Un système de détection de thématiques populaires sur Twitter

Adrien Guille*, Cécile Favre*

*Laboratoire ERIC, Université Lumière Lyon 2
http://mediamining.univ-lyon2.fr/people/guille/egc2014.php
http://mediamining.univ-lyon2.fr/people/guille/twitterstream.php

1 Introduction

Twitter offre des fonctionnalités de microblogging qui sont utilisées par des millions de personnes à travers le monde pour publier des messages courts. Ces personnes créent et partagent de l'information liée à divers types d'évènements, allant d'évènements personnels banals à des évènements importants et/ou globaux, quasiment en temps-réel. L'explosion du nombre d'utilisateurs de ce service de réseautage social a entraîné l'apparition d'un phénomène de surcharge informationnelle. Pour lutter contre cela, il est nécessaire de doter les utilisateurs de moyens leur permettant d'identifier plus facilement les éléments d'information les plus intéressants et de se tenir au courant des derniers évènements significatifs.

L'information brute produite par Twitter est délivrée sous la forme d'un flux de messages courts. Par conséquent la manière dont ceux-ci arrivent au fil du temps recèle une part importante de leur signification. La dynamique temporelle des thématiques les plus populaires sur ces réseaux est constituée d'une succession de focus et dé-focus, autrement dits, une succession de *pics* de popularité (Leskovec et al., 2009). C'est pourquoi de nombreuses approches – allant de méthodes basées sur la fréquence des mots (Benhardus et Kalita, 2013) jusqu'à des méthodes plus complexes reposant sur des modèles de thématiques probabilistes dynamiques (Lau et al., 2012) – ont été proposées dans le but d'identifier ce genre de thématiques. Dans cet article, nous présentons un système implémentant la méthode décrite par Guille et Favre (2014). Contrairement à la majorité des méthodes existantes, celle-ci prend en compte l'aspect social du flux traité en considérant la fréquence de création de liens dynamiques entre utilisateurs. Un utilisateur crée un lien dynamique en insérant une ou plusieurs mentions (*i.e.* « @pseudonyme ») dans un tweet. Ce lien entre utilisateurs est dynamique car lié au contenu du tweet et sa durée de vie. Par ailleurs, cette méthode localise plus précisément dans le temps les thématiques que les méthodes existantes et traite les grands volumes de données plus efficacement que celles à base de modèles de thématiques probabilistes.

La suite de cet article est organisée comme suit. Dans la section 2 nous décrivons le fonctionnement du système puis dans la section 3 nous détaillons le cadre de la démonstration. Enfin dans la section 4 nous concluons.

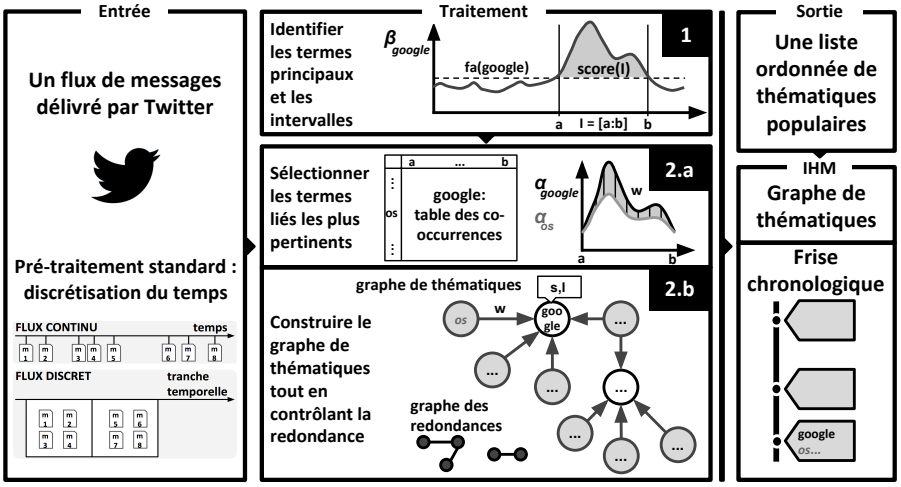

FIG. 1 – *Schéma du fonctionnement du système.*

2 Le système de détection des thématiques populaires

L'objectif du système est d'identifier des thématiques à la fois riches de sens et précisément localisées dans le temps. Son fonctionnement est schématisé par la figure 1.

