A Hybrid Approach for Detecting Influencers in Social Media

Ioannis Partalas*, Cédric Lopez*, Pierre-Alain Avouac*, Matthieu Osmuk*
Domoina Rabarijaona*, Dana Popovici*, Frédérique Segond*

*Viseo R&D
Grenoble, 38000, France
firstname.lastname@viseo.com,
http://www.viseo.com

Résumé. La détection d'influenceurs dans les réseaux sociaux s'appuie généralement sur une structure de graphe représentant les utilisateurs et leurs interactions. Récemment, cette tâche a tenu compte, en sus de la structure du graphe, du contenu textuel généré par les utilisateurs. Notre approche s'inscrit dans cette lignée : des informations sont extraites du contenu textuel par des règles linguistiques puis sont intégrées dans un système d'apprentissage automatique. Nous montrerons le prototype développé et son interface de visualisation qui facilite l'interprétation des résultats.

1 Introduction

An influencer is a person or thing that has the power to affect people, actions or events. Influencer's detection concerns the problem of determining which users have the most influence in a certain social network. Such information is crucial in many research studies such as in sociology and information management domains. Additionally, with the frenetic growth of available data in online social network, being able to analyze and detect influential users becomes crucial as they are susceptible to express their ideas more strongly than other individuals. For example, this information could be used in marketing campaigns in order to maximize their spread (Richardson et Domingos, 2002).

Formally, the task for detecting influent users in a social network, deals with a graph $G = (V, E)$ where V represents the users in the network and E the interactions among them. Apart from the structural information, we also assume that each user produces information as textual content. Such content can induce new interactions between users through new textual content. Therefore, we consider the task of detecting influencers following two ways : analyzing the structure of social networks as well as their textual content.

Our method combines rich linguistic information along with structural properties in order to feed a machine learning model for scoring users. The work presented is part of the SOMA Eurostars project[1] which concerns the enhancement of customer relationship management systems with social media analysis capabilities.

[1] http ://www.somaproject.eu/, SOMA Eurostars program 9292/12/19892

2 Background and Related Work

Usually, influence detection is addressed by analyzing the structure, mainly using graph theory where a plethora of measures exist. In this context, the centrality measures use the structural information in order to identify the most important nodes in a network (Bonacich, 1987). Indicative measures are betweenness centrality and PageRank. Another line of work, employs propagation models which try to specify how actions are propagated across the social network (Kempe et al., 2003). For example, these actions could be the retweets of a post in Twitter.

Finally, several methods try to combine content with structural information in social networks. Weng et al. (2010) alters PageRank in order to favor certain users according to a topic. More recently Katsimpras et al. (2015) proposed a supervised random walk approach towards topic-sensitive influential nodes. Recently, Biran et al. (2012) started to explore the characteristics of communication for influence detection, adopting a machine learning approach based on features such as persuasion, agreement/disagreement, dialog patterns, and sentiments. This approach required the manual annotation of weblogs from LiveJournal and discussion forums from Wikipedia. (Cossu et al., 2016) present an overview of the features that are used to characterize influential users in Twitter.

The originality of our approach is to use linguistic rules in order to extract fine-grained information in the discourse between users. Then, this information is used as attributes in a machine learning model.

3 Influence Detection Tool

The concrete use case for the influence detection toolkit proceeds as follows : 1) The user defines a social-media source (a forum, a social network, *etc.*) to be analyzed, 2) The system collects structured (*e.g.* behavioral information) and unstructured data (texts), 3) A score of influence is attributed to each social user.

The developed system is composed by three main parts regarding the different aspects : data consumption and preprocessing, data analysis and visualization.

Data Wrangling Regarding the collection of data to be analyzed the tool can be plugged with any social media (*e.g.* forums, blogs, Twitter, Facebook, *etc.*) with appropriate collectors and consume the data of interest. Collected data concern structural (*e.g.* gender, location), behavioral (*e.g.* number of tweets) and textual information. Textual part of the collected data s pre-processed by linguistic tools in order to extract the morphosyntactic structure.

Data Analysis As for identifying influencers, the main challenge is the development of a hybrid system, merging linguistic and non-linguistic descriptors. A large part of this task is usually ensured by taking into account non-linguistic information that has already demonstrated good results. The use of linguistic information for detection influencers relies on the assumption that an influencer has a specific behavior which translates into linguistic terms. This is a recent consideration from the linguistic researchers community, and obtained results are encouraging (Rosenthal, 2015).

As for the non-linguistic information, we use relevant statistical information well known in the literature such as the length of messages, the number of messages posted by an author, the number of followers, *etc*. Regarding the linguistic information, we focus on the analysis of various dimensions of the discourse : intonation, writing style, rhetoric, argumentation, speech acts, relation between text and context, *etc*. All the attributes extracted are used to feed a machine learning model which will allow for scoring users based on their influence. To this end, we use state-of-the-art Random Forests mainly for their ability to leverage non-linear interactions among features as well as for their interpretability.

