Évaluation de l'influence dans un réseau multi-relationnel : le cas de Twitter

Lobna Azaza [1,2], **Sergey Kirgizov** [1], **Marinette Savonnet** [1], **Éric Leclercq** [1], **Rim Faiz** [2]

1. *Laboratoire LE2I - UMR 6306 - CNRS - ENSAM*
   Univ. Bourgogne Franche-Comté
   9, Avenue Alain Savary
   F-21078 Dijon - France
   Prenom.Nom@u-bourgogne.fr

2. *Laboratoire Larodec*
   Université de Carthage
   Tunis, Tunisie
   Rim.Faiz@ihec.rnu.tn

RÉSUMÉ. *L'influence sur Twitter est devenue un sujet de recherche important. Certains utilisateurs révèlent plus de capacité que d'autres pour influencer les personnes avec lesquelles ils sont connectés. Ainsi, trouver les utilisateurs les plus influents peut permettre une diffusion efficace de l'information à grande échelle, action très utile dans le marketing ou les campagnes politiques. Dans cet article, nous proposons une nouvelle approche pour l'évaluation de l'influence dans les réseaux multi-relationnels tels que Twitter. Notre méthode est basée sur la règle de combinaison conjonctive de la théorie des fonctions de croyance qui permet de fusionner différents types de relations. Nous expérimentons notre méthode sur des données Twitter collectées lors des élections européennes de 2014 et déterminons les candidats les plus influents.*

ABSTRACT. *Influence in Twitter has become recently a hot research topic. Some users are more able than others to influence peers. Thus, studying most influential users leads to reach a large-scale information diffusion area, something very useful in marketing or political campaigns. In this paper, we propose a new approach for influence assessment on multi-relationnal networks such as Twitter. This is based on the conjonctive combination rule in belief functions theory in order to combine different types of interactions. We experiment the proposed method on a large amount of data gathered from Twitter in the context of the European Elections 2014 and deduce top influential candidates.*

MOTS-CLÉS : *Réseau multi-relationnel, Influence, Fusion d'information, Fonctions de croyance.*

KEYWORDS: *multirelational network, Influence, Information fusion, Belief theory.*

Congrès INFORSID, 34ème édition, Grenoble, 31 mai – 3 juin 2016

1. Introduction

Les réseaux sociaux en ligne tel que *Twitter* rassemblent les personnes et renforcent leurs relations avec de nouvelles formes de coopération et de communication. En raison de son immense popularité, *Twitter* est exploité comme une plate-forme pour le marketing et les campagnes politiques. L'une des caractéristiques de *Twitter* est la diffusion de l'information à travers les liens sociaux. Les liens entre les utilisateurs déterminent le flux de l'information et indiquent ainsi l'influence d'un utilisateur sur un autre. Certains utilisateurs, appelés influents, sont plus capables que d'autres de diffuser des informations à un grand nombre d'utilisateurs. Par conséquent, la détection des utilisateurs influents dans un réseau est une clé du succès pour parvenir à une diffusion d'information à large échelle et à faible coût.

L'influence sur *Twitter* est définie comme la capacité d'un utilisateur à provoquer une action chez un autre utilisateur (Leavitt *et al.*, 2009). Le terme "action"signifie les différentes interactions possibles entre les utilisateurs. Par conséquent, la mesure de l'influence sur *Twitter* n'est pas aussi simple puisque *Twitter* offre plusieurs relations (*retweet*, *réponse*, *mention*, *suivre*) et donc plusieurs formes d'interactions. Un utilisateur peut *suivre* un autre utilisateur, ce qui lui permet de voir les *tweets* et les informations de l'utilisateur qu'il suit. Il est également capable de *retweeter* un *tweet*, ce qui expose ce *tweet* à ses abonnés, qui peuvent aussi le *retweeter*. Un utilisateur peut *mentionner* un autre utilisateur en utilisant le préfixe "@"s'il veut lui adresser le *tweet*, ce même *tweet* pouvant être *retweeté* par un autre utilisateur. Enfin, un utilisateur peut *répondre* à un *tweet* et créer ainsi une conversation avec l'utilisateur du tweet intital. Ces différents types d'interactions sont ce qui fait de *Twitter* un réseau multi-relationnel (Wu *et al.*, 2013; Rodriguez, Shinavier, 2010).

L'évaluation de l'influence pose trois principaux défis. Le premier est la diversité des interactions sur lesquels nous pouvons baser les calculs de l'influence. Il est important de les combiner afin d'établir une mesure générale d'influence qui prend en compte les différents types d'interactions entre les utilisateurs. Le second défi est la considération de l'influence indirecte. L'influence est indirecte lorsqu'elle atteint un utilisateur à travers des utilisateurs intermédiaires. Par exemple, un utilisateur peut *retweeter* un *tweet* d'un autre utilisateur indirectement à travers un utilisateur intermédiaire. Il est donc nécessaire de mesurer l'influence en tenant compte des interactions directes et indirectes dans le réseau. Le troisième défi est relatif à l'incertitude lors de la combinaison d'interactions. Dans le cas des réseaux multi-relationnels, il est difficile d'attribuer des pondérations valuées aux différentes interactions avant de fusionner les données quantitatives.

Dans ce papier, pour mesurer l'influence d'un utilisateur, nous combinons, en tenant compte de l'incertitude dans le processus de la mesure, différentes interactions définies par des experts du domaine étudié (la communication politique dans notre cas). La mesure peut être établie entre un couple d'utilisateurs en tenant compte des différentes interactions entre eux deux, mais aussi étendue à une mesure d'influence globale d'un utilisateur dans le réseau.Pour cela, nous définissons un cadre théorique

sur la base de la règle de combinaison conjonctive de la théorie des fonctions de croyance et la règle de Smets (Smets, 1997) pour la fusion et la combinaison des informations. L'approche proposée est flexible et l'influence indirecte dans le graphe multi-relationnel peut être prise en considération. Une évaluation à travers des expérimentations est proposée, elle est basée sur des données *Twitter* collectées dans le cadre du projet TEE 2014 lors de la campagne pour les élections européennes de 2014.

