WorldSpread: un modèle de propagation de l'information entre populations

Didier Henry*, Erick Stattner*, Martine Collard*

* LAMIA, Université des Antilles, Guadeloupe, France
didier.henry ; erick.stattner ; martine.collard @univ-antilles.fr

Résumé. Les modèles de diffusion proposés dans les médias sociaux reposent pour la plupart sur des hypothèses épidémiologiques et non sur l'observation de données réelles pour décrire les caractéristiques de la diffusion. De tels modèles ne peuvent pas reproduire fidèlement le phénomène de diffusion dans la mesure où ils ne considèrent pas les facteurs observés qui peuvent influencer ce phénomène. Notre approche innove dans le sens où elle se place au niveau des populations des pays et qu'elle consiste à identifier en plus du nombre de populations atteintes, le rayon géographique d'influence autour de ces populations, l'instant de diffusion de l'information, la durée de la diffusion et le pays auquel appartiennent ces populations en connaissant la population à l'origine de l'information et sa thématique.

1 Introduction

Anticiper la réaction des utilisateurs vis-à-vis d'une information dans les médias sociaux peut permettre de prévoir l'ampleur et l'évolution du phénomène de diffusion. Ainsi, grâce à ces prévisions, il serait possible de cibler de façon pertinente des utilisateurs pour améliorer l'image d'une marque d'un point de vue marketing, ou pour atténuer la propagation de rumeurs ou d'une "infox [1]". Cependant, la majorité des travaux de modélisation de la diffusion dans les médias sociaux reposent en majorité sur des hypothèses épidémiologiques [Hethcote (1989); Anderson et May (1992)] et non sur l'observation de données réelles pour décrire les caractéristiques de la diffusion. Dans des approches relativement récentes, des chercheurs [Li et al. (2018); Zhou et al. (2017); Hoang et al. (2016)] ont introduit des modélisations du phénomène de diffusion au niveau microscopique, c'est-à-dire au niveau des utilisateurs, prenant en compte des variables de diffusion observées sur des cas réels de diffusion. Notre approche innove dans le sens où elle se place au niveau des populations des pays et qu'elle consiste à identifier en plus du nombre (N) de populations atteintes, le rayon géographique (R) d'influence autour de ces populations, l'instant de diffusion de l'information (T), la durée de la diffusion (I) et le pays auquel appartiennent ces populations. Le modèle **WorldSpread** que nous proposons permet de décrire le processus de diffusion en fonction de ces variables et d'identifier les populations qui sont atteintes par une information en connaissant la population

1. fausse information

à l'origine de l'information et sa thématique. Nous apportons ainsi avec le modèle World-Spread une contribution à la modélisation du processus de diffusion au niveau macroscopique en prenant en compte la dimension géographique des diffuseurs. Le reste du papier est organisé comme suit. La section 2 présente la méthodologie de collecte et d'extraction des données. La section 3 détaille le modèle WorldSpread proposé. La section 4 est consacrée à la présentation des résultats obtenus. Enfin la section 5 conclut et présente nos travaux futurs.

2 Collecte et extraction de données

Dans notre approche, nous avons utilisé les sujets tendance de Twitter. Un sujet tendance est une information très diffusée dans le média social à un instant précis, souvent représenté par un hashtag (mot ou ensemble de mots précédés du caractère dièse résumant le sujet). La méthodologie que nous proposons s'effectue en trois grandes étapes. Dans un premier temps, nous avons collecté toutes les cinq minutes, la liste des 50 meilleurs sujets tendance pour les 62 pays disponibles dans l'API de Twitter. Nous avons sélectionné les sujets diffusés par au moins 4 populations de pays dans un intervalle de 3 jours suivant la première diffusion. Dans un second temps, nous avons ordonné chronologiquement l'adoption des sujets pour chaque population de pays. Ensuite, nous avons classé les sujets automatiquement en fonction du pays C où le sujet apparaît en premier. De plus, nous avons classé les sujets manuellement selon les neuf thématiques S suivantes : les célébrités, les jeux, les films/TV, la musique, les nouvelles, la politique, le sport, la technologie et les autres. Dans un troisième temps, nous avons extrait 3 paramètres pour chaque population ayant adopté le sujet : le temps d'adoption, le temps d'infection et le rayon d'impact. **Le temps d'adoption** T est le temps écoulé entre l'apparition du sujet dans le pays d'origine et l'apparition de ce sujet dans un autre pays. **Le temps d'infection** I est le temps écoulé entre la première apparition du sujet dans la liste des tendances du pays et le moment où il disparaît de la liste. **Le rayon d'impact** R est la distance géographique entre la capitale du pays à l'origine du sujet et la capitale d'un pays où le sujet apparaît ensuite. Enfin, nous avons utilisé l'outil d'extraction de motifs fréquents SPMF (Fournier-Viger et al. (2016)) afin de découvrir les populations de pays souvent impactées ensemble. Dans notre approche, nous avons utilisé deux jeux de données. Un jeu de données d'observation (n° 1) collecté en juin 2017 est composé d'environ 2900 sujets. Un jeu de données d'évaluation (n° 2) collecté en octobre 2017 contient environ 3000 sujets.

