JAISCR, 2023, Vol. 13, No. 3, pp. 165 – 180
10.2478/jaiscr-2023-0013

A NOVEL APPROACH OF VOTERANK-BASED KNOWLEDGE GRAPH FOR IMPROVEMENT OF MULTI-ATTRIBUTES INFLUENCE NODES ON SOCIAL NETWORKS

Hai Van Pham[1,*], Pham Van Duong[2], Dinh Tuan Tran[3,*], Joo-Ho Lee[4]

[1] *School of Information Communication and Technology, Hanoi University of Science and Technology, Hanoi, Vietnam*

[2] *ICT Department, FPT University, Hanoi, Vietnam*

[3] *College of Information Science and Engineering, Ritsumeikan University, Japan*

[4] *College of Information Science and Engineering, Ritsumeikan University, Japan*

[*] *E-mail: haipv@soict.hust.edu.vn, tuan@fc.ritsumei.ac.jp*

Submitted: 25th December 2022; Accepted: 7th May 2023

Abstract

Recently, measuring users and community influences on social media networks play significant roles in science and engineering. To address the problems, many researchers have investigated measuring users with these influences by dealing with huge data sets. However, it is hard to enhance the performances of these studies with multiple attributes together with these influences on social networks. This paper has presented a novel model for measuring users with these influences on a social network. In this model, the suggested algorithm combines Knowledge Graph and the learning techniques based on the vote rank mechanism to reflect user interaction activities on the social network. To validate the proposed method, the proposed method has been tested through homogeneous graph with the building knowledge graph based on user interactions together with influences in real-time. Experimental results of the proposed model using six open public data show that the proposed algorithm is an effectiveness in identifying influential nodes.

Keywords: Video surveillance, deep learning, Moving object detection

1 Introduction

The rapid growth of data coming from multiple sources will present both opportunities and challenges for researchers in data analysis [1, 2, 3, 4]. The development of the internet leads to abundant utilities and social networks such as Facebook, Twitter, and Instagram. They have been becoming extremely popular with almost all ages [5]. Accordingly, social networks [6] become an important and useful platform for advertising campaigns. However, disseminating information directly to all users on social networks is currently inefficient and costly. It is recommended to target a large number of users to communicate each others, i.e., in order to find a certain group of people called KoLs on social networks to carry out campaigns to promote products in digital marketing to reach as many users as possible at the lowest cost. It is not only advertised with information on social networks, but also huge data in systems such as power grids [7, 8], computer systems [9], transportation systems [10]. The tracking COVID-19 persons [11] are also performed on networks using clustering approaches [12]. Therefore, considering as the rapid spread of information, which affects the scale of the network, is very necessary. For example, spreading fake news in the prevention and control of the Covid-19 epidemic in Vietnam from accounts in social media [13]. In the case of disease transmission, it is imperative to locate and immunize the most influential individuals who can prevent further spread of the virus because of these dangers usually begin with a small node, spread some nodes and coverage spread to the entire network [14, 15]. In a large-scale computer network, it is imperative to design a robust and secure architecture because the actual system is not interrupted suddenly, so creating redundant links with servers in the main system depending on their importance is a fairly effective solution [16]. To solve the problem of identifying influential nodes in social networks, scientists have investigated many algorithms with different approaches. One of the approaches used by many researchers is to determine the set of influential nodes in the network based on ranking the nodes in the network according to the score determined by the measure of the centrality of these nodes.

From the challenges in building the social network graph in particular and determining the influence nodes in the network in general, the paper has presented the process of identifying nodes and ranking the influence nodes in the social network. This paper has presented a novel algorithm for measuring users with these influences on a social network. The suggested algorithm combines Knowledge Graph and the learning techniques based on the vote rank mechanism to reflect user interaction activities on the social network. This proposed model has significant study contributions as follows: (1) construct a homogeneous graph from large data set collected by the method of building knowledge graph based on user interaction; (2) enhance voting rank among neighbors by dealing with voting process in real-time; (3) reduce the computation time of the proposed algorithm while updates some attributed of the nodes in each iteration instead of all the nodes. To validate the proposed method, the proposed method has been tested through homogeneous knowledge graph using open data sets and the case study of real-time user interactions together with influences for demonstrated method's effectiveness.

The remainder of this paper is structured as follows: Literature review and related research is discussed in Section 2. Theoretical determinism, preliminary, and research background is considered in Sections 3. The proposed model with suggested algorithms represents by steps in Section 4 with experimental results. The conclusion and observations provided in Section 5.

2 Literature Review

In related works [17], the investigation focuses on tracking influential nodes based on the decision-making preferences. The investigation has focused on identifying top-k influential nodes in social network [18]. Farzaneh et. al has investigated influence maximization based on community structure for improvement of time efficiency [19]. New techniques as calculating nodes' centrality in graphs can be proposed relations under uncertainty to the edges on social networks [20]. As mentioned in the studies above, many centrality measures have been proposed. Measures in this approach are divided into 3 categories including: centrality measure based on local information, centrality measure based on semi-local information, and centrality measure based on global information. For example, the Degree Centrality (DC) [21] determines the number of neighbors of the node to evaluate the influence. The DC measure assumes that a node with many neighbors will have more influence. The PageRank (PC) [22] algorithm not only considers the number of neighbors but also the influence of the neighbors. The PageRank algorithm determines that a node's influence can be influenced by other nodes connected to it, since nodes connected to influential nodes can also make those nodes more influential. These nodes are centrality measures based on local information, which is a low accuracy, however the approach is matched with large-scale networks. To solve the limitation of local information of the network, some other algorithms measure the influence of nodes based on the global information of the network are proposed such as Betweenness Centrality (BC) [23] and Closeness Centrality (CC) [21]. The BC determines the influence of a node by calculating the number of shortest paths through it. The CC centrality measure determines that the shorter the average distance between a node and other nodes in the network. These measures are improved an accuracy for small networks. Centrality measures based on local information and centrality measures based on global information have the same advantages and limitations as described above. Therefore, centrality measures based on semi-local information. To avoid an imbalance when focusing only on local or global information, it is necessary to take into account both these elements of the network. However, the conventional methods of determining the influence node in the past often have not taken into account the global and local information of the network simultaneously, leading to information loss and then the final result is often greatly affected. Recently, some investigations have proposed combined approaches for the measurements, such as the GSM model [24], gravity model [25], and node's location and neighboring node information [26]. The gravity model is based on Newton's gravity formula, taking into account the influence of the neighbors as well as the path information between the nodes. However, this calculation is difficult in the case of the network becomes more complex. To solve the difficult problem of big data and complex networks, another algorithm is proposed called GSM algorithm. It not only considers the influence of the node itself in the network but also focuses on the global influence of the nodes in order to calculate the influence score of the nodes in the network.

