BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 05/06 pp130–139
Volume 5, Number 2, June 2022
DOI: 10.26599/BDMA.2021.9020024

A Novel Influence Maximization Algorithm for a Competitive Environment Based on Social Media Data Analytics

Jie Tong, Leilei Shi*, Lu Liu*, John Panneerselvam, and Zixuan Han

Abstract: Online social networks are increasingly connecting people around the world. Influence maximization is a key area of research in online social networks, which identifies influential users during information dissemination. Most of the existing influence maximization methods only consider the transmission of a single channel, but real-world networks mostly include multiple channels of information transmission with competitive relationships. The problem of influence maximization in an environment involves selecting the seed node set for certain competitive information, so that it can avoid the influence of other information, and ultimately affect the largest set of nodes in the network. In this paper, the influence calculation of nodes is achieved according to the local community discovery algorithm, which is based on community dispersion and the characteristics of dynamic community structure. Furthermore, considering two various competitive information dissemination cases as an example, a solution is designed for self-interested information based on the assumption that the seed node set of competitive information is known, and a novel influence maximization algorithm of node avoidance based on user interest is proposed. Experiments conducted based on real-world Twitter dataset demonstrates the efficiency of our proposed algorithm in terms of accuracy and time against notable influence maximization algorithms.

Key words: influence maximization; competitive environment; dynamic network

1 Introduction

Online social networks have grown rapidly in recent years, as people around the world are connected well than before via social networks[1]. In recent years, the emergence of Facebook, Twitter, Weibo, and other social platforms have gradually become an integral part of people's lives[2].

Online social networks have turned out to be a hot research topic in recent years, particularly with inextricable links to data mining and machine learning[3–5]. At present, the main research directions include community discovery[6–8], recommendation system[9], and Information dissemination[10], especially influence dissemination is deemed important due to the ongoing trend of dissemination of both genuine and fake news in online social networks at a rapid pace[11]. Influence in this context refers to the ability of a given user and his/her posts to change the cognition of others. It is common that users change their original cognition after being influenced by the dynamics of other users in social networks. Influence in social media is mainly manifested on the word of mouth effect and viral marketing. For instance, finding influential people to promote certain products via social media is common. Herein, the maximum impact can be achieved by realizing the diffusion effect of that user's influence, in such a way to cover the entire possible object among

• Jie Tong, Leilei Shi, and Zixuan Han are with the School of Computer Science and Telecommunication Engineering and Jiangsu Key Laboratory of Security Technology for Industrial Cyberspace, Jiangsu University, Zhenjiang 212013, China. E-mail: 319209944@qq.com; 937109472@qq.com; 823460370@qq.com.

• Lu Liu and John Panneerselvam are with the School of Computing and Mathematical Sciences, University of Leicester, Leicester LE1 7RH, UK. E-mail: l.liu@leicester.ac.uk; J.Panneerselvam@leicester.ac.uk.

* To whom correspondence should be addressed.

Manuscript received: 2021-11-01; accepted: 2021-11-12

the target audience. From a research perspective, the aforementioned scenario is the influence maximization problem of practical application.

Influence maximization problem has attracted the attention of many boffins since it was proposed by Domingos and Richardson[12] in 2001. In 2003, Kemple et al.[13] proved the influence maximization problem as an NP hard problem, and initially proposed a greedy algorithm with good influence range and high time complexity. However, the greedy algorithm has a high time complexity and cannot efficiently solve the problems in large-scale social networks. Moreover, a wide range of well performing heuristic algorithms have been developed in recent years. Wang et al.[14] designed and evaluated a scalable and adjustable heuristic – the Prefix excluding MIA (PMIA) algorithm to deal with the influence maximization problem in a large-scale network comprising millions of nodes. Tian et al.[15] proposed the Hybrid Potential-influence Greedy (HPG) algorithm by exploiting the "accumulation characteristics" of the Linear Threshold (LT) model and incorporating the network topology and the propagation characteristics of social networks. The algorithm is divided into two parts. In the first step, a heuristic algorithm is used to select the nodes characterising higher influence. In the second step, the greedy algorithm is used to find the node with the highest influence. This algorithm is superior to the traditional greedy algorithm in terms of achieved accuracy. Beutel et al.[16] studied the relationship between multiple pieces of information and proposed a new propagation model based on the influence of two kinds of communication information.

At present, existing studies mostly consider the spread of single topic information in social networks. Nevertheless, in reality, multiple competing messages are spread simultaneously in social networks. Therefore, efficiently spreading self-interested information with maximum benefits is crucial whilst resolving maximizing influence problems in a competitive environment[17]. Since Bharathi et al.[18] first proposed the influence maximization problem in a competitive environment in 2007, Bozorgi et al.[19] conducted further detailed research on the same problem from the perspective of community. Various such studies have proved that the fusion algorithm[15] characterizes excellent performance in terms of algorithm efficiency and accuracy, which has become the direction of future research.

When a great deal of competitive information is flooded in the social network, it is easy to cause users to have antipathy, resulting in a large reduction in the effect of publicity. User's diversity in their subject preference of information affects the scope of influence in the social network. Therefore, considering users' interests and preferences in the information transmission model is crucial to depict the real-world characteristics in order to achieve a more accurate transmission effect.

In view of the above problems, the goal of this paper is to maximize the influence that we need in a competitive social environment. The main contributions of this paper are listed below.

