QUAND UN TWEET DETECTE UNE CATASTROPHE NATURELLE...

Benjamin Rosoor (*), Laurent Sebag (*), Sandra Bringay (,***), Pascal Poncelet (***), Mathieu Roche (***)**
contact@webreport.fr {mroche,bringay,poncelet}@lirmm.fr

(*) Web Report, France
(**) Dept. MIAP, Université Montpellier 3, France
(***) LIRMM, CNRS, Université Montpellier 2, France

Mots clefs :
Veille scientifique et technologique, Fouille de Textes, Classification, Recherche d'Information.
Keywords:
Scientific and technical observation, Business Intelligence, Text Mining, Classification, Information Retreieval
Palabras clave :
Escudriñar científico y tecnológico,

Résumé

Nous évoluons dans un monde où l'information est centrale dans la mesure où l'ensemble de nos actions, interactions, personnelles et professionnelles, sont dépendantes des informations à notre disposition. Accéder à une information pertinente, au bon moment, est un enjeu stratégique important pour en faire un bon usage. Ces dernières années, les blogs, les "Statuses" (tweets, statuts Facebook, …) et autres dépêches RSS se sont multipliés car simples à créer et gérer. Ces nouvelles formes de publication ont un potentiel inattendu en termes de veille stratégique. En effet, les professionnels de l'information peuvent les utiliser comme nouvelles ressources documentaires pour y rechercher de l'information. Le travail présenté dans cet article est issu d'une collaboration entre la société Web Report et le laboratoire d'informatique LIRMM. La société développe un outil de veille stratégique pour détecter les informations avant même leur apparition dans les nouvelles des agences de presse en s'appuyant sur la détection automatique de catastrophes à partir de ressources hétérogènes issues du Web. Cet outil repose sur une représentation saltonienne d'un corpus de textes classés en thèmes (inondation, tremblement de terre…). Chaque thème est représenté sous la forme d'un vecteur de mots. Chaque nouveau texte à classer est comparé à ces vecteurs pour identifier le thème le plus proche. Une interface graphique, destinée aux journalistes / documentalistes et associée à cette méthode, a été conçue. Des expérimentations sur des jeux de données réelles soulignent la pertinence de notre proposition et ouvrent de nombreuses perspectives.

We live in a world where information is central because all our personal and professional actions/interactions depend on the available information. The access to the relevant information at the right time is an important issue. In recent years, blogs, tweets, and other RSS news have been developped. These novel forms of publication have an unexpected potential or the Business Intelligence (ie BI) domain. Indeed, professionals can use them as new resources in order to find information. The work presented in this paper is a collaboration between the Web Report company and the LIRMM. The company has developed a BI tool to detect the information before they appear in the agencies. The automatic detection method uses heterogeneous resources from the Web. This tool is based on a "salton" representation of a corpus classified into different topics (flood, earthquake, and so forth). Each category is represented as a vector of words. The new text to be classified is compared to these vectors in order to identify the relevant category. A GUI for journalists / librarians with this method has been developped. Experiments on real data sets show the relevance of our proposal and open up many opportunities.

1 Introduction

Nous évoluons dans un monde où l'information est centrale dans la mesure où l'ensemble de nos actions, interactions, personnelles et professionnelles, sont dépendantes des informations à notre disposition. Accéder à une information pertinente, au bon moment, est un enjeu stratégique important pour en faire un bon usage.

Ces dernières années, le développement du web social et collaboratif 2.0 a rendu les internautes plus actifs au sein des réseaux participatifs. Les blogs pour diffuser son journal intime de manière massive, les tweets pour publier ses faits et gestes en 140 caractères maximum et autres dépêches RSS sont extrêmement répandus. Simples à créer et gérer, ces outils sont utilisés par les internautes, les entreprises ou autres organisations pour communiquer. Ces nouvelles formes de publication s'inscrivent désormais dans une logique d'intelligence collective et de gestion des connaissances, et ont un potentiel inattendu en termes de veille stratégique. En effet, les professionnels de l'information peuvent les utiliser comme nouvelles ressources documentaires pour y rechercher de l'information. Or, ces derniers se confrontent à l'abondance d'informations. Comment effectuer un tri efficace à partir de cette masse de ressources, pour ne conserver que les informations pertinentes en fonction d'une problématique ?

Le travail présenté dans cet article est issu d'une collaboration entre la société Web Report et le LIRMM. La société est spécialisée dans l'animation de communautés, la valorisation d'avis et commentaires d'internautes et consommateurs. Des webmasters éditoriaux rédigent des articles, notes de blog, guides d'achat, brèves et autres dossiers multimédia. La société souhaite désormais développer un outil de veille stratégique pour détecter les informations avant même leur apparition dans les nouvelles des agences de presse. Nous adoptons ici une approche similaire à celle de Google [9] qui a montré un lien entre les requêtes des internautes qui utilisent des termes liés à la grippe et le nombre de personnes présentant les symptômes de cette maladie. Google est ainsi capable de fouiller les requêtes pour suivre l'évolution d'une épidémie en temps réel. De la même manière, nous nous intéressons ici à la détection automatique de catastrophes à partir de ressources hétérogènes issues du web telles que les blogs, les tweets, les dépêches. Notre objectif est de fouiller ces ressources pour détecter une catastrophe en temps réel.

