Prédictions d'activité dans les réseaux sociaux en ligne

François Kawala [1,2], **Ahlame Douzal-Chouakria** [2], **Eric Gaussier** [2], **Eustache Diemert** [1]

1. Laboratoire d'Informatique de Grenoble
Bat. CE4, allée de la Palestine, F-38610, Gières, France
{francois.kawala,alhame.douzal,eric.gaussier}@imag.fr

2. BestofMedia Group
4 rue des méridiens, F-38130, Echirolles, France
{fkawala,ediemert}@bestofmedia.com

RÉSUMÉ. *Les sites dédiés aux réseaux sociaux sont le théâtre de nouveaux phénomènes sociaux. Ainsi certains mot-clés connaissent une augmentation abrupte de leur popularité, caractérisée par un grand nombre de discussions sur un court laps de temps. Ces pics d'activité sont fréquemment qualifiés de "buzz". Dans cet article, nous traitons le problème de la prédiction du volume d'activité relatif à un mot-clé sans connaissances a priori des réseaux sociaux hôtes. Pour ce faire nous proposons une définition des médias sociaux centrée sur les contenus échangés. L'évaluation de notre approche est conduite à échelle industrielle sur deux réseaux sociaux : Twitter, une plate-forme axée sur la spontanéité (Kwak* et al.*, 2010), et Tom's Hardware, un réseau de forums dédiés aux nouvelles technologies. Nous montrons que le volume d'activité associés à un mot-clé peut être précisément prédit.*

ABSTRACT. *Online Platforms dedicated to social networking host new social phenomenons. Thus several keywords may suddenly take an unprecedented importance, reflecting the number of discussions they have raised within a short time period. Such bursts in topic discussions are usually referred to as buzz events. We address in this paper the problem of predicting the activity volume associated to a given keyword without a priori knowledge on the underlying social network. To do so, we propose to define social netowrk on a content-centric way. Our approach is evaluated at "industrial scale" on two different social networks: Twitter, a platform with extremely fast dynamics (Kwak* et al.*, 2010), and Tom's Hardware, a worldwide forum network focusing on new technology. The experiments conducted reveal that it is possible to predict activity volume associated to a keyword in social media with high accuracy.*

MOTS-CLÉS : *Séries temporelles, Réseaux sociaux, Twitter, Buzz, Fouille de données.*

KEYWORDS: *Time-series, Social-network, Twitter, Buzz, Large-scale data-mining.*

---

1. Introduction

A notre connaissance, le phénomène de “buzz” n’est pas formellement défini. Faute de définition indépendante de son contexte d’utilisation, nous nous contentons ici de décrire le phénomène de buzz. Celui-ci est caractérisé par : un thème support (*ie.* un ensemble de mot-clés), une période, et enfin une mesure d’attention. La période du buzz commence lorsque le thème support capte une part suffisante (au regard du contexte d’utilisation) de l’attention disponible sur le réseau social cible. Elle ce termine quand la part d’attention capté par le thème support n’est plus suffisante. On observe qu’un buzz peut se déclarer graduellement ou abruptement. Par exemple, un tremblement de terre est généralement le support d’un buzz abrupt, au contraire, un téléphone haut de gamme largement attendu peut être le support d’un buzz graduel.

Ne pouvant définir un buzz sans connaître le contexte d’utilisation, nous traiterons le problème de prédiction du volume d’activité associé à un mot-clé. Celui-ci peut-être défini informellement comme suit. Considérons le cas d’un opérateur observant un réseau social. Ce dernier observe une multitude d’échanges associés à différent mots-clés. Afin de maximiser son audience, l’opérateur est amené à vouloir déterminer quels mot-clés capteront la majorité de l’attention dans un avenir proche. Cette tâche est définie en fonction d’un réseau social, d’un ensemble de mot-clés et d’un horizon temporel. Pour être utile en pratique, une méthode de prédiction du volume d’ activité doit répondre aux contraintes suivantes: (a) être utilisable pour différents sites dédiés aux réseaux sociaux dans différentes langues, (b) passer à l’échelle pour anticiper la croissance des réseaux sociaux, (c) ne pas se reposer sur une connaissance *a priori* du graphe des utilisateurs du réseau social cible, car cette information est rarement disponible. Nous proposons une méthode qui respecte ces trois contraintes et permet l’extraction massive et automatique d’exemples annotés.

Cet article est structuré comme suit. Nous présenterons tout d’abord les travaux relatifs à la prédiction du volume d’activité. La section 3 présente ensuite un cadre unifié pour l’étude des réseaux sociaux en ligne ainsi que le problème de prédiction. Nous présenterons alors les descripteurs que nous avons défini dans la section 4. Les expériences et les résultats obtenus sont présentés dans la section 5. Enfin nous discuterons notre travail dans la section 6.

