Une approche multidimensionnelle basée sur les comportements individuels pour la prédiction de la diffusion de l'information sur Twitter

Adrien Guille*, Hakim Hacid**, Cécile Favre*

*ERIC, Université Lumière Lyon 2, Lyon, France
{prénom.nom}@univ-lyon2.fr
**Bell Labs France, Alcatel-Lucent, Nozay, France
hakim.hacid@alcatel-lucent.com

Résumé. Aujourd'hui, les réseaux sociaux en ligne sont devenus des outils très puissants de propagation de l'information. Ils favorisent la diffusion rapide à grande échelle de contenu et les conséquences d'une information inexacte voire fausse peuvent alors prendre une ampleur considérable. Par conséquent il devient indispensable de proposer des moyens d'analyser le phénomène de diffusion de l'information dans ces réseaux. De nombreuses études récentes ont traité de la modélisation du processus de diffusion de l'information, essentiellement d'un point de vue topologique et dans une perspective théorique, mais les facteurs impliqués sont encore méconnus. Nous proposons ici une solution pratique dont l'objectif est de prédire la dynamique temporelle de la diffusion au sein de Twitter, basée sur des techniques d'apprentissage automatique. Notre approche repose sur l'inférence de probabilités de diffusion tirées d'une analyse multidimensionnelle des comportements individuels. Les expérimentations menées montrent l'intérêt de la modélisation proposée.

1 Introduction

Les réseaux sociaux en ligne sont des plateformes où l'information est publiée et mise à jour rapidement par des commentaires, des réponses, des transferts, etc. Par conséquent, l'information se déplace, d'un nœud à un autre du réseau, d'une communauté à une autre, etc. C'est le phénomène bien connu de diffusion, ou propagation de l'information, qui a suscité et suscite encore un grand intérêt au sein de la communauté de recherche (Bakshy et al., 2011; Kwak et al., 2010; Saito et al., 2011; Yang et Leskovec, 2010).

Par le passé, la propagation a surtout été étudiée dans le domaine de l'épidémiologie, afin de comprendre le déroulement du processus de diffusion d'un virus selon certaines conditions. Ainsi, la majorité des travaux actuels portant sur la propagation de l'information s'appuient largement sur ces travaux. Mais avec l'émergence des réseaux sociaux en ligne, les mécanismes de diffusion de l'information se sont complexifiés, en raison des spécificités suivantes : (i) la grande ampleur des réseaux, (ii) l'importante diversité des profils des utilisateurs et (iii) la particularité des lois régissant ces structures. Par conséquent, les modèles traditionnels apparaissent comme inefficaces voire dénués de sens. En réponse, nous proposons ici une solution

applicable de modélisation dont l'objectif est de prédire la dynamique temporelle de la diffusion, basée sur des techniques d'apprentissage automatique. Notre approche repose sur l'inférence de probabilités de diffusion tirées d'une analyse multidimensionnelle des comportements individuels et a été expérimentée sur Twitter.

L'article est organisé comme suit. La section 2 présente des travaux en rapport avec les nôtres, selon deux catégories : les modèles pour la diffusion dans les réseaux complexes et les études sur la diffusion de l'information. La section 3 décrit les données utilisées et l'approche que nous proposons. Enfin nous concluons dans la section 4.

2 Travaux connexes

Modèles pour la diffusion dans les réseaux complexes Pour simuler le processus de diffusion dans des réseaux complexes, les modèles de Cascades Indépendantes –IC, i.e Independent Cascades– et à Seuil Linéaire –LT, i.e Linear Threshold– (Newman, 2003) sont généralement utilisés. Ces modèles se basent sur des graphes orientés et requièrent plus ou moins de paramètres. Dans le cas d'IC, les probabilités de diffusion doivent être déterminées pour chaque arc tandis que LT requiert la définition d'un degré d'influence pour chaque arc et un seuil d'influence pour chaque nœud. Avec IC, un utilisateur transmet une information à un voisin avec la probabilité définie sur cet arc alors qu'avec LT, un utilisateur reçoit une information si la somme des influences de ses voisins informés est supérieure à son seuil. L'inconvénient de ces modèles est qu'ils ont la particularité de se concentrer sur les aspect topologiques et de survoler l'aspect temporel. Plus précisément, ils utilisent un axe temporel discret et entrainent des changements d'état synchrones dans le réseau. Pour cette raison, Saito et al. (2010) ont proposé des extensions asynchrones des modèles IC et LT, respectivement AsIC et AsLT, qui requièrent la définition d'un paramètre temporel sur chaque arc du graphe.