Le reste de l'article est organisé comme suit. La section 2 présente un état de l'art. La section 3 décrit notre approche. La section 4 présente les résultats expérimentaux. Et enfin la section 5 conclut le papier.

2. État de l'art

Dans cette section, nous présentons des travaux sur l'évaluation de l'influence dans *Twitter* et rappelons les concepts de base de la théorie des fonctions de croyance sur lesquels se fonde notre approche.

2.1. L'influence dans Twitter

Dans la littérature, plusieurs approches ont été proposées pour classer les utilisateurs selon leur influence. Certaines approches sont basées sur la **topologie du réseau** et les mesures de centralité. D'autres approches ont établi un classement des utilisateurs en utilisant des algorithmes basés sur la **diffusion**. Une autre famille étend les approches topologiques pour assurer la **fusion d'information**. Dans la suite, nous présentons les travaux principaux pour chaque type d'approches.

Pour mesurer l'influence dans *Twitter*, de nombreux critères peuvent être pris en considération. Les auteurs dans (Leavitt *et al.*, 2009) utilisent trois caractéristiques pour mesurer l'influence : le nombre des *réponses*, *retweets* et *mentions*, en plus du nombre d'*abonnés*. Ils donnent des statistiques relatives à ces mesures mais ne proposent pas un score global de l'influence se basant sur toutes les relations prises en compte. (Cha *et al.*, 2010) utilisent les critères nombre d'*abonnés*, de *retweets* et de *mentions*. Ils calculent la valeur de chaque mesure d'influence pour 6 millions d'utilisateurs puis ils les comparent. Pour ce faire, ils trient les utilisateurs en fonction de chaque relation, puis, ils quantifient comment le classement d'un utilisateur varie selon les différentes relations. La corrélation de Spearman est utilisée comme une mesure de la force d'association entre deux ensembles du classement. Ils ont constaté que le nombre d'abonnés représente la popularité d'un utilisateur, mais il n'est pas lié à d'autres relations telles que les *retweets* et les *mentions*. Leur résultat suggère que le nombre d'*abonnés* seul révèle très peu sur l'influence d'un utilisateur. De même, cette méthode ne fournit pas une mesure globale de l'influence.

(Chen *et al.*, 2013) proposent une méthode de classement local, nommée Cluster Rank, prenant en considération le nombre de voisins et leur coefficient de cluster-

ing[1]. (Bakshy *et al.*, 2011) ont suivi une approche différente pour estimer les utilisateurs influents : ils utilisent les cascades de diffusion d'URL raccourcis et considérent que les utilisateurs qui produisent les cascades les plus longues sont les plus influents. (Brown, Feng, 2011) pensent que la localisation d'un nœud dans le réseau peut déterminer son influence. Considérant ce fait, l'algorithme de décomposition *k-shell* peut être utilisé (Seidman, 1983). Son principe est d'attribuer un indice de référence ks pour chaque nœud tel que les nœuds ayant les valeurs les plus faibles sont situés à la périphérie du réseau tandis que les nœuds avec les valeurs les plus élevées se trouvent au centre du réseau, ce sont ces nœuds qui auront le plus d'influence. Les auteurs ont adapté l'algorithme de décomposition *k-shell* aux caractéristiques du réseau *Twitter*. Qasem *et al.* (Qasem *et al.*, 2015) présentent une nouvelle approche pour la détection des utilisateurs influents. L'approche proposée détecte les utilisateurs qui augmentent la taille du réseau social en attirant de nouveaux utilisateurs dans le réseau.

L'inconvénient des algorithmes basés sur la topologie du réseau est de ne considérer que les informations de l'utilisateur, sans considérer l'interaction entre les utilisateurs à travers les séquences des relations. Avec *Twitter*, l'influence d'un utilisateur est impactée par la diffusion de l'information entre les utilisateurs.

D'autres recherches proposent de classer les utilisateurs en utilisant des algorithmes basés sur la **diffusion**, avec l'hypothèse commune selon laquelle un utilisateur est influent s'il pointe vers de nombreux voisins très influents. Dans (Weng *et al.*, 2010), les auteurs proposent TwitterRank, une extension de l'algorithme PageRank (Page *et al.*, 1999), afin de mesurer l'influence des utilisateurs en tenant compte des sujets associés aux tweets. Bien que l'idée soit prometteuse, les résultats expérimentaux montrent qu'il y a des utilisateurs qui *suivent* d'autres utilisateurs sans présence de similarité de sujets entre eux et leurs amis. La méthode a ignoré d'autres critères importants tels que les *mentions* et les *réponses*. Romero *et al.* (Romero *et al.*, 2011) proposent le IP-Algorithme basé sur l'algorithme HITS (Kleinberg, 1999). Les auteurs considérent l'influence comme le niveau de propagation du contenu dans le réseau (*retweets*). De plus, les auteurs estiment que l'influence d'un utilisateur ne dépend pas seulement de la taille de son audience, mais aussi de sa passivité. La passivité d'un utilisateur est le fait qu'il ne transmet pas l'information au réseau. L'algorithme a montré une meilleure précision que d'autres mesures d'influence tels que PageRank, le nombre d'abonnés et le nombre de *mentions*. Bien que la passivité semble être un facteur à prendre en compte dans le calcul de l'influence, ce travail a ignoré d'autres relations importantes telles que la *réponse*. Ashwini *et al.* (Ashwini, M.R., 2015) considèrent que *Twitter* est une plate-forme de diffusion d'information et étudient le problème de l'identification des utilisateurs influents. Ils proposent ProfileRank, un modèle de diffusion d'information basé sur la marche aléatoire qui estime l'influence des utilisateurs et la pertinence du contenu. ProfileRank est fondé sur le principe qu'un utilisateur influent crée du contenu pertinent. La limite de cette ap-

1. En théorie des graphes, le coeffcient de clustering mesure à quel point les voisins d'un sommet sont connectés.

proche est que l'influence est estimée en se basant seulement sur la relation *retweet* et la méthode ignore d'autres relations importantes.