(1) Considering the global network of space and time overheads, we select the local community based on the discrete degree of the local community discovery algorithm.

(2) Considering the competitive environment of information transmission in a real-world environment, potential user influence is improved in our model through the establishment of a novel influence maximization algorithm for a competitive environment.

(3) We conduct experiments to evaluate the performance of our proposed models. Experiments conducted based on a Twitter dataset demonstrate the efficiency and accuracy of our models in influence discovery and influence maximization in a competitive environment.

The rest of this paper is organized as follows. In Section 2, this paper summarizes and analyses existing research on influence maximization. Section 3 describes our local community discovery algorithm of community dispersion. Section 4 describes our computation of the node influence probability, along with our propagation model developed based on the user interest. Our proposed algorithm for the maximization of the topic influence under competitive social environments is also detailed in Section 4. Section 5 compares and analyses the experimental results. Section 6 concludes this paper along with outlining our future work.

2 Relevant Work

2.1 Overlapping communities and non-overlapping communities

In a non-overlapping community, a node can only belong to a single community. There is no intersection between communities. According to graph segmentation in social networks, communities can be regarded as the compact substructure of graphs, thus community segmentation algorithms can be used to solve the problem. The representative algorithms in graph segmentation include the Kernighan-Lin (KL)[20] algorithm and spectral bisection algorithm[21]. While the KL algorithm requires setting up a prior value and its algorithm efficiency is generally low, the spectral bisection algorithm can only divide two communities at a time. These defects limit the practical application of graph segmentation algorithm under harsh conditions, resulting in lower efficiency.

Traditional community detection algorithms[22] believe that nodes can only belong to one community at a time, and further assume that common nodes among communities never exist. However, in reality, human beings are complex and often change positions. Herein, in a real-world social network, people often belong to multiple communities, which depicts the most realistic network structure. Palla et al.[7] believed that the community in a complex network is a fully connected sub-graph in a larger graph, and proposed a community detection algorithm based on the clique theory to solve the overlapping community discovery problem[23].

2.2 Maximization of influence based on user interest topic model

In the traditional influence maximization algorithms[12–18], given the number of nodes, K, in the network, which will be selected as the seeds to maximize the number of nodes that could be affected by the set of seeds. Although the first K nodes that can maximize the influence of the set of seeds can be analyzed, the analysis is conducted on the premise that a given node's influence on other nodes is identical. Thus, specific event information content spread is not considered in traditional algorithms. For completely unrelated event topics, the influence exerted by users is generally fixed. This may not be accurate in real life, which also affects the reliability of the traditional influence maximization algorithms. Therefore, the early influence maximization algorithms cannot meet the needs of social network users. For this reason, researchers introduce the problem of topic information, giving emergence to the topic-based influence maximization algorithm concepts.

2.3 Influence maximization in a competitive environment

Maximizing influence problems in competitive environments can be solved in two ways on the whole. First, from the perspective of self-interested communication, we ignore the influence of competitors and find the seed nodes with the greatest influence of self-interested communication. Second, from the perspective of restraining competitors' information, when competitive information is disseminated, competitors' influence is minimized so as to maximize its own communication influence[17]. Bharathi et al.[18] first proposed and proved that these problems could be solved by a greedy algorithm. Zhu et al.[24] studied the minimization of seed node placement cost in the competitive environment when considering the suppression of competitors' influence spread. Ribeiro and Faloutsos[25] studied the effects of new information on nodes which have been activated by other information in competitive environments under a delayed transmission of competitive environments.

A few communication models for the competitive environment have been developed in recent years. Beutel et al.[16] studied the relationship between multiple pieces of information and proposed to use the correlation coefficient between two different genres of information to do research on the degree of their interaction in the synchronization process. Some researchers have extended the traditional transmission model of using a single piece of information, and put forward the independent cascade model of multiple activities, the suppressed transmission model, and the linear threshold model[19] for competitive environments. Such aforementioned models reasonably extend the communication models for different problem scenarios, but their computation of the node influence probability is relatively simple. Most of such models start from the network or community structure, and seldom consider the content and semantic information of user interaction[26].

2.4 Dynamic social networks

Large-scale social networks are usually dynamic and complex in nature, due to their strong functionality, simplicity, and convenience dot, favored by many network users[17,27,28]. Businesses with a large user base are highly leveraging social networks for various purposes. For instance, Internet is used as a platform to promote commercial products, and social media are increasingly utilized to publicize public opinions as such. In addition, the problem of influence maximization in dynamic complex networks can also be used to predict some epidemic trends. As a result, Large-scale dynamic networks have become a new research hotspot. In comparison with static networks, influence maximization's problem in dynamic networks requires additional attention to the dynamic changes in the network topology, that is, the increasing and decreasing trend of nodes and edges, as well as the changes in their propagation probability.

As we can see from Fig. 1, the influence maximization's problem in a social network refers to the way of selecting a certain number of initial nodes from a known set of nodes and adding the selected nodes to the seed node set. This is followed by carrying out a series of influence propagation by using a specific propagation model to influence as many nodes as possible in the end.

A dynamic social network is a series of static network graphs. A node is a network graph over a continuous period, where edges change over time. At every moment, new nodes are added and old nodes disappear. The edges between the connecting nodes also change dynamically as the nodes are added and/or deleted. Each static network diagram is generated statically to indicate the current network status. For the sake of convenience, we discretize the time, without passing through a small interval. It is only after time t that the network topology changes.

