Evidential Independence Maximization on Twitter Network

Siwar Jendoubi, Mouna Chebbah, Arnaud Martin

Evidential Independence Maximization on Twitter Network

Siwar Jendoubi[1,2], Mouna Chebbah[3], and Arnaud Martin[4]

[1] LISTIC, University Savoie Mont Blanc, 74944 Annecy-le-Vieux, France
[2] LARODEC, University of Tunis, ISG Tunis, 2000 Le Bardo, Tunisie
siwar.jendoubi@univ-smb.fr
[3] LARODEC, Univ. Manouba, ESEN, Tunisie
mouna.chebbah@esen.tn
[4] Univ Rennes 1, CNRS, IRISA, Lannion, France
Arnaud.Martin@univ-rennes1.fr

Abstract Detecting independent users in online social networks is an interesting research issue. In fact, independent users cannot generally be influenced, they are independent in their choices and decisions. Independent users may attract other users and make them adopt their point of view. A user is qualified as independent when his/her point of view does not depend on others ideas. Thus, the behavior of such a user is independent from other behaviors. Detecting independent users is interesting because a part of them can be influencers. Independent users that are not influencers can be directly targeted as they cannot be influenced. In this paper, we present an evidential independence maximization approach for Twitter users. The proposed approach is based on three metrics reflecting users behaviors. We propose an useful approach for detecting influencers. Indeed, we consider the independence as a characteristic of influencers even if not all independent users are influencers. The proposed approach is experimented on real data crawled from Twitter.

Keywords Independence measure, Independence maximization, Theory of belief functions, Twitter social network, Influence.

1 Introduction

Nowadays, most of web users are connected over *online social networks* (*OSN*) like Facebook, Twitter, LinkedIn, *etc.* OSN Users are different and may have distinguishable characteristics. Some of them are active and others are passive. Some of them are dependent on others, thus their choices, points of views and ideas depend on others. Other users are independent and impose their own choices and points of view. These users are independent from others and may be influencing them. Therefore, in this paper, we assume that the independence is a characteristic of influence users in the network. However, we cannot consider all independent users as influencers.

Independent users are more active and they attract others with their activities on OSN. Detecting these users is an interesting task for many companies to

promote their business over OSNs. A part of independent users is influencers. Independent users that are not influencers can be directly targeted as they cannot be influenced. OSN provided a wide spread platform to promote new products and services by several companies. To summarize, companies propagate their new products throw influencers and may also target independent users who are not targeted otherwise.

Previous researches were already interested in measuring the independence of users in OSN. However, the independence was never studied from the influence point of view. Kudelka *et al.* [4] proposed to quantify the dependence between vertices of an OSN considered as a network in the aim of community detection. Chehibi *et al.* [1] proposed a dependence measure for Twitter. Their proposed approach is detailed in this paper from an independence point of view. Indeed, our independence maximization approach uses their independence measure.

Twitter limits the access to its data, thus we cannot obtain all information about all users. Therefore, we propose an approximate estimation using the *theory of belief functions* [2, 6]. It models uncertainty, imprecision, incompleteness, total and partial ignorance. Besides the theory of belief functions provides a mathematical framework for combination [7, 2]. Recently, the theory of belief functions was used to estimate the influence on Twitter. In fact, Jendoubi *et al.* [3] introduce an evidential influence measure for Twitter. Their measure fuses three Twitter metrics to quantify the user's influence: *followers*, *mentions*, *retweets*.

In this paper, we study the independence in OSN from the influence point of view. Then, we consider the influence of independent users in ONS. In fact, the notions of independence and influence were never studied together in the literature. Then, we propose an evidential independence maximization model for Twitter users. The aim is to detect the most independent users that may be influencers. Indeed, we consider the independence useful to characterize influence users. In addition, this hypothesis validates the independence measure proposed in [1] with regards to the influence maximization. Furthermore, we study the independence of Twitter users through a set of experiments.

The sequel of the paper is organized as follows: We detail the approach of estimating users independence in section 2. Then, we detail the independence maximization model in section 3. Finally, before concluding in section 5, we detail experimental results on real data collected from Twitter in section 4.

2 Independence on Twitter

Twitter is an OSN that allows its users to connect to each others through an explicit relation, *i.e. follow* and/or through many implicit relations, *i.e.* a *retweet*, a *mention* or a *citation*.