2. Travaux reliés

Les réseaux sociaux sont omniprésents, à tel point qu’ils sont utilisés pour réaliser des sondages sur diverses problématiques sociétales. Ainsi, des domaines variés ont fait l’objet d’études visant à établir des prédictions à partir de signaux captés sur les réseaux sociaux en ligne : les revenus des blockbusters hollywoodiens (Asur, Huberman, 2010), les épidémies de grippe (Ginsberg *et al.*, 2008 ; Culotta, 2010), les résultats d’élection (Tumasjan *et al.*, 2010). Les signaux captés sont explicites ou implicites. Un message publié au sujet d’un film est une trace explicite, au contraire l’utilisation d’un moteur de recherche au sujet des symptômes d’une maladie est une trace implicite. Des descripteurs génériques, comme la fréquence des observations

(Ginsberg *et al.*, 2008), sont employés en concomitance avec des descripteurs dépendants de l'application : l'analyse des sentiments par exemple. La prédiction en elle-même correspond alors à l'apprentissage d'une application ayant pour ensemble source les descripteurs et pour ensemble d'arrivée les observations tangibles dans le domaine cible. Ces travaux mettent en exergue la difficulté de traiter une grande quantité d'événements distincts tout en considérant plusieurs réseaux sociaux et différentes langues.

Les descripteurs proposés dépendent du temps et nous pouvons les utiliser pour définir une série temporelle multivariée unique. Les événements tangibles sont alors associés à un motif de cette série temporelle multivariée. Plusieurs études (Lehmann *et al.*, 2012 ; Matsubara *et al.*, 2012 ; Yang, Leskovec, 2011) utilisent cette approche. Par exemple (Yang, Leskovec, 2011) propose d'utiliser le clustering spectral à k centroïdes, avec k déterminé par une recherche linéaire maximisant la valeur moyenne de la silhouette ou l'index de Hartigan. Ces études concluent à l'existence d'un nombre réduit de classes de motifs correspondant à des périodes de fort accroissement de l'attention pour un mot-clé, et sont interprétées comme autant d'illustrations des mécanismes sous-jacents à celles-ci. Le clustering pourrait être utilisé pour tenter de répondre à des problèmes de prédiction, mais cet aspect n'est pas développé à notre connaissance.

Récemment de nombreux travaux (Naveed *et al.*, 2011 ; Tsur, Rappoport, 2012 ; Petrovic *et al.*, 2011 ; Guille, Hacid, 2012 ; Zaman *et al.*, 2010 ; Suh *et al.*, 2010) ont été dédiés à Twitter, comme la prédiction des "re-tweets" (*ie.* partager à ses abonnés une information préalablement reçue). Les solutions proposées impliquent la connaissance du graphe des utilisateurs. Cependant cette information est rarement disponible et s'avère difficile à maintenir. C'est fort de ce constat que nous proposons une solution ne nécessitant pas la connaissance du graphe des utilisateurs.

3. Définition du problème

Nous proposons d'étudier différents réseaux sociaux sans tenir compte du graphe des utilisateurs. Pour ce faire nous utilisons un formalisme simplifié. Appliqué aux réseaux sociaux considérés dans cette étude, celui-ci nous permet de définir des descripteurs d'une façon unique. Dans un premier temps nous présenterons ce formalise, après quoi nous définirons formellement le problème de prédiction d'amplitude des volumes d'activité.

3.1. Un formalisme unifié pour les réseaux sociaux

Notre formalisme permet de faire abstraction du graphe des utilisateurs en adoptant un point de vue centré sur les contenus échangés, mais ne permet pas de tenir compte du fait que certaines contributions sont destinées à un utilisateur spécifique au sein d'une discussion. Il est applicable à différents réseaux sociaux, comme Twitter, ainsi que les réseaux de forums ou de blog. Il est défini sur la base des concepts suivants :

DÉFINITION 1. — *(Conteneur unitaire $\langle z, a, \tau \rangle$) Chaque publication d'un auteur $a \in \mathcal{A}$ comportant un mot-clé $z \in \mathcal{Z}$ ayant lieu à un temps $\tau \in \mathcal{T}$ correspond à un conteneur unitaire $\langle z, a, \tau \rangle$.*

DÉFINITION 2. — *(Discussion d_t) Une discussion d_t est une séquence de contributions unitaire ordonnées temporellement:*

$$d_t = \left\{ \langle z^1, a^1, \tau^1 \rangle, \ldots, \langle z^{l_{d_t}}, a^{l_{d_t}}, \tau^{l_{d_t}} \rangle \right\}, \text{ avec } \tau^{l_{d_t}} \leq t$$

$\mathcal{D}$ correspond à l'ensemble de toute les discussions. Une discussion défini un ou plusieurs thèmes.