3 Local Community Discovery Algorithm Based on Community Dispersion

As we can see from Fig. 1, Social networks often include a large number of low-quality and invalid data and nodes. Accurately mapping all these nodes in the network generates a lot of computation overheads and further reduces the efficiency. Therefore, we propose to use the Hyperlink-Induced Topic Search (HITS) based relationship evaluation algorithm to filter out high authorities and hubs.

(1) Data preprocessing based on HITS

The algorithm based on HITS iteratively calculates the centrality of each user and the authority of each post, to obtain high authorities and hubs. Figure 2 shows the link relationship between users and posts and the iteration process. The process is described as follows:

(1) Initialize users' centrality and post authority;

(2) The centrality of the users is the sum of the authority of all the linked posts, and the authority of the post is the sum of the centrality of all the linked users. Calculate users' centrality and post authority in the follwing:

$$\text{hub} = \sum \text{authority} \tag{1}$$

$$\text{authority} = \sum \text{hub} \tag{2}$$

(3) The users' centrality and post authority are normalized,

$$\text{hub} = \frac{\text{hub}}{\sum \text{hub}} \tag{3}$$

$$\text{authority} = \frac{\text{authority}}{\sum \text{authority}} \tag{4}$$

(4) Evaluate whether the users' centrality degree and the post authority degree converge, the convergence will output the result; otherwise continue with Eqs. (2)–(4),

$$\theta a = \frac{\sum_{i=0}^{n} \text{authority}}{n} \tag{5}$$

$$\theta_h = \frac{\sum_{i=0}^{n} \text{hub}}{n} \tag{6}$$

where θ_a and θ_h denote the thresholds after normalization of the posts' authority and users' hub, respectively.

Finally, the users' centrality and post authority are sorted in a descending fashion (high to low) according to the numerical results, and the parameters θ_a and θ_h are defined to represent the critical value of screening, which is used to extract users' interest labels, and other information are filtered out.

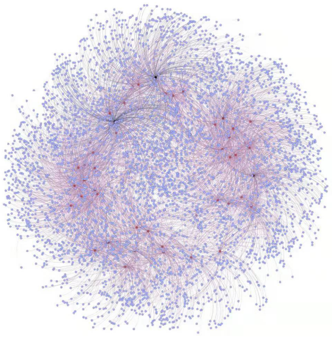


Fig. 1 Nodes and edges in social network.

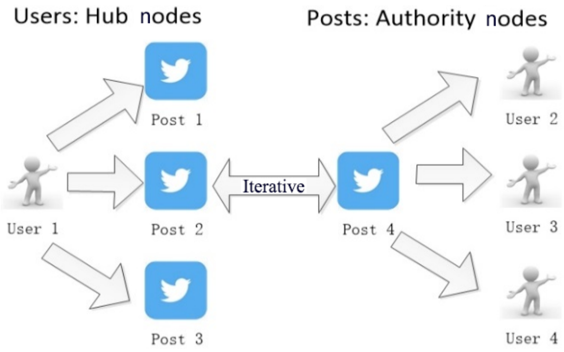


Fig. 2 Schematic diagram of the iterative process of the relationship evaluation algorithm.

(2) Initial local community discovery

Community Dispersion (CD) is a key parameter. The higher the value of CD, the closer the community will be.

$$\text{CD} = \frac{e_i}{e_o} \tag{7}$$

where e_i represents the number of edges in the community, and e_o represents the number of edges in the union of the community (C) and the neighboring nodes set (N). Our local community discovery algorithm is described in Algorithm 1.

Algorithm 1 Local community discovery algorithm

```
Input: Social network $G(V, E)$, the node in $G$, $v_0$, and the neighboring nodes of $v_0$
Output: Community $C$
  1. Initialize $C = v_0$, add $N$ to neighboring nodes of $v_0$, initialize CD to 0, initialize $\text{CD}_{\max}$=0;
  2. Calculate CD in set $N$, $\text{CD}_{v_0}$ ($\text{CD}_{v_0}$ represents the CD value after adding the node $v_0$), select the largest value, record $\text{CD}_{\max} = \text{CD}_{v_0}$, $v_{\text{best}} = v_0$ ($v_{\text{best}}$ represents the suitable value of $v_0$). If $\text{CD}_{\max} > \text{CD}$, execute Eq. (3), otherwise execute Eq. (4);
  3. If $\text{CD} = \text{CD}_{\max}$, $v_{\text{best}}$ will be added to the community set of $N$ and be removed from network $G$, nodes of $N$'s neighbor will be added to the set $N$;
  4. Output the last community $C$.
```

4 Topic-Based Influence Maximization Algorithm in a Competitive Environment

4.1 Information propagation models

In general, online social networks can be abstracted as a directed (undirected) graph representing a set of nodes, where each point represents an individual or an organization[26, 27]. Figure 1 represents a collection of edges, with each edge representing cooperation, friendship, hostility, etc. Each node can have two states as active and inactive. The active node has an effect on the inactive node, and if this effect causes an inactive node to be an active node, the process is called active. When more neighboring nodes of a given node are activated, it is highly likely that the corresponding node will also be activated. The newly activated nodes affect their inactive neighboring nodes. Over time, more and more nodes have a transition from the inactive state to the active state. The whole propagation process is irreversible, that is, a node cannot change from an active state to an inactive state.