Études sur la diffusion de l'information La diffusion de l'information dans des contextes en-ligne variés a généré beaucoup de travaux de recherche. De par son objectif, prédire la dynamique temporelle de la diffusion de l'information, le travail de Yang et Leskovec (2010) est certainement le plus proche de notre proposition. Ils ont étudié la diffusion de hashtags sur Twitter et ont proposé une modélisation où chaque nœud est associé à une fonction d'influence globale relative au temps. Toutefois, il y a une différence substantielle avec notre approche, puisque Yang et Leskovec considèrent une structure implicite entre les acteurs, tandis que dans notre cas, nous voulons tirer parti d'une structure explicite. Plusieurs études se sont intéressées à la prédiction de la diffusion de l'information à travers des réseaux explicites. Citons par exemple la proposition de Saito et al. (2011), qui vise à prédire le graphe de diffusion et est basée sur une fonction paramétrique évaluant la similarité entre utilisateurs, qui sont décrits par un vecteur. Toutefois, les auteurs ne fournissent pas de définition des vecteurs descriptifs. Citons également les travaux de Bakshy et al. (2011) qui portent sur la diffusion d'une URL sur Twitter. À partir d'attributs basiques et de l'influence globale passée des utilisateurs, ils proposent un modèle capable de prédire la taille de la cascade engendrée.

Discussion On remarque que les diverses approches se basent sur des facteurs variés et il ne semble pas encore qu'un consensus émerge à propos des dimensions minimales requises à la

capture du processus de diffusion dans les réseaux sociaux. Notre objectif est de proposer un modèle plus réaliste en adoptant une approche multidimensionnelle. Néanmoins, ajouter des dimensions à un modèle n'est pas trivial, puisque cela peut favoriser le *sur-apprentissage*. Pour pallier à ce problème, on pourra notamment régulariser les modèles d'apprentissage employés.

3 Modèle de diffusion et son apprentissage

Notre approche modélise la diffusion sous forme de cascades et adopte un point de vue centré sur les émetteurs. Nous utilisons le graphe social de Twitter comme base et considérons trois dimensions issues d'une analyse préalable : (i) sémantique, (ii) sociale, et (iii) temporelle. Afin d'exploiter au mieux la troisième dimension, nous adaptons le méta-modèle AsIC, une extension du très courant IC avec un axe temporel rendu continu par l'adjonction d'un paramètre correspondant au délai de transmission sur chaque arc. Les probabilités de diffusion sont inférées pour chaque arc à partir de propriétés locales représentant les trois dimensions et d'un modèle généré par un apprentissage automatique. Nous utilisons un jeu de données collecté par Yang et Leskovec (2011), constitué de 476 millions de tweets ainsi que le graphe d'abonnement de Twitter correspondant (Kwak et al. (2010), 1,47 milliards d'arcs).

3.1 Description du modèle

Twitter fournit un réseau explicite et orienté, basé sur un système d'abonnement. Il est le graphe support à la diffusion. Comme ce réseau a une signification sociale relativement faible (Yang et Counts, 2010), nous créons également un graphe à partir des discussions des utilisateurs que l'on exploite pour l'inférence des probabilités. Le calcul d'une probabilité de diffusion s'appuie sur trois dimensions : sociale, sémantique et temporelle. On note $p_{u_1,u_2}(i,t)$ la probabilité qu'un utilisateur u_1 transmette une information i à un instant t à l'utilisateur u_2. Les attributs dérivés de ces dimensions sont soit des valeurs réelles continues variant entre 0 et 1, soit des booléens, dont nous donnons une description textuelle ci-après.

Dimension sociale : Cette dimension vise à capturer les différentes caractéristiques du réseau social (i.e. nœuds et arcs) et leurs interactions. Cinq propriétés sont retenues :

- *Activité (I) :* un indice d'activité qui exprime l'activité relative de l'utilisateur. Cet attribut est défini comme le nombre moyen de tweets émis par heure borné par 1.
- *Homogénéité Sociale (H) :* un indice portant sur l'arc u_1, u_2 qui reflète la similarité des ensembles d'utilisateurs avec lesquels u_1 et u_2 communiquent. Il est calculé par un indice de Jaccard.
- Le ratio de tweets orientés pour chaque utilisateur (dTR) qui reflète la manière dont il diffuse du contenu (i.e., vers un utilisateur spécifique ou globalement vers une communauté).
- Un attribut booléen pour chaque utilisateur lié à la pratique du *mentioning* dans le but de capturer l'existence éventuelle d'un lien social explicite entre ces utilisateurs.
- Le taux de mentioning (mR) de chaque utilisateur représente leur popularité. Plus la valeur est élevée, plus il reçoit de tweets directement adressés.

