2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST)

2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST) | 978-1-6654-0267-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/IAECST54258.2021.9695654

Research on Influence Maximization of Citation Network from the Perspective of Meme

Yishu Wang
School of Electronic and Information Engineering
Lanzhou Jiaotong University
Lanzhou, China
0619684@stu.lzjtu.edu.cn

Guanghui Yan*
School of Electronic and Information Engineering
Lanzhou Jiaotong University
Lanzhou, China
*Yanguanghui@mail.lzjtu.cn

Zhe Li
School of Electronic and Information Engineering
Lanzhou Jiaotong University
Lanzhou, China
0219654@stu.lzjtu.edu.cn

Ye Lv
School of Electronic and Information Engineering
Lanzhou Jiaotong University
Lanzhou, China
0219656@stu.lzjtu.edu.cn

Abstract—**The influence maximization problem aims to find the most influential node set in the network and make it produce the greatest influence through information dissemination. At present, there are many researches on social networks, but there are few systematic researches on identifying the most influential results in citation networks. This paper introduces the meme to study the mode of knowledge transmission . Firstly, the memes are extracted from abstracts, and then the information transmission model based on memes is proposed. In combination with the network topology, a discount algorithm based on LeaderRank is proposed to obtain the candidate seed node set. On this basis, the greedy algorithm is introduced to effectively screen out the seed node set, and a citation network influence maximization method is designed. Experimental results on five real citation network datasets show that the meme as a unit propagation is more consistent with the characteristics of citation network propagation, and the proposed method is superior to the traditional heuristic method .**

***Keywords-** Citation Networks; Influence Maximization; Propagation Model; Meme; LeaderRank*

I. Introduction

Scientific literature is an important form of knowledge achievements, and the relationship between them constitutes a citation network. It is helpful for the progress of science to find out the most influential scientific literature in the citation network and make it reach the greatest influence through dissemination. However, with the rapid growth of the number of journal literature, how to find the most influential scientific literature in such a large citation network has become a problem. For example, for those who want to know about relevant research in a certain field, it is helpful to quickly find a part of important literature for future research. The above problem can be modeled as the maximization of the influence of citation networks.

Maximization of influence means to make use of limited resources to find the most influential seed node set and maximize the influence scope through information dissemination. Reference [1] first proposed the problem of influence maximization in their research and analysis of the optimal plan of network marketing. Subsequently, Reference [2] proposed two classical propagation models. On this basis, a greedy algorithm based on local optimal solution was designed, which could reach the approximation degree of the global optimal solution. In fact, the greedy algorithm's solution is very close to the optimal solution. However, the computation time will increase greatly with the increase of network size, which makes the greedy algorithm difficult to deal with the problem of maximizing the influence of large-scale network in reality. To solve the problem of excessive time complexity of greedy algorithm, on the one hand, researchers mainly improve greedy algorithm [3-6] by reducing the number of Monte Carlo simulations. On the other hand, heuristic methods are proposed to consider the influence of nodes themselves, such as degree centrality [7], path-based closeness centrality [8] and betweenness centrality [9], etc.

Existing researches on influence maximization focus on the topology of the network, while the semantic information contained by nodes is largely ignored. This observation prompts us to reveal the relationship and influence between citations at the micro level. At present, scholars have studied the transmission characteristics of meme in the network [10-13]. Existing studies have shown that considering the transmission process of memes can reveal the transmission mode of scientific knowledge in a more direct and fine-grained way[14].

To sum up, in order to more accurately describe the influence among citations and obtain a more effective solution to the problem of influence maximization, this paper takes meme as the research unit of citation relationship, and from the perspective of meme, combines discount strategy and greedy thought to obtain a method to maximize the influence of citation network. Finally, the effectiveness of the proposed method is verified by comparative experiments on real citation networks.

978-1-6654-0267-5/21/$31.00 ©2021 IEEE

II. Information Propagation Model Based On Meme

A. propagation probability

Firstly, the meme is extracted from the abstract of each literature through the meme recognition method [15], and the propagation score of each meme is calculated. Further, a calculation method of edge propagation score is proposed.

Definition 1 (Meme propagation score). Meme propagation score $P(m)$ is an indicator to quantify the importance of meme m by the ratio of sticking factor to sparking factor, which can be expressed as:

$$P(m)=\frac{d_{m\to m}}{d_{\to m}+\delta}\Bigg/\frac{d_{m\to\tilde{m}}+\delta}{d_{\to\tilde{m}}+\delta} \qquad (1)$$

Sticking factor quantifies the degree to which a meme is inherited and transmitted in the process of knowledge diffusion, specifically expressed as $d_{m\to m}/d_{\to m}$, where $d_{m\to m}$ refers to the number of citation literatures carrying meme and quoting at least one literature carrying this meme, and $d_{\to m}$ refers to the number of all citation literatures quoting at least one literature carrying this meme. Sparking factor quantifies the degree of variation of a meme in the process of knowledge diffusion, specifically expressed as $d_{m\to\tilde{m}}/d_{\to\tilde{m}}$, where $d_{m\to\tilde{m}}$ represents the number of citation references that carry the meme but do not cite the meme-carrying literature, and $d_{\to\tilde{m}}$ represents the number of all citation references that do not cite the meme-carrying literature. δ is the smoothing factor to prevent the occurrence of zero denominator phenomenon.

