A Centrality for Social Media Users Focusing on Information-Gathering Ability

Mamoru Yamakawa, Keishi Tajima

Kyoto University

Introduction

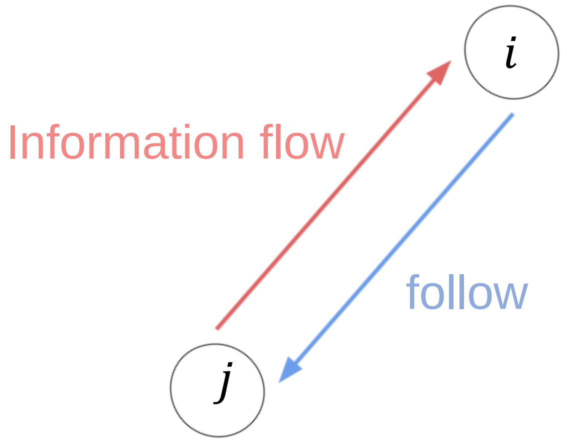
- Most existing SNS user scoring methods use the individual's social influence.
- However, users with a high ability to **collect useful information quickly** are also important in SNS.

➡ Can we devise a scoring method that uses information-gathering ability on Twitter?

Related work (1)

Katz Centrality [Katz 1953]

$$C_{\mathrm{Katz}}(i) = \sum_{k=1}^{\infty} \sum_{j} \alpha^k (A^k)_{ji}$$



- $0 < \alpha < 1/|\lambda_{\max}|$
- $\lambda_{\max}$ is the eigenvalue of A with the largest absolute value

$C_{\mathrm{Katz}}(i)$ gives the sum of the information flowing into i

Related work (1)

Katz Centrality [Katz 1953]

$$C_{\mathrm{Katz}}(i) = \sum_{k=1}^{\infty} \sum_{j} \alpha^k (A^k)_{ji}$$

- $0 < \alpha < 1/|\lambda_{\max}|$
- $\lambda_{\max}$ is the eigenvalue of A with the largest absolute value

Shortcomings:

- Weights of information transmitted by each node are equal.
- Parent node propagates information to all child nodes with probability 1.

Related work (2)

HITS algorithm [Kleinberg 1999]

$$\mathbf{a}^{(0)} = \mathbf{e}$$

$$\mathbf{h}^{(0)} = \mathbf{e}$$

$$\mathbf{a}^{(k)} = A^T \mathbf{h}^{(k-1)}$$

$$\mathbf{h}^{(k)} = A \mathbf{a}^{(k-1)}$$

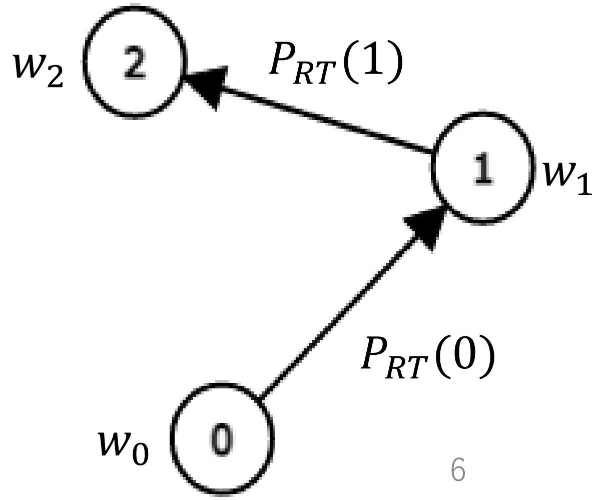
Shortcomings:

- Hub score is based only on the authority score of direct neighbors.
  → No consideration is given to information dissemination through retweets.

Proposed Method

- Node i has a weight w_i representing the quality of information
- Node i forwards Information from the parent node in $P_{RT}(i)$

- α: attenuation rate makes the quality of information smaller for those originating from more distant nodes.