When there is no active node with influence in network G, the propagation process finishes.

In social networks $G = (V, E)$, V is the node set and $E \subseteq V \times V$ is the edge set. In the independent cascade model, edge $(u, v) \in E$ has a probability value $p_{uv} \in [0, 1]$, which represents the probability to activate neigbouring nodes through the edge (u, v). At $t = 0$, select the seed node set (the seed node is in an active state); at any time t, u_i will attempt to activate the nodes next to u_i through edges (u_i, v_j), with the probability of p_{uv}. The process is independent. If the process succeeds, v_j will participate in the next process at $t+1$; if it is not successful, v_j remains inactive at $t+1$. v_j will remain the state that can be activated until no node is activated in the future.

Figure 3 shows that at moment $x-1$, two active nodes, namely c and f, attempt to activate their neighboring nodes at a certain probability. Node c can activate a, b, d, e, f, and g, but c and f have only one chance to activate all of their neighboring nodes. If the activation fails at the first attempt, the activation process cannot continue further. Nodes c and f have a neighboring node e, so e can and only can be activated by one of them. At time x, nodes d and e are successfully activated and they will be added to the set, and the above activation process goes on until no nodes can be influenced by their neighbors.

Definition 1 $G\ (V_t, E_t, T)$ denotes the node where the social network can be activated. E_t represents the set of edges at time t in G_t; and T represents the collection of topics in G. Therefore, the social network can be expressed as $G\ (V_t, E_t, T)$, assuming that the network evolves over time in accordance with some network evolution model. The problem of influence maximization on a dynamically growing network involves solving the problem of selecting the initial nodes in A to maximize the expected influence of its influence transmission on the network at time t on the premise that the current network structure at time t is known. The formula is as follows:

$$A_k^* = \underset{A \in \{S \subset V_t, |S|=k\}}{\text{argmax}} \quad \sigma_{t+\Delta t}(A) \tag{8}$$

where, $V_t = \{V_{1t}, V_{2t}, \ldots, V_{nt}\}$, $|V_t| = n$, σ_t represents the timestamp, S represents the number of nodes which are influenced, and $_{\Delta t}(A)$ represents the change of

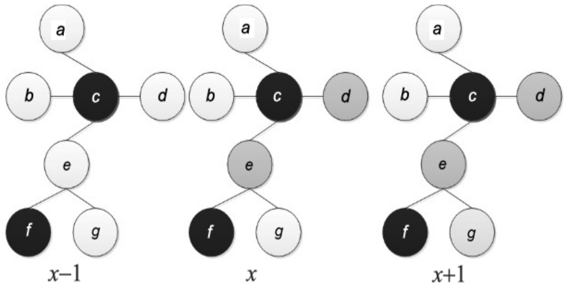


Fig. 3 Independent cascading model propagation process.

influence of A in time variation period. When there is a link between u and v, an edge exists between them, otherwise there is no edge.

4.2 Probability of influence between nodes

A very important step in the process of influence maximization is to measure the influence increment of each node. Correctly measuring the influence increment of each node is crucial to achieve the maximum influence. Highly influential nodes are usually deemed important in the network. The most important statistical measures used to denote the importance of nodes are degree centrality, intermediate centrality, compact centrality, clustering centrality, etc. After computing these statistics for each node, sorting all such nodes based on the measurement statistics is a simple strategy to measure the node importance. However, using only the aforementioned statistics during the node selection process whilst solving the influence maximization problem might not yield the best outcome. This is due to the node overlapping problem, whereby two different nodes with overlapping statistics can have varied impact on the node set. Herein, selecting nodes only based on simple metrics cannot deliver the optimal solution whilst resolving the problem of maximum impact. Thus, efficiently removing the influence of these overlaps during the selection of nodes is a key problem.

Usually, many activation paths exist between nodes, thus computing each of these paths can be a tedious process. Considering the theory of six degrees of segmentation, the connection between nodes usually does not exceed a certain value. Here, it is assumed that node u can activate node v only through the path with the highest activation probability between nodes. Finally, the activation probabilities of each node to all other nodes in the network are combined to approximate the influence range of the nodes.

Latent Dirichlet Allocation (LDA)[29] model can be used to obtain the topic-based influence probability distribution among nodes[17]. Considering the transmission of information B as an example, the calculation of the probability of p_{uv}^{B} are as follows:

$$p_{uv}^{B}=\sum_{j=1}^{k} p(\lambda_B|z_j)p(u,v|z_j) \tag{9}$$

where $p(\lambda_B|z_j)$ represents the probability that information B is the topic z_j, and $p(u,v|z_j)$ represents the probability that a node influences another node under the topic.

Therefore, in the network $G(V,E)$, calculating the influence of node v under information B is as follows:

$$\mathrm{Inf}^{B}(v)=\sum_{i=1}^{k}\mathrm{hub}(v)\cdot p(\lambda_B|z_i) \tag{10}$$

where $\mathrm{hub}(v)$ represents the hub of node v, and $p(\lambda_B|z_i)$ represents the probability that information B is topic z_i.