In this paper, we propose to study the users behavior throw the implicit relations. Thus, a retweet is an information tweeted by a user from the tweets of another user connected with him. The number of retweets reflects the amount that a user adopts opinions of others. A mention is a message directly sent to

another specific user to communicate with him. Finally, a citation is the fact that a user cites other users in their tweets.

Thus, retweets, mentions and citations reflect amounts of adoption of others ideas by a specific user. In this paper, we propose to estimate degrees of independence between Twitter users. A user of Twitter u is independent from another user v when information provided by u are not affected by the information produced by v. When a user u is independent from v, the number of times that u retweets, mentions and cites v is quite small.

Therefore, we propose to estimate the independence degrees of Twitter users based on their numbers of follows, retweets, mentions and citations.

A user u of Twitter is independent from another user v if u is following v and u does not frequently retweet tweets of v or/and, u does not frequently mention v in his tweets.

Figure 1 summarizes the approach of user's independence estimation on Twitter proposed in [1]. The approach is in three steps:

Step 1 Weights estimation: w define a weight for each implicit relation: retweet, mention and citation. Thus, we define 3 weights (w_r, w_m, w_c), such that w_r is the weight of retweets, w_m is the weight of mentions and w_c is the weight of citations.

Step 2 Mass functions estimation: a mass function is estimated from each weight. Each mass function reflects the degree of belief on the users independence from the 3 (incomplete) collected information. We define 3 mass functions (m_r, m_m, m_c), such that m_r, m_m and m_c reflect the degree of belief on the users independence knowing the weight of retweets, mentions and citations.

Step 3 Independence degree estimation: mass functions m_r, m_m and m_c are combined in order to deduce independence degrees by considering the 3 aspects of retweets, mentions and citations.

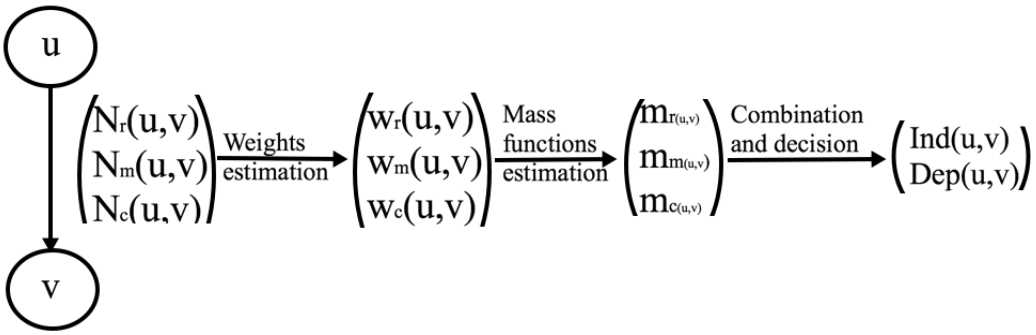

Fig. 1. Independence estimation

2.1 Step 1: Weights estimation

Let $G = (V, E)$ be an OSN such that V is the set of nodes, E is the set of links, $u \in V$ is a follower of $v \in V$ on Twitter. A user u following a user v can retweet,

mention or/and cite v. The number of retweets, mentions or/and citations may indicate the independence or the dependence of u on v. Thus, a vector of weights (w_r, w_m, w_c) is assigned to each link (u, v) as shown in figure 1.

The weights w_r, w_m and w_c of the link $(u, v) \in E$ are computed as follows:

$$w_i(u, v) = \frac{N_i(u, v)}{NT_i(u)} \tag{1}$$

such that $i = \{r, m, c\}$. Thus:

1. The retweet weight $w_r(u, v)$, is the number of times that u has retweeted v's tweets ($N_r(u, v)$) proportioned by the total number of u's retweet ($NT_r(u)$).
2. The mention weight $w_m(u, v)$, is the number of times that u has mentioned v ($N_m(u, v)$) proportioned by the total number of mentions of u ($NT_m(u)$).
3. The citation weight $w_c(u, v)$, is the number of times that u has cited v ($N_c(u, v)$) proportioned by the total number of u's citations ($NT_c(u)$).

