Identify Influential Spreaders in Online Social Networks Based on Social Meta Path and PageRank

Vang V. Le[1], Hien T. Nguyen[1](✉), Vaclav Snasel[2], and Tran Trong Dao[3]

[1] Faculty of Information Technology, Ton Duc Thang University,
Ho Chi Minh City, Vietnam
{levanvang,hien}@tdt.edu.vn

[2] Department of Computer Science, VSB-Technical University of Ostrava,
Ostrava, Czech Republic
vaclav.snasel@vsb.cz

[3] Division of MERLIN, Ton Duc Thang University, Ho Chi Minh City, Vietnam
trantrongdao@tdt.edu.vn

Abstract. Identifying "influential spreader" is finding a subset of individuals in the social network, so that when information injected into this subset, it is spread most broadly to the rest of the network individuals. The determination of the information influence degree of individual plays an important role in online social networking. Once there is a list of individuals who have high influence, the marketers can access these individuals and seek them to impress, bribe or somehow make them spread up the good information for their business as well as their product in marketing campaign. In this paper, according to the idea "Information can be spread between two unconnected users in the network as long as they both check-in at the same location", we proposed an algorithm called SMPRank (Social Meta Path Rank) to identify individuals with the largest influence in complex online social networks. The experimental results show that SMPRank performs better than Weighted LeaderRank because of the ability to determinate more influential spreaders.

Keywords: Influential spreader · LeaderRank · Random walk · PageRank · Social meta path

1 Introduction

Today, the online social network such as Facebook, Twitter becoming a popular channel for transmission of information such as news, brochures, and marketing, ... The booming in the number of OSN users poses a major challenge is how information can be spread to the users in the most effective and optimal way with a fixed cost. One way to do is to find users who have the greatest degree of spread (influential spreaders) and inject information into these people to get the benefit, information from them will be widely spread in online social networks and lead to the most effective marketing result.

H.T. Nguyen and V. Snasel (Eds.): CSoNet 2016, LNCS 9795, pp. 51–61, 2016.
DOI: 10.1007/978-3-319-42345-6_5

Given a network G(V, E) - V is the user set and E is the edge set of G which represents the connections between users in G. X is a subset of V and a function $influence(X)$ is the influence function which maps the seed user set X to the number of users influenced by users in X. Identifying influential spreaders aims at selecting the optimal subset $X^\star$ which contains n seed users to maximize the propagation of information across the networks.

$$\begin{aligned} & X^\star = argmax_{X \subseteq V} influence\,(X) \\ & |X| = n \end{aligned} \tag{1}$$

How to determine efficiently the individuals who have the highest degree of influence in social networks is a major challenge up to the present [4–9,14]. Recently, Lu et al. [10] proposed an algorithm LeaderRank to identify influential spreaders in directed network which is a simple variant of PageRank. The authors said that the connection matrix between individuals (adjacency matrix) in social networks is relatively sparse and they introduced the concept "ground node" (an additional node) and create virtual connections from the ground node to all existing nodes in social networks and set the weight of virtual edge a value of 1. This approach has limited success in shortening the convergence time when running the PageRank algorithm to determine the ranking of the node. However, it has one drawback is whether individuals who have more fans or less fans then receives the same weight value of 1 from the ground node and this slightly estate reasonable. Li et al. [1] proposed the Weighted LeaderRank algorithm, an improvement of standard LeaderRank by allowing nodes with more fans get more scores from the ground node. Weighted LeaderRank is a straightforward and efficient method, however, it is less relevant to real network in which the information diffusion depends not only on the network structure but also the network behavior. In fact, when applying the Weighted LeaderRank to actual dataset (Twitter), the obtained result is not the most influential spreaders.

In this paper, we further improve the Weighted LeaderRank algorithm by applying the definition of social meta path which introduced by Zhan et al. [3]. Our approach, which called SMPRank is the hybrid method of Weighted LeaderRank method and a part of social meta path. The experiments on the real social network (Twitter) show that the SMPRank can considerably improve the spreadability of the original Weighted LeaderRank. Our approach is based on the idea:

(1) Typically, information can only spread from a user to another user if and only if they are connected to each other (friends or following). However, our approach assumes that even if there is no direct connection to each other, the information is still able to exchange if they both check-in at the same location (by talking directly).