The formula for calculating the influence increment of node V is defined as:

$$\Delta\mathrm{Inf}^{B}(v)=\delta(S\cup\{v\})-\delta(S) \tag{11}$$

In the process of maximization of influence, the selection of the initial node is related to the direction and scope of influence of the subsequent information communication, so it is very important to find the initial communication node with the highest influence. In some existing influence maximization algorithm[26–28,30], the initial transmission nodes are derived from the entire social network by incrementally mining the largest node. Although the influence of these nodes is high in the entire network society, their corresponding influence for a particular topic may not be necessarily higher, and the influence of each node to calculate incremental time efficiency can be low.

In the first stage, we only statically selected $\lceil\beta\times k\rceil$ nodes with large influence as influence nodes, without considering the characteristics of information transmission in social networks. Therefore, the nodes acquired in the first stage are considered as the initial transmission nodes, and the information transmission in the social network is simulated through the model. Then, the greedy strategy is adopted to iteratively mine the $k-\lceil\beta\times k\rceil$ nodes with the largest increment of the topic influence as the remaining influence nodes. Among them, the largest increment of topic influence refers to the largest difference between the scope of influence of the current collection and the scope of influence before the addition of U. The specific algorithm flow is shown in Algorithm 2.

5 Experiment

5.1 Experimental dataset

In order to test the effectiveness of the algorithm, we choose the data of the mainstream social network Twitter for testing. The Twitter dataset contains 38 584 users and 1 568 130 tweets, obtained during November 2020. The experiment is carried out on a PC with Intel Core i5-8259U CPU 2.30 GHz and 8 GB memory.

Algorithm 2 Influence maximization algorithm based on dynamic network in competitive environment

```
Input: Social network $G(V, E)$, the nodes in $A$, node set $S_A$, nodes number $K$, and parameter $\beta$
Output: Information $B$ and node set $S_B$
1. Initialize the influence node set $S_B$;
2. The authority of the remaining nodes set is less than $\theta_a$ after the elimination of nodes, $\theta_h$ is taken as the threshold of hub in the node mining set $U$;
3. Calculate the influence of all nodes in $U$ and sort them, and select the node with $\lceil \beta \times k \rceil$ before the median rank as the node in $S_B$;
4. Set the node in $S_B$ obtained in the previous step in an active state, and the node in $U - S_B - S_A$ set in an inactive state. The nodes in $S_B$ are taken as the initial propagation nodes to propagate information through the model, and the influence range $\delta(S_B)$ of the set $S_B$ is obtained;
5. All the nodes in the collection of $U - S$ influence increment value $\Delta\text{Inf}^B(u) = \delta(S_B \cup \{u\}) - \delta(S_B)$;
6. Select the influence incremental node $U$ to join the set $S$, and record the biggest increment $\max\delta(), u = \{v \mid \max\{\Delta\text{Inf}^B(u)\}, v \in U - S_B - S_A\}$, $\max\delta() \leftarrow \Delta\text{Inf}^B(u)$, $S_B \leftarrow S_B \cup \{u\}$. Then the new $S_B$ is continued to propagate information through the propagation model and a new influence range $\delta(S_B)$ is obtained.
7. For any node $v$ in $U - S_B - S_A$, if $\Delta\text{Inf}^B(v) > \max\delta()$, then $\Delta\text{Inf}^B(v)$ is the maximum increment of subject influence according to the submodule, and there is no need to calculate the increment for other nodes to update $\max\delta() \leftarrow \Delta\text{Inf}^B(v)$, $S_B \leftarrow S_B \cup \{v\}$.
8. Repeat Step 7 ($K - \lceil \beta \times k \rceil - 1$) times, and the cycle ends.
9. Output influence node set $S_B$.
```

5.2 Determination of parameters

In order to determine the effects of the regulating parameter β in the algorithm, the influence range of node sets under different β and K, and the running time of the algorithm are analyzed experimentally. In order to reduce the experimental bias, average values are obtained after repeating the experiments 100 times. For any node number K, influence $\beta = 1$ is obtained as the worst case, but the running time is short due to the fact that when $\beta = 1$, the algorithm follows a heuristic approach, runs so fast, and ignores the propagation characteristics of the social network, Therefore, the influence range is small when $\beta = 1$. When $\beta = 0$, the influence range is large, but the computation time is relatively long. When $\beta = 0.5$, the influence range under some K values is almost close to that of under $\beta = 0$, and the running time is also relatively short. Therefore, $\beta = 0.5$ is selected as the best parameter value of the algorithm after comprehensively considering the running time and the influence range, as we can analyse from Figs. 4 and 5.

5.3 Comparison with other algorithms

In Table 1, we evaluate the performance of the algorithm against CELF[26], MixGreedy[27], Random, PageRank[31], Degree, Influence Maximization algorithm of Node Avoidance (IMNA)[17], and Influence Maximization algorithm based on Dynamic network in Competitive Environment (IMDCE). The efficiency of the influence maximization algorithm is usually determined based on the running time and scope. Running time represents the efficiency of the algorithm. Running scope represents the effect of the algorithm.