2.2 Step 2: Mass functions estimation

Weights computed in step 1 may induce to some degree of belief on the users independence. Thus, a mass function is built from each weight. Let $\mathcal{I} = \{D, I\}$ be the frame of discernment of the independence where D is the hypothesis that users are dependent and I is the hypothesis that users are independent. Mass functions are estimated as follows:

$$\begin{cases} m^{\mathcal{I}}_{i_{(u,v)}}(\{D\}) = \alpha_{i_u} \times w_i(u, v) \\ m^{\mathcal{I}}_{i_{(u,v)}}(\{I\}) = \alpha_{i_u} \times (1 - w_i(u, v)) \\ m^{\mathcal{I}}_{i_{(u,v)}}(\{D, I\}) = 1 - \alpha_{i_u} \end{cases} \tag{2}$$

such that $i = \{r, m, c\}$. Thus:

1. The mass function $m^{\mathcal{I}}_{r_{(u,v)}}$ is deduced from the retweet weight $w_r(u, v)$. Note that $\alpha_{r_u} = \dfrac{NT_r(u)}{T_u}$ is a discounting coefficient that takes into account the total number of tweets T_u. The estimation of the mass function $m^{\mathcal{I}}_{r_{(u,v)}}$ is more reliable when the number of retweets is big enough in comparison with the total number of tweets.
2. The mass function $m^{\mathcal{I}}_{m_{(u,v)}}$ is deduced from the mention weight $w_m(u, v)$ and $\alpha_{m_u} = \dfrac{NT_m(u)}{T_u}$ is a discounting coefficient that takes into account the total number of tweets quoted by u with respect to the total number of tweets of u.
3. The mass function $m^{\mathcal{I}}_{c_{(u,v)}}$ is deduced from the citation weight $m_c(u, v)$ and where $\alpha_{c_u} = \dfrac{NT_c(u)}{T_u}$ is a discounting coefficient that takes into account the total number of tweets of u mentioning v with respect to the total number of tweets of u.

2.3 Step 3: Independence degree estimation

Mass functions $m^{\mathcal{I}}_{r_{(u,v)}}$, $m^{\mathcal{I}}_{m_{(u,v)}}(D)$ and $m^{\mathcal{I}}_{c_{(u,v)}}$ are combined with Dempster's rule of combination as follows:

$$m^{\mathcal{I}}_{(u,v)} = m^{\mathcal{I}}_{r_{(u,v)}} \oplus m^{\mathcal{I}}_{m_{(u,v)}} \oplus m^{\mathcal{I}}_{c_{(u,v)}} \tag{3}$$

Finally, degrees of independence $Ind(u,v)$ and dependence $Dep(u,v)$ corresponds to pignistic probabilities computed from the combined mass function $m^{\mathcal{I}}_{(u,v)}$ such that:

$$\begin{cases} Dep(u,v) = BetP(D) \\ Ind(u,v) = BetP(I) \end{cases} \tag{4}$$

The independence degree $Ind(u,v)$ is non-negative, it is either positive or null. It lies in the interval $[0,1]$. When $Ind(u,v) = 1$, u is totally independent from v; $Ind(u,v) = 0$ implies that u is totally dependent of v. Decision is made according to the maximum of pignistic probabilities. If $Dep(u,v) \geq Ind(u,v)$ then u is dependent on v, in the opposite case, if $Ind(u,v) > Dep(u,v)$, u is independent from v.

3 Independence Maximization

The independence measure can be considered as an influence measure. In fact, social influencers are characterized by their independence from the other users. Then, we propose to validate the proposed independence measure by using it to detect influencers, we call this task independence maximization. The maximization of the user's independence in this paper is similar to the problem of influence maximization presented in [3]. In fact, we can maximize the independence through a maximization model that was defined for the influence, we just need to replace the influence measure with an independence measure that has the same mathematical properties which are the monotonicity and the submodularity.

To maximize the independence in the network, we define the amount of independence of a set of nodes, S, on the network. It is the total independence given to S from all users in the network. First, we estimate the independence of S to a user v as follows:

$$Ind\,(S,v) = \begin{cases} 1 & \text{if } v \in S \\ \displaystyle\sum_{u \in S} \sum_{x \in IN(v) \cup v} Ind\,(u,x) \times Ind\,(x,v) & \text{otherwise} \end{cases} \tag{5}$$

where $Ind\,(v,v) = 1$ and $IN\,(v)$ is the set of in-neighbors of v, *i.e.* the set of nodes linked to v through a directed link having v as destination. Next, we define the independence spread function that estimates the amount of independence of S on the network as follows:

$$\sigma(S) = \sum_{v \in V} Ind(S, v) \tag{6}$$

We are looking for S on the network that maximizes $\sigma(S)$, *i.e.* $\underset{S}{\operatorname{argmax}}\, \sigma(S)$.