Dans cet article, nous présentons une méthode de détection automatique de signaux faibles (tâche de veille technologique) à partir d'un ensemble de messages courts (français et anglais). Cette méthode repose sur une représentation saltonienne d'un corpus de textes classés en thèmes (inondation, tornade, etc). Chaque thème est représenté sous la forme d'un vecteur de mots. Chaque nouveau texte à classer est comparé à ces vecteurs pour identifier le thème le plus proche. Notre contribution se décline selon deux axes : (1) identifier des catastrophes à partir d'un ensemble de tweets ; (2) classer des tweets selon le type de la catastrophe. La méthode est présentée dans la section 2. Dans la section 3, nous présentons l'interface graphique de l'outil. Enfin, nous avons réalisé des expérimentations sur des jeux de données réelles qui sont présentées dans la section 4. Ces expérimentations soulignent la pertinence de notre proposition et ouvrent de nombreuses perspectives détaillées dans la section 5.

2 Etat de l'art

Le succès des réseaux sociaux ne fait plus aucun doute et leurs taux d'activité ont atteint des niveaux sans précédent. Des centaines de millions d'internautes sont inscrits dans ces réseaux. Ils échangent via des forums, maintiennent des blogs, racontent leurs dernières pensées, humeurs ou activités en quelques mots... Le développement des outils mobiles tels que les téléphones portables, permettant de contribuer à ces réseaux de n'importe quel endroit, a favorisé l'émergence de ces nouvelles pratiques. Twitter[1] est l'un de ces réseaux. Il permet aux internautes de « microblogguer », c'est-à-dire d'envoyer des messages courts, des « tweets » de 140 caractères uniquement et de lire les messages d'autres utilisateurs. Un exemple de tweet est donné dans la Figure 1. Les internautes restent ainsi connectés à leurs amis, famille ou collègues via le réseau [16]. En 2010, plus de 15 millions d'utilisateurs utilisent Twitter et plus de 6M de tweets sont produits chaque jour.

Very heavy rain in Incheon; the tropical storm has arrived *(http://twitter.com/melbuckpitt/status/20779466592)*
there's a tropical storm here in the phils, its already raining hard! *(http://twitter.com/bottle26/status/18424113721)*
Eek! Tornado sirens!! *(http://twitter.com/CraveSome_Alex/status/16878531196)*

Figure 1 : Exemples de tweet

D'après Sakaki *et al.* [19] la caractéristique essentielle des tweets est leur production en temps réel. Contrairement aux blogs qui sont mis à jours au mieux quotidiennement, les tweets sont produits plusieurs fois dans la journée, parfois dès la survenue d'un événement. Les internautes savent ce que les autres membres du réseau viennent de faire, font ou vont faire en temps réel car ceux-ci rapportent les événements majeurs qui les concernent, qu'ils soient d'ordre personnel (e.g. soirée) ou public (e.g. bouchon, séisme). Comme Sakaki *et al.*, nous allons dans cet article chercher à exploiter le fait que les tweets contiennent des informations les plus « à jour » pour détecter des catastrophes.

Dernièrement, Twitter a mis en place des APIs qui permettent de rechercher des messages selon différents critères et d'y associer des informations comme la date, des informations personnelles sur l'auteur du message issues de son profil (nom, bibliographie, goûts, etc), sa situation géographique… Ces flux de documents représentent une incroyable richesse en termes d'exploration automatique.

Une première application associée à ces données consiste à détecter automatiquement et à analyser en temps réel des sujets émergents, des histoires intéressantes qui font le « buzz » sur le réseau. Il s'agit d'étudier ici des tendances apparaissant au cours du temps. D'après Mathioudakis *et al.* [15], les tendances sont liées à des événements nouveaux assez importants pour attirer l'attention d'une grande partie des internautes. Pour les journalistes et autres analystes, détecter ces tendances le plus tôt possible puis suivre leur évolution représentent une information capitale. Par exemple, détecter qu'une partie des internautes s'intéresse à la dernière explosion d'un puit de pétrole et suivre l'évolution des avis de ces internautes sur cette explosion jusqu'à l'oubli de cette information peut être très utile pour une compagnie pétrolière. Dans le contexte du web en général, nous pouvons citer les travaux de [1,2,14] qui proposent des outils pour suivre la popularité de sujets sur le net. Le développement d'outils commerciaux tels que Alltop[2], Radian 6[3], Scoutlabs[4],

[1] http://twitter.com/
[2] http://alltop.com/
[3] http://www.radian6.com/
[4] http://scoutlabs.com/

Sysomos[5], Thoora[6] soulignent les intérêts économiques des entreprises pour ce type de méthodes. Certains auteurs se sont focalisés sur le cas particulier de Twitter. C'est le cas de Mathioudakis *et al.* [15] qui s'intéressent plus spécifiquement à la détection de tendances dans les flux Twitter. Ils identifient des mots-clés qui apparaissent soudainement dans les tweets à un taux anormalement élevé, regroupent les tweets qui contiennent ces mots-clés en tendances puis associent à ces ensembles de messages des indicateurs (comme le nombre de messages postés dans les dernières 24h). Dans le même esprit, l'outil Twitscoop[7] permet de suivre les derniers scoops parmi les posts des internautes, c'est-à-dire les sujets qui ont attiré l'attention du plus d'internautes. D'autres types d'applications ont été envisagées à partir de ces tweets. Par exemple, en se fondant sur les travaux de [6], des chercheurs de Harvard et de la Northeastern University de Boston[8] ont mesuré l'évolution de l'humeur de la population américaine sur trois années à partir des messages postés sur Twitter, offrant ainsi des indicateurs pertinents aux politiques. Barbosa *et al.* [3] proposent également une approche pour détecter automatiquement les sentiments dans les messages. Comme les internautes donnent souvent leurs opinions [17] dans les messages à propos des personnalités, produits ou sociétés, l'application commerciale est immédiate. Dans le contexte de la santé, Kostkova *et al.* [12] montrent l'intérêt de suivre les messages à propos de la grippe pour un système d'alerte efficace de la maladie et une meilleure compréhension de l'évolution de celle-ci. Récemment, Boyd *et al.* [5] ont travaillé sur l'activité appelée « retwit » qui consiste à faire suivre les messages d'autres utilisateurs, un peu comme lorsque l'on fait suivre un mail. Certains internautes cherchent à faire « retwitter » leurs messages car cela signifie qu'ils sont appréciés, qu'ils apportent une information récente, inédite ou encore insolite. Un indicateur de pertinence fondé sur le nombre de tweets générés par un seul tweet a été proposé. Le site TweetMeme[9] [14] permet d'identifier les tweets retwités les plus populaires. Toutes ces applications soulignent l'intérêt d'utiliser les tweets dans les tâches de détection automatique de signaux faibles.

