Une méthode pour la 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

Résumé. L'explosion du volume de messages échangés via Twitter entraîne un phénomène de surcharge informationnelle pour ses utilisateurs. Il est donc crucial de doter ces derniers de moyens les aidant à filtrer l'information brute, laquelle est délivrée sous la forme d'un flux de messages. Dans cette optique, nous proposons une méthode basée sur la modélisation de l'anomalie dans la fréquence de création de liens dynamiques entre utilisateurs pour détecter les pics de popularité et extraire une liste ordonnée de thématiques populaires. Les expérimentations menées sur des données réelles montrent que la méthode proposée est capable d'identifier et localiser efficacement les thématiques populaires.

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 réseau 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. 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 est constituée d'une succession de focus et dé-focus, autrement dits, une succession de *pics* de popularité. C'est pourquoi de nombreuses approches – allant de méthodes basées sur la fréquence des mots jusqu'à des méthodes plus complexes reposant sur des modèles de thématiques probabilistes dynamiques – ont été proposées dans le but d'identifier ce genre de thématiques. Ces méthodes reposent sur des stratégies variées de détection des pics et produisent des résultats très différents. Nos travaux s'intéressent au filtrage et à l'identification de thématiques à partir de l'information contenue dans un flux de messages produits par Twitter afin de, entre autres, fournir une vue rétrospective des thématiques les plus populaires ou bien recommander des éléments d'information intéressants en temps réel. Une bonne solution doit satisfaire deux critères : d'une part les thématiques identifiées doivent être précisément localisées dans le temps et intelligibles et d'autre part, la méthode doit pouvoir passer à l'échelle et traiter de grands

volumes de données. Afin de conserver une complexité temporelle raisonnable, beaucoup de méthodes existantes assimilent une thématique à un simple mot. C'est le cas par exemple d'approches basées sur l'analyse de la fréquence des mots telles que la méthode « Peaky Topics » (Shamma et al., 2011) ou la méthode basée sur l'indice MACD (Rong et Qing, 2012). Un pic de popularité se traduit par l'augmentation soudaine de la fréquence d'un mot. Cette définition n'est pas toujours appropriée à cause de l'ambiguïté possible. Pour pallier à ce problème, Benhardus et Kalita (2013) proposent d'étudier des n-grammes de mots, mais les n-grammes ne peuvent capturer les relations entre des mots trop éloignés dans le corps d'un message et sont très sensibles au bruit. Afin d'identifier des thématiques plus explicites, de nombreuses méthodes se basant sur des modèles probabilistes ont été développées, telles que OLDA (Al-Sumait et al., 2008) ou Online-LDA (Lau et al., 2012). Un pic de popularité se traduit alors par une variation soudaine de la distribution des thématiques. Cependant, l'introduction de la dimension temporelle dans ces modèles augmente la complexité des mécanismes d'inférence mis en œuvre (il faut notamment faire en sorte que les thématiques qui évoluent peu dans le temps restent comparables au sein du modèle, et également faire en sorte que le modèle ait un niveau de sensibilité constant de sorte qu'il puisse détecter de nouvelles thématiques au fil du temps), ce qui limite leur capacité de passage à l'échelle. Par ailleurs, Aiello et al. (2013) ont montré que les méthodes à base de modèles probabilistes dynamiques ne sont pas efficaces sur des flux sociaux trop hétérogènes au sein desquels de nombreux éléments d'information sont relatés simultanément. Qui plus est, la très vaste majorité des méthodes existantes ignore un aspect important de ces flux, précisément leur aspect social. En effet, un message ne se limite pas à un simple contenu textuel et il est notamment possible d'y insérer une ou plusieurs « mentions » (à l'aide de la syntaxe « @pseudonyme » dans le corps des messages). Lorsque l'auteur d'un message insère une mention, il crée en réalité un lien dynamique vers un autre utilisateur. Ce lien est considéré comme dynamique puisque sa création est datée et liée à un contenu particulier, celui du message. Les mentions permettent aux utilisateurs d'exprimer leur volonté d'engager la discussion à propos du contenu du message avec la ou les personnes ciblées, et traduisent donc l'intérêt qu'ils portent à la thématique liée. À notre connaissance, les travaux menés par Takahashi et al. (2011) sont les seuls à intégrer cette caractéristique. La méthode qu'ils proposent repose sur une modélisation probabiliste du comportement de chaque utilisateur du réseau en terme de création de liens dynamiques. En détectant les points de rupture par rapport à ces comportements standards, il est possible d'identifier des thématiques émergentes, une thématique étant définie comme un simple mot. Outre cette définition qui limite l'intérêt des résultats obtenus, la méthode souffre de la complexité de la phase d'apprentissage du modèle puis de détection qui rend quasiment impossible sa mise en œuvre dans des conditions réelles (*i.e.* un réseau comportant un grand nombre d'utilisateurs). Globalement, il apparaît nécessaire de développer des méthodes mieux adaptées. Le reste de cet article est organisé comme suit. Dans la section suivante, nous présentons une nouvelle méthode pour la détection de thématiques populaires sur Twitter, puis nous présentons les résultats obtenus dans la section 3.