Visualization The visualization module is Web-based which allows straight-forward accessibility. Specifically, one can segment the social users according to key-terms or topics of interest which will allow a fine-grained view on the set of influencers. Of course one can have global view of the detected influencers with different types of visualization. Figure 1 presents a screen of the visualization module where the top 20 users are presented in form of bubbles according to their score of influence. Figure 2 presents the interaction among users in a graph for a certain discussion. Users with higher score of influence are represented with bigger circles.

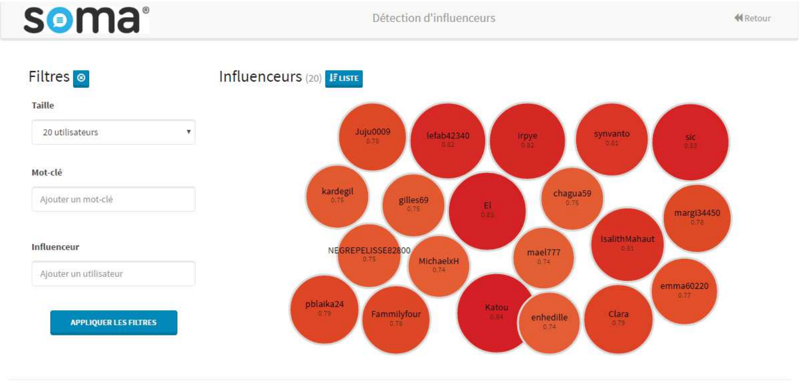

Taille: 20 utilisateurs

FIG. 1 – *Visualization of users using bubbles.*

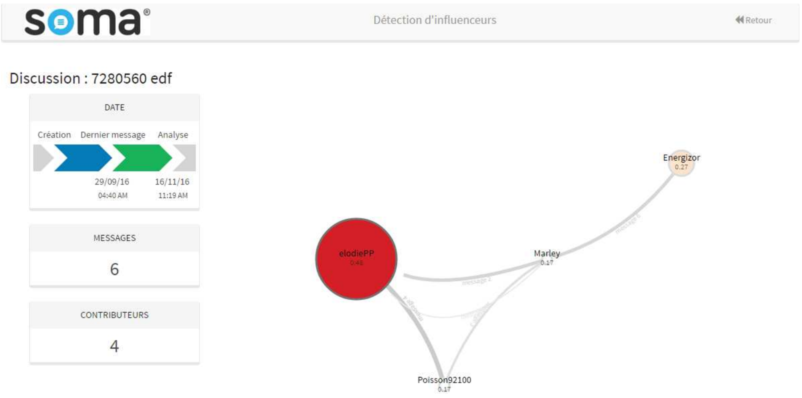
Acknowledgments

This work has been supported by the SOMA Eurostars program 9292/12/19892.

Références

Biran, O., S. Rosenthal, J. Andreas, K. McKeown, et O. Rambow (2012). Detecting influencers in written online conversations. In *Proceedings of the Second Workshop on Language in Social Media*, pp. 37–45. Association for Computational Linguistics.

Bonacich, P. (1987). Power and Centrality : A Family of Measures. *American Journal of Sociology 92*(5), 1170–1182.



Discussion : 7280560 edf

DATE

29/09/16 04:40 AM

16/11/16 11:19 AM

MESSAGES

6

CONTRIBUTEURS

4

FIG. 2 – *The interaction graph of users throughout a discussion.*

Cossu, J.-V., V. Labatut, et N. Dugué (2016). A review of features for the discrimination of twitter users : application to the prediction of offline influence. *Social Network Analysis and Mining 6*(1), 25.

Katsimpras, G., D. Vogiatzis, et G. Paliouras (2015). Determining influential users with supervised random walks. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, New York, NY, USA, pp. 787–792. ACM.

Kempe, D., J. Kleinberg, et E. Tardos (2003). Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pp. 137–146. ACM.

Richardson, M. et P. Domingos (2002). Mining knowledge-sharing sites for viral marketing. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 61–70. ACM.

Rosenthal, S. (2015). *Detecting Influencers in Social Media Discussions*. Ph. D. thesis, Columbia University.

Weng, J., E.-P. Lim, J. Jiang, et Q. He (2010). Twitterrank : Finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pp. 261–270. ACM.

Summary

Detecting influencers in social networks generally relies on a graph structure representing the users and their interactions. Recent approached take into account, in addition to the structure of the graph, the textual content generated by the users. Along this line, our approach uses information extracted from the textual content by linguistic rules and then integrated into a machine learning system. In this work we present the developed prototype along with the visualization used in order to facilitate the interpretation of the results.