DÉFINITION 3. — *(Fonctions d'auteur, de thème et d'affichage) Nous définissons trois fonctions pour obtenir des informations au sujet d'une discussion $d_t \in \mathcal{D}$:*

1. $\texttt{authors} : \mathcal{D} \mapsto \mathcal{A}^n$ *qui permet d'obtenir l'ensemble des auteurs impliqués dans d_t:* $\texttt{authors}(d_t) = \{a \in \mathcal{A} \mid \langle z, a, \tau \rangle \in d_t\}$*;*

2. $\texttt{topics} : \mathcal{D} \mapsto \mathcal{Z}^m$ *qui fournit l'ensemble des mots-clés abordés dans d_t:* $\texttt{topics}(d_t) = \{z \in \mathcal{Z} \mid \langle z, a, \tau \rangle \in d_t\}$*;*

3. $\texttt{displays} : \mathcal{D} \mapsto \mathbb{N}$ *qui retourne le nombre de fois que d_t à été affichée (ou vue) par un utilisateur.*

Ces fonctions permettent de définir l'ensemble $\mathcal{D}_{t,z}$ des discussions comportant le mot-clé z au temps t, ainsi que l'ensemble des auteurs $\mathcal{A}_{t,z}$ ayant interagi dans une discussion comportant le mot-clé z au temps t.

$$\begin{aligned} \mathcal{D}_{t,z} &= \{d_t \in \mathcal{D} \mid z \in \texttt{topic}(d_t)\} \\ \mathcal{A}_{t,z} &= \{\texttt{authors}(d_t) \mid d_t \in \mathcal{D}_{t,z}\} \end{aligned}$$

DÉFINITION 4. — *(Fonction d'incrémentation) La fonction d'incrémentation $\mathcal{D} \times \langle \mathcal{Z}, \mathcal{A}, \mathcal{T} \rangle \mapsto \mathcal{D}$ est utilisée pour ajouter un conteneur unitaire à une discussion existante:*

$$\texttt{update}(d_t, \langle z, a, \tau \rangle) = d_t \cup \langle z, a, \tau \rangle$$

Nous pouvons maintenant illustrer ce formalisme avec les deux réseaux sociaux que nous ciblons dans ce travail : Twitter et les forums de Tom's hardware. Twitter est une plate-forme de micro blogging dans laquelle les utilisateurs peuvent échanger des contenus textuels, les "tweets", de taille limitée. La relation d'abonnement (*follows*) est essentielle : en effet, chaque utilisateur est notifié de tous les messages publiés par les utilisateurs auxquels il est abonné. Un utilisateur peut effectuer des "re-tweets", action par laquelle il transmet à ses abonnés un message d'un autre utilisateur. Il est également possible pour un utilisateur de répondre à un message existant, ce qui à pour effet de notifier l'auteur du premier message. Considérons le formalisme précédent : dans ce cas, un "tweet" correspond à un conteneur unitaire, et une discussion est un ensemble de "tweets" associés au moyen de "re-tweets" et de réponses. Ainsi la fonction d'incrémentation correspond indistinctement à ces deux actions.

Les forums, quant à eux, font partie des plus anciens moyens d'échange sur le web. Aussi, bien que faisant l'objet de peu d'attention, ces derniers restent très largement utilisés. Un forum est constitué de "threads" chacun étant une collection de messages émis par les utilisateurs. Lorsque l'un d'eux veut échanger au sujet d'un thème qui n'est abordé dans aucun des "threads" existants, il en démarre un nouveau. Il permet ainsi aux autres utilisateurs de contribuer à la discussion sur ce nouveau thème. Un "thread" correspond à une discussion, et le fait d'y contribuer correspond à la fonction d'incrémentation.

3.2. Le problème de prédiction

Supposons que pour un mot-clé z nous observons au temps t un vecteur à m dimensions $X(z,t)$ constitué par un ensemble de descripteurs (décrits dans la section 4). Pour un intervalle temporel $[t,t']$, ces observations sont groupées dans une série temporelle multivariée $\boldsymbol{X}(z,t \rightarrow t') = \{X(z,t), X(z,t+1), \cdots, X(z,t'-1), X(z,t')\}$ et le problème qui nous intéresse est celui de prédire la valeur d'une variable cible $Y(z,t' \rightarrow t'+\delta)$ qui représente le volume d'activité associée au mot-clé z pendant l'intervalle temporel $[t',t'+\delta]$. Dans certains cas, Y peut être observé sur $[t,t']$ et correspond alors à une dimension du vecteur X ; nous souhaitons alors prédire ses futures valeurs connaissant les anciennes ainsi que celles des autres $(m-1)$ variables.