Si les intérêts associés à l'exploitation automatique des tweets sont assez évidents, la mise en place de programmes permettant de réaliser cette tâche est difficile. En effet, la prise en compte des gros volumes de données et de leur dynamique nécessite la mise en œuvre d'algorithmes efficaces, minimisant le nombre de passes à effectuer sur les données pour assurer le passage à l'échelle. Par ailleurs, contrairement aux documents textuels traditionnels, les descripteurs pertinents des tweets ne sont pas nécessairement les mots-clés d'un « dictionnaire classique ». Par exemple la présence d'URL peut se révéler particulièrement pertinente pour classer/discriminer des tweets [8]. De plus, des émoticônes (suites de quelques caractères représentant une émotion), présents dans les tweets, peuvent aussi révéler un certain sentiment qui peut être détecté automatiquement [7]. Ceci montre que le traitement de ces textes courts nécessite au préalable la mise en place de techniques d'extraction automatique de descripteurs pertinents souvent spécifiques aux tweets.

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 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 [19] qui détecte les tremblements de terre au Japon via les tweets et dont est issu le site Toretter[10]. Sakaki *et al.* considèrent les tweets comme autant de capteurs produisant des informations sensorielles. L'originalité de notre approche décrite dans la section suivante repose sur l'intégration d'approches de TAL (Traitement Automatique du Langage) combinées à des méthodes de fouille de données pour affiner les résultats de détection automatique de signaux faibles.

[5] http://www.sysomos.com/
[6] http://www.thoora.com/
[7] http://www.twitscoop.com/
[8] http://www.ccs.neu.edu/home/amislove/twittermood/
[9] http://tweetmeme.com/
[10] http://toretter.com/

3 Méthodes et outils

Le système proposé dans le cadre du projet Langma repose sur une méthode automatique consistant, dans un premier temps, à représenter les tweets sous forme vectorielle. La méthode mise en œuvre se décompose en quatre phases que nous détaillons ci-dessous.

- **Phase 1 :** Acquisition d'un corpus. La première étape est une phase d'acquisition de corpus afin d'obtenir les données d'apprentissage utiles à notre système. Le corpus est un ensemble de textes homogènes sur le fond (même thème global) et la forme (même format des documents).

- **Phase 2 :** Représentation du corpus. Une fois le corpus acquis, ce dernier sera représenté de manière vectorielle. Chaque texte sera considéré comme un sac de mots. Dans cette représentation dite « Saltonienne » un traitement préalable consiste à éliminer les mots outils (préposition, articles, etc). Chaque mot présent dans le corpus représentera la base sur laquelle nous nous appuierons. Deux types de représentations peuvent alors être effectuées : une représentation booléenne (présence/absence des mots) et/ou fréquentielle (nombre d'occurrences des mots dans chaque document). Notons que d'autres types de représentations numériques peuvent être appliqués. Par exemple, la mesure TF-IDF [10] consiste à calculer l'importance d'un mot dans un document relativement à une collection. Ainsi, un mot présent dans tous les documents d'une collection aura un poids moindre. Dans la suite, nous allons appliquer la représentation saltonienne à partir d'un corpus constitué des deux documents suivants (*cf.* Figure 2) :

Doc 1: *Avec les importants développements technologiques du web, les travaux concernant la veille technologique sont importants.*
Doc 2 : *Le web joue donc un rôle essentiel.*

Figure 2 : Exemple de corpus

La représentation saltonienne du corpus est donnée sous forme booléenne dans la Figure 3 et sous forme fréquentielle dans la Figure 4 :

	importants	*développements*	*technologiques*	*web*	*travaux*	*concernant*	*veille*	*technologique*	*joue*	*Rôle*	*essentiels*
Doc 1	1	1	1	1	1	1	1	1	0	0	0
Doc 2	0	0	1	1	0	0	0	0	1	1	1

Figure 3 : Représentation booléenne

	importants	*développements*	*technologiques*	*web*	*travaux*	*concernant*	*veille*	*technologique*	*joue*	*Rôle*	*essentiels*
Doc 1	2	1	1	1	1	1	1	1	0	0	0
Doc 2	0	0	1	1	0	0	0	0	1	1	1

Figure 4 : Représentation fréquentielle

Notons que pour éviter d'obtenir des vecteurs trop creux, une phase d'élagage consistant à ne prendre en compte que les mots présents un minimum de fois peut être appliquée [18]. Avec ces représentations, les mots d'une même famille peuvent être rassemblés (mots singuliers/pluriels, féminins/masculins, mots conjugués, etc). La représentation canonique des mots permet un regroupement de ces derniers comme le montre l'exemple ci-dessous qui rassemble les mots technologiques et technologique (*cf.* Figure 5).

	important	*développement*	*technologique*	*web*	*travail*	*concerner*	*veille*	*joue*	*Rôle*	*essentiel*
Doc 1	2	1	2	1	1	1	1	0	0	0
Doc 2	0	0	1	1	0	0	0	1	1	1

Figure 5 : Représentation fréquentielle avec lemmatisation

D'autres types de préparation des données textuelles consistent à ne considérer que la forme radicale de chaque mot [13]. Ces mots, lemmes ou radicaux seront appelés les descripteurs textuels des documents. Notons qu'un filtrage linguistique peut aussi être appliqué. Ce dernier consiste à effectuer un filtrage morphosyntaxique pour ne conserver que les mots ayant une étiquette grammaticale spécifique [4]. Enfin des connaissances expertes permettent d'accorder un poids plus important aux mots du domaine (dans notre cas, le domaine des catastrophes naturelles).