As different from the traditional approach that only determines the influence of the node by the centrality measure, another approach based on the voting mechanism is also being developed by many researchers. In related works of influential nodes [27], the VoteRank algorithm is based on the voting mechanism to identify influential nodes in the network. As doing the similar approach [28], the proposed EnRenew algorithm with the entropy information of the nodes to consider root nodes with its improvement of performance. In addition, the research proposed to improve the VoteRank algorithm called VoteRank++ [29]. Moreover, for social networks, the information on this platform is difficult to collect and represent because of its security characteristics [30, 31]. Therefore, we need to build a method to be able to collect and represent this data so that it is suitable for the algorithms to solve the problem. From the above studies, it can be seen that the centrality is an important measure to determine the influential nodes and the influence of the neighboring nodes to it are also the factors that make that node become popular more influence. In addition, the links between neighbors also have an impact on the propagation ability of a node, the more connections there are between neighboring nodes, the higher the influence of that node will be. Besides, combining different attributes and building weights for each attribute also makes determining the influence node set more accurate [32, 33]. Data on social networks is represented in a very complicated way, since it is difficult to apply algorithms by dealing with huge data [34]. In practice of the studies [35, 3], personal privacy, i.e., people hiding their information, makes it difficult to follow users on social networks and reduce the complexity of the graph. Duong et. al have investigated multiple Attributes by influence ranking of nodes [36] and VoteRank++ to find optimal influential nodes in social networks [37]. These studies have focused limited data sets with static environment of the social networks.

3 Preliminaries

3.1 Modeling social network data into a homogeneous graph

This investigation has proposed modeling social network data into a homogeneous graph by developing an interaction-based knowledge graph.

User's relationship

Assume that the relationship between user v_i and user v_j if user v_j has influenced with, commented on or reacted to any of user v_i's comments/posts. e_{ij} is the symbol for the connection between v_i and v_j as given by Eq.(1).

$$e_{ij} = \{r, c, s\}_{ij}, e_{ij} \in R^p \tag{1}$$

where r, c, v are the number of user interactions such as reactions, shares, and comments of user v_j on posts of user v_i, respectively.

Interaction-based knowledge graph The interaction-based Knowledge Graph [38, 39, 40] is denoted as $G = (V,E)$ where $V = \{v_i\},(i = 1,...,N_V)$ is a set of users who are friends or followers and interact with each other. And $E = \{e_ij \in R^p\},(i, j = 1,...,N_E)$ are the edges that indicate user interaction activities. The information gained from social networks is shown as interaction activity. In addition, the interaction-based knowledge graph enables us to understand behaviors and actions of specific users. Users with substantial interactions are intuitively influenced users. In actuality, even there are no substantial interactions, certain users are strongly impacted. Furthermore, edges only show the link between users without regard to the importance of the edge. A weighted function can be also mapped as defined by:

$$e_ij = \phi(r,c,s) \tag{2}$$

where

$$\phi(r,c,s) = \alpha r+\beta c+\gamma s \tag{3}$$

with $\phi = \{\alpha,\beta,\gamma\}$ are hyper-parameters learned or trained by using Machine Learning tools. For example, these function are such as linear, or logistic regression, etc.

3.2 Typical centrality Measures

Degree centrality

DC [21] is a concept as definition for the number of edges, interacting with the number of edges node. $DC(i)$ of node i is expressed by:

$$DC(i) = k(i) = \sum_j a_ij \tag{4}$$

where $k(i)$ represents a degree of node i.

Mixed core, degree and entropy (MCDE) algorithm

The MCDE algorithm [2] considers a combination of indices of node position in the network, node's order coefficient, and that node's entropy. This algorithm is used to rank the nodes in the network. MCDE of node v is calculated by:

$$MCDE(v) = \alpha KS(v)+\beta DC(v)+\lambda Entropy(v) \tag{5}$$

$$Entropy(v) = -\sum_{i=0}^{KS_{max}} p_i * \log_2 p_i \tag{6}$$

$$p_i = \frac{the\,number\,of\,neighbors\,of\,node\,v\,in\,the\,i-th\,KS}{DC(v)} \tag{7}$$

To explain the MCDE algorithm, we have considered and calculated the MCDE measure for the nodes included in the graph network. As considering the example with node 1, we have $KS(1) = 3, DC = 7, p$. According to the formulas, we can apply the MCDE algorithm to the remaining nodes.

EnRenew

These two concepts are fed into the complex network in order to calculate node importance [28]. The information entropy of any node v can be calculated in

$$E_v = \sum_{u\in\Gamma_v} H_uv = \sum_{u\in\Gamma_v} -p_uv.\log p_uv \tag{8}$$

with $p_{uv} = \frac{d_u}{\sum_{l\in\Gamma_v} d_l}$ and Γ_v is the set of neighbors of node v, d_u is the degree centrality of the node u, H_{uv} is the ability to propagate information from node u to node v. After each high impact node is selected, the algorithm improves the information entropy of all nodes in its local range according to the formula

$$H_{u^{l-1}u^l} - = \frac{1}{2^{l-1}} \cdot \frac{H_{u^{l-1}u^l}}{E_k} \tag{9}$$

where k is the average of the degree centralityity of the entire network, $\frac{1}{2^{l-1}}$ is the coefficient decreases.

VoteRank++

With the VoteRank++ approach [29], nodes with different degree centrality will have influence scores voted on by different node. Like VoteRank algorithm, VoteRank++ is divided into 4 main stages

Initialization: The influence score vs_v for node v is value 0, initialized by default, and the score of voting of each node va_v is calculated by the formula:

$$va_v = \log \frac{k_v}{k_max} \tag{10}$$

Voting: While voting is conducted during this phase. Each node v receives a vote equal to the total of the voting ability scores of its neighbors. This is considered as the score of influence of that node, as expressed by

$$VP_u v = \frac{k_v}{\sum_{w \in \Gamma_v} k_w} \tag{11}$$

$$vs_v = \sqrt{\mid \Gamma_v \mid \sum_{u \in \Gamma_v} VP_u v.va_u} \tag{12}$$

The node with the highest influence score is considered as the node that propagates the influence. In addition, the selected node will not be allowed to joint as subsequent voting rounds by setting that node's voting ability score to 0.

Update: The VoteRank++ algorithm improves the reduction of the voting ability score to two-hop influence on its neighbors by the formula:

$$va_v = \begin{cases} \lambda.va_v & ,v\ is\ a\ first-order\ neighbor \\ \sqrt{\lambda}.va_v & ,v\ is\ a\ second-order\ neighbor \end{cases} \tag{13}$$

where parameter $\lambda \in [0,1]$

Iteration: The voting and updating steps will be repeated until l influence nodes are selected, where l is a predefined constant.

Information Entropy

Information entropy [28] is a well-known notion in information theory. In general, the more unpredictable or random an occurrence is, the more data it contains. Information entropy is expressed by the following.

$$H(X) = H(x_1, x_2, \ldots, x_n) = -\sum_{i=1}^{n} p(x_i).\log p(x_i) \tag{14}$$

where $p(x_i)$ is the probability of event x_i and $X = \{x_1, x_2, \ldots, x_n\}$ is set of possible events. One of the most important uses of information entropy in the field of social science is the entropy weighting technique. The entropy weighting approach is frequently applied to calculate weights of qualities in multi-attribute decision-making situations. It is also used to evaluate the associated nodes in social networks, due to its good performance. Assume there are n qualities to take into account. The weight of attribute i abbreviated w_i, is calculated as follows:

$$w_i = \frac{1 - H_i}{\sum_1^n (1 - H_i)} \tag{15}$$

where $H_i (i = 1, 2, .., n)$ represent each attribute of the information entropy.