Definition 2 (edge propagation score). In network $G=(V,E)$, for any two nodes (v_i,v_j) with a connected edge e_{ij}, the sum of the propagation score of all common memes in the two nodes is the propagation score of the edge, denoted as $p(e_{ij})$, and defined as:

$$p(e_{ij})=\begin{cases}\sum p(m_{ij}),\ i\neq j\\ 0\ ,\ others\end{cases} \qquad (2)$$

Since the edge propagation score varies greatly between different nodes, it is normalized to obtain the propagation probability p_{ij} between nodes, which can be expressed as:

$$p_{ij}=\begin{cases}\dfrac{p(e_{ij})-\min p(e_{ij})}{\max p(e_{ij})-\min p(e_{ij})}\ ,\ e_{ij}\in E\\ 0\ ,\ others\end{cases} \qquad (3)$$

Where, $p_{ij}\in(0,1)$, $\min p(e_{ij})$ is the smallest edge propagation score, and $\max p(e_{ij})$ is the largest propagation score.

B. Meme cascading model

On the basis of obtaining the propagation probability, we propose Meme cascading model (MC). Each node has two states, 0 state or 1 state. When $t=0$, The state of the initial propagation source node is set to 1, and other nodes are attempted to be activated. At any time of $t\geq 1$, the newly activated node tries to activate its neighbor node with probability p_{ij}. If the activation is successful, the node status is 1, and the status of the unactivated node is still 0, until the nodes in the network reach stability, and no new nodes are activated, and the propagation ends. It should be noted that each newly activated node has only one chance to activate its neighbor nodes, and once it fails to activate, it will not be activated further.

III. Influence Maximization Algorithm Based On Discount Strategy And Greedy Thought

In this section, the research scheme to solve the problem of influence maximization under the cascading model of network meme is presented. We divided the research scheme into two stages: the stage of candidate seed selection and the stage of seed node selection. Firstly, an algorithm based on LeaderRank discount, LRDiscount, was proposed to screen candidate seed nodes, and then CELF algorithm was used to screen seed nodes among candidate seed nodes, so as to obtain an influence maximization algorithm combining discount strategy and greedy thought. Compared with the traditional heuristic algorithm, the two-step strategy proposed in this paper combines the time advantage of the heuristic algorithm and the quality advantage of the greedy algorithm, and at the same time effectively alleviates the problem of influence overlap.

A. Mining candidate seed nodes

In fact, the influence of a literature is not only reflected in the number of papers it cites, but also in the quality of the references.LeaderRank algorithm is similar to this idea.

In a directed no power network, we add a background node and connect it to all nodes in the network.The algorithm first assigns LR value of 1 unit to N nodes other than node bg, and the LR value of background node bg is 0. Then, the LR value of this 1 unit is equally distributed to the outgoing neighbor nodes directly connected to it. After the iteration, the LR value $LR_{bg}(t_c)$ of background node g is equally divided among all nodes in the network, and for the final LR value of node v_i , we thus obtain:

$$LR_i=LR_i(t_c)+\frac{LR_{bg}(t_c)}{n} \qquad (4)$$

To take into account the reverse links between nodes, we redefine the iteration process of LeaderRank. Since each node has a certain influence, when they are all selected as seed nodes, the combined influence scope will be much smaller than the sum of the influence scope of each node, which is caused by the overlap of influence. In order to alleviate the influence overlap, a discount algorithm based on LeaderRank, LRDiscount, is proposed. LRDiscount algorithm discounts the influence of neighbor nodes of each candidate seed node, that is, in the reverse network, discounts the influence of each node pointing to the candidate seed node, and the discounted influence is expressed as:

$$INF_i = INF_i - INF_i \frac{\sum_{j\in S} w(i,j)}{\sum_{j=1}^{n+1} w(i,j)} \qquad (5)$$

Where S is the candidate seed node set, and $\sum_{j\in S} w(i,j) \Big/ \sum_{j=1}^{n+1} w(i,j)$ represents the ratio of the number of candidate seed nodes in the neighbor nodes of node v_i to the number of all neighbor nodes.

B. Mining seed nodes

Kempe [2] proved the influence maximization problem is an np-hard optimization problem, but high time complexity, so we use the improved greedy algorithm CELF in screening of candidate seed node set seed nodes.