- **Phase 3 :** Représentation vectorielle des thèmes. Chaque document du corpus est associé à un thème par des experts du domaine (dans notre cas, différentes catastrophes naturelles telles que les tremblements de terre, les inondations, etc). Ainsi, pour constituer la base d'apprentissage, nous pouvons former un vecteur moyen par thème. Une telle représentation sur la base de la mise en place d'un vecteur moyen permet une comparaison plus efficace lors de la phase suivante du processus. En effet, les algorithmes de type « K plus proches voisins » [4,11] se révèlent coûteux car ils demandent une comparaison de nombreux vecteurs.

- **Phase 4 :** Classification d'une nouvelle dépêche. Dans la dernière phase propre à la classification, nous calculons la similarité entre les vecteurs moyens et un nouveau document afin d'associer ce dernier à un thème. Ainsi, lorsqu'une dépêche partage souvent les mêmes descripteurs (par exemple, les mots) avec un thème, celle-ci est automatiquement associée au thème en question. Pour calculer cette similarité, nous appliquons la mesure cosinus [10] bien connue en Recherche d'Information (RI). Celle-ci permet de comparer la « proximité » entre les vecteurs (dans notre cas, thème vs dépêche à classer).

L'ensemble de ce processus développé dans le cadre du projet Langma est rigoureusement évalué dans la section suivante.

4 Expérimentations

4.1 Classification globale : Identification des catastrophes parmi un ensemble de tweets

Pour distinguer des textes évoquant des catastrophes parmi un ensemble de tweets, nous nous sommes appuyés sur un corpus en anglais réparti en deux classes : documents (tweets) traitant d'une catastrophe, documents issus d'un thème quelconque. Un exemple de tweet est donné dans la Figure 6.

#Earthquake of M 5.0, near the south coast of western Honshu, Japan http://bit.ly/90qpAy

Figure 6 : Exemple de tweet

Dans ce protocole expérimental, nous avons utilisé les mots radicalisés. Par ailleurs, certains mots ont été filtrés manuellement afin d'éliminer le bruit susceptible de dégrader les tâches de classification. Pour évaluer les résultats, nous utilisons un corpus de test constitué de 135 textes dont 74 sur le thème des catastrophes naturelles. Ceci nous permet d'estimer les résultats sur une partie des données qui ne sont pas utilisées lors de la phase d'apprentissage (construction du modèle). L'évaluation s'appuie sur les mesures de précision et rappel bien connues dans le domaine de la Fouille de Données. Ces résultats sont ensuite combinés avec la mesure de F-Score qui donne une estimation générale de la méthode testée (*cf.* Figure 7).

Précision = Nombre documents pertinents retrouvés / Nombre documents retrouvés

Une précision de 100% signifie que tous les textes retrouvés sont pertinents.

Rappel = Nombre documents pertinents retrouvés / Nombre documents pertinents

Un rappel de 100% signifie que tous les textes pertinents ont été retrouvés.

F-Score = 2× Précision × Rappel / (Précision + Rappel)

Figure 7 : Mesures d'évaluation : Précision, Rappel et F-Score

Dans ces expérimentations, lorsqu'un document possède une valeur de similarité (cosinus) supérieure à S, le document est associé à la thématique des catastrophes naturelles. La Table 1 donne les résultats de la classification globale pour différentes valeurs de S. Les résultats montrent qu'avec un seuil assez élevé (supérieur à 0.5), nous obtenons une précision de très bonne qualité (100% avec S=0.5). Ainsi, tous les tweets prédits comme des textes relatifs aux catastrophes traitent réellement de ce thème. Cependant, de nombreux tweets de cette thématique ont été omis (rappel faible). Ainsi, il est préférable de diminuer le seuil, ce qui fait décroître le score de précision mais nettement augmenter le rappel. Dans nos expérimentations, le bon compromis semble être un seuil de S=0.2 qui fournit un excellent rappel mais une précision de bonne qualité également. Notons que d'autres expérimentations effectuées à partir d'un corpus de test constitués de près de 400 tweets confirment ces résultats.

S	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

Table 1 : Résultats sur la catégorisation générale des tweets

4.2 Classification spécialisée : Identification des différents types de catastrophes naturelles

Pour classer les tweets en différentes catégories de catastrophes, nous avons travaillé sur des textes associés aux catégories : inondation (173 textes), tremblement de terre (454), marée noire (472), tempête (505) et tornade (427). Le but de ce travail est d'étudier la performance d'une classification fine qui permet de distinguer automatiquement les différents types de catastrophes. Ces textes ont été préparés de la même manière que les textes utilisés pour la classification générale (*cf.* section précédente). Pour évaluer les résultats de la classification par thèmes spécifiques, nous calculons l'exactitude (Accuracy) de la méthode selon la formule donnée dans la Figure 8.