Dans des travaux récents, la **fusion d'information** est considérée afin de contourner les limitations des méthodes existantes. Dans (Simmie *et al.*, 2013), les auteurs proposent la combinaison de deux modèles pour classer les utilisateurs influents : l'algorithme PageRank et HMC (Modèle de Markov Caché). Ils ont construit un HMM pour observer l'évolution de l'influence à travers le temps et utilisent les trois relations *retweet*, *mention* et *réponse*. Le modèle est évalué sur une enquête considérée comme une réalité du terrain. Le modèle proposé diffère des autres par la combinaison de trois relations. Toutefois, puisque le but est de classer l'influence des utilisateurs, l'influence d'un utilisateur donné ne révèle pas d'informations sur son degré d'influence (forte ou faible influence), le résultat du modèle est utile uniquement pour le classement des utilisateurs.

Aucun travail de recherche existant ne prend en compte la combinaison de plusieurs relations avec de l'incertitude. Or, il nous parait important, pour mesurer l'influence, de tenir compte des degrés d'incertitude sur les poids attribués aux différentes interactions selon leur importance. Dans cet objectif, nous proposons l'utilisation de la théorie des fonctions de croyance. Dans des recherches récentes, la théorie des fonctions de croyance est exploitée pour mesurer l'influence dans des réseaux pondérés (Cai *et al.*, 2013; Wei *et al.*, 2013) et complexes (Mo *et al.*, 2015) avec l'objectif commun de modifier les mesures de centralité existantes. Au meilleur de notre connaissance, ceci est la première fois que la théorie des fonctions de croyance est exploitée pour mesurer l'influence sur le réseau *Twitter* avec des patterns d'interactions au lieu des mesures de centralité.

2.2. Théorie des fonctions de croyance

La théorie des fonctions de croyance est considérée comme un outil général pour le raisonnement avec incertitude, et a été reliée à d'autres cadres tels que les théories des probabilités, des possibilités et des probabilités imprécises (Denoeux, Masson, 2012). La théorie des fonctions de croyance, aussi connue comme la théorie de l'évidence ou théorie de Dempster-Shafer, a été d'abord introduite par A. Dempster dans le contexte de l'inférence statistique, et a été developpée plus tard par G. Shafer comme un outil général pour la modélisation de l'incertitude épistémique (Kotz, N. L. Johnson eds., 1982).

Dans les paragraphes suivants, nous allons rappeler les concepts de base de la théorie des fonctions de croyance. Soient Ω un ensemble fini et 2^{Ω} l'ensemble de tous les sous-ensembles de Ω. Dans le contexte de la théorie de Dempster-Shafer, Ω est souvent appelée un cadre de discernement. La masse m est une fonction $m : 2^{\Omega} \rightarrow [0, 1]$ tel que :

$$\sum_{X \in 2^{\Omega}} m(X) = 1 \text{ and } m(\emptyset) = 0 \tag{1}$$

La masse $m(X)$ exprime la part de la croyance qui supporte le sous-ensemble X de Ω, $m(\emptyset) = 0$ car nous considérons que le cadre de discernement est exhaustif et exclusif (hypothèse du monde clos).

La théorie des fonctions de croyance permet, non seulement la représentation de la connaissance partielle, mais aussi la fusion de l'information (Nimier, Appriou, 1995). La fusion d'information est réalisée par la règle de combinaison conjonctive (Smets, 1997), elle suppose que toutes les sources sont fiables et consistantes. Considérant deux fonctions de masse m_1 et m_2, la règle de combinaison conjonctive est définie par :

$$(m_1 \textcircled{\cap} m_2)(C) = \sum_{A \cap B = C} m_1(A) m_2(B), \quad A, B, C \in 2^{\Omega} \tag{2}$$

Afin de prendre une décision, nous essayons de sélectionner l'hypothèse la plus probable, ce qui peut être difficile à réaliser directement avec les bases de la théorie des fonctions de croyance où les fonctions de masse sont données, non seulement pour les singletons, mais aussi pour les sous-ensembles du cadre de discernement. Ils existent plusieurs solutions pour assurer la prise de décision au sein de la théorie des fonctions de croyance, la plus connue est la probabilité pignistique (Smets, 1989). Contrairement aux fonctions de masse qui sont définies sur 2^{Ω}, la probabilité pignistique est une mesure de probabilité définie sur Ω. La probabilité pignistique a été proposée dans le modèle des croyances transférables (Smets, Kennes, 2008). Elle est basée sur deux niveaux : le "niveau crédal"où les croyances sont représentées par des fonctions de croyance et le "niveau pignistique"où les croyances sont utilisées pour prendre la décision et représentées comme des fonctions de probabilité appelées probabilités pignistiques et notées *bet* définies par :

$$\mathrm{bet}(x) = \sum_{x \in X \subseteq \Omega} \frac{m(X)}{|X|} \tag{3}$$

3. Approche proposée

L'objectif de notre approche est de trouver des critères de manifestation de l'influence et de pondérer ces critères. Afin de mesurer l'influence d'un utilisateur, nous utilisons la théorie des fonctions de croyance pour effectuer la fusion des informations issues des différentes formes d'interactions (patterns d'interaction directs ou indirects). La figure 1 donne un aperçu des différentes étapes de l'approche proposée. D'abord, l'information de *Twitter* est modélisée dans un graphe en sélectionnant les relations et les patterns pertinents puis le choix des degrés d'influence et l'initialisation des masses de croyance sont réalisés, cette étape dépend du domaine étudié. Dans le niveau crédal, nous associons les différentes fonctions de masse à chaque relation et pattern puis nous les combinons pour obtenir la masse de croyance de l'influence. Dans le niveau pignistique, nous calculons la probabilité pignistique afin de prendre la décision sur

le degré d'influence d'un nœud. Dans cette section, nous détaillons chaque étape du processus de l'évaluation.