3 Le modèle WorldSpread

3.1 Formulation du problème

Dans la suite de ce papier, nous appelons sujet, une information relative à une thématique, qui est susceptible de se propager. Par exemple le sujet #AgentsofSHIELD correspond à une série télévisée et sa thématique est "films/TV". Étant donné C le pays d'origine du sujet, S la thématique du sujet et $t = [t_0, t_1, ..., t_m]$ l'ensemble des instants d'apparition des sujets dans de nouveaux pays relatifs à C et S, où t_0 est l'instant d'apparition dans le pays source. Notre objectif est ici double. Premièrement, il s'agit de définir le nombre de pays diffusant le sujet aux différents instants t_j, c'est-à-dire une liste $n = [n_0, n_1, ..., n_m]$ où n_j est le nombre de

nouveaux pays diffuseurs au temps t_j. Deuxièmement, il s'agit d'identifier également les noms des pays diffusant le sujet aux différents instants t_j, c'est-à-dire une liste $Lc = [Lc_1, ..., Lc_m]$ contenant, pour chaque instant, les pays qui diffusent l'information.

3.2 Description du modèle WorldSpread

Le modèle que nous proposons repose sur la formalisation de nos observations sur des données réelles. Ainsi, à un pays source C et une thématique S sont associés N_{total} le nombre total de pays ayant diffusé la thématique ainsi que les 4 listes suivantes :

— $T = [T_1, T_2, ..., T_k]$ la liste des instants de la diffusion où un ou plusieurs nouveaux pays ont adopté la thématique S,
— $N = [N_1, N_2, ..., N_k]$ la liste des nombres de pays ayant adopté la thématique S à chaque instant T_j,
— $I = [I_1, I_2, ..., I_k]$ la liste des temps d'infection relatifs aux pays ayant adopté la thématique S à chaque instant T_j,
— $R = [R_1, R_2, ..., R_k]$ la liste des rayons d'adoption (distance géographique par rapport au pays C) relatifs aux pays ayant adopté la thématique S à chaque instant T_j.

En plus de ces variables, nous générons également le réseau du pays source du sujet qui repose sur un graphe dirigé $G'_{S,C}$, où S est la thématique du sujet. La construction de ce graphe s'effectue en deux étapes. Premièrement, pour chaque pays C, nous générons en fonction de la thématique S le **graphe local** dirigé $G_{S,C} = (\mathcal{C}_S, \mathcal{E}_S)$ où $\mathcal{C}_S = \{C, C_{1_S}, C_{2_S}, ..., C_{n_S}\}$ est l'ensemble des noeuds représentant les pays qui apparaissent fréquemment dans la diffusion provenant de C pour la thématique S et $\mathcal{E}_S = \{E_{0_S}, E_{1_S}, ..., E_{n-1_S}\}$ est l'ensemble des liens dirigés de C aux noeuds de $\mathcal{C}_S$. Il existe un lien dirigé d'un noeud C_1 vers un noeud C_2 si le sujet se diffuse de C_1 à C_2 directement. Deuxièmement, nous construisons pour chaque pays C selon la thématique S un **graphe global** dirigé $G'_{S,C} = (C'_S, E'_S)$ comme l'union des G_{S,C_n} pour chaque C_n de $\mathcal{C}_S$. L'ensemble des noeuds C'_S est l'union de l'ensemble des noeuds de C_S avec l'ensemble des noeuds des graphes locaux des noeuds des $G_{S,C_{k_S}}$ pour tout C_{k_S} de $\mathcal{C}_S$ et l'ensemble des arêtes dirigées E'_S est l'union de l'ensemble des arêtes des $G_{S,C_{k_S}}$ pour tout C_{k_S} de $\mathcal{C}_S$.

L'objectif du modèle est de décrire l'évolution de la diffusion d'un sujet de manière à pouvoir identifier : le moment où une population est susceptible de diffuser le sujet, le laps de temps pendant lequel elle va diffuser le sujet, la distance géographique qui sépare la population de celle à la source du sujet, et le pays auquel appartient la population.