Exactitude = Nombre de documents bien classés / Nombre de documents total

Figure 8 : Mesure d'évaluation : Exactitude

La Table 2 donne les résultats de la classification (matrice de confusion). L'exactitude de la méthode est de 0.811 ce qui est un résultat de bonne qualité. Cependant, la matrice de confusion montre que le thème des Inondations est assez difficile à prédire. En effet, les textes issus de ce thème sont souvent confondus avec le thème des Tempêtes. Ceci semble assez cohérent car le vocabulaire de ces deux thèmes est assez proche.

Classe Réelle /Classe Prédite	Inondation	Tremblement de terre	Marée noire	Tempête	Tornade
Inondation	78	0	0	**95**	0
Tremblement de terre	36	**383**	7	15	13
Marée noire	30	0	**350**	88	4
Tempête	12	0	0	**487**	6
Tornade	40	7	7	23	**350**

Table 2 : Résultats sur la catégorisation spécialisée des tweets

Après avoir évalué la qualité du traitement automatique mené dans le cadre du projet Langma, la section suivante décrit l'interface graphique de notre système.

5 Interface graphique

Les journalistes et documentalistes du projet Langma disposent d'une interface graphique en mode Web pour gérer les principales actions décrites dans cet article. De manière plus précise, trois interfaces associées aux types d'actions à mener sont proposées : initialisation, gestion et consultations des données. Les sous-sections suivantes décrivent ces interfaces graphiques qui sont manipulées par différents types d'utilisateurs.

5.1 Interface d'initialisation

Les utilisateurs ont accès à une interface en mode Web sécurisée pour gérer les éléments nécessaires à l'analyseur sémantique. Dans un premier temps, le système permet le traitement d'un corpus d'apprentissage : acquisition d'un corpus d'apprentissage, vectorisation et normalisation de celui-ci. Ceci permet d'obtenir un « dictionnaire » de référence pour chaque thème défini.

La deuxième étape consiste à tester un corpus (corpus de test) constitué d'éléments connus (association ou non des tweets à une catastrophe naturelle). Ceci permet de mesurer les critères de précision et rappel (*cf.* section précédente) et de construire une matrice de confusion afin d'évaluer la performance globale du système.

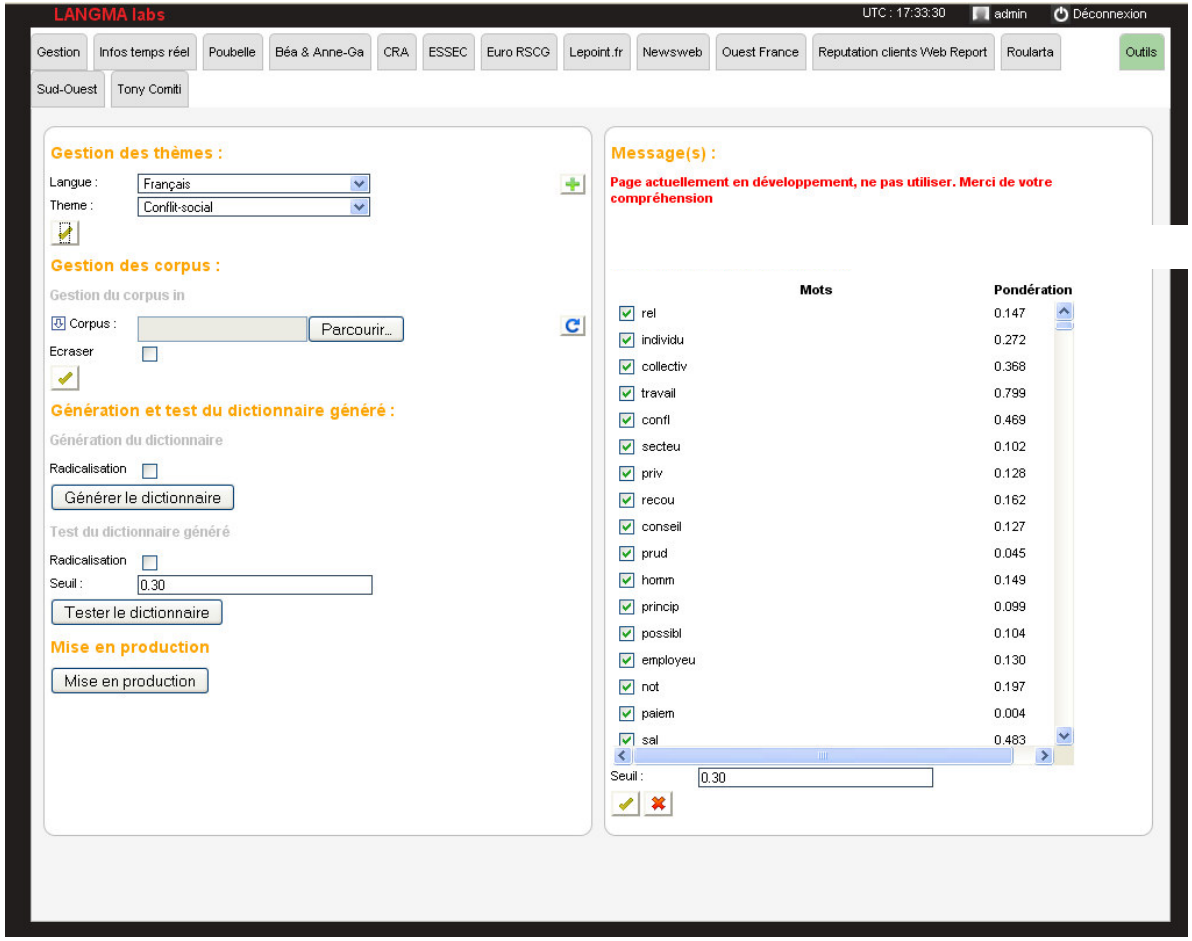

Figure 9 : interface d'initialisation

5.2 Interface graphique de gestion

Les journalistes et documentalistes disposent d'une interface graphique en mode Web pour gérer les sources : flux RSS, tweets, status, Facebook, sites web. Ces données sont aspirées à fréquence régulière paramétrable (toutes les minutes à toutes les heures).