Through the above calculation, we obtain the algorithm LDGIM based on LeaderRank discount strategy and greedy thought, In the citation network with n nodes and m edges, the time complexity of LDGIM is $O(m + ckrm)$.Where c is the number of candidate seed nodes, k is the number of seed nodes, and r is the number of Monte Carlo simulations.

IV. Using the Template

In this section, we compare and verify the proposed meme cascading model applicable to citation network with the independent cascading model, and meanwhile analyze and compare the proposed method with the existing heuristic method on the meme cascading model. The seed node was selected as the initial active node, and the influence spread after propagation was used as the evaluation criterion. The greater the influence spread, the better the algorithm performance.

A. Datasets

In our experiment, we use five citation network datasets from the American Physical Society (APS), Database Systems and Logic Programming (DBLP), and ArnetMiner's published paper Annotate data source that appear in different domains.Table 1 shows the basic statistics from the above dataset.

TABLE I. Experiments Datasets

Dataset	Nodes	Edges
Annotate	805	1227
APS2000-2004	69506	253388
APS2005-2009	83866	328869
APS2010-2014	87334	383095
DBLP2000	192670	419030

B. Experimental results and analysis

Experiment 1 (Influence spread of LDGIM on two models). The seed nodes obtained from LDGIM on the five data sets are propagated through independent cascade model and meme cascade model respectively, and the differences of influence spread under the two models were compared. As shown in Fig.1, the horizontal axis represents the number of seed nodes, and the vertical axis represents the influence spread. With the same number of seed nodes, it is obvious that the meme cascading model has a wider influence spread than the independent cascading model ($p = 0.01$, $p = 0.02$), and the independent cascading model with a higher propagation probability has a wider influence spread than the independent cascading model with a lower propagation probability. Besides, We found that the transmission rate of LDGIM method on the meme cascade model was faster than that on the independent cascade model ($p = 0.01$, $p = 0.02$). It can be seen that taking meme as the propagation unit of citation network is more consistent with the characteristics of citation network, and the meme cascading model can more accurately reflect the propagation mode of citation network.

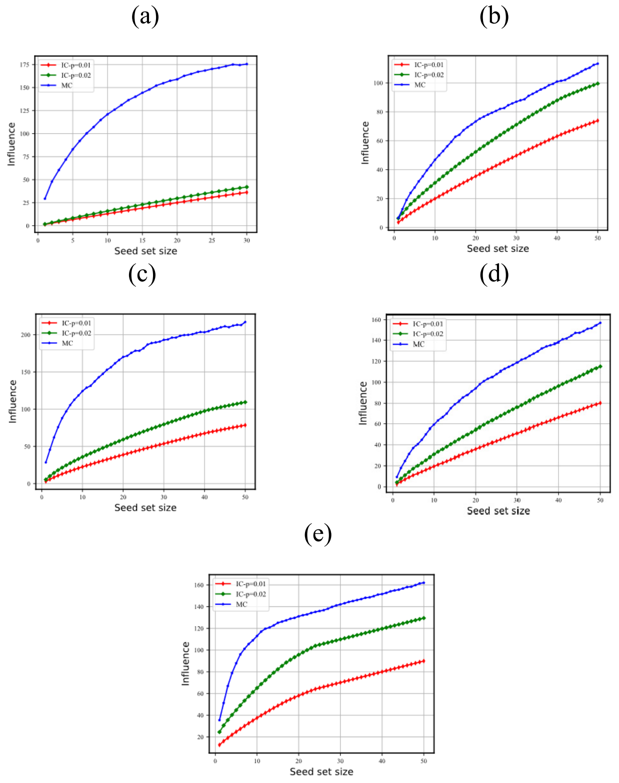

Figure 1. Influence spread of the two models on different data sets. **a** Annotate dataset **b** APS2000-2004 **c** APS2005-2009 **d** APS2010-2014 **e** DBLP2000

Experiment 2 (Influence Diffusion of Different Methods through Meme Cascade Model). The initial seed nodes selected by LDGIM and the baseline method propagate through the meme cascade model on the five data sets respectively. The experimental results are shown in Fig.2, where the horizontal axis represents the number of seed nodes, and the vertical axis represents the influence spread. On the whole, compared with the baseline algorithm, LDGIM has a wider influence spread, faster transmission rate and more stable propagation.