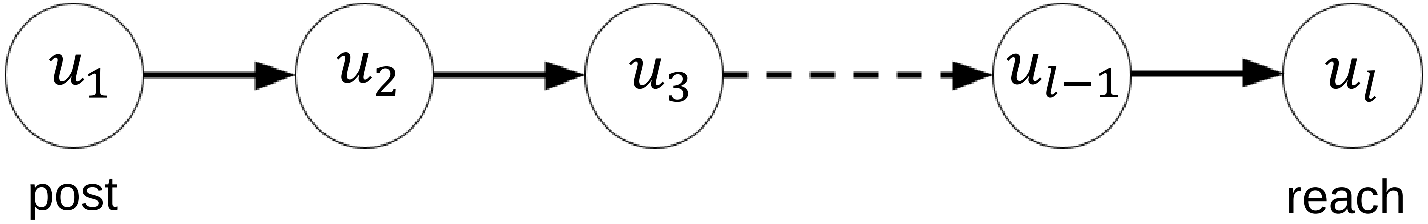
Proposed Method

$P_{\mathrm{RT}}(i)$: information forwarding probability of i

$P_{\mathrm{RT}}(p)$: information propagating probability through path p

- we calculate the probability that information posted by u_1 reaches u_l through a $(l-1)$-hop path $p = u_1, u_2, \ldots, u_l$, denoted by $P_{\mathrm{RT}}(p)$, by the formula below :

$$P_{\mathrm{RT}}(p) = P_{\mathrm{RT}}(u_2) \times P_{\mathrm{RT}}(u_3) \times \cdots \times P_{\mathrm{RT}}(u_{l-1})$$



Probability of Multi-hop Propagation

Adjacency matrix weighted by user retweet probability P_{RT} :

$$P = \begin{pmatrix} A_{11}P_{\mathrm{RT}}(1) & \cdots & A_{1n}P_{\mathrm{RT}}(1) \\ \vdots & \ddots & \vdots \\ A_{n1}P_{\mathrm{RT}}(n) & \cdots & A_{nn}P_{\mathrm{RT}}(n) \end{pmatrix}$$

- $P_{ij} = A_{ij} \cdot P_{RT}(i)$
- $\frac{(P^l)_{ij}}{P_{\mathrm{RT}}(i)} = \sum_p P_{\mathrm{RT}}(p)$
  - p is l-hop paths from i to j

Information-Gathering Ability Including Self-Originated Information

We first define the metric including self-originated information, denoted by $IGC_+(i)$ (Information-Gathering Centrality including self-originated information), as follows:

$$IGC_+(i) = \sum_{l=1}^{\infty} \sum_{j=1}^{n} \left(\alpha^{l-1} \frac{(P^l)_{ji}}{P_{RT}(j)} w_j \right)$$

IGC_+ gives sum of the information flowing into i through retweets

Information-Gathering Ability Including Self-Originated Information

We first define the metric including self-originated information, denoted by $IGC_+(i)$ (Information-Gathering Centrality including self-originated information), as follows:

$$IGC_+(i) = \sum_{l=1}^{\infty} \sum_{j=1}^{n} \left(\alpha^{l-1} \frac{(P^l)_{ji}}{P_{RT}(j)} w_j \right)$$

- There are many ways to define the value of w_i
  - PageRank
  - Topic-sensitive PageRank
  - etc.