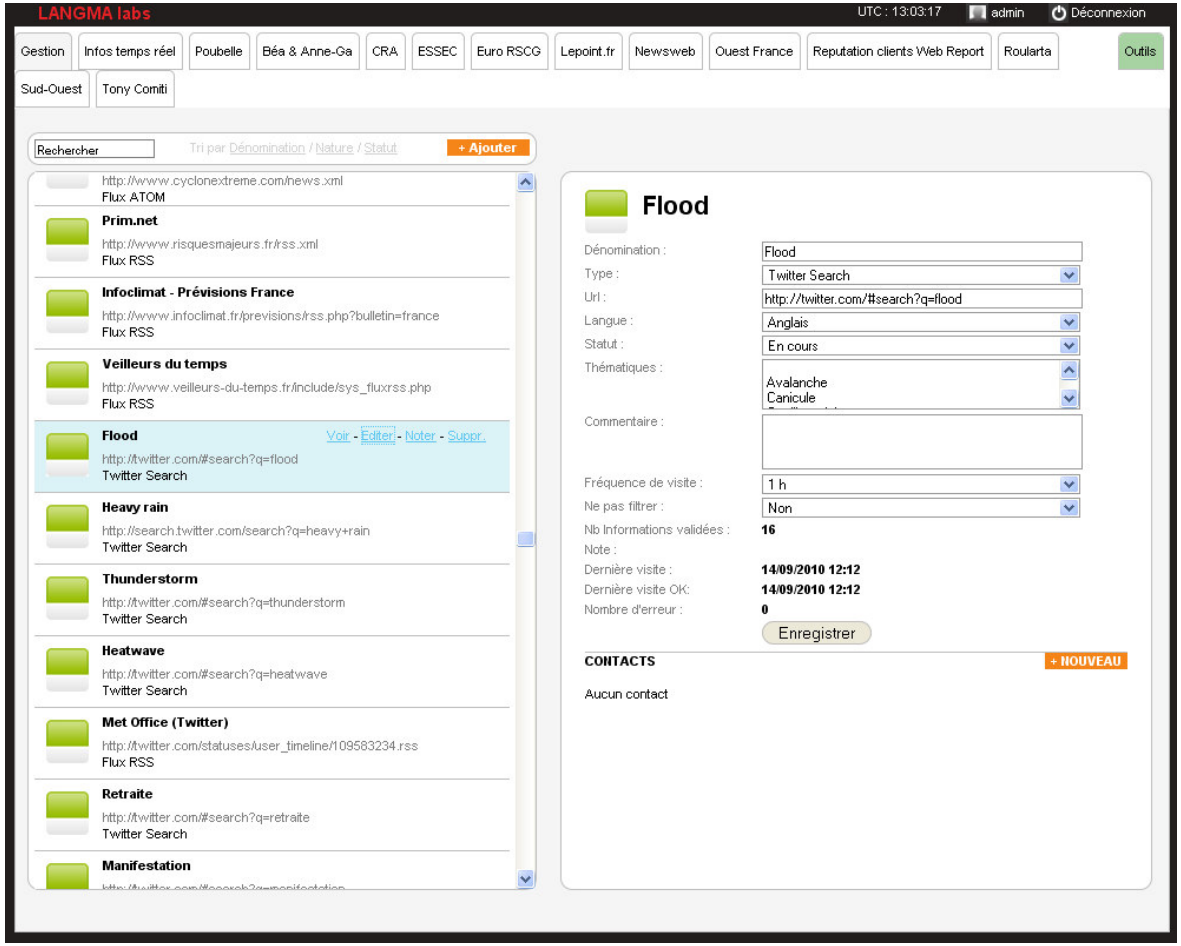

Figure 10 : interface de gestion des sources

Les informations issues des sources sont automatiquement filtrées par les méthodes d'analyse et de classification décrites dans cet article. Ces sources peuvent alors être sélectionnées et vérifiées par les journalistes avant d'être classées, réécrites et/ou diffusées aux agences de presse.