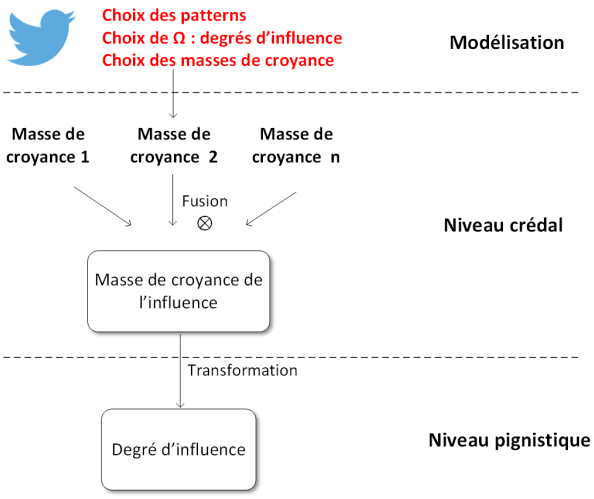

Figure 1. Étapes de l'approche proposée

3.1. Graphe de l'influence

Les réseaux sociaux sont généralement modélisés comme un graphe (Barnes, 1969) représenté par $G = (V, E)$ comprenant un ensemble V de sommets ou nœuds et un ensemble E d'arcs ou de liens. Dans le réseau *Twitter*, le graphe est hétérogène puisque nous avons différentes relations entre les nœuds et différents types de nœuds. Par exemple, il peut exister une relation *Suivre* entre deux utilisateurs, une relation *Retweet* entre un *tweet* et un utilisateur. Afin de modéliser cette hétérogénéité, un graphe multi-relationnel peut être utilisé (Rodriguez, Shinavier, 2010). Comme nous souhaitons évaluer l'influence d'un utilisateur sur d'autres, nous limitons le graphe à des nœuds homogènes (les utilisateurs), ainsi, nous travaillons sur un graphe multiplexe (Kanawati, 2015). Dans un graphe multiplexe, l'ensemble des liens E est divisé en classes disjointes : $E = \bigcup_{r \in R} E_r$, où R est l'ensemble de types de relations possibles. Par exemple, dans *Twitter* nous pouvons considérer :

R = {*Retweet*, *Mention*, *Réponse*, *Suivre*}

Nous définissons un pattern d'interaction p comme une séquence de relations, par exemple un *Retweet* d'une *Réponse* ou *Retweet* d'un *tweet* avec une *Mention*. Soit P l'ensemble des patterns d'interaction qui ont été identifiés pour modéliser l'influence dans un domaine spécifique, cet ensemble peut être donné par les chercheurs en sciences sociales par exemple. Notons par $R = R \bigcup P$ l'ensemble des relations y compris les patterns d'interaction.

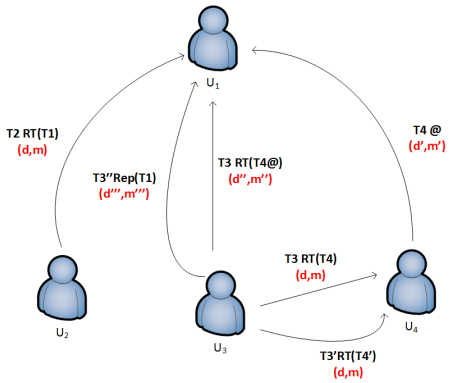

Figure 2. Graphe de l'influence

Dans ce contexte, nous introduisons le graphe de l'influence (Figure 2) comme un graphe multiplexe étiqueté. Les nœuds représentent les utilisateurs, et les liens sont les différentes relations entre eux. Par exemple, le lien entre u_4 et u_1 signifie que le *tweet* $T4$ émis par u_4 mentionne u_1, le lien étiqueté $T3RT(T4@)$ correspond au *tweet* $T3$ de u_3 qui est un *Retweet* (RT) du *tweet* $T4$ de u_4 qui mentionnait u_1, ce qui représente le pattern d'interaction *retweet* d'une *mention* entre les utilisateurs u_3 et u_1 effectué à travers l'utilisateur u_4. Nous trouvons d'ailleurs dans le graphe le tweet $T3$ de u_3 vers u_4. Les liens sont aussi étiquetés avec les degrés d'influence d (par exemple, Faible, Moyenne, Forte) et les masses de croyance m qui dépendent du type de la relation.

3.2. Fusion des masses dans le graphe de l'influence

En se basant sur la théorie de Dempster-Shafer discutée dans la section 2, nous fusionnons différentes fonctions de masse définies dans le graphe multiplexe.

Soit Ω un ensemble ordonné des degrés d'influence possibles :

$$\Omega = \{\text{Très Faible}, \text{Faible}, \text{Assez Moyenne}, \text{Moyenne}, \text{Assez forte}, \\ \text{Forte}, \text{Très Forte}, \text{Extrêmement Forte}\} \tag{4}$$

Dans la théorie générale de Dempster-Shafer, 2^{Ω} est utilisé comme domaine des fonctions de masse, dans notre approche, nous utilisons seulement un sous-ensemble Λ de 2^{Ω}, précisément :

$$\Lambda = \{\text{Très Faible}, \text{Faible}, \text{Assez Moyenne}, \text{Moyenne}, \text{Assez Forte}, \\ \text{Forte}, \text{Très Forte}, \text{Extrêmement Forte}, \Omega\} \tag{5}$$

Alors, les fonctions de masse sont définies comme suit : $m : \Lambda \rightarrow [0, 1]$