Le problème décrit correspond à un problème de régression de séries temporelles multivariées, l'objectif étant de trouver une fonction f, dans une famille $\mathcal{F}$, qui associe la variable à prédire aux variables observées:

$$Y(z,t' \rightarrow t'+\delta]) \approx f(\boldsymbol{X}(z,t \rightarrow t'), \boldsymbol{\Omega})$$

avec $\boldsymbol{\Omega}$ l'ensemble des paramètres de f. Ce problème peut être résolu en apprenant une fonction de régression sur un ensemble d'apprentissage. Nous décrivons dans la section 4 comment produire cet ensemble d'apprentissage, ainsi que les descripteurs utilisés pour définir $\boldsymbol{X}$.

4. Approche

Le problème de prédiction du volume d'activité associé à un mot-clé dans les réseaux sociaux en ligne est récent et nous n'avons connaissance d'aucun travail établissant une référence pour les performances attendues sur cette tâche. De plus, les résultats publiés dans ce domaine portent sur une faible quantité d'exemples, souvent extraits et annotés manuellement.

Dans cette section nous présentons un processus qui décrit les échanges des utilisateurs d'un réseau social selon un ensemble de descripteurs génériques au regard de l'application escomptée, et permet d'extraire automatiquement depuis ceux-ci des périodes susceptibles de précéder une période d'accroissement du volume d'activité associé à un mot clé. Ainsi il permet de produire les ensembles d'entraînement et de test nécessaire à l'apprentissage de f. Nous décrivons tout d'abord les descripteurs

utilisés avant d'expliciter la méthode de construction des ensembles d'entraînement et de test.

4.1. Descripteurs génériques

L'évolution du volume d'activité associé à un mot clé est étroitement lié à l'intérêt que portent les utilisateurs aux thèmes discutés. Il est parfois possible de mesurer directement la popularité d'un mot clé, par exemple en comptant le nombre de fois que les utilisateurs ont visionné les discussions l'abordant. Nous appelons ND le descripteur associé à cette quantité. Lorsque cette dernière n'est pas publiquement disponible, ce qui peut être le cas pour garantir l'anonymat des utilisateurs, il est possible d'utiliser la quantité de discussions utilisant le mot-clé cible à la seule condition que les contributions sont publiques. Nous appelons NAD cette quantité. Ces deux descripteurs peuvent être utilisés pour définir le descripteur cible mentionné précédemment dans la section 3. Nous utiliserons NAD uniquement lorsque ND n'est pas observable.

Comme le suggère notre formalisme (*cf.* section 3.1), nous pouvons définir deux classes de descripteurs : ceux centrés discussion et ceux centrés utilisateurs.

Descripteurs centrés discussion. Ces descripteurs caractérisent l'activité associée à un mot-clé z en utilisant les discussions utilisant ce mot-clé.

1. *Nombre de discussions actives (*NAD*).* Ce descripteur mesure la quantité de discussions actives portant sur le mot-clé z jusqu'au temps t :

$$\text{NAD}(t,z) \quad = \quad |\{d_t \in \mathcal{D}_{t,z} \mid \exists \langle z, a, \tau \rangle \in d_t \wedge \tau = t\}|$$

2. *Nombre d'affichages (*ND*).* Ce descripteur correspond au nombre de fois qu'un utilisateur quelconque a affiché une discussion portant sur le mot-clé z;

3. *Nombre de créations de discussions (*NCD*).* Ce descripteur mesure le nombre de discussions créées au temps t et qui traitent de z : $\text{NCD} = |\mathcal{D}_{t,z} \setminus \mathcal{D}_{t-1,z}|$;

4. *Degré de nouveauté (*BL*).* Le degré de nouveauté d'un mot-clé z à un temps t correspond au rapport entre NCD et NAD : $\text{BL}(t,z) = \frac{NCD(t,z)}{NAD(t,z)}$;

5. *Nombre de conteneurs unitaire (*NAC*).* Ce descripteur mesure le nombre total de conteneurs unitaires existant sur le réseau social et appartenant a des discussions utilisant sur le mot-clé z créées avant t : $\text{NAC}(t,z) = \sum_{d_t \in \mathcal{D}_{t,z}} l_{d_t}$;