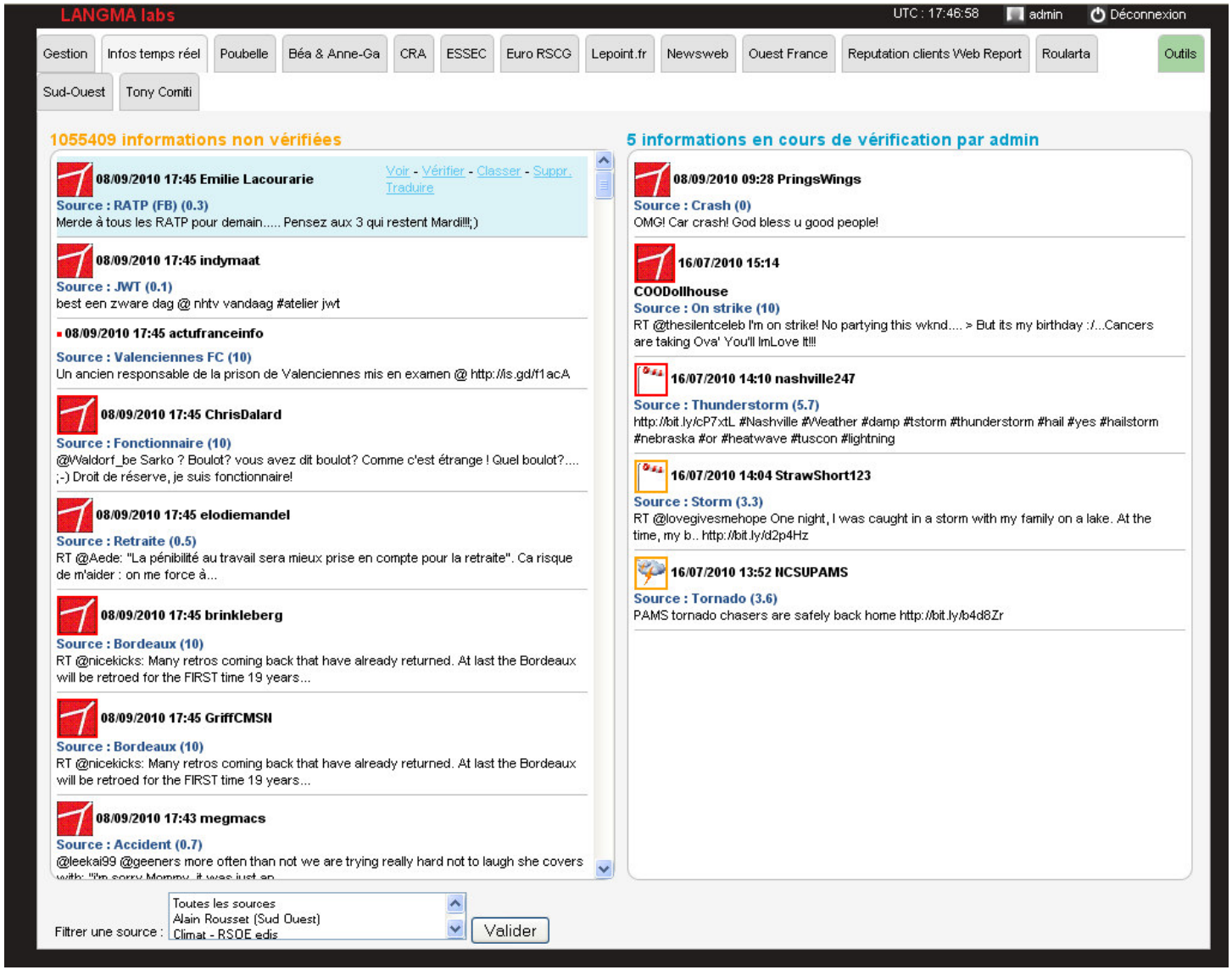

Figure 11 : interface de gestions des informations

Des pictogrammes associés aux différentes thématiques, favorisent l'identification de la thématique détectée par l'analyseur sémantique. L'écran est rafraîchi automatiquement toutes les minutes sans rechargement de la page : les nouvelles informations viennent s'empiler dans la colonne « Informations non traitées/vérifiées ».

5.3 Interface de consultation

Enfin, la dernière interface web sécurisée qui est proposée dans le cadre du projet Langma est dédiée aux clients. Elle donne un accès aisé aux informations diffusées par notre système.