(2) Even between connected users, the information may be spread stronger between users who often communicate to each other and weaker between users who rarely communicate to each other. For instance, A and B are two followers of C, usually each 10 tweets C writes then A retweets 5 and B retweets 3 mean that information may be spread from C to A stronger than from C to B.

The main contribution of our research is the improvement of the accuracy, our influential spreaders obtained from SMPRank is closer to observed dataset than the result obtained from Weighted LeaderRank. The remaining parts of this paper are organized as follows. We summary the related work in Sect. 2. In Sects. 3, we introduce the proposed SMPRank method. Experiments are given in Sect. 4. Finally, we conclude the paper in Sect. 5.

2 Related Work

Identifying the most influential spreaders in a network is critical for ensuring efficient diffusion of information. For instance, a social media campaign can be optimized by targeting influential individuals who can trigger large cascades of further adoptions. This section presents briefly some related works that illustrate the various possible ways to measure the influence of individuals in the online social network.

Cataldi et al. [12] propose to use the well known PageRank algorithm [11,13] to calculate the influence of individuals throughout the network. The PageRank value of a given node is proportional to the probability of visiting that node in a random walk of the social network, where the set of states of the random walk is the set of nodes. It directly applies the standard random walk process to determine the score of every node. Accordingly, the score of each node in the network will be calculated step by step from t_0 to t_n. At the time t_i, the score of node u will be calculated based on the score of u and the score of u's neighbors in the previous step t_{i-1}. The random walk can be described by an iterative process as formulate (2). In that: $S_u(t_i)$ is the score of node u at the time t_i, $w_{v,u}$ is the weight of connection from v to u, it has value of 1 if existing a connection from v to u and opposite it has value of 0.

$$S_u\left(t_i\right) = \sum_{v \in Neighbor(u)} \frac{w_{u,v}}{outdeg(v)} * S_u(t_{i-1}) \tag{2}$$

Recently, Lu et al. [10] proposed an algorithm LeaderRank to identify influential spreaders in directed network which is a simple variant of the algorithm PageRank [2]. To reduce the convergence time of PageRank, it adds an additional node called ground node, by creating many virtual connections from real nodes to ground node it improves the sparseness of original connection matrix. The Fig. 1 demonstrates the LeaderRank by set the value of 1 to all virtual connections from real nodes to ground node and vice versa.

Li et al. [1] proposed the Weighted LeaderRank algorithm, an improvement of standard LeaderRank by allowing nodes with more fans get more scores from the ground node. Instead of setting the value of 1 to all virtual connections, Weighted LeaderRank sets the difference values to difference virtual connections. The virtual connections from ground node to high in-degree real node will get higher weight value compare with the virtual connections from ground node to low in-degree real node. For example, in the Fig. 2, the connection from ground

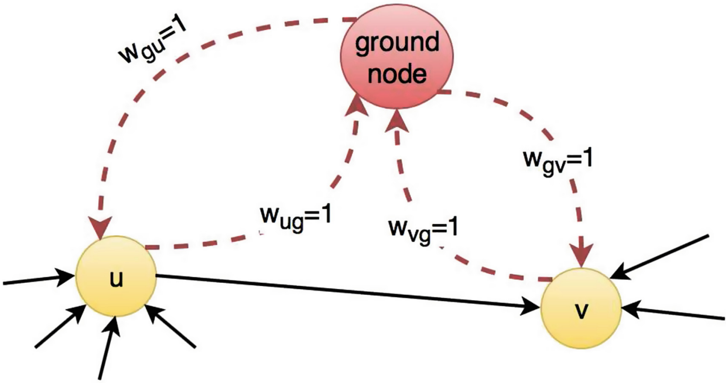

Fig. 1. An example of LeaderRank algorithm

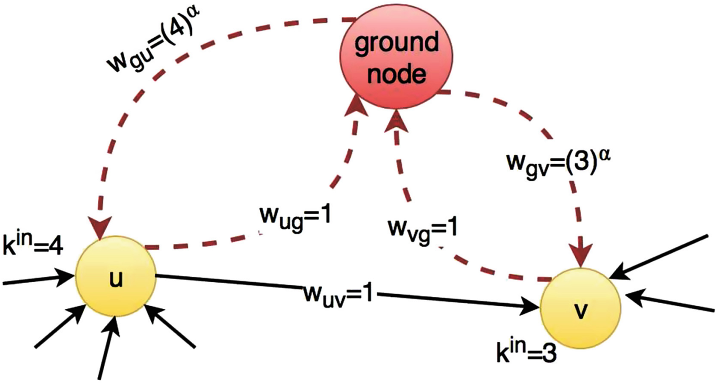

Fig. 2. An example of Weighted LeaderRank algorithm

node g to node u will get higher weight value than the connection from ground node g to node v because u's in-degree is higher than v's in-degree.