(1) Accuracy analysis of the algorithm

In our experiment, a total number of 1–50 seed nodes is selected, and the average value of the experimental results of 20 000 Monte Carlo simulation propagation process is considered as the final influence range. From Figs. 6 and 7, the CELF algorithm can effectively eliminate the nodes in a social network that cause diminishing marginal benefits and reduce the propagation times between nodes. In comparison with

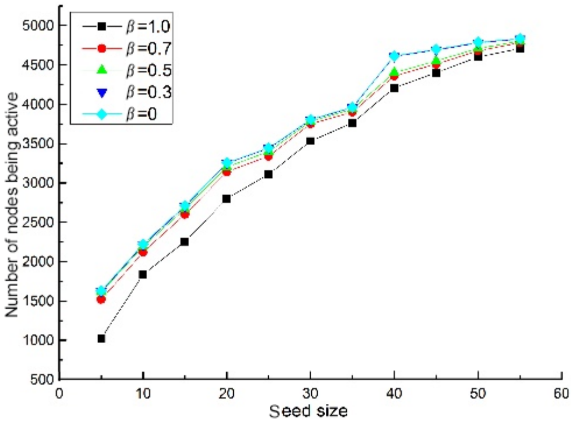


Fig. 4 Node influence range under different parameters.

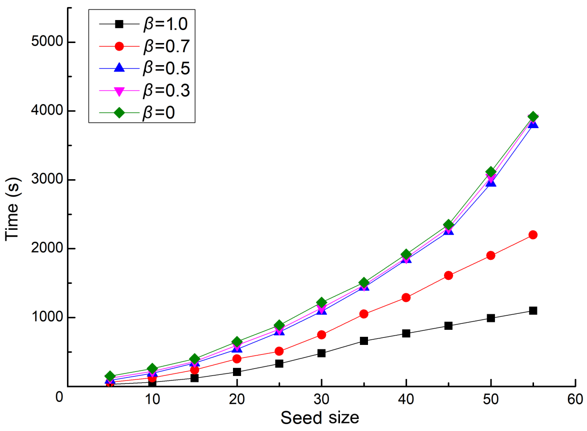


Fig. 5 Running time under different parameters.

Table 1 Comparison algorithms and descriptions in the experiment.

Algorithm	Description
CELF	An algorithm that eliminates nodes in a social network and causes diminishing marginal benefits
Random	An algorithm for removing edges in a random way from a network that is not successfully activated
MixGreedy	The greedy algorithm of CELF algorithm that is performed after the first round of NewGreedy algorithm
PageRank	An algorithm used by Google to identify the rank/importance of web pages
Degree	A heuristic algorithm for selecting k nodes with maximum degree
IMNA	An algorithm for selecting seed nodes by node avoidance in a competitive environment

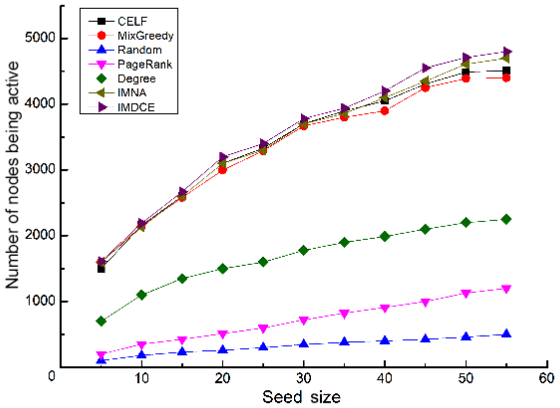


Fig. 6 Comparison of different algorithms on influence maximization.

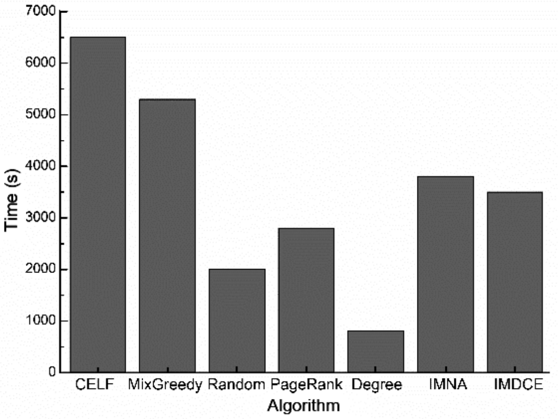


Fig. 7 Comparison of the running times of different algorithms.

the simple greedy algorithm, the CELF algorithm is 700 times faster in terms of the algorithm efficiency and achieves the best propagation range, while the MixGreedy algorithm performs slightly worse. This is because of the fact that both algorithms use the greedy approach, where the optimal influence range is obtained through multiple Monte Carlo simulations. The Random algorithm does not consider the influence propagation, but just randomly selects several nodes as the initial nodes for propagation. The Degree algorithm only considers the topology of the network; hence its propagation range is poor. PageRank algorithm uses the PageRank value of all the web pages obtained by an offline calculation, whereby effectively reducing the amount of calculation when compared with an online query, and further reducing the query response time. People's queries usually have topic characteristics, but the PageRank algorithm ignores the topic relevance, thus resulting in lower relevancy and subjectivity of results.

IMDCE performs better than the other algorithms as it incorporates the communication effect and works based on user interest subject. Thus, it can identify the user interest, according to the interested topic selection influence transmission ability of the user as a seed node. Moreover, to join the competition factors, the IMDCE algorithm reduces the impact of competitive information and the data pretreatment, and filters out the low influence of social network users, whereby improving the seed node recognition accuracy, and improving the influence scope.

In comparison with IMNA, IMDCE considers the heuristic of calculating the total influence of the nodes to select the initial seed nodes, and then adopts the greedy strategy to complete the final acquisition of the seed nodes. The seed selection strategy is optimized to obtain a better influence range of the seed nodes more effectively. Therefore, it can obtain a good influence range of seed nodes.