Information-Gathering Ability Including Self-Originated Information

If α satisfies the condition, We can simplify the computation of $\overrightarrow{IGC_+} = (IGC_+(1), \dots, IGC_+(n))^{\mathrm{T}}$ using the inverse matrix:

$$\overrightarrow{IGC_+} = (E - \alpha P^T)^{-1} P^T \overrightarrow{w_p}$$

- $w_p(j) = w_j / P_{\mathrm{RT}}(j)$
- $\overrightarrow{w_p} = \left(w_p(1), \dots, w_p(n)\right)^T$

Elimination of Self-Originated Information

- $IGC_{+}(i)$ includes information originating i itself.
- $IGC_{\text{self}}(i)$: the amount of information originating i and recieved by i.
- $IGC(i)$: our proposed metric

$$IGC(i) = IGC_{+}(i) - IGC_{\text{self}}(i)$$

$$IGC_{\text{self}}(i) = \sum_{t=1}^{\infty} \alpha^{t-1} \frac{(P^t)_{ii}}{P_{RT}(i)} w_i$$

Elimination of Self-Originated Information

As in IGC_+, We can efficiently the compute $\overrightarrow{IGC}$ using inverse matrix and Hadamard product:

$$\overrightarrow{IGC} = \overrightarrow{IGC_+} - \frac{1}{\alpha}\left(E \otimes \left(\left(E - \alpha P^T\right)^{-1} - E\right)\right)\overrightarrow{w_p}$$

- $\otimes$ is Hadamard product
  - $A \otimes B = (a_{ij} b_{ij})$.

Metric for Users Retweeting Useful Information

IGC_{rt}: Another metric based on the amount information that

- i collects and
- Forward to its followerd

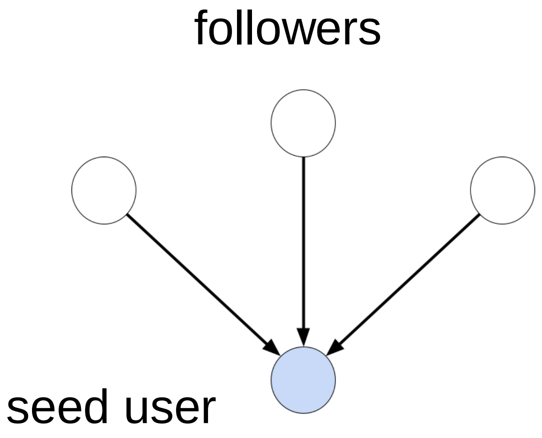
$$\overrightarrow{IGC_{\mathrm{rt}}} = \overrightarrow{IGC} \otimes \overrightarrow{P_{\mathrm{RT}}}$$

Experiment

To compare the proposed metric with several existing metrics, we conducted experiments on two datasets collected from Twitter.

- The node dataset consist of seed user, and their followers and followees.
- The edge of dataset is the existing follow-relation between node pairs in dataset.

	Dataset 1	Dataset 2
seed user	@univkyoto	@A_I_News
number of nodes	40,691	32,739
number of edges	509,978	456,483
average P_{RT}	2.58e-06	9.43e-07

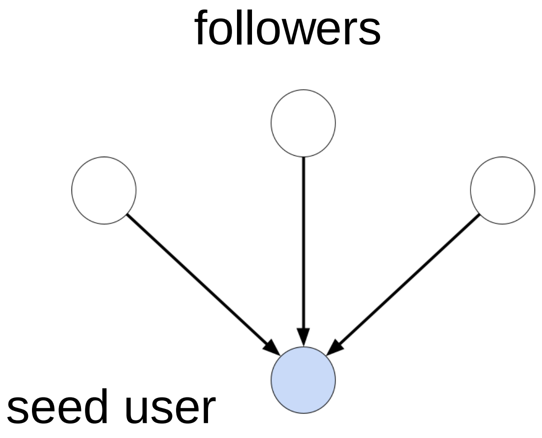


Experiment

To compare the proposed metric with several existing metrics, we conducted experiments on two datasets collected from Twitter.

- To estimate $P_{RT}(i)$, we collected the total number of tweets so far and the 100 most recent tweets for each account.
- In addition, collected the number of followers and following relationships for each account.