4 Proposed Model

The proposed model is enable to design Voterank-based knowledge graph for improvement of multi-attributes influence nodes on social networks is divided in four steps as follows:

Step 1: Construct weights for attributes

Step 1.1: Calculate attribute measure score

To combine three attributes including $I_{DC}(v)$ indicates the influence of the node's degree and the degree of its neighbors v, $I_{KS_im}(v)$ respectively, we denote the influence of the node's k-shell and the k-shell of its neighbors v and $I_C(v)$ represent the influence of clustering coefficient in hierarchical nodes. Hence, v is calculated by Eq. (16), Eq. (17) and Eq. (18).

$$I_D C(v) = DC(v) + \sum_{u \in \Gamma_v} DC(u) \tag{16}$$

$$I_{KS\_im}(v) = KS\_im(v) + \sum_{u \in \Gamma_v} KS\_im(u) \tag{17}$$

$$I_{C_v} = e^{-C_v}. \sum_{u \in \Gamma_v^2} C_u \tag{18}$$

Step 1.2: Build a weighting formula with Entropy Information

Let w_1, w_2, w_3 be considered a weight corresponding to $I_{DC}(v)$,$I_{KS\_im}(v)$ and $I_C(v)$, respectively. They are calculated using entropy information [28] as follow:

Firstly, a decision matrix with its values $I_{DC}(v)$,$I_{KS\_im}(v)$ and $I_C(v)$ for all nodes presented in the social network.

$$D = \begin{bmatrix} I_{DC}(1) & \ldots & I_{DC}(n) \\ I_{KS\_im}(1) & \ldots & I_{KS\_im}(n) \\ I_C(1) & \ldots & I_C(n) \end{bmatrix} \tag{19}$$

To normalize the multi-attribute decision matrix is the next step. Matrix D is normalized to matrix R since each index has a multiple dimension as follows.

$$R = \begin{bmatrix} r_{11} & \ldots & r_{1n} \\ r_{21} & \ldots & r_{2n} \\ r_{31} & \ldots & r_{3n} \end{bmatrix}, \qquad r_{ij} = d_{ij} / \sqrt{\sum_{j=1}^{n} (d_{ij})^2} \tag{20}$$

Secondly, calculating information entropy of each rating metric is given by Eq. (14), the information entropy of index j is expressed by:

$$E_i = \frac{-1}{\ln n} \sum_{j=1}^{n} r_{ij} \ln r_{ij}, \qquad i, j = 1, 2, 3 \tag{21}$$

Finally, to determine the weight of each metric, the weights of index j is calculated by:

$$w_i = \frac{1 - E_i}{2 - \sum_{i=1}^{3} E_i}, \qquad i = 1, 2, 3 \tag{22}$$

Step 2: Calculate voting ability score The voting ability score (va_v) is obtained by considering the order coefficient, which is based on the k-shell decomposition as well as the clustering coefficient of the node along with its neighbors. The voting ability of node v is calculated mathematically as

$$va_v = \log(1 + \frac{score_v}{score_max}) \tag{23}$$

Note that $score_v$ of each node v is calculated by:

$$score_v = w_1 I_{DC}(v) + w_2 I_{KS\_im}(v) + w_3 I_C(v) \tag{24}$$

where w_1, w_2, w_3 is a weight corresponding to the index as calculated by Eq. (22). Each node's state is represented as a tuple (vs_v, va_v), where vs_v is the voting score of node v as determined by its neighbors, and va_v is the voting score of node v as determined by its neighbors.

Step 3: Calculate influence score by voting score

A voting score called a voteRank is used to determine the of a node as an influence score, as defined by the following.

$$vs_v = \sqrt{|\Gamma_v| .va_v + \sum_{u \in \Gamma_v} w_{uv}.va_u} \tag{25}$$

where w_{uv} is a weight as calculated the distance of nodes between u and v, $w_{uv} = 1$ in unweighted networks and Γ_v is the neighbor set of node v.

Step 4: Update the voting score

The updated voting score of the neighbor nodes is expressed by the Eq. (13). The voting score between a node's first-order and second-order neighbors is reduced if it is determined as influential with the shortest distances.

Step 5: Repeat voting score and influence score

Repeat steps 1, 2 and 3 until l influential nodes, where l is a constant.

4.1 Performance metrics

The susceptible infected recovery (SIR) [41, 42] model is a conventional infectious disease model as a descriptive information transfer. The SIR model consists of all nodes divided into three categories as follows: the susceptible, the infected, and the recovered nodes. Susceptible node $(S(t))$ signifies the number of nodes that are diseased-susceptible at time t, Infected node $(I(t))$ is the number of those infected, and recovered node $(R(t))$ is the number of those recovered at this time. Every susceptible node has communicated by contracting all these nodes by infected at each time interval with probability β. Additionally, every infected node recovers with the probability μ at every time step to become a recovered node.

Figure 1 illustrates the parameters to adjust the SIR model. Specifically, to consider the influence of a node in the graph, the SIR model uses the input of three factors including: the node determines the influence, the probability of infection and recovery. Initially, it is an initialize the node to determine the influences in state I and the remaining nodes in state S. Nodes in state S belong to node I's neighborhood possibly to get an information with infection probability β. At the same time of propagation, nodes in state I switch to state R with a recovery probability Γ that the node has no collection of being propagated again. The propagation can be stoped when the graph reaches the "stable" state. Hence, there are no more nodes in the graph in state I. To limit the randomness of the model, the experiments of the model with the results are averaged over the total number of experiments. Note that the value of β plays a significant role in controlling the propagation speed.

The infected scale $F(t)$ [43]: In the process of information diffusion according to the SIR model, the number of nodes that are recovered in changes over time throughout the network. At any time t, the infected scale $(F(t))$ is the total number of nodes in the social network that have propagated the information and the number of nodes recovered up to

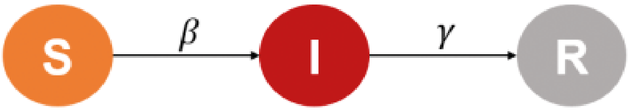


Figure 1. Diagram representing the SIR model's factors

time t divided by the total number of nodes in the entire network. Propagation scale is the most realistic measure of the effectiveness of an influence algorithm in the network over time. The formula for determining the infected scale $F(t)$ is given by:

$$F(t) = \frac{n_{I(t)} + n_{R(t)}}{n} \tag{26}$$

where $n_I(t), n_R(t)$ and n are respectively the number of infected nodes at time t, the number of recovered nodes up to time t and the total number of nodes in the network.

The final infected scale $F(t_c)$ [44]: is an important metric that corresponds to the degree of influence that information eventually propagates in the network. It is the small number of nodes in the network, which corresponds to the number of nodes in the network that have propagated information then recovered the information diffusion simulation of the SIR model. Assume that the SIR simulation model terminates at time t_c, then the measure $F(t_c)$ lists all the nodes in the network that have propagated information then recovered in the network. The following formula is developed to determine the final infected scale $F(t_c)$ as given by

$$F(t_c) = \frac{n_{R(t)}}{n} \tag{27}$$

where $n_R(t)$ and n are the number of recovered nodes and the total number of nodes in the network, respectively.