6. *Degré d'attention (*AS*).* Soit $\rho(t,z) = \text{NAC}(t,z)$ ou $\rho(t,z) = \text{NAD}(t,z)$ ou $\rho(t,z) = \text{ND}(t,z)$ (ρ est toujours une mesure de l'attention capté par le mot-clé z sur un réseau social). Le degré d'attention est défini pour le mot-clé z au temps t par : $\text{AS}(t,z) = \rho(t,z) \;/\; \sum_{z' \in \mathcal{Z}} \rho(t,z')$;

7. *Longueur moyenne d'une discussion (*ADL*).* Ce descripteur mesure directement la longueur moyenne des discussion associées au mot-clé z : $\text{ADL}(t,z) =$

$\sum_{d_t \in \text{AD}(t,z)} l_{d_t}$ / $\text{NAD}(t,z)$ (voir (Adamic *et al.*, 2008) pour plus de détails au sujet de cette quantité).

Descripteurs centrés utilisateurs. Nous pouvons également qualifier l'activité associée à un mot-clé en considérant les utilisateurs qui interagissent à son sujet :

1. *Nombre d'utilisateurs (*NA*).* Ce descripteur, défini par $\text{NA}(t,z) = |\mathcal{A}_{t,z}|$, mesure le nombre d'auteurs participant aux discussions associées au mot-clé z au temps t. C'est une mesure directe de l'influence d'un mot-clé ;

2. *Nombre de nouveaux utilisateurs (*AI*).* Afin de mesurer le nombre d'utilisateurs commençant à participer aux discussions associés au mot-clé z au temps t, nous utilisons : $\text{AI}(t,z) = |\mathcal{A}_{t,z} \setminus \mathcal{A}_{t-1,z}|$;

3. *Nombre moyen de participants (*AT*).* Ce descripteur est défini par : $\text{AT}(t,z) = \frac{\sum_{d_t \in \mathcal{D}_{t,z}} \texttt{authors}(d_t)}{\text{NAD}(t,z)}$. Il mesure le nombre moyen d'utilisateurs impliqués dans une discussion associée au mot-clé z.

Enfin, nous utilisons, en complément de ces descripteurs, leurs différences d'ordre 1, par exemple $\text{NAD}(t+1,z) - \text{NAD}(t,z)$, en vue de capturer la dynamique des phénomènes observés.

4.2. Processus d'extraction et d'annotation automatique de périodes de fort intérêt

Les périodes de fort intérêt associées à un mot-clé, exception faite des réactions à des surprises (voir (Lehmann *et al.*, 2012)), sont isolées et précédées par une phase d'accroissement de l'activité. Il est ainsi raisonnable de ne pas considérer les mot-clés dont l'activité est décroissante ou stable. Au contraire, si l'activité associée à un mot-clé croît durant la période $[t, t']$, nous devons le considérer comme un candidat potentiel pour présenter une période de fort intérêt. Pour formaliser les notions d'activité stable ou tendance croissante nous utilisons une méthode de segmentation non supervisée. La période $[t, t']$ fait ainsi l'objet d'une segmentation en deux sous-séquences. La différence de la moyenne des deux sous-séquences est utilisée pour décider si l'activité sur la période $[t, t']$ est croissante. Le processus de segmentation non supervisée de séries temporelles *CuSum*, décrit dans (Bissell, 1969), s'est révélé efficace dans différents domaines (Sibanda, Sibanda, 2007 ; Yi *et al.*, 2006) et c'est sur celui-ci que nous nous appuyons pour segmenter les séquences et juger de la croissance ou non d'un thème. Le processus d'extraction de candidats correspond aux étapes suivantes :

1. Pour chaque mot-clé z, à chaque pas t, nous considérons une période de taille fixe $\beta = t' - t$. Nous observons la popularité de z pendant cette période $Y(z, t \to t')$ (rappelons que Y est une série temporelle uni-variée correspondant au descripteur ND ou à défaut NAD) ;

2. $Y(z, t \to t')$ est alors divisé en deux sous-séquences à l'aide de *CuSum*. Si la segmentation n'aboutit pas, le niveau d'activité est considéré comme stable et se poursuit à l'étape 1 pour le temps $t+1$. Sinon, le point de segmentation identifié β_0

(avec $\beta_0 < \beta$) est utilisé pour définir les deux sous-séquences dont la moyenne est donnée par la fonction : $\mu(Y(z, \tau \rightarrow \tau'))$, valeur moyenne de Y pour le mot-clé z sur l'intervalle de temps $[\tau, \tau']$):

- Si $\mu(Y(z, t + \beta_0 \rightarrow t + \beta)) < \varphi \times \mu(Y(z, t \rightarrow t + \beta_0))$ alors la période est considérée comme décroissante ou stable pour le mot-clé z, et celle-ci ne sera pas utilisée ; le processus reprend à l'étape 1 pour le temps $t + 1$. φ permet ici de qualifier l'accroissement minium attendu sur la période $[t, t']$; fixer $\varphi = 1$ correspond à considérer les périodes stables comme des périodes de fort intérêt potentiels. La valeur utilisée dans notre étude est fixée à $\varphi = 1.25$ (*cf.* section 5) ;