Entrée. Le système traite un flux de messages produit par Twitter. Le vocabulaire des termes employés dans ces messages est noté V. L'axe temporel est discrétisé en partitionnant les messages dans des tranches temporelles de même durée (*cf.* la figure 1 pour une illustration de ce pré-traitement). Ce pré-traitement est commun à toutes les méthodes existantes.

Sortie. Le système génère une liste de thématiques, ordonnées selon leur popularité. Une thématique est définie par un terme principal, une liste pondérée de termes liés et un intervalle temporel. Par exemple, la thématique : {["google", {("chrome",0.8), ("os", 0.8), ("desktop", 0.75) }], ['19/11/09' ;'20/11/09']}, capture l'évènement créé par la sortie de Google Chrome OS le 19 novembre 2009.

Traitement. La tâche d'identification des thématiques populaires est décomposée en 3 problèmes : (1) l'identification des termes principaux et des intervalles temporels, lesquels sont associés à un score de popularité ; (2.a) la sélection de termes liés pertinents ; (2.b) la construction du graphe des redondances et du graphe de thématiques, duquel est extrait la liste finale de thématiques. La méthode se déroule comme suit. Tout d'abord, le problème (1) est résolu pour chaque terme appartenant au vocabulaire V à travers l'analyse de l'anomalie dans la fréquence de création de liens dynamiques. Ensuite, pour chaque couple de terme principal et intervalle temporel, le problème (2.a) est résolu afin d'identifier l'ensemble pondéré de termes liés. Chaque thématique ainsi constituée est insérée dans le graphe de thématiques si elle n'est pas redondante avec une autre thématique déjà présente (2.b). Les redondances constatées sont modélisées par un second graphe, qui permet d'identifier les thématiques à fusionner à la fin du processus, avant d'extraire la liste des thématiques populaires qui sera retournée à l'utilisateur.

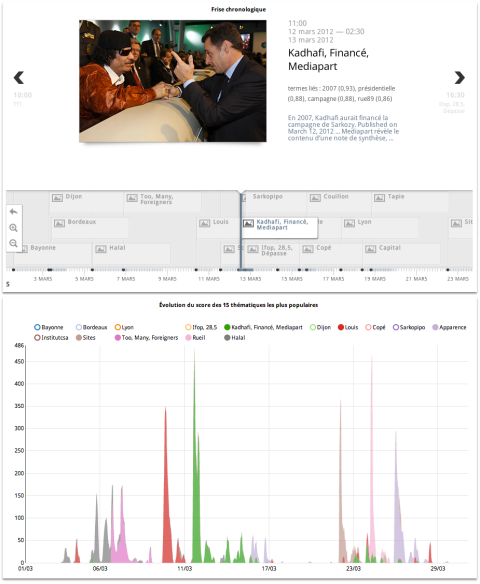

FIG. 2 – *En haut, la frise chronologique générée automatiquement. L'utilisateur navigue dans le temps à l'aide du ruban. Il obtient des détails lorsqu'il sélectionne une thématique. En bas, les scores de popularité, où l'aire de couleur verte correspond à la thématique sélectionnée dans la frise.*

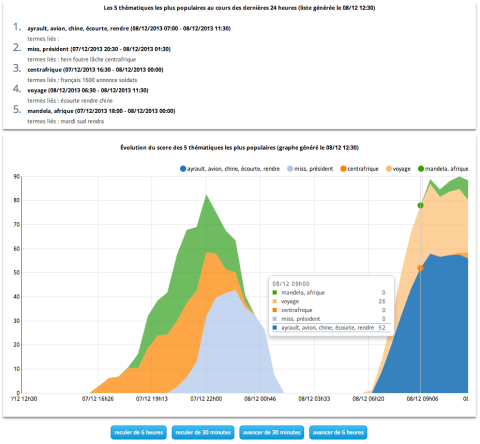

FIG. 3 – *En haut est présentée la liste des 5 thématiques les plus populaires du 7 décembre à 12h30 au 8 décembre 2013 à 12h30. L'évolution de leur score dans le temps est présentée en dessous. L'utilisateur peut parcourir l'historique complet des thématiques détectées depuis le lancement du système en décembre 2013.*