The methods we have just described above exist a drawback that they only exploit the structure (topology) of the network, and ignore other important properties, such as nodes' features and the way they interact with other nodes in the network.

Zhan et al. [3] proposed a new model M&M to resolve the Aligned Heterogeneous network Influence maximization (AHI) problem. The explosion of online social networks lead to a person can participate and have multiple accounts on different online social networks. Information can be spread not only on internal network but also it can be exchanged together between difference networks. If a user A participate onto two online social networks X and Y simultaneously, the information A received on the network X can be forwarded to the network Y this means that information can be spread through difference channels: internal and external channel. Through this idea, the author proposed a definition of path, meta path and social meta path.

3 Proposed Model

Typically, information can only spread from user A to user B if and only if A and B are connected to each other (friends or following). However, in our approach we assume that even if there is no direct connection with each other, the information is still able to spreading from A to B (i.e., A and B check-in at the same location on the same event, A is the host of the event and B is the client that attends the event - information will spread from A to B). The Fig. 3 demonstrates the idea of our algorithm, the actual network doesn't have a direct connection from node v to node u but it may exist a hidden connection from v to u (represented by dotted line) through another channel such as v and u check-in the same location on the same event.

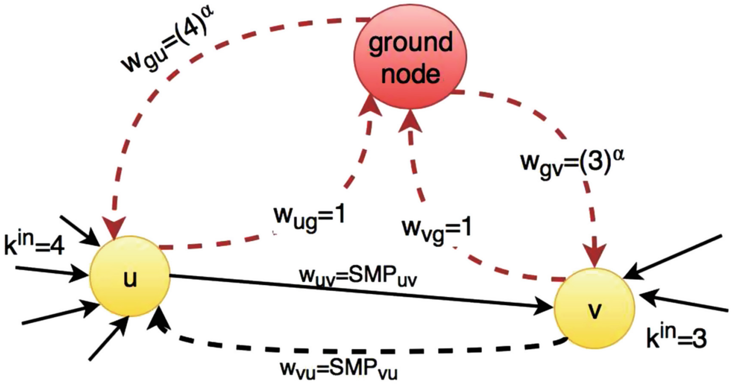

Fig. 3. An example of SMPRank algorithm

In this paper, we will follow the definitions of concepts Social Meta Path proposed in [3]. The Fig. 4 illustrates the schema of Twitter network which we chose to do the experiment. Depend on the network schema, we select 3 social meta paths as below:

(1) Follow
MP^1: $User \xrightarrow{follow} User$

(2) Co-location check-in
MP^2: $User \xrightarrow{write} Tweet \xrightarrow{checkin} Location \xrightarrow{checkin^{-1}} Tweet \xrightarrow{write^{-1}} User$

(3) Re-tweet
MP^3: $User \xrightarrow{write} Tweet \xrightarrow{retweet} Tweet \xrightarrow{write^{-1}} User$