5 Results and Discussions

5.1 Data sets

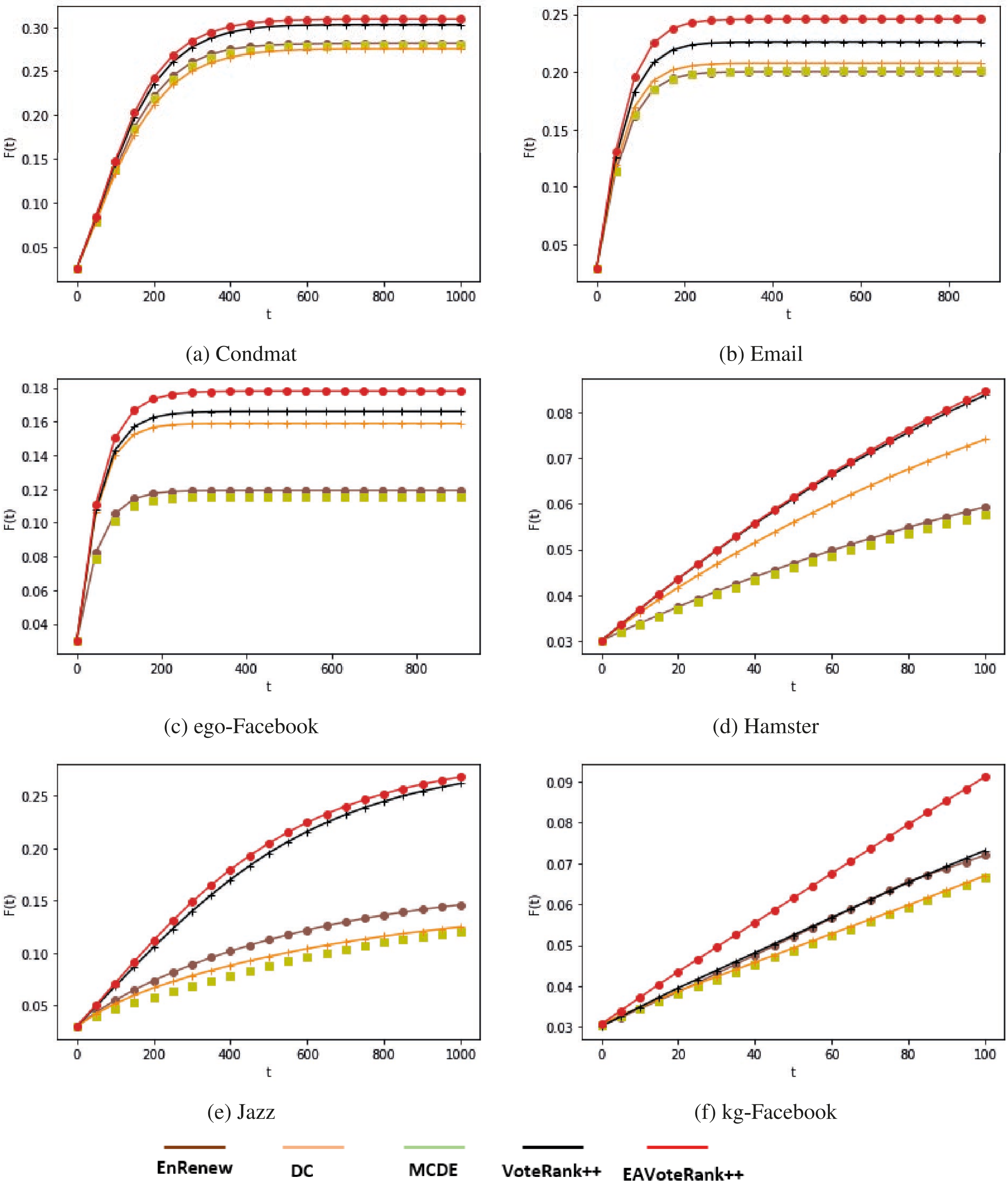
5.1.1 Modeling crawled data into a homogeneous graph

This investigation has selected the social networking platform Facebook to collect data and construct in graphs. This is a platform that is widely used and regularly updated with statuses and interactions. The proposed model has applied to the application of Scrapy tool to collect data sets using Python programming language to build graphs between users on social networks. Data collected from individuals includes [35]: user information, information about articles as well as interactions, comments, and shares of user's posts. This data is graphically represented and depicts user activities with posts by other users. However, this data is not suitable for algorithms to determine the influence node because of heterogeneous graph. To transform into a homogeneous graph, the study has applied to a data modeling method based on user interaction, as described in the previous section.

5.1.2 Data sets description

The investigation has performed in experiments with the proposed algorithm on six available data sets to gauge its effectiveness, as follows:

- Jazz: The dataset represents the interactions in the network of Jazz musician. With the nodes in the graph representing Jazz musicians and the edges in the graph representing the relationship among musicians if they play together in a band [45].
- Email: A network of email exchange among students of the University of Roviraa Virgilli, Spain. With the nodes in the graph representing email addresses and the edges in the graph representing the relationship among these email addresses if there are exchangeable among them.[46].
- Brightkite: Users of a network on the Brightkite social network are based on these locations. Nodes in the graph represent accounts on the Brightkite social network and the edges in the graph representing the relationship among accounts if they check in the same location. [47].
- Condmat: Co-author network among researchers on the topic of Condensed Matter.



(a) Condmat

(b) Email

(c) ego-Facebook

(d) Hamster

(e) Jazz

(f) kg-Facebook

Figure 2. The infection scales $F(t)$ against time t on the networks

Nodes in the graph represent the researchers and the edges in the graph represent the relationship among researchers if they are the same author as their major [48].

- Ego-Facebook: Dataset is a set of users on a social network containing information about the user's relationship on Facebook. Nodes in the graph represent accounts on the social network Facebook and the edges in the graph in the relationship among accounts if they are friends relationship to each other. [49].
- Kg-Facebook: Data set of users collected on the social network Facebook using knowledge graph method based on user interactions. Nodes in the graph represent accounts on the social network -Facebook and the edges in the graph represent the relationship among accounts, described by the method in Section 3.1

Network	N	M	$\langle k \rangle$	k_{max}
Jazz	198	2,742	8.94	100
Email	1,133	4,451	9.62	71
Brightkite	58,228	214,078	7.35	1,134
Condmat	23,133	63,479	8.08	281
ego-Facebook	4,039	88,234	43.69	1,045
kg-Facebook	369,310	458,994	2,49	30,259

Table 1. The basic topological characteristics of the networks

The data is publicly uploaded at the author's github link[1]. In Table 1, N and M are the number of nodes and connections, respectively and k_{max} and k denote the maximal and average degree of the network, respectively. The above networks belong to the public network dataset often used in the problem of determining the affected node, fully meeting the diverse requirements of different types of networks in reality. According to the above table, the data set meets a variety of network types in practice, with the kg-Facebook network having the largest number of nodes and relationships.

