big Data, Tweeter.

Search and view events

Start date: 2016-07-03 19:00:02 Country: France
End date: 2016-07-03 23:55:02 City: Paris

Type: natural disaster cybercrime football

Search

Event	Category	Date and time
goal	football	2016-07-03T21:44:00.000Z
yellow	football	2016-07-03T21:36:00.000Z
goal	football	2016-07-03T21:16:00.000Z
goal	football	2016-07-03T20:44:00.000Z
penalty	football	2016-07-03T20:20:00.000Z
goal	football	2016-07-03T20:20:00.000Z
goal	football	2016-07-03T20:13:00.000Z

The map shows a red pin indicating the location of an event in Paris, France, with labels for nearby cities like Brussels, London, and Amsterdam.

FIGURE 1 – Exemple de l'interface utilisateur, permettant la visualisation des données.

1 Introduction

Lors de la dernière décennie, la quantité d'information disponible sur Internet a radicalement explosé d'une part par le passage au numérique de la quasi-totalité des médias traditionnels, mais aussi par l'explosion de l'usage des réseaux sociaux et des sites d'informations non officiels de type blogs. Cette explosion à l'échelle mondiale donne théoriquement accès à une palette gigantesque d'informations multilingues mais noie pratiquement l'utilisateur final dans un excès d'information. Face à l'explosion

des flux et des sources d'information, de nouveaux outils et services voient le jour pour les exploiter. L'outil présenté dans cet article en fait partie.

L'objectif de l'outil est de surveiller l'actualité, en particulier autour de différents domaines témoin incluant les événements sportifs ou les catastrophes naturelles. L'outil utilise une approche polylingue, c'est-à-dire qu'il traite de manière native dans les différentes langues concernées (soit le français, l'anglais et l'arabe). Cet outil est l'aboutissement du projet POPYRUS (Plateforme Adaptative PolYlingue de suRveillance de FIUx Spécialisés). Ce dernier s'intéresse à la recherche d'approche polylingues pour améliorer la précision des modules de fouilles documentaires.

2 Système de détection

Le système effectue un premier filtrage en utilisant des mots-clés relativement courants. Puis, les tweets ainsi filtrés sont classés suivant les différents événements auxquels ils appartiennent. Ces derniers peuvent n'appartenir à aucun événement et pourront être rejetés par le système de classification. L'approche de classification est fondée sur les réseaux de neurones convolutionnels tels que décrits par (Kim, 2014). Une fois la classification faite, nous étudions l'évolution de la fréquence des tweets répondant à chacun des événements, telle que décrite dans (Earle *et al.*, 2012). Les événements ainsi collectés sont stockés dans une base de données, de même que les identifiants des tweets associés à ces mêmes événements.

2.1 Géolocalisation

Afin de géolocaliser les événements, nous utilisons d'une part les informations contenues dans les tweets, lorsque ceux-ci sont géolocalisés et, d'autre part, nous utilisons une détection d'entités nommées, lorsque la géolocalisation du tweet n'est pas présente ou non-pertinente. En effet, lorsqu'un utilisateur tweete un message à propos d'un match de foot depuis son lieu d'habitation, la géolocalisation associée à ce tweet n'est pas pertinente. Par contre, la géolocalisation d'un tweet émis par un utilisateur proche d'un épicentre d'un tremblement de terre est pertinente.

2.2 Visualisation des données

La figure 1 présente un exemple de visualisation des données collectées. Ainsi, l'événement et ses caractéristiques (date, lieu, contexte, *etc.*) peuvent être affichés et visualisés sur une carte. Les tweets correspondant aux événements peuvent également être affichés dans une autre partie de l'interface, permettant une validation humaine de la détection automatique.

Remerciements

Cette réalisation a été financée à travers le projet DGA-RAPID 2014 N° 1429060465 POPYRUS

Références

EARLE P. S., BOWDEN D. C. & GUY M. (2012). Twitter earthquake detection : earthquake monitoring in a social world. *Annals of Geophysics*, **54**(6).

KIM Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 1746–1751, Doha, Qatar.