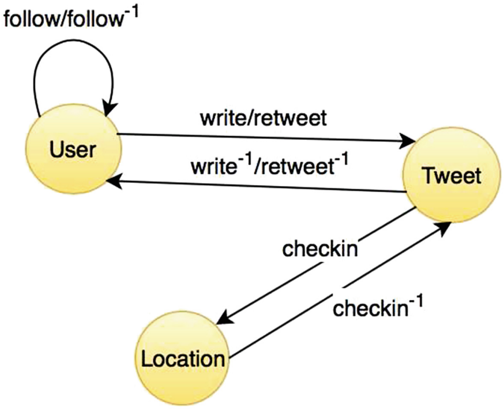

Fig. 4. The network schema of Twitter

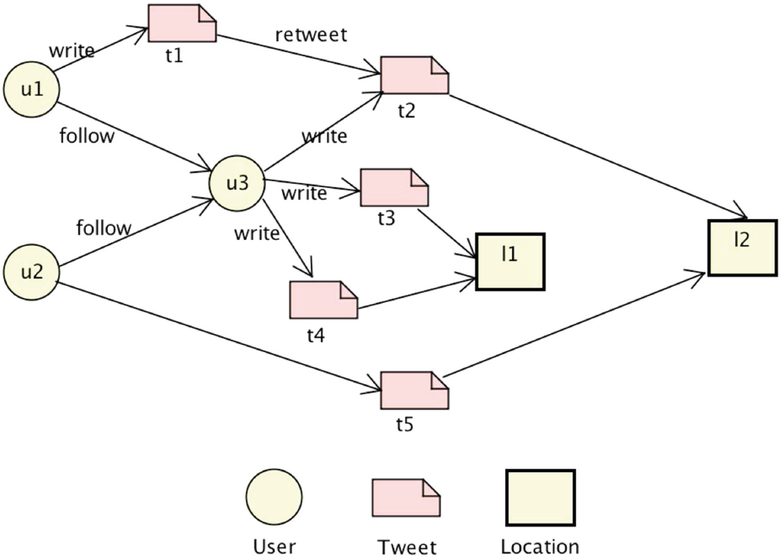

Fig. 5. An example of Twitter network with User, Tweet, and Location

Based on the social media path information, we calculated the value of $\theta^i_{u,v}$ based on Formula (3). In which u, v are vertices of the network, i is in [1, 3] represents the three types of social meta paths selected above. The values of $\theta^i_{u,v}$ represent the power of information transmission from vertex u to vertex v through the i^{th} social meta path channel.

Applying the formula (3) to the example in the Fig. 5 we get the values as shown in Table 1.

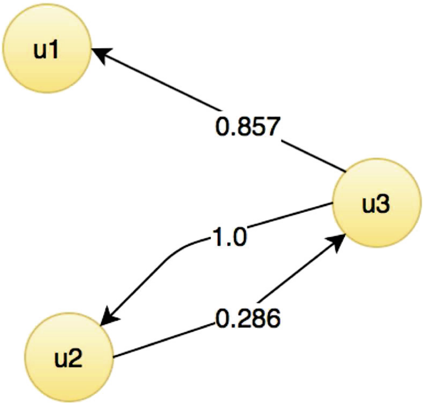

Fig. 6. Re-draw the network base on Table 2

$$\theta^i_{u,v} = \frac{2 * |MP^i_{u,v}|}{|MP^i_{(u,)}| + |MP^i_{(,v)}|} \tag{3}$$

After obtaining the value which represent the power of information transmission from user u to user v in each channel (each social meta path) individually, we will calculate the aggregation weight $w(u, v)$ based on Formula (4). In which α^i is the ratio of each type of meta social path, the greater value of α^i then the information will likely spread greater through the i^{th} social meta path. The value of $w(u, v)$ represents the degree of information that can be transmitted from u to v (u, v is not necessarily to be a friend of each other).

Table 1. The value of $\theta^i_{u,v}$ for example in the Fig. 5

	u_1	u_2	u_3
u_1		$\theta^1 = 0,\ \theta^2 = 0,\ \theta^3 = 0$	$\theta^1 = 0,\ \theta^2 = 0,\ \theta^3 = 0$
u_2	$\theta^1 = 0,\ \theta^2 = 0,\ \theta^3 = 0$		$\theta^1 = 0,\ \theta^2 = 2,\ \theta^3 = 0$
u_3	$\theta^1 = 1,\ \theta^2 = 0,\ \theta^3 = 1$	$\theta^1 = 1,\ \theta^2 = 2,\ \theta^3 = 0$	

In the experimental process, our team selected the optimal value for α^1, α^2, α^3 respectively 5, 1, 1.

$$w(u, v) = \frac{\sum_{i=1}^{3} \alpha^i * \theta^i_{u,v}}{\sum \alpha^i} \tag{4}$$

Applying Formula (4) to the example in the Fig. 5 along with value of $\theta^i_{(u,v)}$ calculated in Table 1 we will calculate the value of $w(u, v)$ as shown in Table 2.