Table 1. Definition of the operation ⓐ

ⓐ	**T.Faible**	**Faible**	**A.Moyenne**	**Moyenne**	**A.Forte**	**Forte**	**T.Forte**	**E.Forte**	**Ω**
T.Faible	Faible	A.Moyenne	Moyenne	A.Forte	Forte	T.Forte	T.Forte	E.Forte	T.Faible
Faible	A.Moyenne	A.Moyenne	Moyenne	A.Forte	Forte	T.Forte	T.Forte	E.Forte	Faible
A.Moyenne	Moyenne	Moyenne	A.Forte	Forte	T.Forte	T.Forte	T.Forte	E.Forte	A.Moyenne
Moyenne	A.Forte	A.Forte	Forte	Forte	T.Forte	T.Forte	T.Forte	E.Forte	Moyenne
A.Forte	Forte	Forte	T.Forte	T.Forte	T.Forte	T.Forte	T.Forte	E.Forte	A.Forte
Forte	T.Forte	T.Forte	T.Forte	T.Forte	T.Forte	E.Forte	E.Forte	E.Forte	Forte
T.Forte	T.Forte	T.Forte	T.Forte	T.Forte	T.Forte	E.Forte	E.Forte	E.Forte	T.Forte
E.Forte	E.Forte	E.Forte	E.Forte	E.Forte	E.Forte	E.Forte	E.Forte	E.Forte	E.Forte
Ω	T.Faible	Faible	A.Moyenne	Moyenne	A.Forte	Forte	T.Forte	E.Forte	Ω

Pour chaque type de relation $r \in R$, une fonction de masse m_r est associée. Afin d'estimer le degré d'influence d'un nœud spécifique u, nous prenons en compte la structure locale du graphe de l'influence autour du nœud u et nous combinons les fonctions de masses de croyance des liens incidents en utilisant une version modifiée de la règle de combinaison conjonctive (2) :

$$(m \otimes m')(z) = \sum_{y \,ⓐ\, x = z} m(x)m'(y), \quad x, y, z \in \Lambda \tag{6}$$

ⓐ est une opération symétrique ⓐ : $\Lambda \times \Lambda \to \Lambda$, le tableau 1 représente un exemple de l'opération ⓐ. Cette fonction assure notre hypothèse : plus nous combinons des relations relatives à un utilisateur, plus il prend de l'importance en influence.

Comme plusieurs relations peuvent exister entre le nœud u et ses voisins, nous désignons par I_r l'ensemble de tous les liens ayant le type de relation r et nous obtenons l'ensemble des fonctions de masse suivant : $\{m_{r,i} : r \in R, i \in I_r\}$. Nous combinons les fonctions de masse afin d'obtenir une masse de croyance globale correspondante au degré d'influence du nœud u. L'ordre des combinaisons pouvant affecter nos résultats, nous devons choisir un ordre pour être consistant. Afin de simplifier les expressions nous écrivons $\bigotimes_{i \in \{1,2,3\}}$ au lieu de $m_1 \otimes m_2 \otimes m_3$. Ainsi, nous considérons l'ordre des combinaisons suivant :

1. Pour un type de relation donné r, nous combinons les masses des relations de type r afin d'obtenir r-pré-résultat avec $\hat{m}_r$ défini comme suit : $\hat{m}_r = \bigotimes_{i \in I_r} m_{r,i}$
2. Après nous combinons tous les r-pré-résultats en utilisant : $\bigotimes_{r \in R} \hat{m}_r$

En fonction de l'opération ⓐ, une telle procédure peut finalement converger vers une certaine masse stationnaire.

Une fois que nous avons la masse de croyance globale sur un certain nœud, nous utilisons une version modifiée de la probabilité pignistique définie dans l'équation 3 afin de prendre la décision à propos du degré de l'influence du nœud. Dans notre cas les masses de croyance sont définies sur Λ et la probabilité pignistique est calculée en répartissant uniformément la masse de Ω sur tous les autres éléments de Λ :

$$\mathrm{bet}(x) = m(x) + \frac{m(\Omega)}{|\Omega|}, \quad x \in \Omega \tag{7}$$

Le code source est disponible sur github : https://github.com/kerzol/Influence-assessment-in-twitter.

3.3. Illustrations

Afin d'illustrer notre méthode, nous considérons les fonctions de masse suivantes associées à la relation *retweet* et au pattern de diffusion *retweet* d'une *mention* :

$$\text{Retweet} \mapsto \begin{cases} m_{\text{Retweet}}(\text{Faible}) = 0.4 \\ m_{\text{Retweet}}(\Omega) = 0.6 \end{cases} \qquad \text{RTmention} \mapsto \begin{cases} m_{\text{RTmention}}(\text{Moyenne}) = 0.7 \\ m_{\text{RTmention}}(\Omega) = 0.3 \end{cases}$$

Les masses de croyance $m_{\text{Retweet}}(\Omega)$ et $m_{\text{RTmention}}(\Omega)$ représentent l'ignorance partielle. Nous avons affecté une masse de croyance plus importante au pattern d'interaction *Retweet* d'une *Mention* car nous considérons que l'existance de ce pattern est très significative en terme d'influence.

Cas : Retweet + Retweet d'une Mention

Après initialisation des masses de croyance pour les relations, nous suivons le processus de l'approche proposée pour mesurer l'influence obtenue suite à la combinaison d'un *Retweet* avec le pattern *retweet* d'une *Mention*. D'abord nous utilisons l'opération ⓐ donnant les correspondances entre les degrés d'influence (tableau 1), après nous calculons la combinaison conjonctive. La fonction de masse combinée des deux relations est donnée dans le tableau 2:

Tableau 2. Combinaison d'un Retweet avec un Retweet d'une Mention

$\otimes$	$Faible$	Ω
	0.4	0.6
$Moyenne$	$Assez\ Forte$	$Moyenne$
0.7	0.28	0.42
Ω	$Faible$	Ω
0.3	0.12	0.18