2 Méthode proposée

L'objectif de la méthode est d'identifier des thématiques à la fois riches de sens et précisément localisées dans le temps, tout en tenant compte de l'aspect social du flux de messages.

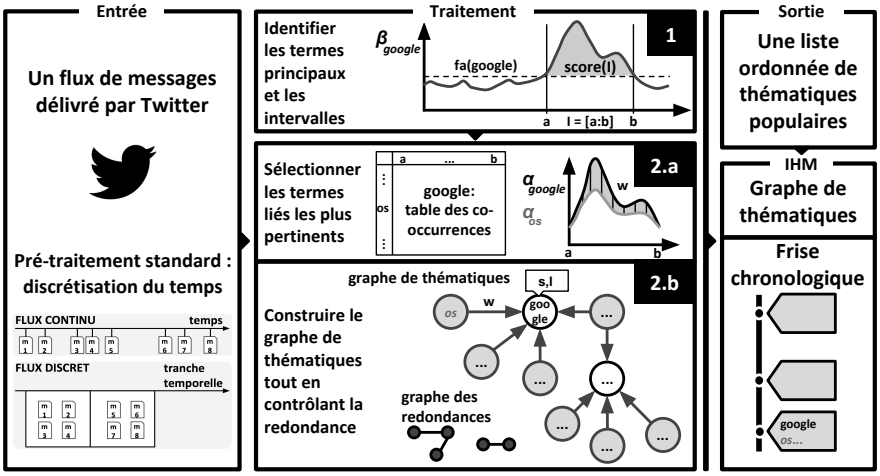

FIG. 1 – *Grands principes de la méthode proposée.*

Entrée. Nous traitons un flux produit par Twitter contenant M messages. Le vocabulaire des termes employés dans ces messages est noté V. Nous discrétisons l'axe temporel en partitionnant les messages en n tranches temporelles de même durée (*cf.* la figure 1 pour une illustration de ce pré-traitement). Cette étape de pré-traitement est commune à toutes les approches de détection de thématiques temporelles citées précédemment. On note $\alpha_t(i)$ la fonction qui donne le nombre de messages inclus dans la *i*^{ème} tranche temporelle qui contiennent au moins une occurrence du terme *t*. $\beta_t(i)$ désigne la fonction qui donne le nombre de messages inclus dans la *i*^{ème} tranche temporelle qui contiennent au moins une « mention » et le terme *t*. Les séries temporelles correspondantes sont notées α_t et β_t.

Sortie. La méthode 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.

Grands principes de la méthode. Nous décomposons la tâche d'identification des thématiques populaires en 3 problèmes : (1) l'identification des termes principaux et des intervalles temporels, chaque couple étant associé 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. Ensuite, pour chaque couple de terme principal et intervalle temporelle, 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. La figure 1 décrit le déroulement de la

méthode ainsi que les IHM permettant de visualiser les thématiques identifiées.

2.1 Identification des termes principaux et des intervalles temporels

L'objectif de cette première étape est de produire une liste ordonnée de thématiques, chacune étant définie par un terme principal, un intervalle temporel et un score caractérisant sa popularité. Notre hypothèse est que la fréquence de création de liens dynamiques (*i.e.* fréquence des « mentions ») liée à une thématique est un meilleur indicateur du niveau d'attention qu'elle reçoit de la part des utilisateurs que sa fréquence d'apparition globale. Nous utilisons donc cette mesure pour localiser dans le temps les thématiques et estimer leur niveau de popularité.