Based on the result in Table 2, we re-draw the network as shown in the Fig. 6. Next step, we apply the algorithm Weighted Rank Leader in the [1] and

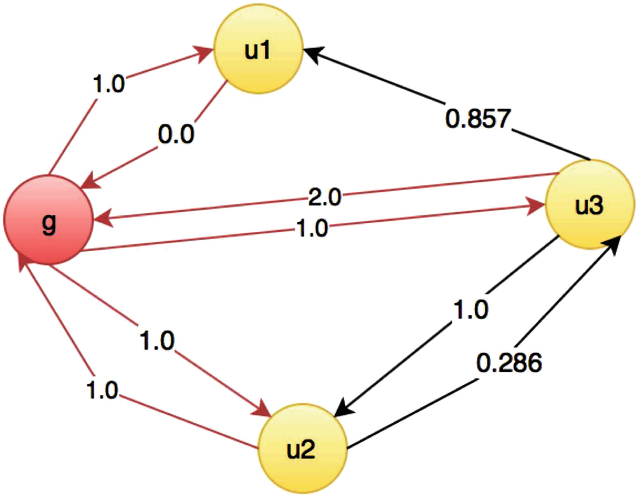

Fig. 7. Re-draw the network after add the ground node

Algorithm 1. Calculate Weight

```
function CalculateWeight($G, MP$)    ▷ Where G - Input network, MP - Meta-path values
    $k \leftarrow 3$
    $\alpha \leftarrow [5,\ 1,\ 1]$
    for $u \in$V do
        for $v \in$V do
            $w_{u,v} \leftarrow 0$
            for $i = 1$ to $k$ do
                $\theta^i_{u,v} \leftarrow \frac{2 * MP^i_{u,v}}{MP_{(u,)} + MP_{(,v)}}$
                $w_{u,v} \leftarrow w_{u,v} + \alpha^i * \theta^i_{u,v}$
            end for
        end for
    end for
end function
```

Table 2. The value of $w(u, v)$ for example in the Fig. 5

	u_1	u_2	u_3
u_1	0	0	0
u_2	0	0	$2/7 = 0.286$
u_3	$6/7 = 0.857$	$7/7 = 1$	0

Table 3. The value of $w(u, v)$ with ground node for example in the Fig. 5

	g	u_1	u_2	u_3
g	0	1	1	1
u_1	0	0	0	0
u_2	1	0	0	0.286
u_3	2	0.857	1	0

proceed adding a ground node (virtual node) along with the virtual edges which connecting the ground node to existing other nodes (real nodes) in the network. The weight of virtual connections (virtual edges) from real node (u) to ground node (g) and vice versa are calculated according to the principle (5)

$$\begin{aligned} w(u, g) &= 1 \\ w(u, g) &= k_u^{out}. \end{aligned} \tag{5}$$

Apply above principle to example in the Fig. 5 along with $w(u, v)$ in Table 2 we will calculate the final weight matrix as shown in Table 3 (Fig. 7).

Finally, after obtaining the weight $w(u, v)$ of all edges in the network (which has an additional ground node virtual node), we proceed to run the PageRank algorithm and obtain the ranking list which represents the ordering of user's influential degree in the network. The users who have higher ranking value will have greater impact in the network.

4 Experiments

To validate the effectiveness of our SMPRank algorithm, we run the experiments on real datasets of Twitter social network. We use and extend the dataset of Jure Leskovec which published on website: http://snap.stanford.edu/data/egonets-Twitter.html The original dataset contains only the users and the connections between users (following relationship). We extended the original dataset by collecting all the tweets of users (each users we collect the maximum 3,200 tweets). For each tweet collected in the previous step, we proceed to gather information such as: number of likes (favorite), number of user retweet, along with the number of followers of their retweet users (Table 4).

Table 4. The statistic of real Twitter dataset

Number of nodes	Number of edges	Number of tweets
76,120	55,458,375	2,941,374

We divide the collected dataset into two parts, the first part contains only tweets written before 30/12/2015 (for running the algorithm), the second part consists of tweets written after 30/12/2015 (for testing the effectiveness of the algorithm). Run the SMPRank and Weighted LeaderRank algorithm on first part dataset we have the output $Rank_{SMP}$ and $Rank_{WL}$

$$Influence(u) = \frac{\sum_{t \in tweets(u)} Infection(t)}{|tweets(u)|}. \tag{6}$$

We use Formula (6) to calculate the actual influence degree of each user in the network. In which, $Influence(u)$ is the influence rate of user u, $tweets(u)$ is the set of tweets written by the user u, $|tweets(u)|$ is the number of tweets written by the user u, $Infection(t)$ is calculated according to Formula (7).

$$Infection(t) = infect\_rate * |follower(u^t)| + \sum_{t_i \in tweets(t)} infect\_rate * |follower(u^{t_i})| \tag{7}$$

In Formula (7), t is a tweet, u^t is the user who write the tweet t, $Infection(t)$ is the number of users who saw the tweet t (seen times) which obtained from formula (7), $retweet(t)$ is the set of tweets that are retweeted from tweet t, $infect\_rate$ (in range [0, 1]) represents the rate of information diffusion. For instance, $infect\_rate = 0.5$ means that if a user has 10 followers then every tweet written by this user will have 5 followers see that tweet.