Dimension sémantique / thématique : Au-delà de la structure du réseau, nous considérons le contenu échangé pour mieux comprendre et capturer les raisons de la diffusion. Cette dimen-

sion capture la relation entre les utilisateurs et les contenus. Elle est actuellement représentée par un booléen pour chaque couple utilisateur-sujet. C'est-à-dire que pour une information donnée, il indique si l'utilisateur a évoqué le sujet antérieurement.

Dimension temporelle : Cette dimension vise à capturer les dynamiques individuelles. Une journée est partitionnée en six blocs de quatre heures et le pourcentage de tweets émis durant chaque période est calculé et stocké dans un vecteur 6-dimensionnel noté V, tel que $\sum_{i=0}^{5} V^i = 1$. Cela nous permet de connaître l'intensité de l'activité d'un utilisateur à un instant t.

3.2 Inférence des probabilités de diffusion

Une fois l'espace de représentation défini, nous cherchons à établir un modèle capable d'en inférer les probabilités de diffusion. Pour ce faire, nous avons adopté une démarche fondée sur des méthodes d'apprentissage automatique. Nous avons calculé les attributs par rapport à un mois d'activité et extrait des exemples de diffusion au cours du mois suivant.

C4.5	**Perceptron Linéaire**	**Perceptron Multi-couche**	**Régression Bayésienne**
91%	85%	86%	85%

TAB. 1 – *Performances des classifieurs en cross-validation (5 folds)*

Trois algorithmes (pouvant fonctionner de manière probabiliste) ont été évalués pour une tâche de classification supervisée avec un attribut de classe binaire {diffusion, non-diffusion} : un *arbre de décision C4.5*, *Perceptrons linéaire et multi-couches (14 couches cachées)*, et une *régression logistique Bayésienne*. Les résultats d'une cross-validation sont donnés par le tableau 1.

À première vue, nous pouvons constater que tous les classifieurs ont des taux d'erreur inférieurs à 15%. Nous constatons également que le Perceptron linéaire a des performances équivalentes au Perceptron avec 14 couches cachées. Ceci suggère que la probabilité de diffusion peut être vue comme une combinaison linéaire des variables. L'arbre de décision obtient le taux d'erreur le plus bas, mais son modèle est plus spécifique du fait de l'algorithme de partitionnement sur lequel il repose. Nous nous concentrons par la suite sur le perceptron li-

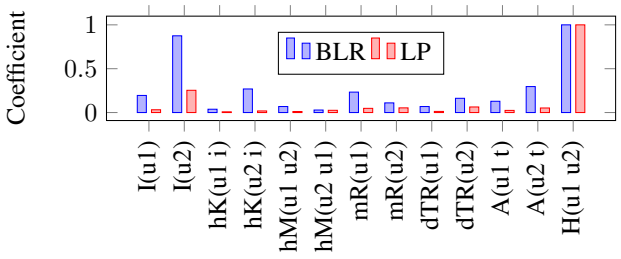

FIG. 1 – *Comparaison des valeurs normalisées des coefficients du Perceptron linéaire et de la régression logistique Bayésienne.*

néaire (LP) et la régression logistique Bayésienne (BLR) et comparons leurs coefficients dans la figure 1. Il s'avère qu'il y a une tendance commune vis-à-vis de l'importance accordée aux divers attributs par le Perceptron et la régression logistique. En particulier, les deux classifieurs associent le coefficient le plus important à l'homogénéité sociale. La régression logistique présente un aspect plus équilibré que le perceptron, qui accorde une importance relativement plus grande à homogénéité sociale. L'aspect plus nivelé de la régression logistique Bayésienne est dû au fait que nous régularisons les coefficients grâce à l'utilisation d'une "fonction pondérée de log-vraisemblance" (Mitchell, 1997) qui limite le sur-apprentissage. Pour cette raison, nous avons retenu ce modèle pour inférer les probabilités de diffusion. La régression logistique bayésienne induit la forme paramétrique suivante pour la distribution $P(Y|V)$ (avec Y ={diffusion, non-diffusion} et V le vecteur J-dimensionnel d'attributs) :

$$P(Y = \text{diffusion}|V) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^{J} w_i V_i)} \tag{1}$$

$$P(Y = \text{non-diffusion}|V) = \frac{\exp(w_0 + \sum_{i=1}^{n} w_i V_i)}{1 + \exp(w_0 + \sum_{i=1}^{J} w_i V_i)} \tag{2}$$