- Sinon, $\boldsymbol{X}(z, t \rightarrow t+\beta)$ est extrait comme une période à étudier. Le processus reprend alors à l'étape 1 au temps $t' = t + \beta$. De cette façon, nous assurons que les exemples extraits ne se chevauchent pas. Cette condition est nécessaire pour que les exemples soient indépendamment et identiquement distribués (i.i.d.). Il s'agit d'un postulat de nombreuses méthodes de classification et régression, dont celles que nous utilisons dans ce travail.

Le processus ci-dessus permet de considérer uniquement les périodes les plus susceptibles d'aboutir à une période de fort intérêt, qui sont sont annotées par la moyenne des valeurs de Y sur la période $[t', t' + \delta]$ (avec $t' = t + \beta$). L'ensemble des exemples annotés $\mathcal{A}$ est donc défini comme suit:

$$\mathcal{A} = \{(\boldsymbol{X}(z, t \rightarrow t'), \mu(Y(z, t' \rightarrow t' + \delta))), \boldsymbol{X}(z, t \rightarrow t') \text{ période à étudier}\}$$

5. Expérimentation

En vue d'évaluer l'efficacité de l'approche proposée, nous conduisons des expériences sur deux réseaux sociaux distincts, ayant des spécificités marquées : Twitter (Kwak *et al.*, 2010). et Tom's Hardware, un forum dédié aux nouvelles technologies (BestofMedia, 2013).

5.1. Spécificités des réseaux sociaux

Tom's Hardware (TH) et Twitter (TW) se distinguent nettement par les aspects suivants :

1. Les contributions collectées sur TW sont rédigées en anglais, en français et en allemand. En revanche les contributions en Anglais ne sont pas collectées sur TH ;

2. Pour TW nous ne pouvons pas estimer la popularité directement, en conséquence nous utilisons le descripteur NAD comme descripteur cible. Au contraire, la popularité est directement estimée pour TH ou le nombre d'affichages est disponible (descripteur ND) ;

3. TW présente un degré de réactivité plus élevé que TH (Kwak *et al.*, 2010 ; Lin, Mishne, 2012) : 80% des re-tweets ont lieu dans la journée suivant le tweet initial. Au contraire, pour TH, les réponses à un thread ont généralement lieu dans la semaine suivant sa création ;

4. La communauté de TW est plus grande que celle de TH car elle compte plus de 500 millions de visiteurs par mois, contre 41 millions pour TH.

Pour ce travail nous avons étudié un total de 6671 thèmes en rapport aux nouvelles technologies sur les deux réseaux sociaux présentés précédemment. Ces données sont mises à disposition dans la collection UCI ML data-sets: *http://archive.ics.uci.edu/ml/*, où des informations détaillées sont également fournies (en particulier sur le découpage entraînement/test).

5.2. Choix des paramètres et validation de l'approche

Le processus d'extraction automatique des exemples est configuré pour tenir compte des spécificités de l'objectif de la prédiction, et de la source des observations. Les paramètres correspondants sont : la durée en jours d'un pas de temps, la taille des périodes considérées lors de la segmentation (paramètre β), l'accroissement pour les périodes candidates (paramètre φ), et enfin l'horizon temporel auquel la prédiction est réalisée (paramètre δ). Pour tenir compte de la différence entre les dynamiques propres à chaque source, le pas de temps a été ajusté à une journée pour TW et à une semaine pour TH. Le paramètre β est fixé à 7 pour les deux sources ; le processus d'extraction traitera donc des périodes d'une semaine dans le cas de TW et deux mois pour TH. φ est ici fixé à 1.25. Enfin, les prédictions sont réalisées à deux pas de temps ($\delta = 2$), ce qui correspond à deux jours pour TW et deux semaines pour TH. Le processus décrit dans la section 4 aboutit à l'extraction de 140 707 périodes candidates pour TW et 7 905 pour TH.