The independence maximization is an NP-Hard problem. Besides, the function $\sigma(S)$, is monotone and sub-modular. Then, a greedy-based solution can provide a good approximation of the optimal independence users set S. In this case, the cost effective lazy-forward algorithm (CELF) [5] is adapted to maximize the independence. Furthermore, it is a two pass maximization algorithm that is about 700 times faster than the greedy algorithm.

4 Experiments

In our experiments, we crawled the Twitter network using the streaming API on 20/01/2018. We obtained 54960 users, 686542 implicit relations between them (retweet, mention and citation) and 352420 tweets. Next, we used the independent maximization model introduced in the previous section to detect influencers in the collected network. We fixed the number of the detected nodes (size of S) to 100.

We study the independent maximization model according to for criteria of the detected nodes which are the number of accumulated mentions #Mention, the number of accumulated retweets #Retweet, the number of accumulated tweets #Tweet and the number of accumulated citations #Citation. Indeed, these criteria are considered as quality indicators of detected nodes. Then, higher their values are, better the quality of detected nodes is.

Figure 2 presents the obtained results according to the four fixed criteria, *i.e.* #Mention, #Retweet, #Tweet and #Citation receptively. According to Figure 2, the detected users (horizontal axis) have a good quality especially in terms of #Mention, #Retweet and #Tweet. In fact, the detected users have more than 4500 accumulated mentions, about 1200 accumulated retweets and more than 1800 accumulated tweets. These observations mean that the detected users are active in the network in terms of tweets. Also, their content is frequently propagated (retweeted). Besides, they are frequently mentioned in others tweets. Whereas, we notice a less important number of citations of the detected users. In fact, we have 18 accumulated citation which is relatively small compared to #Mention, #Retweet and #Tweet. We think that this is a result of weakness of the proportion of the citations in the data.

These observations confirm the assumption introduced in the previous section, then we can deduce that influencers are characterized by their independence from the other users in the network. In fact, the detected users using the proposed independence maximization model have a good quality according to the chosen criteria which confirms that they are influencers in the network.

In this paper, the main purpose is to validate the independence measure through detecting influencers. In fact, the independence is one important characteristic of influencers. The experiments presented in this section confirm this

fact. Indeed, the detected users have a good quality according to the chosen criteria. However, the independence itself is not sufficient as an influence measure and we can obtain better results by fusing it with other influence behaviors in the network like the user's position for example.

5 Conclusion

In this paper, we study the independence of Twitter users proposed in [1] from the influence point of view. Furthermore, we propose an independence maximization model that can be useful to detect influencers. In fact, a common property of social influencers is their independence from the other users in the network. Then, we use an independence measure to estimate the user's influence and to detect a set of influencers that maximizes the global independence in the network. Next, we experiment the proposed solution on real world data collected from Twitter and we study the quality of selected users according to their #Mention, #Retweet, #Tweet and #Citation.

In future works, we will study in a more refined way the notions of influence, dependence and independence to compare them. Besides, we will search to define an influence measure that fuses the user's independence with other influence behaviors like the user's activities and position.

References

1. Chehibi, M., Chebbah, M., Martin, A.: "independence of sources in social networks". In: Proceedings of the 17th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), Cádiz, Spain, June 11th – 15th. Springer. pp. 291–303 (2018)
2. Dempster, A.P.: Upper and lower probabilities induced by a multiple valued mapping. The Annals of Mathematical Statistics (1967)
3. Jendoubi, S., Martin, A., Liétard, L., Ben Hadji, H., Ben Yaghlane, B.: Two Evidential Data Based Models for Influence Maximization in Twitter. Knowledge-Based Systems (2017)
4. Kudelka, M., Drázdilová, P., Ochodkova, E., Slaninová, K., Horak, Z.: "local community detection and visualization: Experiment based on student data". In: Proceedings of the Third International Conference on Intelligent Human Computer Interaction (IHCI 2011), Prague, Czech Republic, August, 2011. Springer. pp. 291–303 (2011)
5. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: Proceedings of KDD'07. pp. 420–429 (August 2007)
6. Shafer, G.: A mathematical theory of evidence. Princeton University Press (1976)
7. Smets, P.: The combination of evidence in the transferable belief model. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(5), 447–458 (1990)