4 Evaluation du modèle

4.1 Evalutation quantitative

Afin d'évaluer l'"erreur" ou la distance entre les valeurs obtenues sur le jeu d'observation et le jeu d'évaluation, nous nous sommes focalisés sur l'évolution des nombres de pays diffusant l'information au cours du temps et nous avons observé les différences relatives entre les courbes de diffusion. En considérant le cas (États-Unis,Sport), les valeurs obtenues sont présentées dans la figure 1.

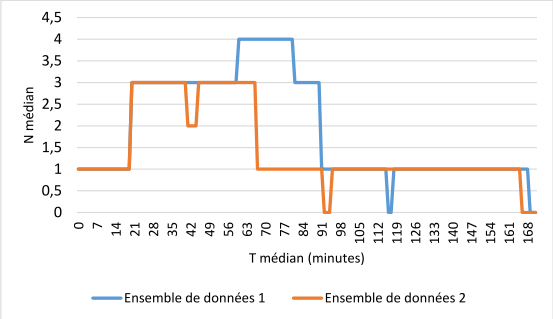

FIG. 1: Comparaison des courbes de diffusion entre le modèle et les données réelles pour le couple (États-Unis,Sports).

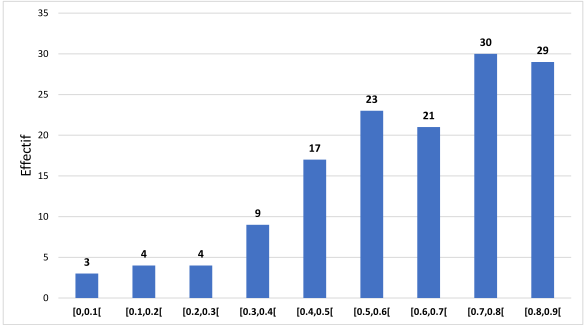

FIG. 2: Distribution de l'erreur du modèle.

Pour estimer numériquement les distances entre les valeurs obtenues sur le jeu 1 et le jeu 2 pour un couple (C,S) donné, nous calculons une erreur relative comprise dans l'intervalle [0-1] donnée par la formule :

$$\Delta_N = \frac{\sum_{i=0}^{n} \frac{|N_1(i) - N_2(i)|}{|max(N_1(i), N_2(i))|}}{n+1}$$

Ainsi, plus la valeur est proche de 0, et plus le modèle décrit fidèlement le phénomène. Nous avons observé la distribution de l'erreur pour les 140 couples (C,S) dont nous disposons d'au moins un exemple à la fois dans le jeu de données 1 et dans le jeu de données 2 (voir figure 2). Ainsi, nous avons remarqué que nous obtenons une erreur strictement inférieure à 0.6 pour 43% de ces couples.

L'erreur semble plus importante lorsque le nombre d'exemples de diffusion est faible. En effet, en observant le nombre moyen d'exemples en fonction de l'erreur, nous avons constaté que moins il y a d'exemples, plus l'erreur est élevée (voir figure 3).

4.2 Evalutation qualitative

Décrire de manière numérique la diffusion d'un sujet en connaissant le pays source C et sa thématique S peut s'avérer utile pour estimer la vitesse de propagation à d'autres populations.

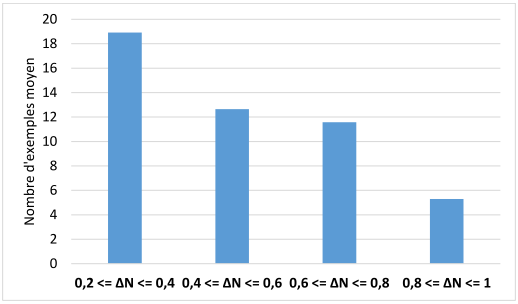

FIG. 3: Distribution de l'erreur en fonction du nombre moyen d'exemples.