5.3. Prédiction de l'amplitude des périodes de fort intérêt

Nous utilisons les Regression Random Forests avec validation croisée répétée cinq fois sur l'ensemble des exemples pour construire notre fonction de prédiction. Dans une telle situation il est courant d'évaluer les performances de prédiction en utilisant le coefficient de détermination R^2. Dans le cas de TH, celui-ci vaut 0.972, et 0.942 pour TW. Ces deux valeurs ne suffisent toutefois pas à déterminer les capacités prédictives de cette approche car le coefficient R^2 est, par définition, sensible à la présence de données aberrantes (*ie.* outliers). Nous proposons donc de calculer les erreurs relatives de prédiction ($|\tilde{Y} - Y|/Y$) et examinons leurs distributions pour les exemples correspondant à une période de fort intérêt. Ces résultats, présentés en figure 1, montrent qu'environ 12,000 exemples (20% du total) appartiennent à la première classe, l'écart entre la valeur prédite et la valeur observée pour ces exemples étant d'au plus 4%. De plus, pour 80% des exemples, l'écart entre prédiction et observation correspond à moins de 25% de la valeur observée. Ceci montre que le prédicteur est capable de fournir des estimations précises et valide l'approche suivie.

La figure 2 montre enfin l'importance de chacun des descripteurs. Dans le cas de TW les descripteurs suivants : NAC, NCD, ainsi que NAD sont importants pour prédire

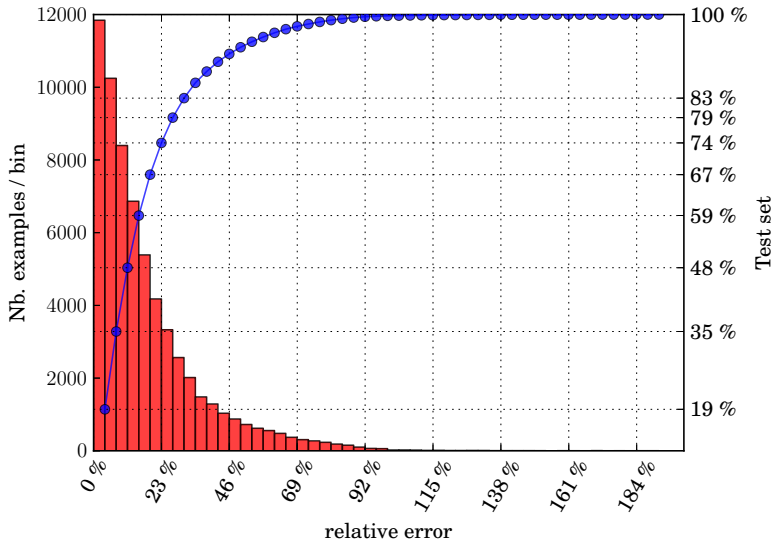

FIGURE *1. Distribution des erreurs relatives pour* TW *(une classe ≈ 4% d'erreur relative)*

qu'un mot-clé va connaître une période de fort intérêt. Dans le cas de TH, la prédiction peut être réalisée à partir du seul descripteur cible ND.

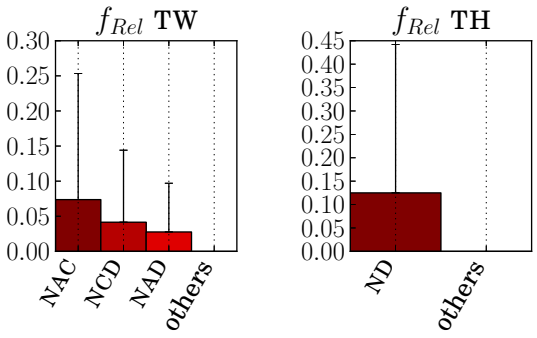

FIGURE *2. Importance des descripteurs (Gini impurity) pour la prédiction de l'amplitude des périodes de fort intérêt (*TW *à gauche,* TH *à droite) ; les barres d'erreur correspondent à l'écart-type*

6. Conclusion

Nous avons proposé dans ce travail une formalisation du problème de prédiction du volume d'activité associé à un mot-clé ainsi qu'une méthode pour le résoudre. Notre approche complète les travaux précédents sur des problématiques proches (comme par exemple (Kleinberg, 2003) ou (Petrovic *et al.*, 2011)) sur plusieurs aspects :

– nous avons défini un formalisme original, qui permet de décrire des réseaux sociaux d'une manière unifiée, sans avoir besoin de définir explicitement le graphe des utilisateurs, une information rarement disponible ;

– à partir de ce formalisme, nous avons défini un ensemble de descripteurs génériques qui peuvent être employés pour décrire l'activité sur les réseaux sociaux et qui ne dépendent pas de considérations *ad hoc* liées à l'application visée ;

– nous avons proposé une méthode d'extraction et d'annotation automatique de périodes de fort intérêt. Cette méthode, en plus d'être non supervisée, est capable de traiter les grandes quantités de données produites par les réseaux sociaux actuels. C'est, à notre connaissance, la première fois qu'une telle méthode est proposée. Le jeu de données que nous avons conçu à partir de Twitter contient environ 140 000 exemples, pour 588 millions de contributions produites par 48 millions d'utilisateurs. Le jeu de données issu de Tom's hardware contient, quant à lui, près de 8 000 exemples pour 474 000 contributions produites par 105 000 utilisateurs. Nous avons mis à disposition ces données dans la collection UCI ML: *http://archive.ics.uci.edu/ml/* ;

– enfin, nous avons montré que le problème de prédiction de l'amplitude du volume d'activité peut être résolu, d'une manière efficace, en utilisant les *Random Forest*. A notre connaissance, il s'agit de la première démonstration de ce type à cette échelle.