Nous obtenons : $m(\text{Faible}) = 0.12 \quad m(\text{Moyenne}) = 0.42$

$m(\text{Assez Forte}) = 0.28 \quad m(\Omega) = 0.18$

Finalement, pour prendre la décision sur le degré d'influence, nous calculons la probabilité pignistique en utilisant l'équation 7 (Tableau 3). Par exemple, pour le degré Faible, nous procédons comme suit pour obtenir la probabilité pignistique :

$$\text{bet}(\text{Faible}) = m(\text{Faible}) + \frac{m(\Omega)}{|\Omega|} = 0.12 + \frac{0.18}{8} = 0.1425$$

Tableau 3. Probabilité pignistique dans le cas d'un Retweet suivi d'un Retweet d'une Mention

Très Faible	0.0225
Faible	0.1425
Assez Moyenne	0.0225
Moyenne	0.4425
Assez Forte	0.3025
Forte	0.0225
Très Forte	0.0225
Extrêmement Forte	0.0225

On peut conclure que le degré d'influence d'un *Retweet* suivi d'un *Retweet* d'une *Mention* est Moyenne puisqu'il a la plus grande probabilité pignistique soit 0.4425.

4. Expérimentations et résultats

Les travaux de recherche menés se déroulent dans le cadre du projet TEE 2014 dont l'intitulé exact est "*Twitter* aux élections européennes : Une étude contrastive internationale des utilisations de *Twitter* par les candidats aux élections au Parlement Européen en mai 2014". Ce projet international, mené par la Maison des Sciences de l'Homme (MSH) de Dijon, réunit près de 45 chercheurs (majoritairement des politologues, sociologues, chercheurs en communication) de 10 laboratoires de recherche répartis dans 5 pays européens (France, Allemagne, Belgique, Italie et Espagne). L'objectif global de ce projet est d'observer et d'analyser la communication des politiques sur Twitter durant les élections européennes de mai 2014 dans les 5 pays d'étude.

4.1. Description des données

Pour collecter les informations de *Twitter*, nous avons utilisé notre outil *SNFreezer*[2] (Leclercq *et al.*, 2015). Trois types d'informations (généralisées sous le terme "source") peuvent être pris en paramètre dans cette collecte : des comptes utilisateurs, des *hashtags* et des mots ou phrases. Ces différentes sources ont été choisies par les politologues, et nous retrouvons parmi elles les noms des principaux candidats, leurs

2. https://github.com/SNFreezer

comptes *Twitter*, et les *hashtags* relatifs à ces candidats, leurs partis, ou plus généralement à l'élection étudiée. L'objectif de la collecte est de capter les *tweets* mentionnant les utilisateurs désignés, ceux contenant un certain *hashtag*, mot ou phrase, ou encore les *tweets* envoyés par les utilisateurs spécifiés. En plus, nous collectons les informations sur ces *tweets* tels que les *tweets retweetés*, les utilisateurs *mentionnés* dans les *tweets* et les *réponses* aux *tweets*. La collecte sur 50 jours consécutifs nous a permis de disposer de 37 millions de *tweets*.

4.2. Expérimentations et résultats

L'objectif de nos expérimentations est de mesurer l'influence des candidats sur *Twitter*. Afin d'étudier leur influence, nous affectons des masses aux relations considérés, puis pour chaque candidat, nous combinons les masses de croyances en itérant sur le nombre de *retweets*, *mentions* et *réponses*.

Le choix et l'affectation des masses dans l'étape de l'initialisation sont une question importante lorsque nous traitons de données réelles. Dans certains domaines tels que la politique, les utilisateurs ont un très grand nombre de relations. Avec une initialisation avec les mêmes valeurs que celles utilisées dans la section illustration, l'influence converge pour tous les candidats vers le plus haut degré d'influence Extrêmement Forte après un nombre d'itérations restreint ($\simeq$ 45 itérations), ce qui ne nous a pas permis de pouvoir comparer l'influence des candidats. Pour règler cette question, nous effectuons une mise à l'échelle et nous utilisons les affectations de masses suivantes :

$$\textit{Retweet} \mapsto \begin{cases} m_{\text{Retweet}}(\text{Très Faible}) = 0.55 \cdot 10^{-3} \\ m_{\text{Retweet}}(\Omega) = 1 - 0.55 \cdot 10^{-3} \end{cases}$$

$$\textit{Mention} \mapsto \begin{cases} m_{\text{Mention}}(\text{Très Faible}) = 0.45 \cdot 10^{-3} \\ m_{\text{Mention}}(\Omega) = 1 - 0.45 \cdot 10^{-3} \end{cases}$$

$$\textit{Réponse} \mapsto \begin{cases} m_{\text{Réponse}}(\text{Très Faible}) = 0.45 \cdot 10^{-3} \\ m_{\text{Réponse}}(\Omega) = 1 - 0.45 \cdot 10^{-3} \end{cases}$$

Nous avons appliqué notre approche sur le corpus français comprenant 616 candidats et 4 millions de *tweets*. Le tableau 4 montre les résultats obtenus avec notre approche pour trois candidats français Marine Le Pen, Florian Philippot et Jean-Luc Mélenchon. Nous pouvons conclure que le degré d'influence dans *Twitter* pour le candidat Marine Le Pen est Extrêmement Forte avec la masse de croyance de 0.8173448. Les résultats fournissent non seulement le degré d'influence d'un candidat, mais aussi donnent par les masses sur les différents degrés d'influence une indication de la croyance que nous avons dans les résultats donnés.

