Unsupervised Scalable Statistical Method for Identifying Influential Users in Online Social Networks

A. Azcorra[1,4], L.F. Chiroque[4], R. Cuevas[*,1,3] , A. Fernández[1,4], H. Laniado[2] ,R. Lillo[2,3], J. Romo[2,3], C. Sguera[2,3]

Dpto. Ingeniería Telemática UC3M [1], Dpto. Estadística UC3M [2], IBiDAT UC3M [3], IMDEA Networks [4]

[*] Corresponding Author e-mail: rcuevas@it.uc3m.es

This summary refers to the paper "*AZCORRA, Arturo, et al. Unsupervised scalable statistical method for identifying influential users in online social networks. Scientific reports, 2018, vol. 8, no 1, p. 6955.*"

I. Summary

Online Social [1]Networks (OSNs) such as Facebook, Twitter, or Google+ have rapidly become the most used online services, through which billions of users intensively interact every day. This makes OSNs an invaluable resource for sectors like advertising, marketing, or politics, which can use them for collecting information and launching campaigns. A challenging important problem is the identification of influential OSNs users, which can be leveraged by the abovementioned actors for, e.g., advertising a product, propagating a message, or improving the image of a company. The research community has devoted significant effort in characterizing influential OSNs users. However, most existing works define a priori the properties that identify influential users, and then use mechanisms based on that definition to find them. These supervised techniques have two main drawbacks. First, they require considerable manual analysis of the problem and the data for the definition of properties. Second, their effectiveness is fully tied to the definition: if such definition is inaccurate or unsuitable in a given context, the results would be likewise inaccurate or unsuitable.

In this paper, we present a new unsupervised method, that we call Massive Unsupervised Outlier Detection (MOUD), for supporting the identification of influential users in OSNs. MOUD is based on outlier detection in the area of Functional Data Analysis (FDA), and it scales to its application in OSNs with millions of users. MUOD considers the characteristics of a user in the form of signal. Each point in the x-axis is one of the user's characteristics and the correspondent value for each characteristic is represented in the y-axis. MUOD identifies three types of outliers: (a) Magnitude outliers, whose associated signals present a magnitude significantly different from the mass of users; (b) Amplitude outliers, whose associated signals present an amplitude significantly different from the mass of users; (c) Shape outliers, whose associated signals present a shape significantly different from the mass of users. Finally, by considering the intersection of these three sets of outliers, MUOD provides a total of 7 differentiated outlier classes.

Our trials with real data sets prove that MOUD is as effective in the identification of outliers as the best state-of-the-art FDA methods, while its much higher computational efficiency allows to apply it to much larger scale problems, including the large data scale of current OSNs. We have applied MOUD to a dataset including a complete snapshot of the social graph of Google+ (400 million nodes) as well as the overall public activity of this OSN in its two first years of operation. In particular, our goal is to test the ability of MOUD as a support algorithm in the identification of influential users without pre-defining a target profile. The obtained results confirm the applicability of our methodology in practice; since it is capable of finding separate sets of outliers that include different types of influential users based on their capacity to generate engagement, attract followers or their infection capabilities, in the considered large-scale OSN. Hence our proposed method offers unsupervised support to identifying influential users in OSN in those cases where there is not a predefined type of outlier. In turn, MOUD opens alternative paths for the exploration of interesting entities in other online systems, like Social Media or Online Advertising. Additionally, MOUD could also be applied to the identification of relevant nodes in big data problems from other disciplines (neuroscience, immune interactions, …).

Acknowledgments

This work has been partially supported by the EU throughout the SMOOTH project (786741) and the Spanish Government through the 5Gcity project (TEC2016-76795-C6-3-R).