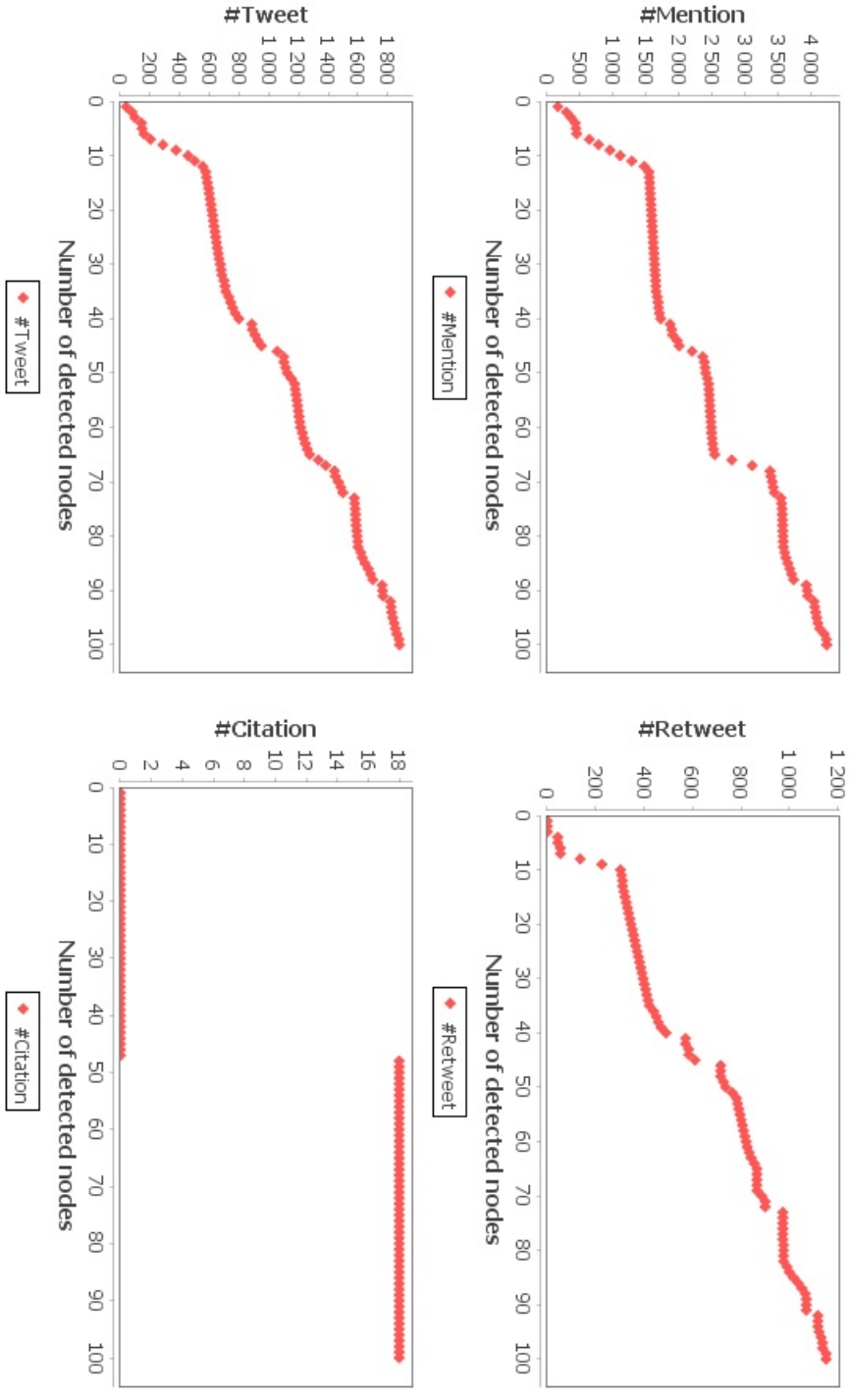

Fig. 2. Detected users using the proposed independence measure according to #Mention, #Retweet, #Tweet and #Citation