Le modèle est implémenté sous la forme d'un moteur de prédiction qui produit des séries temporelles représentant le niveau d'activité d'une communauté (i.e. le volume de tweets) pour un sujet d'information donné. Le moteur requiert quatre paramètres : (i) une communauté, (ii) un sujet, (iii) un ensemble $n << |U|$ d'utilisateurs initialement informés, et (iv) une formalisation du paramètre de délai de transmission r_{u_1,u_2}. La communauté représente le groupe de personnes que l'on veut étudier, décrit par ses deux graphes (à savoir le graphe d'abonnement et le graphe des discussions entre utilisateurs) et les attributs de chaque utilisateur. Le sujet de l'information se traduit par un ensemble de mots-clés devant être contenu dans les messages. La figure 2 montre un résultat obtenu avec le moteur sur deux communautés distinctes de 25000 et 40000 utilisateurs, avec $r_{u_1,u_2} = (1 - I(u_2)) \times 10$.

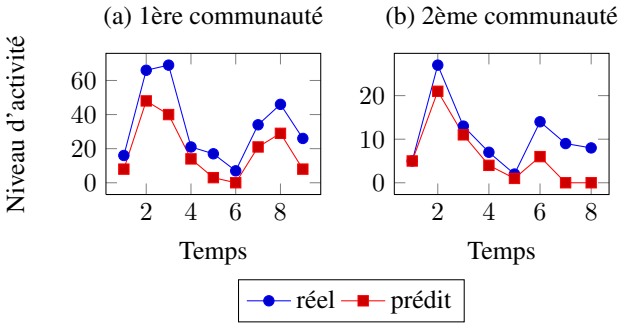

FIG. 2 – *Comparaison de l'activité réelle et prédite pour une information sur l'iPhone.*

4 Conclusion

Avec pour objectif la prédiction de la dynamique temporelle, nous avons proposé dans cet article un modèle de diffusion concret basé sur Twitter. Un ensemble d'attributs, résultant d'une observation en profondeur de données réelles et relevant de trois dimensions –sociale, sémantique et temporelle– est incorporé au modèle. Celui-ci est une adaptation du méta-modèle théorique *AsIC* et se base sur des techniques d'apprentissage automatique, i.e. une régression logistique Bayésienne, pour inférer les probabilités de diffusion entre les nœuds du réseau. Une série d'expérimentations a été réalisée et a permis d'évaluer les performances du modèle et d'identifier les valeurs optimales de ses paramètres. Les résultats montrent que notre objectif initial est atteint (prédire la dynamique de diffusion) et soutiennent notre hypothèse initiale, à savoir que la diffusion d'une information dépend de propriétés locales au graphe support.

Références

Bakshy, E., J. M. Hofman, W. A. Mason, et D. J. Watts (2011). Everyone's an influencer : quantifying influence on twitter. WSDM'11, pp. 65–74.

Kwak, H., C. Lee, H. Park, et S. Moon (2010). What is Twitter, a social network or a news media ? WWW'10, pp. 591–600.

Mitchell, T. M. (1997). *Machine learning*. McGraw Hill series in computer science.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review 45*, 167–256.

Saito, K., M. Kimura, K. Ohara, et H. Motoda (2010). Selecting information diffusion models over social networks for behavioral analysis. PKDD'10, pp. 180–195.

Saito, K., K. Ohara, Y. Yamagishi, M. Kimura, et H. Motoda (2011). Learning diffusion probability based on node attributes in social networks. ISMIS'11, pp. 153–162.

Yang, J. et S. Counts (2010). Predicting the speed, scale, and range of information diffusion in twitter. ICWSM'10, pp. 355–358.

Yang, J. et J. Leskovec (2010). Modeling information diffusion in implicit networks. ICDM'10, pp. 599–608.

Yang, J. et J. Leskovec (2011). Patterns of temporal variation in online media. WSDM'11, pp. 177–186.

Summary

Online social networks facilitate the rapid and large-scale propagation of content and the consequences of an inaccurate or false information can then take considerable proportions. Therefore it is essential to provide means to analyze the phenomenon of information dissemination in such networks. In this paper we propose a practical solution which aims to predict the temporal dynamics of diffusion in Twitter, based on machine learning techniques. Experimental results on real data show the interest of the proposed approach.