	Dataset 1	Dataset 2
seed user	@univkyoto	@A_I_News
number of nodes	40,691	32,739
number of edges	509,978	456,483
average P_{RT}	2.58e-06	9.43e-07



Experiment

We compared the node ranking by our IGC and IGC_{rt} with the ranking by the following existing metrics:

- P_{RT}
- In-degree (d^{-})
- Out-degree (d^{+})
- IGC (w_i: PR)
- IGC_{rt} (w_i: PR)
- Katz Centrality(Katz)
- Hub score(Hub)
- PageRank(PR)

Experiment 1

IGC has strong positive correlation with Katz, but their correlation is smaller than the correlation between IGC and the hub score.

- Node weight w_j and retweet probability P_{RT} certainly makes IGC different from Katz.

	P_{RT}	d^-	d^+	IGC	IGC_{rt}	Katz	Hub	PR
P_{RT}	—	−0.11	−0.51	−0.48	0.16	−0.47	−0.48	−0.14
d^-	−0.11	—	0.34	0.18	−0.06	0.31	0.26	0.89
d^+	−0.51	0.34	—	0.76	0.33	0.88	0.85	0.34
IGC	−0.48	0.18	0.76	—	0.37	0.77	0.87	0.20
IGC_{rt}	0.16	−0.06	0.33	0.37	—	0.36	0.36	−0.04
Katz	−0.47	0.31	0.88	0.77	0.36	—	0.87	0.31
Hub	−0.48	0.26	0.85	0.87	0.36	0.87	—	0.27
PR	−0.14	0.89	0.34	0.20	−0.04	0.31	0.27	—

Experiment 2

To compare the ranking by the hub score and the ranking by IGC in more details, we show their top 10 users in Dataset 1.

In addition to the metrics from Experiment 1, we use the $\overline{\mathrm{f.PR}}$

- the average PageRank values of the followees of the user

Experiment 2

The users with the highest hub scores have high out-degree values (d^+) , and high values in the column $\overline{\text{f.PR}}$.

Hub	P_{RT}	d^-	d^+	IGC	IGC_{rt}	Katz	PR	$\overline{\text{f.PR}}$
1	39412	2	1	4	24532	1	4	0.385
2	30784	49	2	162	15993	3	224	0.139
3	28591	490	7	300	13731	2	1402	0.039
4	36761	58	11	5160	23869	8	438	0.017
5	27127	159	5	5225	14223	5	584	0.036
6	30478	96	4	5121	17271	12	193	0.049
7	32345	47	13	5172	19547	11	366	0.014
8	25973	91	10	688	11804	6	581	0.023
9	26821	683	9	2574	13114	16	1500	0.024
10	27008	748	12	303	12189	15	1318	0.022

Experiment 2

By contrast, the users with the highest IGC scores do not necessarily have high values for d^+ and $\overline{\text{f.PR}}$.

IGC	P_{RT}	d^-	d^+	IGC_{rt}	Katz	Hub	PR	$\overline{\text{f.PR}}$
1	35175	1200	354	19679	928	460	303	0.001
2	38853	3231	537	23637	647	337	1918	0.001
3	40444	307	289	26571	444	324	299	0.001
4	39412	2	1	24532	1	1	4	0.385
5	31209	112	31	14885	13	11	607	0.012
6	31028	1532	567	14718	1130	758	108	0.001
7	34157	9076	1255	18927	1439	795	3562	0.000
8	31389	691	244	15295	326	310	826	0.002
9	29945	323	59	13799	47	65	539	0.006
10	32659	7652	1014	17056	1799	742	3856	0.000

Conclusion

- We proposed a new centrality metric for social media users, focusing on information-gathering ability of users.
  - assigning different importance weight and different forwarding probability to each node.

- We show that we can compute our metrics efficiently.

- We compared the rankings generated by our metrics and the existing metrics on two social graphs obtained from Twitter.

- The result shows that the rankings by our metrics do not coincide with the rankings by existing metrics.

Future work

IGC in reversed edge graph

- IGC is expected to measure information distributing ability through multi-hop information propagation
- The comparison of IGC and PageRank would be an interesting