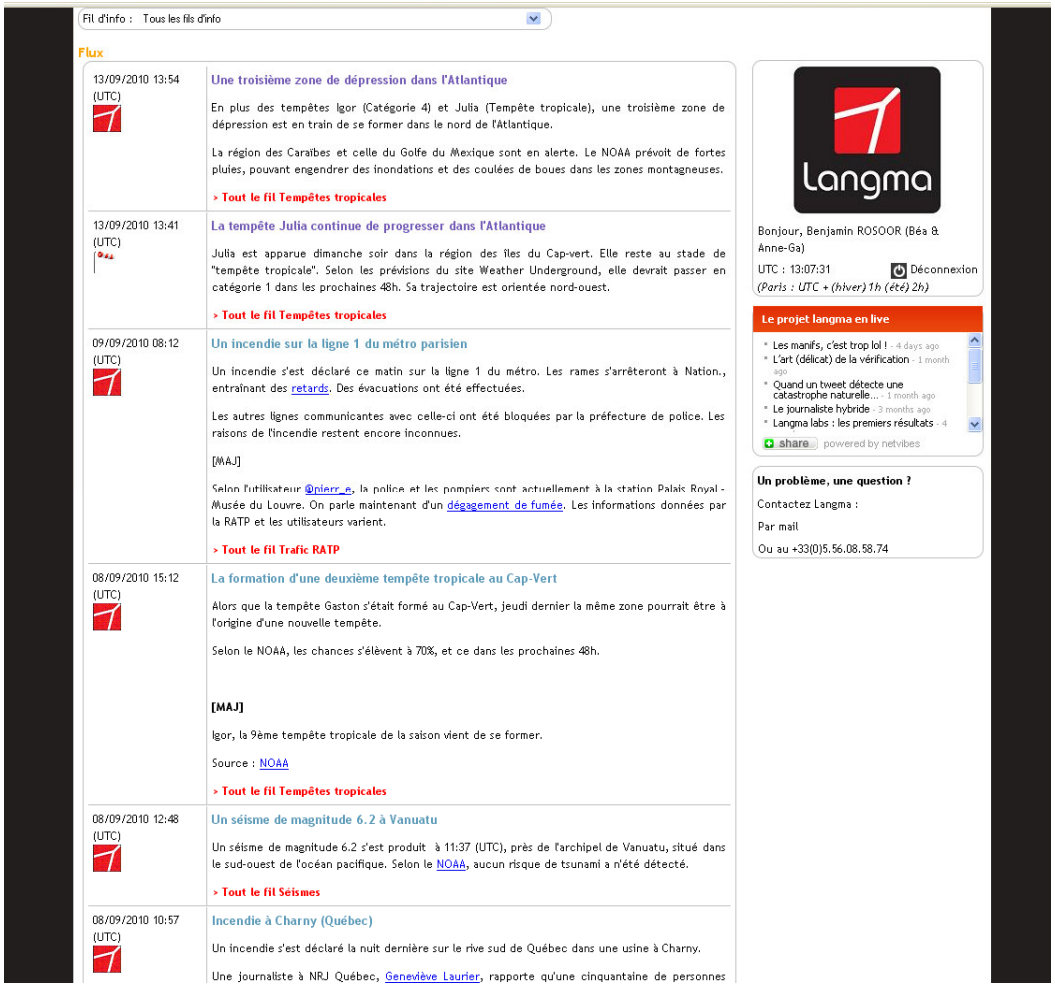

Figure 12 : interface de consultation client

6 Conclusions et perspectives

Dans cet article, nous avons décrit une méthode destinée à des journalistes / documentalistes dans leur tâche de veille, et qui exploite les nouvelles publications massives de type "Statuses" (tweets, statut Facebook, etc). Cette méthode permet d'identifier des catastrophes à partir d'un ensemble de publications et de les classer en différentes catégories. Nous avons implémenté une interface reposant sur cette méthode et réalisé des expérimentations sur des jeux de données réelles. Ces expérimentations ouvrent de nombreuses perspectives comme la prise en compte d'outils linguistiques plus complexes ou l'introduction de connaissances expertes pour définir de manière plus précise les vecteurs représentant les thèmes et ainsi améliorer le processus de classification. Par ailleurs, nous pourrons utiliser des descripteurs plus fins spécifiques aux tweets [7,8].

Références

[1] ANGEL A., KOUDAS N., SARKAS N., SRIVASTAVA D., *What's on the grapevine?* Proceedings of SIGMOD, 2009

[2] BANSAL N., KOUDAS N., *BLOGSCOPE: A system for online analysis of high volume text streams*, Proceedings of WebDB, 2007

[3] BARBOSA L., FENG J., *Robust Sentiment Detection on Twitter from Biased and Noisy Data*, Proceedings of COLING, 2010

[4] BAYOUDH I., BÉCHET N., ROCHE M., *Blog classification: Adding Linguistic Knowledge to Improve the K-NN Algorithm*. Proceedings of IIP'08 (International Conference on Intelligent Information Processing), Springer IFIP, p68-77, 2008

[5] BOYD D., GOLDER S., LOTAN G., *Tweet, tweet, retweet: Conversational aspects of retweeting on twitter*. Proceedings of HICSS-43, 2010

[6] CHA M., HADDADI H., BENEVENUTO F., GUMMADI K.P., *Measuring User Influence in Twitter: The Million Follower Fallacy*, Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM), Washington, DC, 2010

[7] DAVIDIV D., TSUR O., RAPPOPORT A., *Enhanced Sentiment Learning Using Twitter Hashtags and Smileys*, Proceedings of COLING, 2010

[8] DUAN Y., JIANG L., QIN T., ZHOU M., SHUM H., *An Empirical Study on Learning to Rank of Tweets*, Proceedings of COLING, 2010

[9] GINSBERG J., MOHEBBI M.H., PATEL R.S., BRAMMER L., SMOLINSKI M.S., BRILLIANT L., *Detecting influenza epidemics using search engine query data*, Nature, p1012-1014, 2009

[10] HOTHO A., NÜRNBERGER A., PAASS G., *A Brief Survey of Text Mining*, LDV Forum 20(1), p19-62, 2005

[11] JIANG L., CAI Z., WANG D., JIANG S., *Survey of Improving K-Nearest-Neighbor for Classification*, Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery, Volume 1, p679-683, 2007

[12] KOSTKOVA P., QUINCEY E., JAWAHEER G., *The potential of Twitter for early warning and outbreak detection*, Proceedings of ECCMID, 2010

[13] PORTER M.F., *An algorithm for suffix stripping*, Program. 14(3) p130–137, 1980

[14] LESKOVEC J., BACKSTROM L., KLEINBERG J.M.. *Meme-tracking and the dynamics of the news cycle*, Proceedings of KDD, 2009

[15] MATHIOUDAKIS M., KOUDAS N., *TWITTERMONITOR: trend detection over the twitter Stream*, Proceedings of SIGMOD Conference, p.1155-1158, 2010

[16] MILSTEIN S., CHOWDHURY A., HOCHMUTH G., LORICA B., MAGOULAS R.. *Twitter and the micro-messaging revolution: Communication,connections, and immediacy*, O'Reilly Media, 2008

[17] O'CONNOR B., BALASUBRAMANYAN R., ROUTEDGE B., SMITH N., *From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series*, Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM). Washington, DC, 2010

[18] ROCHE M., KODRATOFF Y., *Pruning Terminology Extracted from a Specialized Corpus for CV Ontology Acquisition*, Proceedings of onToContent'06) - OTM'06, Springer-Verlag, LNCS, p1107-1116, 2006

[19] SAKAKI T., OKAZAKI M., MATSUON Y., *Earthquake shakes Twitter users: real-time event detection by social sensors*, Proceedings of the 19th international conference on World wide web (WWW), p.851–860, New York, NY, USA, 2010