Applying Formula (6) to all users on the test data (part 2 of the dataset) we calculated the influence's values of all users, then the actual user's ranking ($Rank_{Actual}$) will be determined based on the strategy: users who have higher influence value will have higher ranking. We compare the SMPRank and Weighted LeaderRank by measuring the Pearson correlation coefficient of each pair ($Rank_{SMP}$, $Rank_{Actual}$) and ($Rank_{WL}$, $Rank_{Actual}$). The empirical data at Table 5 show that the results of SMPRank ranking better than Weighted LeaderRank because of higher correlation coefficient value.

Table 5. The Pearson correlation coefficient comparing between Weighted LeaderRank, SMPRank and the ground truth ranking.

	Actual ranking
Weighted LeaderRank	0.713
SMP rank	0.852

5 Conclusion

Weighted LeaderRank is an efficient method, however, it calculates user's ranking only based on the network structure and ignores the behavior of users (write tweets, retweet, check-in). In this paper, we further improve the Weighted LeaderRank algorithm by apply the definition of social meta path which introduced by Zhan et al. [3]. Typically, information can only spread from user A to user B if and only if A and B are connected to each other (friends or following). However, our approach assumes that even if there is no direct connection to each other, the information is still able to exchange if they both check-in at the same location (by talking directly). Our approach, which called SMPRank is the hybrid method of Weighted LeaderRank method and social meta path. Experiments on the real social network (Twitter) show that the SMPRank can considerably improve the degree of spreadability of the original Weighted LeaderRank.

References

1. Li, Q., Zhou, T., Lü, L., Chen, D.: Identifying influential spreaders by weighted LeaderRank. Phys. A Stat. Mech. Appl. **404**, 47–55 (2014)
2. Zhang, T., Liang, X.: A novel method of identifying influential nodes in complex networks based on random walks. J. Inf. Comput. Sci. **11**(18), 6735–6740 (2014)
3. Zhan, Q., Zhang, J., Wang, S., Yu, P.S., Xie, J.: Influence maximization across partially aligned heterogenous social networks. In: Cao, T., Lim, E.-P., Zhou, Z.-H., Ho, T.-B., Cheung, D., Motoda, H. (eds.) PAKDD 2015. LNCS, vol. 9077, pp. 58–69. Springer, Heidelberg (2015)
4. Zhou, T., Fu, Z.-Q., Wang, B.-H.: Epidemic dynamics on complex networks. Prog. Nat. Sci. **16**(5), 452–457 (2006)
5. Lü, L., Zhou, T.: Link prediction in complex networks: a survey. Phys. A Stat. Mech. Appl. **390**, 1150–1170 (2011)
6. Lu, L., Chen, D.-B., Zhou, T.: The small world yields the most effective information spreading. New J. Phys. **13**, 123005 (2011)
7. Doerr, B., Fouz, M., Friedrich, T.: Why rumors spread so quickly in social networks. Commun. ACM **55**, 70–75 (2012)
8. Aral, S., Walker, D.: Identifying influential and susceptible members of social networks. Science **337**, 337–341 (2012)
9. Silva, R., Viana, M., Costa, F.: Predicting epidemic outbreak from individual features of the spreaders. J. Stat. Mech. Theor. Exp. **2012**, P07005 (2012)
10. Lu, L., Zhang, Y.-C., Yeung, C.H., Zhou, T.: Leaders in social networks, the delicious case. PLoS One **6**, e21202 (2011)
11. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. **30**, 107–117 (1998)
12. Cataldi, M., Di Caro, L., Schifanella, C.: Emerging topic detection on Twitter based on temporal and social terms evaluation. In: Proceedings of the Tenth International Workshop on Multimedia Data Mining, MDMKDD 2010, pp. 4–13 (2010)
13. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: bringing order to the web. In: WWW 1998, pp. 161–172 (1998)
14. Kempe, D., Kleinberg, J., Tardos, E.: 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 2003, pp. 137–146. ACM, New York (2003)