5.2 Experimental results

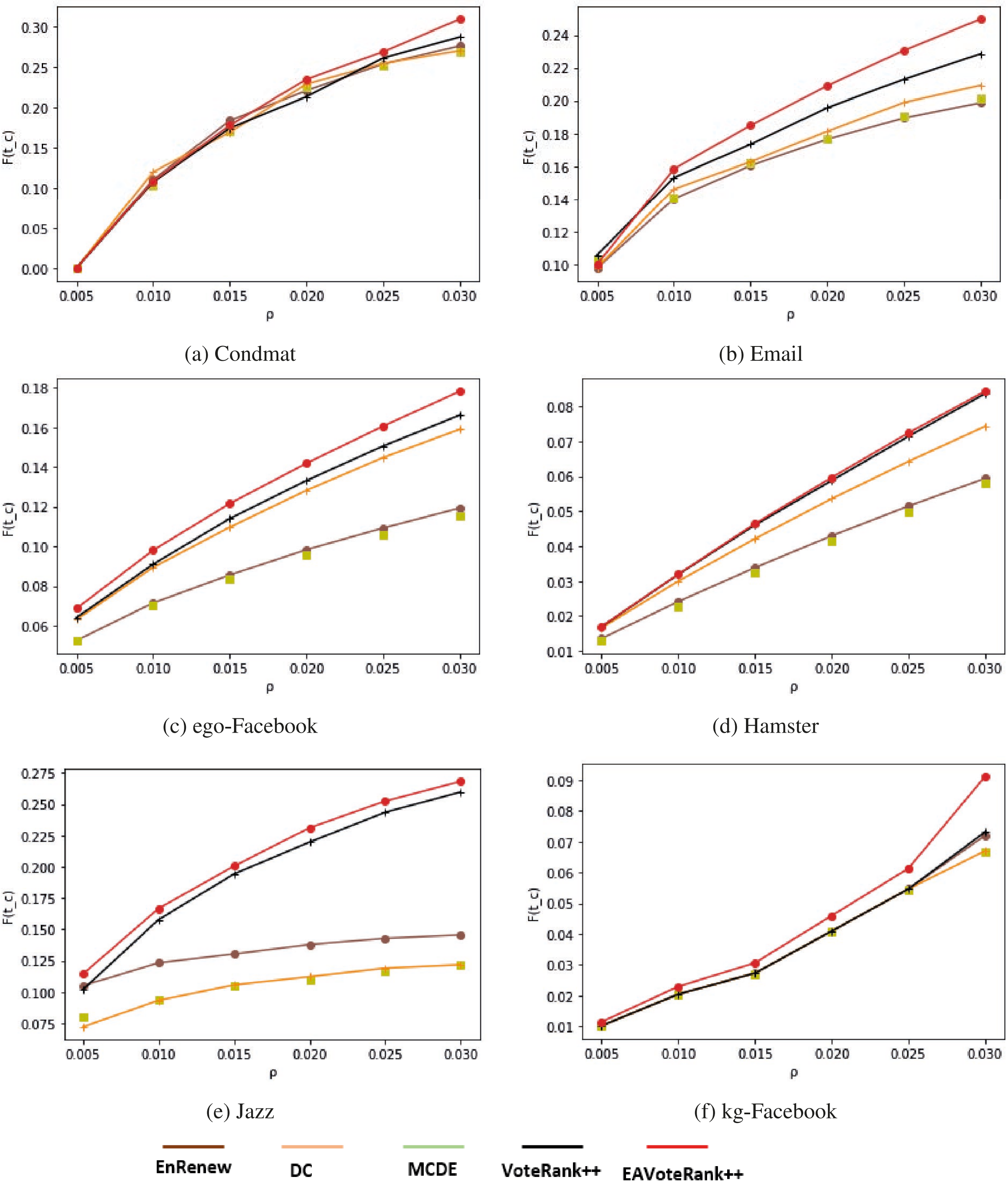
The ability of a set of significant determined nodes to distribute information was simulated in the research using the SIR propagation model. The index of the average shortest path length between spreaders are both used in the study as indicators of how decentralized the cluster of nodes is determined to be distributed. The study also selected 4 other baseline methods including DC [21], MCDE [2], EnRenew [28] and VoteRank++ [29]. These methods are widely used and highly effective in current studies.

- DC defined as the number of edges occurring on a node is known as the number of edges node.
- The MCDE algorithm considers a combination of indices of node position in the network, node's order coefficient, and that node's entropy.
- EnRenew algorithm aimed to identify a set of influential nodes via information entropy by calculated as initial spreading ability with information entropy, and then select the node with the largest information entropy and renovate its l-length reachable nodes' spreading ability by an attenuation factor.
- VoteRank algorithm improves on two issues. (1) the voting power of a node is related to the initialization process; (2) each node can vote for its neighbors in various approaches during voting process.

As shown in the experimental results in Figure 2, Figure 3, Figure 4, Experimental results show that brown line represents the results of the EnRenew algorithm, the orange line represents the results of the DC algorithm, and the green line represents the results of the MCDE algorithm, the black line represents the result for the VoteRank++ algorithm, and the red line represents the result for the proposed algorithm EAVoteRank++. With the experimental results, the experiments has been determined 3% of the most influential nodes in the graph as the node corresponding to $\rho = 0.03$ for the amendment in the propagation that contains the final infected scale $F(t_c)$ and the infected scale $F(t)$.

In the experiments, Figure 2 shows the experimental results with the scale evaluation index $F(t)$ for each time t regarding to the SIR propagation model. With the aid of this index, we have assessed the spread of impact for influence nodes produced by each algorithm. These results are averaged for the numbers of 1000 separate trials across