Tableau 4. Résultats pour 3 candidats français influents

	M. Le Pen	**F. Philippot**	**J.L. Mélenchon**
Ω	0	0	0
Très Faible	0	0.000011065	0.000030278
Faible	0	0.00007295998	0.0001832843
Assez Moyenne	0	0.0007035528	0.001403947
Moyenne	0	0.003033557	0.004954501
Assez Forte	0	0.008340205	0.01247841
Forte	0	0.02191526	0.02977818
Très Forte	0.1826552	0.5830090	0.7960571
Extrêmement Forte	0.8173448	0.3829144	0.1551143

4.3. Vers le classement de l'influence des utilisateurs

L'approche proposée peut être aussi exploitée pour classer les utilisateurs selon leur influence. Afin de classer les candidats selon leur influence et en partant de nos résultats, nous procédons comme suit :

1. Pour chaque candidat nous prenons le degré d'influence ayant la masse de croyance maximale (par exemple, pour Marine Le Pen nous choisissons Extrêmement Forte)

2. Nous classons les candidats selon leur "degré d'influence maximal"

3. Si deux candidats ont le même "degré d'influence maximal", nous comparons les masses de croyance du plus haut degré d'influence en utilisant l'ordre suivant pour les degrés d'influence : $\Omega < Très\ Faible < Faible < Assez\ Moyenne < Moyenne < Assez\ Forte < Forte < Très\ Forte < Extrêmement\ Forte$

Nous procédons ainsi puisqu'il est injuste de classer les candidats selon la masse de croyance maximale qu'ils ont dans les différents degrés. Nous pouvons avoir un utilisateur plus influent qu'un autre même s'il a une masse de croyance plus faible que lui sur le même degré. Ceci est dû au fait que, les masses de croyance du plus haut degré d'influence suivant ont augmenté et sont devenues assez importantes. Par exemple, Influence(Florian Philippot) = Influence(Jean-Luc Mélenchon) = Très Forte, et $m_{\text{Philippot}}(\text{Extrêmement Forte}) > m_{\text{Mélenchon}}(\text{Extrêmement Forte})$ bien que le candidat Florian Philippot a une masse de croyance sur le degré Très Forte plus faible que la masse de croyance du candidat Jean-Luc Mélenchon sur le même degré (voir tableau 4), il est classé avant Jean-Luc Mélenchon (tableau 5) puisqu'il a une masse de croyance plus élevée sur le degré Extrêmement Forte. Nous effectuons alors la procédure de la combinaison pour tous les candidats et déduisons leur classement selon leur degré d'influence. Les résultats sont présentés dans le tableau 5.

Le tableau 6 présente le classement obtenu en utilisant les critères utilisés par (Cha *et al.*, 2009). Ces critères sont le nombre de Retweets, Mentions et Réponses. Les résultats présentés ne montrent pas l'influence globale dans le réseau puisque nous trouvons différents classements pour chaque type de relation. Alors que notre méthode (Tableau 5) nous permet d'avoir un classement unique qui tient compte de

Tableau 5. Candidats français les plus influents selon notre approche

Classement	Candidats	Degré d'influence	Masse de croyance
1	Marine Le Pen	Extrêment Forte	0.8173448
2	Florian Philippot	Très Forte	0.5830090
3	Jean-Luc Mélenchon	Très Forte	0.7960571
4	Christine Boutin	Très Forte	0.9796956
5	Aymeric Chauprade	Très Forte	0.4171324655
6	Nicolas Dupont-Aignan	Très Forte	0.5293170700
7	José Bové	Très Forte	0.2925722297
8	Geoffroy Didier	Moyenne	0.2092645352
9	Raquel Garrido	Moyenne	0.2048485
10	Marielle De Sarnez	Assez Moyenne	0.2074260

Table 6. Candidats français les plus influents selon les différentes relations et le degré de centralité

Classement	*Retweet*	*Mention*	*Réponse*	*Degré de centralité*
1	Marine Le Pen	Marine Le Pen	Christine Boutin	Marine Le Pen
2	Florian Philippot	Christine Boutin	Marine Le Pen	Christine Boutin
3	Jean-Luc Mélenchon	Jean-Luc Mélenchon	Florian Philippot	Florian Filippot
4	Aymeric Chauparde	Florian Philippot	Jean-Luc Mélenchon	Jean-Luc Mélenchon
5	François Asselineau	Nicolas Dupont-Aignan	Louis de Gouyon Matigon	Nicolas Dupont-Aignan
6	Corinne Morel-Darleux	José Bové	Nicolas Dupont-Aignan	Aymeric Chauparde
7	Nicolas Dupont-Aignan	Aymeric Chauparde	Jean-Sébastien Herpin	José Bové
8	Louis Aliot	Raquel Garrido	Julien Rochedy	Geoffroy Didier
9	Denis Payre	Jérome Lavrilleux	Geoffroy Didier	Raquel Garrido
10	Yannick Jadot	Marielle de Sarnez	Louis Aliot	Yannick Jadot

tous les critères considérés. La dernière colonne du tableau 6 montre le classement des candidats selon leur degré de centralité calculé en utilisant le nombre de voisins de chaque candidat dans le réseau. Le degré de centralité permet, pour chaque candidat, d'avoir un classement global sans indication sur leur degré d'influence contrairement à nos résultats présentés dans le tableau 5, l'influence mesurée est globale en prenant en compte les relations possibles dans la même mesure.

5. Conclusion

Dans cet article, nous avons proposé une approche pour l'évaluation de l'influence sur le réseau social *Twitter*. Cette approche répond à des limites des systèmes existants tels que la prise en compte de la combinaison des relations et l'incertitude engendrée par la fusion d'informations. Dans notre approche, nous avons proposé un graphe de l'influence nous permettant de considérer les différentes relations dans le réseau (*Retweet*, *Mention* et *Réponse*) ainsi que les séquences possibles de relations. En se basant sur la théorie des fonctions de croyance, nous avons établi une mesure d'influence globale pour les utilisateurs par combinaison des différentes relations. Nous avons expérimenté l'approche sur des données *Twitter* collectées dans le cadre du projet de recherche TEE 2014. Pour renforcer l'approche proposée nous souhaitons enrichir le modèle et développer d'autres patterns d'interaction en col-

laboration avec les politologues du projet TEE 2014. Nous souhaitons intégrer dans notre approche les *hashtags* et les *favoris* et ainsi pouvoir traiter des patterns plus complexes. Cependant, l'extraction des patterns du jeu de données nécessite des structures de données adaptées pour concevoir une application utilisable par les politologues et permettant d'évaluer et d'affiner leur modèle de l'influence. Par ailleurs, la méthode de classement des utilisateurs sera améliorée, par exemple avec des intervalles de confiance et nous allons comparer les résultats obtenues avec les résultats obtenus à partir des algorithmes connus dans la littérature tels que TwitterRank et HITS.