Nous prévoyons dans un futur proche d'améliorer les capacités prédictives de notre solution en combinant des informations en provenance de différents réseaux sociaux. Nous voulons également conduire des expériences avec des méthodes d'apprentissage définis pour traiter les données temporelles comme les SVMs associés à des noyaux de type Dynamic Time Warping kernels (Gudmundsson *et al.*, 2008).

Bibliographie

Adamic L. A., Zhang J., Bakshy E., Ackerman M. S. (2008). Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on world wide web*, p. 665–674.

Asur S., Huberman B. A. (2010). Predicting the future with social media. *CoRR*, vol. abs/1003.5699.

BestofMedia. (2013, avril). *Tom's Hardware forums.* Web. http://www.tomshardware.com/forum/

Bissell A. F. (1969). Cusum techniques for quality control. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 18, n° 1, p. 1–30. http://dx.doi.org/10.2307/2346436

Culotta A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages. In *Kdd workshop on social media analytics.*

Ginsberg J., Mohebbi M., Patel R., Brammer L., Smolinski M., Brilliant L. (2008). Detecting influenza epidemics using search engine query data. *Nature*, vol. 457, n° 7232, p. 1012–1014.

Gudmundsson S., Runarsson T. P., Sigurdsson S. (2008, juin). Support vector machines and dynamic time warping for time series. In *Ijcnn'08: Ieee international joint conference on neural networks*, p. 2772–2776. IEEE. http://dx.doi.org/10.1109/ijcnn.2008.4634188

Guille A., Hacid H. (2012). A predictive model for the temporal dynamics of information diffusion in online social networks. In *Proceedings of the 21st international conference companion on world wide web*, p. 1145–1152.

Kleinberg J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, vol. 7, n° 4, p. 373–397.

Kwak H., Lee C., Park H., Moon S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on world wide web*, p. 591–600.

Lehmann J., Gonçalves B., Ramasco J., Cattuto C. (2012). Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on world wide web*, p. 251–260.

Lin J., Mishne G. (2012). A study of "churn" in tweets and real-time search queries. In J. G. Breslin, N. B. Ellison, J. G. Shanahan, Z. Tufekci (Eds.), *Icwsm.* The AAAI Press.

Matsubara Y., Sakurai Y., Prakash B., Li L., Faloutsos C. (2012). Rise and fall patterns of information diffusion: Model and implications. In *Proceedings of the 18th acm sigkdd international conference on knowledge discovery and data mining*, p. 6–14.

Naveed N., Gottron T., Kunegis J., Che Alhadi A. (2011). Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proc. web science conf.*

Petrovic S., Osborne M., Lavrenko V. (2011). Rt to win! predicting message propagation in twitter. *5th ICWSM*.

Sibanda T., Sibanda N. (2007, 03 novembre). The CUSUM chart method as a tool for continuous monitoring of clinical outcomes using routinely collected data. *BMC Medical Research Methodology*, vol. 7, p. 46+. http://dx.doi.org/10.1186/1471-2288-7-46

Suh B., Hong L., Pirolli P., Chi E. H. (2010). Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Social computing (socialcom), 2010 ieee second international conference on*, p. 177–184.

Tsur O., Rappoport A. (2012). What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In *Proceedings of the fifth acm international conference on web search and data mining*, p. 643–652.

Tumasjan A., Sprenger T. O., Sandner P. G., Welpe I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the fourth international aaai conference on weblogs and social media*, p. 178–185.

Yang J., Leskovec J. (2011). Patterns of temporal variation in online media. In *Proceedings of the fourth acm international conference on web search and data mining*, p. 177–186.

Yi G., Coleman S., Ren Q. (2006, août). CUSUM method in predicting regime shifts and its performance in different stock markets allowing for transaction fees. *Journal of Applied Statistics*, vol. 33, n° 7, p. 647–661. http://dx.doi.org/10.1080/02664760600708590

Zaman T. R., Herbrich R., Van Gael J., Stern D. (2010). Predicting information spreading in twitter. In *Workshop on computational social science and the wisdom of crowds, nips.*