[1] https://gitlab.com/duongpv.hust/social-network-datasets.git



(a) Condmat

(b) Email

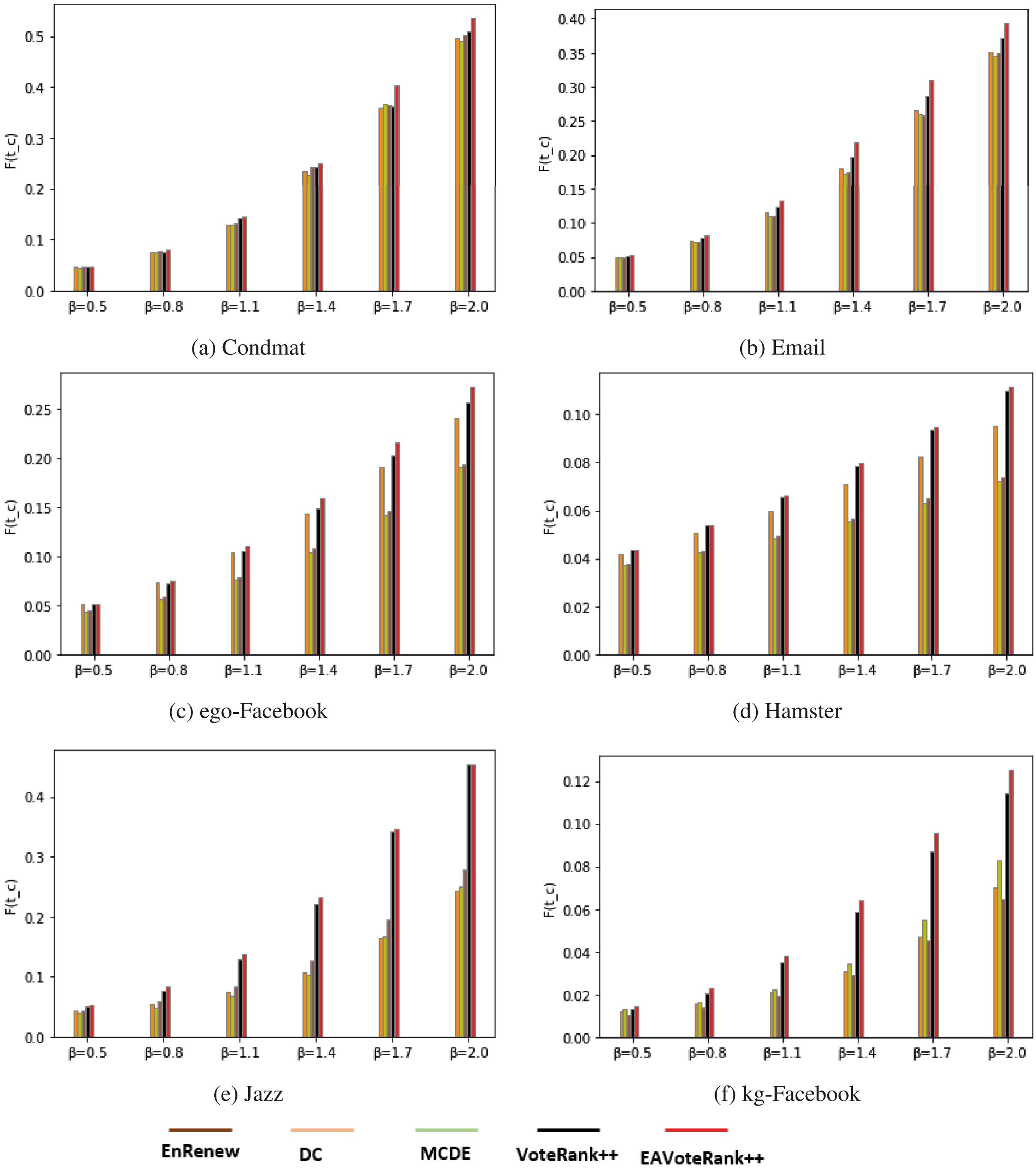
(c) ego-Facebook

(d) Hamster

(e) Jazz

(f) kg-Facebook

Figure 3. Evaluation of $F(t_c)$ in the final infection scales with ratios of initially infected nodes ρ



(a) Condmat

(b) Email

(c) ego-Facebook

(d) Hamster

(e) Jazz

(f) kg-Facebook

Figure 4. The ratios of an infect rate β and the final ration in infection scale $F(t_c)$

data sets with 100 times over two sizable data sets. With the selection of a variety of affected nodes for each algorithm, the number of affected nodes are also different. From the results in Figure 2, it can see that with the same number of initial seed nodes. The EAVoteRank++ algorithm as shown in red line achieves the highest propagation results tested all the experimental data. As for the results of the DC and EnRenew algorithms, these algorithms have the worst performance since the reason can be seen that these two algorithms only consider 1 aspect of the node's influence. For weighted and unweighted datasets (kg-Facebook), the EAVoteRank++ recommendation algorithm still shows its superiority. At the first times $t = 0$, in most datasets, the difference between algorithms is not much because the selection of influence nodes is now of the same number. From this point forward, the number of nodes affected by the seed nodes selected by the EAVoteRank++ algorithm has been performced larger than that of other methods. By taking advantage of the semi-local measurements for the combination with the voting mechanism, the algorithm has been reduced by the computational cost, which is also the factor that makes the EAVoteRank++ algorithm highly efficient as well as the calculation speed.

For various ratios of infected nodes in the beginning, the final effect scales $F(t_c)$ are shown in Figure 3. The experimental results using averaged infection rate $\beta = 1.5$ and 1000 separate tests for setting up the environment. The purpose of this experiment is to examine the effects of different techniques and beginning spreader numbers on the ultimate ranges. Unquestionably, more initial spreaders cover a larger region. The fact that the proposed algorithm consistently ranks top and sees its ultimate influence expand under the same starting spreader ratio is evidence that this algorithm performs impressively on these networks. Furthermore, information can move quickly throughout the network with nodes based on the proposed algorithm. This result demonstrates more clearly the efficiency of the proposed algorithm compared with other baseline algorithms. The results of most of the above experiments on six data sets are better than those of other deterministic algorithms.

The proposed algorithm gives a list and ranks the influential nodes with high influence, proving the effectiveness of the algorithm on network graph data types. With the large data set of Kg-Facebook, the difference among influences of the seed nodes performed by the algorithms is relatively small since the nodes in this network often focus on a group of buttons. While the number of seed nodes is small, the results of all methods are relatively similar. The range of influence in the seed nodes proposed by EAVoteRank++ can expand and propagate in the whole network when ρ is larger. It shows that the improvements of EAVoteRank++ effectively to diffuse information with the nodes as the initial number of seed nodes is larger. For data sets either Jazz or Hamster, the resulting difference is not much since the number of nodes in the relationship of the nodes is small, leading to the propagation of the method with its efficiency.

The spread process is significantly impacted by the infection rate in the SIR model. No matter whatever source spreaders are used, information cannot be disseminated successfully when is little. While is huge, the network can transmit information swiftly. In this experiments, the results are the average of 1000 independent runs with $\rho = 0.03$ except kg-Facebook. The results of kg-Facebook are the average of 100 independent runs due to the large number of nodes. The final effect scales $F(t_c)$ with various infection rates are shown in Figure 4 for various methodologies. Figure 4 clearly shows that the proposed algorithm is capable of achieving the largest spread scale, or one that is very near to it, under a variety of conditions, particularly on the networks of ego-Facebook, kg-Facebook, Email, Jazz, and Condmat. It is indicated that the proposed algorithm has a greater capacity for generalization than the baseline technique.

Experimental results of the proposed algorithm ruining with six data sets are better than those of other deterministic algorithms. The proposed algorithm gives a list and ranks the influential nodes with high influence, tested on many cases with different probabilities in the SIR model. This proves the efficiency of the algorithm on network graph data types.

Conclusion

In summary, experimental results have figured out the influence node, as presented in the construction of a homogeneous graph from social network data using the method of building a knowledge graph based on interaction. This paper has considered a new algorithm in the direction of voting approach to identify identify important nodes in the social network while consolidating multiple attributes, taking into account the extent and location of the nodes in the network and their neighborhoods. The algorithm also proposed to weight each of the above attributes with the entropy theoretical model to determine the point of each node to achieve the most optimal efficiency. We have applied voting score by combining the node's direct influence score with the node's indirect influence score.

In order to develop high efficiency in the future, the algorithm performs appropriate and effective improvements. Either social network data or complex networks are also becoming more complex. Further investigation will improve the proposed algorithm on networks with a large number of nodes with these relationships performs on dynamic networks to match the actual data. To consider reducing the cost of calculating metrics and techniques for weighting with corresponding measures. Take advantage of the homogeneous graph is modeled by using network embedding technique to represent the node features for enhancement of the proposed algorithm together with a Deep Learning model.

