Évaluation comparative d'algorithmes de centralité pour la détection d'influenceurs

Kévin Deturck*
Damien Nouvel**
Frédérique Segond***

*Viseo Technologies 4, avenue Doyen Louis Weil 38000 Grenoble
INaLCO-ERTIM 2, rue de Lille 75007 Paris
kevin.deturck@viseo.com
**INaLCO-ERTIM 2, rue de Lille 75007 Paris
damien.nouvel@inalco.fr
***Viseo Technologies 4, avenue Doyen Louis Weil 38000 Grenoble
frederique.segond@viseo.com

L'influence, d'un point de vue social, peut être définie comme le pouvoir d'un individu qui mobilise des individus cibles pour des actions concrètes ou une opinion donnée. Détecter automatiquement les influenceurs dans les réseaux sociaux fournit des points d'entrée efficaces afin d'avoir un impact dans le cadre d'applications comme la diffusion ciblée d'une information de santé publique, la promotion d'un produit ou encore la réputation en ligne. Nos travaux s'intègrent au projet européen SOMA [1] qui a pour but d'enrichir les connaissances client de systèmes CRM [2] par de l'information issue de médias sociaux ; l'influence en fait partie afin d'évaluer l'impact potentiel d'un client par exemple sur un produit donné.

Il est possible d'analyser un réseau social selon sa structure ou le contenu diffusé, ces deux aspects étant les marqueurs de relations interpersonnelles essentielles à l'influence. Les approches qui analysent le contenu s'intéressent au texte en tenant compte de marqueurs discursifs, comme l'argumentation, qui tendent à influencer (Rosenthal et al., 2012). La structure du réseau est utilisée par les mesures de centralité, qui identifient les nœuds dominants d'un réseau d'après leurs liens. Elles sont beaucoup utilisées parce qu'elles ne requièrent principalement qu'un graphe d'interactions et sont suffisamment variées pour modéliser différentes modalités d'influence (Benyahia et Largeron, 2015). Cette variété requiert de pouvoir appréhender les meilleurs cas d'usage des différentes mesures en les évaluant et en les comparant notamment sur des données aux applications diverses (Ghazzali et Ouellet, 2017).

Nous voulons déterminer les mesures de centralité les plus aptes à identifier les influenceurs en les confrontant à des données réelles portant différentes catégories d'information sur un ensemble d'utilisateurs. Nous utilisons un corpus constitué dans le cadre de la compétition RepLab 2014 (Amigó et al., 2014) qui présente plus de 7000 comptes Twitter [3] manuellement annotés selon qu'ils sont influenceurs ou non. Nous avons sélectionné six algorithmes qui sont le *Degré entrant* comme Baseline, l'*Intermédiarité* qui mesure si un individu fait office de

1. http ://www.somaproject.eu/soma-project/
2. « Customer Relationship Management »
3. http ://www.twitter.com

passage obligé, la *Proximité* qui valorise l'accessibilité, *PageRank* et *LeaderRank* qui prennent en compte la valeur des liens, le dernier se voulant une adaptation du premier aux réseaux sociaux, enfin *Hits* distingue les autorités et les relais, les uns étant particulièrement suivis par les autres. Pour un échantillon de 50 comptes du corpus RepLab avec la même proportion (1/3) d'influenceurs que dans le corpus original, nous extrayons deux types d'interaction (Retweet et suivi) et utilisons deux modes d'attribution (lien, pondération des liens). Nous construisons ainsi des graphes aux sémantiques différentes (par exemple un graphe de suivi et un graphe de suivi pondéré par le nombre de Retweets) nous permettant d'analyser la performance des algorithmes à l'aune des catégories d'information extraites. Nous avons aussi pu comparer les algorithmes sélectionnés avec les systèmes de la compétition qui se différencient par leur besoin de supervision. Comme la compétition, nous utilisons MAP pour évaluer la qualité des classements d'utilisateurs en influence produits par les algorithmes par rapport à la référence binaire.

Nous observons des résultats meilleurs (moyenne de 5 %) sur les graphes de suivi que sur les graphes de Retweet, ce que nous expliquons par une audience particulièrement importante en quantité (nombre de liens) et qualité (valeur des liens) pour les influenceurs, donnant l'avantage aux algorithmes Page Rank, Leader Rank et Hits pour lesquels le score de chaque nœud est fonction du nombre de ses liens et de leur poids. Nous avons obtenu le meilleur résultat sur un graphe de suivi avec l'algorithme Hits à 52

Références

Amigó, E., J. Carrillo-De-albornoz, I. Chugur, A. Corujo, J. Gonzalo, E. Meij, M. De Rijke, et D. Spina (2014). Overview of RepLab 2014 : Author profiling and reputation dimensions for Online Reputation Management. *CEUR Workshop Proceedings 1180*, 1438–1457.

Benyahia, O. et C. Largeron (2015). Mesure d'influence via les indicateurs de centralité dans les réseaux sociaux. In *EGC*, pp. 469–470.

Ghazzali, N. et A. Ouellet (2017). *Comparative Study of Centrality Measures on Social Networks*. Springer International Publishing.

Rosenthal, S., J. Andreas, et O. Rambow (2012). Detecting Influencers in Written Online Conversations. *Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012)* (Lsm), 37–45.

Summary

We evaluate the effectiveness of centrality measures on Twitter data in detecting influencers. We apply these measures on graphs representing different user interactions from the RepLab 2014 corpus. We compare them between themselves, against RepLab systems and highlight influencer characteristics.