(2) Time analysis of the algorithm

From Fig. 7, the CELF algorithm has the longest running time, because of its greedy approach requiring an iterative calculation of the gain function, while the MixGreedy algorithm is superior to CELF, due to the removal of a lot of calculations in the beginning, thus characterizing an improved algorithm efficiency. Since the initial node selection uses the Random approach, and the Degree algorithm only considers the structure, their running time is relatively fast. Their running time is better than PageRank, since IMDCE pre-processes the data at the beginning in the initial stage, screens out high authorities and high hubs, reduces a mass of data, and then improves the operating efficiency of the algorithm. Therefore, with the continuous expansion of the scale of social networks, the advantages of our proposed algorithm have been demonstrated.

Because the IMDCE algorithm adopts heuristics to obtain the initial nodes and then greedily selects seed nodes, when compared with the IMNA algorithm, the IMDCE algorithm can reduce the running time and time complexity.

6 Conclusion and Future Work

This paper mainly studied information influence maximization in a competitive environment. Considering mass data in social networks, the HITS algorithm is used to pre-process the data. The influence maximization algorithm is proposed by considering the user's interest and preference, and the interest topics of the interaction content between nodes. Experiments conducted based on real-world Twitter datasets show that our proposed algorithm can achieve better results in terms of accuracy and computation time when compared with other existing algorithms in most cases.

Of course, there are still many areas worth further study and improvement in our proposed algorithm. For example, the activation probability between nodes has a variety of influencing factors. Efficiently integrating these factors can help to calculate the activation probability between nodes more accurately. Existing studies have not considered the concept of time in social networks. As a future research direction, we will study incorporating the effect of time while resoling topic-based influence maximization. Furthermore, we plan to optimize the running efficiency of the algorithm proposed in this paper.

Acknowledgment

The work was supported by the National Natural Science Foundation of China (Nos. 61502209 and 61502207).

References

[1] L. L. Shi, Y. Wu, L. Liu, X. Sun, and L. Jiang, Event detection and identification of influential spreaders in social media data streams, *Big Data Mining and Analytics*, vol. 1, no. 1, pp. 34–46, 2018.

[2] L. L. Shi, L. Liu, Y. Wu, L. Jiang, M. Kazim, H. Ali, and J. Panneerselvam, Human-centric cyber social computing model for hot-event detection and propagation, *IEEE Trans. Comput. Soc. Syst.*, vol. 6, no. 5, pp. 1042–1050, 2019.

[3] H. Lu, S. X. Liu, H. Wei, and J. J. Tu, Multi-kernel fuzzy clustering based on auto-encoder for fMRI functional network, *Expert Syst. Appl.*, vol. 159, p. 113513, 2020.

[4] A. Monney, Y. Z. Zhan, Z. Jiang, and B. B. Benuwa, A multi-kernel method of measuring adaptive similarity for spectral clustering, *Expert Syst. Appl.*, vol. 159, p. 113570, 2020.

[5] S. N. Tang, S. Q. Yuan, and Y. Zhu, Data preprocessing techniques in convolutional neural network based on fault diagnosis towards rotating machinery, *IEEE Access*, vol. 8, pp. 149487–149496, 2020.

[6] L. L. Shi, L. Liu, Y. Wu, L. Jiang, J. Panneerselvam, and R. Crole, A social sensing model for event detection and user influence discovering in social media data streams, *IEEE Trans. Comput. Soc. Syst.*, vol. 7, no. 1, pp. 141–150, 2020.

[7] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[8] S. Gregory, Finding overlapping communities in networks by label propagation, *New J. Phys.*, vol. 12, p. 103018, 2010.

[9] S. Y. Shih, M. Lee, and C. C. Chen, An effective friend recommendation method using learning to rank and social influence, in *Proc. PACIS 2015*, Singapore, 2015, pp. 242–250.

[10] P. Kim and S. Kim, Detecting overlapping and hierarchical communities in complex network using interaction-based edge clustering, *Phys. A Stat. Mech. Appl.*, vol. 417, pp. 46–56, 2015.

[11] J. Ge, L. L. Shi, Y. Wu, and J. Liu, Human-driven dynamic community influence maximization in social media data streams, *IEEE Access*, vol. 8, pp. 162238–162251, 2020.

[12] P. Domingos and M. Richardson, Mining the network value of customers, in *Proc. 7^{th} ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2001, pp. 57–66.

[13] D. Kempe, J. Kleinberg, and É. Tardos, Maximizing the spread of influence through a social network, in *Proc. 9^{th} ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Washington, DC, USA, 2003, pp. 137–146.

[14] C. Wang, W. Chen, and Y. J. Wang, Scalable influence maximization for independent cascade model in large-scale social networks, *Data Min. Knowl. Discov.*, vol. 25, no. 3, pp. 545–576, 2012.

[15] J. T. Tian, Y. T. Wang, X. J. Feng, A new hybrid algorithm for influence maximization in social networks, (in Chinese), *Chin. J. Comput.*, vol. 10, no. 3, pp. 1956–1965, 2011.

[16] A. Beutel, B. A. Prakash, R. Rosenfeld, and C. Faloutsos, Interacting viruses in networks: Can both survive? in *Proc. 18^{th} ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Beijing, China, 2012, pp. 426–434.