Interfaces utilisateur. En plus de la liste de thématiques triées par popularité décroissante, le système permet à l'utilisateur d'explorer les résultats de trois autres manières : (i) en navigant dans le graphe de thématiques, (ii) en parcourant la frise chronologique et (iii) en explorant le graphique interactif des scores en fonction du temps.

3 Cadre de la démonstration

Le système est capable de traiter à la fois des données statiques, afin par exemple de fournir une vision rétrospective des thématiques marquantes durant une période d'observation, et des données dynamiques, autrement dit le flux délivré en temps réel par Twitter. Dans ce cas, le système met à jour le modèle de façon incrémentale et il peut alors être utilisé pour suivre les thématiques les plus populaires en temps réel. Nous décrivons ci-après les données qui seront traitées lors de la démonstration.

Données statiques. Deux corpus seront analysés : le premier correspond à l'intégralité des tweets publiés par plus de 50.000 utilisateurs anglophones d'octobre à décembre 2009 (Yang et Leskovec, 2011). Le second correspond à des tweets en français publiés durant la campagne

électorale ayant précédée l'élection présidentielle de 2012. Chaque corpus contient plusieurs millions de tweets.

Données dynamiques. Le système interroge le flux public de Twitter et collecte en permanence des tweets francophones mentionnant François Hollande. Les thématiques populaires sont mises à jour chaque fois qu'une nouvelle tranche temporelle de tweets est disponible.

Scénario. Le scénario de la démonstration consistera à analyser ces données à l'aide du système, en faisant varier ses paramètres et en explorant les résultats avec ses interfaces. L'efficacité de la méthode implémentée permet de traiter plusieurs millions de tweets en moins d'une minute avec un PC standard. La figure 2 montre un extrait des résultats/interfaces générés à partir du corpus statique francophone. La figure 3 montre un extrait des résultats/interfaces générés à partir des données collectées en temps réel par le système.

4 Conclusion

Nous avons présenté un système de détection de thématiques populaires sur Twitter. Les URL données dans l'en-tête de cet article permettent le téléchargement du prototype et la consultation des thématiques détectées en continu à partir du flux public de Twitter. Afin de faciliter encore plus sa réutilisation, le système sera intégré à la plateforme d'analyse et de fouille de données sociales SONDY (Guille et al., 2013).

Remerciements. Ces travaux ont été partiellement financés par l'ANR et le projet Imagi-Web (contrat ANR-2012-CORD-002-01).

Références

Benhardus, J. et J. Kalita (2013). Streaming trend detection in twitter. *IJWBC 9*(1), 122–139.

Guille, A. et C. Favre (2014). Une méthode pour la détection de thématiques populaires sur Twitter. In *Actes de la conférence EGC*.

Guille, A., C. Favre, H. Hacid, et D. Zighed (2013). SONDY : An open source platform for social dynamics mining and analysis. In *Proc. of the SIGMOD conference*, pp. 1005–1008.

Lau, J. H., N. Collier, et T. Baldwin (2012). On-line trend analysis with topic models : #twitter trends detection topic model online. In *Proc. of the COLING conference*, pp. 1519–1534.

Leskovec, J., L. Backstrom, et J. Kleinberg (2009). Meme-tracking and the dynamics of the news cycle. In *Proc. of the SIGKDD conference*, pp. 497–506.

Yang, J. et J. Leskovec (2011). Patterns of temporal variation in online media. In *Proc. of the WSDM conference*, pp. 177–186.

Summary

With the ever-growing amount of messages exchanged via Twitter, there is an increasing interest in filtering this information, which is delivered under the form of a stream of messages. In this paper, we present a system for detecting popular topics in Twitter. The system can be applied to static corpora and can also handle the live Twitter stream.