Néanmoins, cette première description ne permet pas de cibler précisément les populations susceptibles de diffuser une information. Ainsi, dans un second temps, nous nous sommes intéressés à l'identification des pays diffusant un sujet au cours du temps toujours en prenant en compte le pays source C et la thématique S. L'objectif de notre approche est double. Premièrement, il s'agit de proposer un ensemble de pays $P = \{C_1, C_2, ...C_k\}$ allant probablement diffuser le sujet dont la thématique est S et le pays source est C. Deuxièmement, il est question de suggérer un ordre chronologique O dans l'adoption du sujet par les pays de cet ensemble. Pour chaque exemple de diffusion du jeu de donnée 2, nous avons calculé le rappel A_P et la précision B_P :

— $A_P = \frac{\text{nombre de pays de } P \text{ correctement identifiés}}{\text{nombre de pays de } P} * 100$

— $B_P = \frac{\text{nombre de pays de } P \text{ correctement identifiés}}{\text{nombre de pays total de pays diffusant}} * 100$

De plus nous avons mesuré le taux de bonnes chronologies en utilisant :

$$H = \frac{\text{nombre de chronologie correctes identifiées}}{\text{nombre d'exemples de diffusion}} * 100$$

Pour estimer la qualité de l'identification des pays du modèle, nous avons calculé A_P, B_P pour chaque exemple de diffusion relatif au top 10 des couples (C,S) contenant le plus d'exemples de diffusion. Pour un couple (C,S) donné, nous avons calculé H ainsi que μ_A, μ_B respectivement la moyenne des A_P, B_P. Ensuite, nous avons calculé M_H, M_A et M_B respectivement les moyennes des H, μ_A et μ_B pour les 10 couples (C,S). En évaluant le modèle sur le jeu de donnée 2, nous avons noté que $M_A = 69.7\%$, $M_B = 66.04\%$ et $M_H = 64,5\%$. Ainsi, en moyenne $\frac{2}{3}$ de l'ensemble des pays diffusants sont identifiés correctement par WorldSpread. De plus, en moyenne dans $\frac{2}{3}$ des cas l'ordre chronologique proposé est correctement identifié.

5 Conclusion et perspectives

Dans ce travail, nous avons introduit WorldSpread, un modèle original et innovant qui permet de décrire simplement la diffusion entre des populations sur le média social Twitter à la fois sur le plan quantitatif et sur le plan qualitatif, en connaissant le pays source C et la thématique S de l'information. Notre approche repose sur l'utilisation de deux jeux de données réelles collectées à trois mois d'intervalle et contenant chacun près de 3000 exemples de diffusion regroupés manuellement dans 9 thématiques. Nous avons constaté que le modèle permet

d'établir une description assez précise du phénomène de diffusion notamment pour les couples (C,S) comptant le nombre d'exemples le plus important. Une perspective intéressante serait d'intégrer en plus dans le modèle WorldSpread le nombre d'utilisateurs diffusant l'information au sein des populations concernées. Décrire le processus en prenant en compte cette variable supplémentaire pourrait permettre d'estimer le nombre d'utilisateurs impactés selon le pays en prenant en compte la population source et la thématique de l'information. D'un point de vue marketing, la stratégie à adopter serait de répandre une information au sein d'une population réceptive afin de maximiser sa diffusion. Du point de vue d'un modérateur, un tel modèle pourrait permettre de cibler les populations à informer afin d'atténuer la propagation de rumeurs ou de fausses informations.

Références

Anderson, R. M. et R. M. May (1992). *Infectious diseases of humans : dynamics and control*. Oxford university press.

Fournier-Viger, P., J. C.-W. Lin, A. Gomariz, T. Gueniche, A. Soltani, Z. Deng, et H. T. Lam (2016). The spmf open-source data mining library version 2. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 36–40. Springer.

Hethcote, H. W. (1989). Three basic epidemiological models. In *Applied mathematical ecology*, pp. 119–144. Springer.

Hoang, B.-T., K. Chelghoum, et I. Kacem (2016). A learning-based model for predicting information diffusion in social networks : Case of twitter. In *Control, Decision and Information Technologies (CoDIT), 2016 International Conference on*, pp. 752–757. IEEE.

Li, C.-T., Y.-J. Lin, et M.-Y. Yeh (2018). Forecasting participants of information diffusion on social networks with its applications. *Information Sciences 422*, 432–446.

Zhou, Y., B. Zhang, X. Sun, Q. Zheng, et T. Liu (2017). Analyzing and modeling dynamics of information diffusion in microblogging social network. *Journal of Network and Computer Applications 86*, 92–102.

Summary

Dissemination models proposed in social media are based for the most part on epidemiological hypotheses and not on the observation of real data to describe the characteristics of the diffusion. Such models can not faithfully reproduce the phenomenon of diffusion because they do not consider the observed factors that may influence this phenomenon. Our approach is innovative because we consider the populations of the countries and that it consists in identifying in addition to the number of populations reached, the geographical radius of influence around these populations, the moment of diffusion of the information, the duration of the diffusion and the country to which these populations belong knowing the population at the origin of the information and its theme.