[17] J. M. Chen, L. L. Shi, L. Liu, A. O. Ayorinde, R. B. Zhu, and J. Panneerselvam, User interest communities influence maximization in a competitive environment, in *Proc. 2020 16^{th} Int. Conf. on Mobility, Sensing and Networking*, Tokyo, Japan, 2020, pp. 614–621.

[18] S. Bharathi, D. Kempe, and M. Salek, Competitive influence maximization in social networks, in *Proc. 3^{rd} Int. Workshop on Web and Internet Economics*, Berlin, Germany, 2007, pp. 306–311.

[19] A. Bozorgi, S. Samet, J. Kwisthout, and T. Wareham, Community-based influence maximization in social networks under a competitive linear threshold model, *Knowl. Based Syst.*, vol. 134, pp. 149–158, 2017.

[20] B. W. Kernighan and S. Lin, An efficient heuristic procedure

for partitioning graphs, *Bell Syst. Tech. J.*, vol. 49, no. 2, pp. 291–307, 2014.

[21] Y. C. Wei and C. K. Cheng, Ratio cut partitioning for hierarchical designs, *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.*, vol. 10, pp. 911–921, 1991.

[22] L. Jiang, L. L. Shi, L. Liu, J. J. Yao, B. Yuan, and Y. J. Zheng, An efficient evolutionary user interest community discovery model in dynamic social networks for internet of people, *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9226–9236, 2019.

[23] S. C. Liu, F. X. Zhu, and L. Gan, A label-propagation-probability-based algorithm for overlapping community detection, (in Chinese), *Chin. J. Comput.*, vol. 39, no. 4, pp. 717–729, 2016.

[24] Y. Zhu, D. Li, and Z. Zhang, Minimum cost seed set for competitive social influence, in *Proc. of IEEE International Conference on Computer Communications*, San Francisco, CA, USA, doi: 10.1109/INFOCOM.2016.7524472.

[25] B. Ribeiro and C. Faloutsos, Modeling website popularity competition in the attention-activity marketplace, in *Proc. 8th ACM Int. Conf. on Web Search and Data Mining*, Shanghai, China, 2015, pp. 389–398.

[26] A. Beutel, B. A. Prakash, R. Rosenfeld, and C. Faloutsos, Interacting viruses in networks: Can both survive? in *Proc. 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Beijing, China, 2012, pp. 426–434.

[27] C. W. Tsai, Y. C. Yang, and M. C. Chiang, A genetic NewGreedy algorithm for influence maximization in social network, in *Proc. 2015 IEEE Int. Conf. on Systems, Man, and Cybernetics*, Hong Kong, China, 2016, pp. 2549–2554.

[28] L. L. Shi, L. Liu, Y. Wu, L. Jiang, and J. Hardy, Event detection and user interest discovering in social media data streams, *IEEE Access*, vol. 5, pp. 20953–20964, 2017.

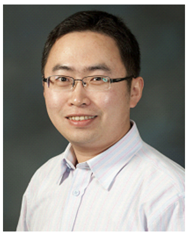
[29] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent dirichlet allocation, *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[30] N. Trivedi and A. Singh, Efficient influence maximization in social-networks under independent cascade model, *Proced. Comput. Sci.*, vol. 173, pp. 315–324, 2020.

[31] A. N. Langville and C. D. J. T. M. I. Meyer, Google's pagerank and beyond, *The Science of Search Engine Rankings*, vol. 30, no. 1, pp. 68–69, 2011.

Jie Tong received the BEng degree from Nantong University, Nantong, China in 2019. He is currently a master student at the School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China. His research interests include social networking, community discovery and influence maximization.

Leilei Shi received the BEng degree in computer science and technology from Nantong University, Nantong, China in 2012, the MEng degree in computer application technology from Jiangsu University, Zhenjiang, China in 2015, and the PhD degree in computer application technology from Jiangsu University, Zhenjiang, China in 2020. He is currently a lecturer at the School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang, China. His research interests include event detection, data mining, and social computing.



Lu Liu received the PhD degree in electronic engineering from University of Surrey, Guildford, UK in 2007, the MEng degree in data communication system from Brunel University, Uxbridge, UK in 2003, and the BEng degree from South-Central Minzu University, Wuhan, China in 2002. He is a professor of computing and mathematical sciences at University of Leicester, Leicester, UK, and an adjunct professor in Jiangsu University, Zhenjiang, China. He is a fellow of British Computer Society (BCS). His research interests are in areas of cloud computing, social computing, service-oriented computing, and peer-to-peer computing.

John Panneerselvam received the PhD degree in computing in 2018 and the MEng degree in advanced computer networks in 2013, from the University of Derby, Derby, UK. He is a lecturer in computing at the University of Leicester, UK. He is an active member of IEEE and British Computer Society (BCS). He has won the best paper award in IEEE International Conference on Data Science and Systems, Exeter, UK in 2018. His research interests include cloud computing, fog computing, Internet of Things (IoT), big data analytics, opportunistic networking, and P2P computing.

Zixuan Han received the BEng degree in computer science and technology from Jilin Normal University, Siping, China in 2018, and the MEng degree in computer application technology from Jilin Normal University, Siping, China in 2021. He is currently a PhD candidate at the School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang, China. His research interests include data mining, social networks, and cloud computing.