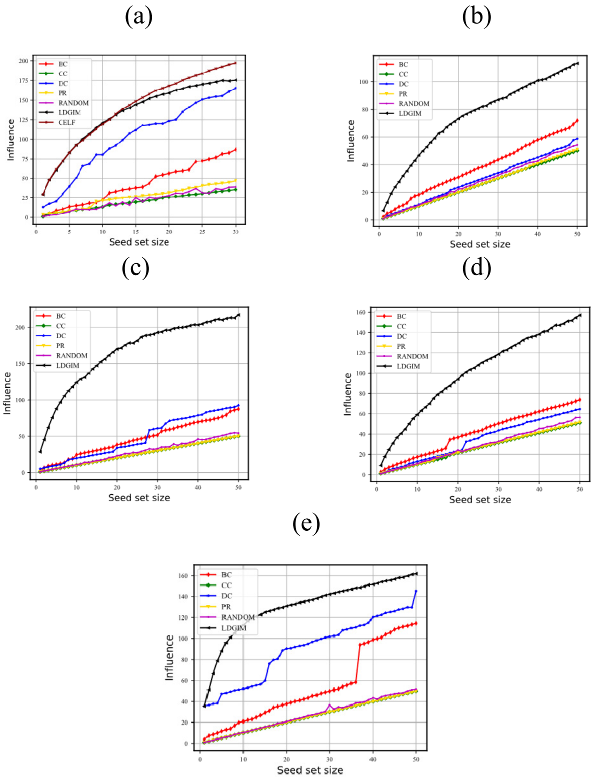

Figure 2. The influence spread of different algorithms on different data sets. **a** Annotate **b** APS2000-2004 **c** APS2005-2009 **d** APS2010-2014 **e** DBLP2000

V. Conclusions

Our paper studies the problem of maximizing the influence of citation networks. We explored a solution to the problem of maximizing the influence of citation networks from a meme perspective. Firstly, a meme cascading model is constructed in this scheme. On this basis, the solution of influence maximization problem is divided into two stages: the candidate stage of selecting candidate seed nodes by LRDiscount and the greedy stage of selecting seed nodes by CELF algorithm. In our study, two possible extensions are as follows: First, the meme cascade model conforms to the knowledge transfer in citation networks and is applicable to most citation networks. so it is applicable to most citation networks. Secondly, LRDiscount provides LeaderRank discount strategy, which is suitable for a wide range of dissemination models. Experimental results show that the proposed meme cascade model is more consistent with the propagation characteristics of citation networks, and the LDGIM method is superior to other heuristic methods. In future research, the influence of semantic information of nodes on seed node mining will be considered.

Acknowledgment

This work is supported by National Natural Science Foundation under Grant No. 62062049, Science Project of Gansu Province under Grant No. 20JR5RA390 and Science and technology project of Lanzhou city under Grant No. 2021-1-150.

References

[1] Richardson, M. and P. Domingos. Mining knowledge-sharing sites for viral marketing. in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. 2002. Edmonton, Alberta, Canada: Association for Computing Machinery.

[2] Kempe, D., J. Kleinberg and É. Tardos. Maximizing the spread of influence through a social network. in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. 2003. Washington, D.C.: Association for Computing Machinery.

[3] Tang, Y., X. Xiao and Y. Shi. Influence maximization: near-optimal time complexity meets practical efficiency. in Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. 2014. Snowbird, Utah, USA: Association for Computing Machinery.

[4] Borgs, C., et al. Maximizing social influence in nearly optimal time. in Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms. 2014: SIAM.

[5] Goyal, A., W. Lu and L.V.S. Lakshmanan. CELF++: optimizing the greedy algorithm for influence maximization in social networks. in Proceedings of the 20th international conference companion on World wide web. 2011. Hyderabad, India: Association for Computing Machinery.

[6] Leskovec, J., et al. Cost-effective outbreak detection in networks. in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. 2007. San Jose, California, USA: Association for Computing Machinery.

[7] Bonacich, P., Factoring and weighting approaches to status scores and clique identification. The Journal of Mathematical Sociology, 1972. 2(1): p. 113-120.

[8] Freeman, L., Centrality in social networks: Conceptual clarification. Social Networks, 1979. 1.

[9] Freeman, L., A Set of Measures of Centrality Based on Betweenness. Sociometry, 1977. 40(1): p. 35-41.

[10] Sun, X. and K. Ding, Identifying and tracking scientific and technological knowledge memes from citation networks of publications and patents. Scientometrics, 2018. 116(3): p. 1735-1748.

[11] Bauckhage, C., K. Kersting and F. Hadiji. Mathematical Models of Fads Explain the Temporal Dynamics of Internet Memes. in International Conference on Weblogs & Social Media. 2013.

[12] Spitzberg, B.H., Toward A Model of Meme Diffusion (M~3D). Communication theory, 2014. 24: p. 311-339.

[13] Leskovec, J., L. Backstrom and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. 2009. Paris, France: Association for Computing Machinery.

[14] Liang, Z., et al., Knowledge Diffusion Pattern Analysis Based on Knowledge Meme Cascade Networks. Information Studies:Theory & Application, 2020. 43(04): p. 40-46+39.

[15] Kuhn, T., M. Perc and D. Helbing, Inheritance Patterns in Citation Networks Reveal Scientific Memes. SSRN Electronic Journal, 2014.