Acknowledgement

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.05-2019.316.

References

[1] R. M. C. J. Bond, Fariss, jason j. jones, adam di kramer, cameron marlow, jaime settle, james h. fowler. 2012. a 61-million-person experiment in social influence and political mobilization, Nature 489 295–298.

[2] A. Sheikhahmadi, M. A. Nematbakhsh, Identification of multi-spreader users in social networks for viral marketing, Journal of Information Science 43 (3) (2017) 412–423.

[3] H. V. Pham, D. N. Tien, Hybrid louvain-clustering model using knowledge graph for improvement of clustering user's behavior on social networks, in: The International Conference on Intelligent Systems & Networks, Springer, 2021, pp. 126–133.

[4] H. P. Van, N. D. Khoa, Applied multivariate regression model for improvement of performance in labor demand forecast, in: B. Unhelker, H. M. Pandey, G. Raj (Eds.), Applications of Artificial Intelligence and Machine Learning, Springer Nature Singapore, Singapore, 2022, pp. 645–654.

[5] X. T. Dinh, H. V. Pham, Social network analysis based on combining probabilistic models with graph deep learning, in: Communication and Intelligent Systems, 3rd Edition, Vol. 204, Springer, Singapore, 2021, Ch. 12, pp. 975–986.

[6] B. Liu, X. L. Yu, S. Chen, X. Xu, L. Zhu, Blockchain based data integrity service framework for iot data, in: 2017 IEEE International Conference on Web Services (ICWS), IEEE, 2017, pp. 468–475.

[7] W.-L. Fan, X.-M. Zhang, S.-W. Mei, S.-W. Huang, Vulnerable transmission line identification considering depth of k-shell decomposition in complex grids, IET Generation, Transmission & Distribution 12 (5) (2018) 1137–1144.

[8] Y. Yang, T. Nishikawa, A. E. Motter, Small vulnerable sets determine large network cascades in power grids, Science 358 (6365) (2017) eaan3184.

[9] H.-J. Li, H. Li, C. Jia, A novel dynamics combination model reveals the hidden information of community structure, International Journal of Modern Physics C 26 (04) (2015) 1550043.

[10] X.-F. Wang, X. Li, G.-R. Chen, Network science: an introduction, Beijing: Higher Education Press 4 (2012) 95–142.

[11] H. V. Pham, Q. H. Nguyen, The clustering approach using som and picture fuzzy sets for tracking influenced covid-19 persons, in: N. H. T. Dang, Y.-D. Zhang, J. M. R. S. Tavares, B.-H. Chen (Eds.), Artificial Intelligence in Data and Big Data Processing, Springer International Publishing, Cham, 2022, pp. 531–541.

[12] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: Structure and dynamics, Physics reports 424 (4-5) (2006) 175–308.

[13] S. Kumar, B. Panda, Identifying influential nodes in social networks: Neighborhood coreness based voting approach, Physica A: Statistical Mechanics and its Applications 553 (2020) 124215.

[14] Q. Shang, B. Zhang, H. Li, Y. Deng, Identifying influential nodes: A new method based on network efficiency of edge weight updating, Chaos: An Interdisciplinary Journal of Nonlinear Science 31 (3) (2021) 033120.

[15] Pham, N. Van H, T. Quoc H., T. Van P, Phuong, The proposed context matching algorithm and its application for user preferences of tourism in covid-19 pandemic, in: International Conference on Innovative Computing and Communications. Lecture Notes in Networks and Systems, Spinger, 2022, pp. 285–293.

[16] Y. Yang, X. Wang, Y. Chen, M. Hu, C. Ruan, A novel centrality of influential nodes identification in complex networks, IEEE Access 8 (2020) 58742–58751.

[17] Q. Zhang, X. Li, Y. Fan, Y. Du, An sei 3 r information propagation control algorithm with structural hole and high influential infected nodes in social networks, Engineering Applications of Artificial Intelligence 108 (2022) 104573. https://doi.org/10.1016/j.engappai.2021.104573

[18] Y. Wang, H. Li, L. Zhang, L. Zhao, W. Li, Identifying influential nodes in social networks: Centripetal centrality and seed exclusion approach, Chaos, Solitons & Fractals 162 (2022) 112513. https://doi.org/10.1016/j.chaos.2022.112513

[19] F. Kazemzadeh, A. Asghar Safaei, M. Mirzarezaee, S. Afsharian, H. Kosarirad, https://www.sciencedirect.com/science/article/pii/S0925231223002084 Determination of influential nodes based on the communities' structure to maximize influence in social networks, Neurocomputing 534 (2023) 18–28. https://doi.org/https://doi.org/10.1016/j.neucom.2023.02.059 https://www.sciencedirect.com/science/article/pii/S0925231223002084

[20] A. Zareie, R. Sakellariou, https://www. sciencedirect.com/science/article/pii/S0306437923000157 Centrality measures in fuzzy social networks, Information Systems 114 (2023) 102179. https://doi.org/https://doi.org/10.1016/j.is.2023.102179 https://www.sciencedirect.com/science/article/pii/ S0306437923000157

[21] J. Zhang, Y. Luo, Degree centrality, betweenness centrality, and closeness centrality in social network, in: 2017 2nd international conference on modelling, simulation and applied mathematics (MSAM2017), Atlantis Press, 2017, pp. 300–303.

[22] Y.-H. Eom, D. L. Shepelyansky, Opinion formation driven by pagerank node influence on directed networks, Physica A: Statistical Mechanics and its Applications 436 (2015) 707–715.

[23] M. Lei, K. H. Cheong, Node influence ranking in complex networks: A local structure entropy approach, Chaos, Solitons & Fractals 160 (2022) 112136.

[24] A. Ullah, B. Wang, J. Sheng, J. Long, N. Khan, Z. Sun, Identification of nodes influence based on global structure model in complex networks, Scientific Reports 11 (1) (2021) 1–11.

[25] Z. Li, T. Ren, X. Ma, S. Liu, Y. Zhang, T. Zhou, Identifying influential spreaders by gravity model, Scientific reports 9 (1) (2019) 1–7.

[26] A. Zareie, A. Sheikhahmadi, K. Khamforoosh, Influence maximization in social networks based on topsis, Expert Systems with Applications 108 (2018) 96–107.

[27] J.-X. Zhang, D.-B. Chen, Q. Dong, Z.-D. Zhao, Identifying a set of influential spreaders in complex networks, Scientific reports 6 (1) (2016) 1–10.

[28] C. Guo, L. Yang, X. Chen, D. Chen, H. Gao, J. Ma, Influential nodes identification in complex networks via information entropy, Entropy 22 (2) (2020) 242.

[29] P. Liu, L. Li, S. Fang, Y. Yao, Identifying influential nodes in social networks: A voting approach, Chaos, Solitons & Fractals 152 (2021) 111309.

[30] S. Kumar, D. Lohia, D. Pratap, A. Krishna, B. Panda, Mder: modified degree with exclusion ratio algorithm for influence maximisation in social networks, Computing 104 (2) (2022) 359–382.