Calcul du score de popularité. Nous définissons d'abord la fréquence de mentions attendues en chaque tranche temporelle pour le terme t : $fa(t)$ (eq. 1). Nous exprimons ensuite l'anomalie de la fréquence des mentions liées à ce terme à la $i^{ème}$ tranche temporelle selon la formule donnée par l'équation 2.

$$fa(t) = \frac{\sum_{i=1}^{n} \beta_t(i)}{n} \qquad (1) \qquad\qquad anomalie(t,i) = \beta_t(i) - fa(t) \qquad (2)$$

Nous étendons cette notion d'anomalie à un intervalle pour calculer le score de popularité d'un terme t durant un intervalle temporel $I = [a; b]$. Nous définissons ce score comme étant l'aire algébrique sous la fonction d'anomalie sur l'intervalle $[a; b]$, laquelle est obtenue en intégrant la fonction discrète calculant l'anomalie, ce qui équivaut à une simple somme :

$$score(t, I) = \int_a^b anomalie(t, i)\,\mathrm{d}i = \sum_{i=a}^{b} anomalie(t, i)$$

Identification de la période de popularité. Identifier l'intervalle I durant lequel un terme t était le plus populaire revient à trouver l'intervalle qui maximise la valeur de $score(t, I)$. Or, nous venons de montrer que ce score est obtenu en sommant la fonction d'anomalie. Par conséquent, identifier l'intervalle le plus populaire revient à résoudre un problème du type « sous-séquence contiguë de somme maximale » (*SSCSM*). Nous résolvons ce problème de type *SSCSM* à l'aide de l'algorithme en temps linéaire décrit par Bentley (1984).

2.2 Sélection de termes liés pertinents

Afin de préciser une thématique décrite par un terme principal t et un intervalle temporel I, nous sélectionnons un ensemble de termes liés S. Afin de limiter la surcharge d'information pour l'utilisateur, il faut que cet ensemble soit d'une taille raisonnable et qu'il contienne des termes pertinents durant l'intervalle temporel.

L'ensemble des termes liés potentiels est réduit aux p termes les plus co-occurrents avec t durant I. Pour sélectionner les termes les plus pertinents parmi ceux-ci, nous proposons de calculer un poids $w_q \in [0; 1]$ pour chaque terme $t'_q \in S$ basé sur la corrélation entre la dynamique temporelle de t et t'_q durant l'intervalle I. Seuls les termes dont le poids dépasse un certain seuil noté θ sont conservés. Les paramètres p et θ sont fixés par l'utilisateur de la méthode. Nous estimons la corrélation entre les séries α_t et $\alpha_{t'_q}$ à l'aide du coefficient récemment proposé par Erdem et al. (2012). Ce coefficient, développé à l'origine par les auteurs pour analyser des données boursières *non-stationnaires*, est particulièrement adapté aux données que nous

traitons. Par soucis de concision, nous donnons directement la formule d'approximation de la corrélation entre la dynamique du terme t et le terme lié t'_q durant l'intervalle $I = [a; b]$:

$$\rho_{Ot,t'_q} = \frac{\sum_{i=a+1}^{b} A_{t,t'_q}}{(b-a-1)A_t A_{t'_q}} \quad \text{où} \quad \begin{aligned} A_{t,t'_q} &= (\alpha_t(i) - \alpha_t(i-1))(\alpha_{t'_q}(i) - \alpha_{t'_q}(i-1)) \\ A_t^2 &= \frac{\sum_{i=a+1}^{b}(\alpha_t(i) - \alpha_t(i-1))^2}{b-a-1} \\ A_{t'_q}^2 &= \frac{\sum_{i=a+1}^{b}(\alpha_{t'_q}(i) - \alpha_{t'_q}(i-1))^2}{b-a-1} \end{aligned}$$

La preuve que $|\rho_O| \leqslant 1$ est donnée par Erdem et al. (2012). Afin d'obtenir une valeur comprise entre 0 et 1, nous définissons le poids du terme t'_q : $w_q = \frac{\rho_{Ot,t'_q}+1}{2}$.

2.3 Construction des graphes de thématiques et des redondances

Afin de générer l'ensemble final des thématiques retourné à l'utilisateur, nous construisons deux structures de graphe : le graphe de thématiques et le graphe des redondances. Le premier est un graphe orienté composé de nœuds appartenant à deux classes : les nœuds représentant les termes principaux, lesquels sont annotés par un intervalle et un score, et les nœuds représentant les termes liés. Ces derniers sont connectés à l'aide d'arcs pondérés, dirigés vers les termes principaux. Le second est un simple graphe non-orienté servant à représenter la redondance entre certaines thématiques. Chaque thématique $T = (t, I, S)$ générée par la résolution des deux problèmes précédemment décrits est insérée dans le graphe des thématiques si elle n'est pas jugée redondante avec une thématique déjà présente. Une thématique T_1 est jugée redondante avec T_2 si le terme principal t_1 serait mutuellement connecté avec le terme t_2 et si l'intersection entre I_1 et I_2 mesurée par $\frac{|I_1 \cup I_2|}{\min(|I_1|,|I_2|)}$ est importante (*i.e.* dépasse un seuil $\sigma < 1$). Dans le cas où la thématique à insérée T_1 est jugée redondante, sa définition est mise de côté et une arrête liant t_1 et t_2 est ajoutée au graphe des redondances. Les thématiques étant ordonnées selon leur score de popularité, l'utilisateur peut paramétrer le nombre k de thématiques qu'il souhaite, et seules le k plus populaires et non-redondantes lui seront présentées. Une fois les deux graphes construits, l'identification des thématiques qui peuvent être fusionnées ensemble consiste en l'identification des composantes connexes au sein du graphe des redondances. Cela se fait en temps linéaire à l'aide de l'algorithme décrit par Hopcroft et Tarjan (1973). Le terme principal de la thématique fusionnée devient l'agrégation des termes principaux, et seuls les p termes liés avec les p plus grands poids sont conservés. En parcourant les nœuds de la classe principale du graphe de thématiques on reconstruit les thématique une par une à partir des annotations du nœud principal et des termes liés qui y sont connectés.