Bibliographie

Ashwini S. S., M.R. S. (2015). Profile ranking using user influence and content relevance with classification using sentiment analysis. *International Journal of Computer Science and Mobile Computing*, vol. 4, p. 1075–1080.

Bakshy E., Hofman J. M., Mason W. A., Watts D. J. (2011). Everyone's an Influencer: Quantifying Influence on Twitter. In *Proceedings of the fourth acm international conference on web search and data mining*, p. 65–74. New York, NY, USA, ACM. Retrieved from http://doi.acm.org/10.1145/1935826.1935845

Barnes J. A. (1969). Graph Theory and Social Networks: A Technical Comment on Connectedness and Connectivity. *Sociology*, p. 215-232.

Brown P. E., Feng J. (2011). Measuring user influence on twitter using modified k-shell decomposition. In *Fifth international aaai conference on weblogs and social media*, p. 18-23.

Cai G., Daijun W., Yong H., Sankaran M., Yong D. (2013). A modified evidential methodology of identifying influential nodes in weighted networks. *Physica A: Statistical Mechanics and its Applications*, vol. 392, n° 21, p. 5490 - 5500. Retrieved from http://www.sciencedirect .com/science/article/pii/S0378437113005773

Cha M., Haddadi H., Benevenuto F., Gummadi K. (2010). Measuring user influence in twitter: The million follower fallacy. In *4th international aaai conference on weblogs and social media (icwsm).*

Cha M., Mislove A., Gummadi K. P. (2009). A Measurement-driven Analysis of Information Propagation in the Flickr Social Network. In *Proceedings of the 18th international conference on world wide web*, p. 721–730. New York, NY, USA, ACM.

Chen D.-B., Gao H., Lü L., Zhou T. (2013). Identifying Influential Nodes in Large-Scale Directed Networks: The Role of Clustering. *PLoS ONE*, vol. 8, n° 10.

Denoeux T., Masson M.-H. E. (2012). Belief Functions: Theory and Applications. In *Proceedings of the 2nd international conference on belief functions, 9-11 may 2012*, p. 444.

Kanawati R. (2015). Multiplex network mining: a brief survey. *IEEE Intelligent Informatics Bulletin*.

Kleinberg J. M. (1999, September). Authoritative sources in a hyperlinked environment. *J. ACM*, p. 604–632.

Kotz S., N. L. Johnson eds. W. (1982). Belief functions. *Encyclopedia of Statistical Sciences 1 209*.

Leavitt A., Burchard E., Fisher D., Gilbert S. (2009, September). The Influentials: New Approaches for Analyzing Influence on Twitter. *Webecology Project*.

Leclercq E., Savonnet M., Grison T., Kirgizov S., Basaille I. (2015). SNFreezer: a Platform for Harvesting and Storing Tweets in a Big Data Context. In A. Frame, A. Mercier, G. Brachotte, C. Thimm (Eds.), *Twitter and the european parliamentary elections: researching political uses of microblogging*, p. 1–16. DE, Peter Lang.

Mo H., Gao C., Deng Y. (2015, April). Evidential method to identify influential nodes in complex networks. *Systems Engineering and Electronics, Journal of*, vol. 26, n° 2, p. 381-387.

Nimier V., Appriou A. (1995). Utilisation de la théorie de dempster-shafer pour la fusion d'informations. *GRETSI, Groupe d'Etudes du Traitement du Signal et des Images*, p. 137-140.

Page L., Brin S., Motwani R., Winograd T. (1999). The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th international world wide web conference*, p. 161–172.

Qasem Z., Jansen M., Hecking T., Hoppe H. (2015). On the detection of influential actors in social media. In *Signal-image technology and internet-based systems (sitis), 2015 11th international conference on*, p. 421–427.

Rodriguez M. A., Shinavier J. (2010). Exposing multi-relational networks to single-relational network analysis algorithms. *Journal of Informetrics*, vol. 4, n° 1, p. 29–41.

Romero D. M., Galuba W., Asur S., Huberman B. A. (2011). Influence and passivity in social media. In *Proceedings of the 20th international conference companion on world wide web*, p. 113–114.

Seidman S. B. (1983). Network structure and minimum degree. *Social Networks*, vol. 5, n° 3, p. 269 - 287.

Simmie D., Vigliotti M., Hankin C. (2013). Ranking twitter influence by combining network centrality and influence observables in an evolutionary model. In *Signal-image technology internet-based systems (sitis), 2013 international conference on*, p. 491-498.

Smets P. (1989). Constructing the pignistic probability function in a context of uncertainty. In *Uai*, vol. 89, p. 29–40.

Smets P. (1997). Imperfect Information: Imprecision and Uncertainty. In A. Motro, P. Smets (Eds.), *Uncertainty management in information systems*, p. 225-254.

Smets P., Kennes R. (2008). The transferable belief model. In *Classic Works of the Dempster-Shafer Theory of Belief Functions*, p. 693-736.

Wei D., Deng X., Zhang X., Deng Y., Mahadevan S. (2013). Identifying influential nodes in weighted networks based on evidence theory. *Physica A: Statistical Mechanics and its Applications*, vol. 392, n° 10, p. 2564 - 2575. Retrieved from http://www.sciencedirect.com/science/article/pii/S0378437113001076

Weng J., Lim E.-P., Jiang J., He Q. (2010). TwitterRank: Finding Topic-sensitive Influential Twitterers. In *Proceedings of the third acm international conference on web search and data mining*, p. 261–270. New York, NY, USA, ACM.

Wu Z., Yin W., Cao J., Xu G., Cuzzocrea A. (2013). Community detection in multi-relational social networks. In *Web Information Systems Engineering – WISE 2013*, p. 43-56.