[31] S. Samanta, V. K. Dubey, B. Sarkar, Measure of influences in social networks, Applied Soft Computing 99 (2021) 106858.

[32] X.-H. Yang, Z. Xiong, F. Ma, X. Chen, Z. Ruan, P. Jiang, X. Xu, Identifying influential spreaders in complex networks based on network embedding and node local centrality, Physica A: Statistical Mechanics and its Applications 573 (2021) 125971.

[33] J. Zhao, T. Wen, H. Jahanshahi, K. H. Cheong, The random walk-based gravity model to identify influential nodes in complex networks, Information Sciences 609 (2022) 1706–1720.

[34] H. V. Pham, D. H. Thanh, P. Moore, Hierarchical pooling in graph neural networks to enhance classification performance in large datasets, Sensors 21 (18) (2021) 6070.

[35] Q. M. Tran, H. D. Nguyen, T. Huynh, K. V. Nguyen, S. N. Hoang, V. T. Pham, Measuring the influence and amplification of users on social network with unsupervised behaviors learning and efficient interaction-based knowledge graph, Journal of Combinatorial Optimization (2021) 1–27.

[36] P. Van Duong, X. T. Dinh, L. H. Son, P. Van Hai, Enhancement of gravity centrality measure based on local clustering method by identifying influential nodes in social networks, in: S.-H. Wang, Y.-D. Zhang (Eds.), Multimedia Technology and Enhanced Learning, Springer Nature Switzerland, Cham, 2022, pp. 614–627.

[37] P. Van Duong, T. M. Dang, L. H. Son, P. Van Hai, Enhancement of voting scores with multiple attributes based on voterank++ to identify influential nodes in social networks, in: A. L. Pinto, R. Arencibia-Jorge (Eds.), Data and Information in Online Environments, Springer Nature Switzerland, Cham, 2022, pp. 242–257.

[38] L. T. H. Lan, T. M. Tuan, T. T. Ngan, N. L. Giang, V. T. N. Ngoc, P. Van Hai, et al., A new complex fuzzy inference system with fuzzy knowledge graph and extensions in decision making, Ieee Access 8 (2020) 164899–164921.

[39] C. K. Long, P. Van Hai, T. M. Tuan, L. T. H. Lan, P. M. Chuan, L. H. Son, A novel fuzzy knowledge graph pairs approach in decision making, Multimedia Tools and Applications (2022) 1–30.

[40] D. N. Tien, H. P. Van, Graph neural network combined knowledge graph for recommendation system, in: International Conference on Computational Data and Social Networks, Springer, 2020, pp. 59–70.

[41] S. Kumar, A. Panda, Identifying influential nodes in weighted complex networks using an improved wvoterank approach, Applied Intelligence 52 (2) (2022) 1838–1852.

[42] C. Alger, K. Todd, The sir model of disease spread (2015).

[43] H. Ahmadi Beni, A. Bouyer, Identifying influential nodes using a shell-based ranking and filtering method in social networks, Big Data 9 (3) (2021) 219–232.

[44] M. Bahutair, Z. Al Aghbari, I. Kamel, Noderank: Finding influential nodes in social networks based on interests, The Journal of Supercomputing 78 (2) (2022) 2098–2124.

[45] P. Gleiser, L. Danon, Advances in compl, Sys 6 (2003) 565–573.

[46] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, A. Arenas, Self-similar community structure in a network of human interactions, Physical review E 68 (6) (2003) 065103.

[47] E. Cho, S. A. Myers, J. Leskovec, Friendship and mobility: user movement in location-based social networks, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011, pp. 1082–1090.

[48] J. Leskovec, J. Kleinberg, C. Faloutsos, Graphs over time: Densification and shrinking diameters, arXiv preprint physics/0603229 (2008).

[49] B. Viswanath, A. Mislove, M. Cha, K. P. Gummadi, On the evolution of user interaction in facebook, in: Proceedings of the 2nd ACM workshop on Online social networks, 2009, pp. 37–42.

Van Hai Pham is received Doctor of Engineering degree (Ph.D.) at Ritsumeikan University (Japan). He is an Associate Professor at School of Information and Communication Technology, Hanoi University of Science and Technology. His major fields include Artificial Intelligence, Knowledge Based, Big data, Soft Computing, Rule-based Systems and Fuzzy Systems. His publications are over 150 International journals and conferences in ISI/Scopus indices. He also serves as Chairs and Co-chairs of organized several sessions at international conferences such as KSE 2019, KSE 2017, KSE 2015, SOICT 2014. He can be contacted at email: haipv@soict.hust.edu.vn.
https://orcid.org/0000-0001-8325-1662

Pham Van Duong received B.E and M.E degrees with major in Computer Science at School of Information and Communication Technology, Hanoi University of Science and Technology, Vietnam in 2020 and 2022. From 2020 to now, he has been working as a Data Scientist at CMC Applied Technology Institute. Since 2022, he has been a lecturer at Information and Communication Technology Department, FPT University. His research interests include Big Data Processing, Graph Data Mining and Deep Learning.
https://orcid.org/0009-0000-8871-3969

Dinh Tuan Tran received the B.E., M.E., and Ph.D. degrees in Information Science and Engineering from Ritsumeikan University, Japan, in 2012, 2016, and 2019, respectively. From 2019 to 2020, he worked as a Postdoctoral Researcher with the College of Information Science and Engineering at Ritsumeikan University, Japan. He is currently an Assistant Professor in the College of Information Science and Engineering at Ritsumeikan University, Japan. Between 2017 and 2019, he was a short-term visiting researcher with the Technical University of Munich (Munich, Germany, 2017), the Budapest University of Technology and Economics (Budapest, Hungary, 2018), the University of Auckland (Auckland, New Zealand, 2018), the University of Greenwich (London, UK, 2018), the Cardiff

University (Cardiff, UK, 2018), and the University of Melbourne (Melbourne, Australia, 2019). His research interests include machine learning, image processing, computer vision, robot vision, reinforcement learning, imitation learning, human process modeling, and medical imaging.
https://orcid.org/0000-0001-7443-9102

Joo-Ho Lee received the B.E. and M.E. degrees from Korea University, Seoul, Korea, in 1993 and 1995, respectively, and the Ph.D. degree from The University of Tokyo, Tokyo, Japan, in 1999, all in electrical engineering. He is currently a Professor in Dept. of Information Science and Engineering at Ritsumeikan University, Shiga, Japan. From 1999 to 2003, he was a JSPS Postdoctoral Researcher at the Institute of Industrial Science, The University of Tokyo. From 2003 to 2004, he was Research Associate at Tokyo University of Science, Japan and from 2004 he joined Ritsumeikan University as an Associate Professor. From 2008 to 2009, he was a Visiting Scholar at the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. In 2017, he was a Research Professor at Dept. of Mechanical Engineering, Korea University, Seoul, Korea. His research interests include intelligent environments, intelligent robots, computer vision, machine learning, and medical/healthcare applications. He is a Senior Member of IEEE and a Member of the RSJ, JSME, SICE, HIS, IEICE, KROS, and IEEJ.
https://orcid.org/0000-0003-1015-5615