3 Évaluation de la méthode

La méthode a été expérimentée sur un flux d'1.5M de tweets publiés en novembre 2009. La précision a été mesurée selon la définition donnée par Weng et Lee (2011) (*i.e.* la fraction de thématiques retournées qui correspondent à des événements réalistes) pour des valeurs de k allant de 15 à 35. Le tableau 1 donne ces mesures, tandis que le tableau 2 présente un extrait des thématiques identifiées. On remarque que la mise en valeur des termes principaux facilite la compréhension des thématiques (#3 veterans, en référence au « Veterans' Day » le 11

paramètres : $\theta = 0.7, \sigma = 0.85, p = 8$		
k = 15	k = 25	k = 35
P@15 : 80%	P@25 : 76%	P@35 : 77.1%
46 sec.	54 sec.	57 sec.

TAB. 1 – *Précision et temps de calcul de la méthode proposée, pour un nombre de thématiques (k) allant de 15 à 35.*

Intervalle	Thématique
du 10, 03h00 au 12, 08h00	(3) **veterans** : served(0.8) country(0.8) military(0.7) happy(0.7)
du 27, 05h00 au 29, 06h00	(7) **tiger, woods** : accident(0.9) car(0.8) crash(0.8) injured(0.8) seriously(0.8) hospital(0.8)
du 13, 11h30 au 15, 05h30	(32) **water** : moon(0.9) nasa(0.8) found(0.8) significant(0.7) amount(0.7)

TAB. 2 – *Trois des thématiques identifiées : la position en terme de popularité est donnée devant le terme principal (en gras).*

novembre aux USA, #7 tiger, woods (composé donc issu d'une fusion), victime d'un accident, #32 water, *i.e.* l'eau trouvée sur la Lune par la NASA). L'hypothèse selon laquelle β (fréquence des mentions) est un meilleur indicateur de popularité qu'α a été vérifiée expérimentalement, puisqu'une version de la méthode se basant exclusivement sur α a obtenu une précision@n systématiquement inférieure et a globalement détectée les thématiques avec du retard.

Références

Aiello, L. M., G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, Y. Kompatsiaris, et A. Jaimes (2013). Sensing trending topics in twitter. *IEEE TMM 15*(6), 1–15.

AlSumait, L., D. Barbará, et C. Domeniconi (2008). On-line LDA : Adaptive topic models for mining text streams with applications to topic detection and tracking. In *ICDM*, pp. 3–12.

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

Bentley, J. (1984). Programming pearls : algorithm design techniques. *CACM 27*(9), 865–873.

Erdem, O., E. Ceyhan, et Y. Varli (2012). A new correlation coefficient for bivariate time-series data. In *MAF*, pp. 58–73.

Hopcroft, J. et R. Tarjan (1973). Algorithm 447 : efficient algorithms for graph manipulation. *CACM 16*(6), 372–378.

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

Rong, L. et Y. Qing (2012). Trends analysis of news topics on twitter. *IJMC 2*(3), 327–332.

Shamma, D. A., L. Kennedy, et E. F. Churchill (2011). Peaks and persistence : modeling the shape of microblog conversations. In *CSCW*, pp. 355–358.

Takahashi, T., R. Tomioka, et K. Yamanishi (2011). Discovering emerging topics in social streams via link anomaly detection. In *ICDM*, pp. 1230–1235.

Weng, J. et B.-S. Lee (2011). Event detection in twitter. In *ICWSM*, pp. 401–408.

Summary

We propose a new method for detecting popular topics on Twitter. It mainly relies on the modeling of the anomaly in the frequency of creation of dynamic links between users.