Journal of Electronic Research and Application, 2023, Volume 7, Issue 2
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Integrating Multiple Linear Regression and Infectious Disease Models for Predicting Information Dissemination in Social Networks

Junchao Dong[1], Tinghui Huang[2]*, Liang Min[3], Wenyan Wang[4]

[1]School of Computer Engineering, Guilin University of Electronic Technology, Beihai 536000, Guangxi Zhuang Autonomous Region, China

[2]School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, Guangxi Zhuang Autonomous Region, China

[3]The Youth Innovation Team of Shaanxi Universities, Xi'an Jiaotong University City College, Xi'an 710018, Shaanxi Province, China

[4]School of Computer Sciences, Universiti Sains Malaysia, Gelugor 11800, Penang, Malaysia

***Corresponding author:** Tinghui Huang, glhth@guet.edu.cn

Abstract: Social network is the mainstream medium of current information dissemination, and it is particularly important to accurately predict its propagation law. In this paper, we introduce a social network propagation model integrating multiple linear regression and infectious disease model. Firstly, we proposed the features that affect social network communication from three dimensions. Then, we predicted the node influence via multiple linear regression. Lastly, we used the node influence as the state transition of the infectious disease model to predict the trend of information dissemination in social networks. The experimental results on a real social network dataset showed that the prediction results of the model are consistent with the actual information dissemination trends.

Keywords: Social networks; Epidemic model; Linear regression model

Online publication: May 31, 2023

1. Introduction

With the development of internet technology, social network has become the mainstream medium of information dissemination. In view of the huge impact of information dissemination in social networks on people's life and social development, the analysis of information dissemination in social networks has garnered widespread attention from many researchers. They have attempted to mine, quantify, and predict the spread of information in social networks. These works add to our understanding of the topology, characteristics, and rules of communication of social network from the perspective of information dissemination. It has wide application prospects in various fields, including enterprise marketing, public opinion control, and social business recommendation. Therefore, it is of great research and application value to quantify the node influence, determine the cause of information dissemination in social networks, and to explore the regularities of information dissemination in social networks.

Information dissemination in social networks mainly involves node influence, the maximization of influence, the interpretation and prediction of the propagation law, *etc*. Among them, the interpretation and

prediction of the propagation law involve the independent cascade model, the linear threshold model, the classification model, the game theory model, and the epidemic model. Since the process of information dissemination in social networks is similar to the spread of disease, most of the information dissemination models are derived from the classical susceptible-infected (SI), susceptible-infected-susceptible (SIS), susceptible-infected-recovered (SIR), susceptible-decreasingly infectious-recovered (SDIR), and susceptible-exposed infectious-recovered-susceptible (SEIRS) epidemic models. Tsur *et al*. [1] used linear regression to predict the influence range of tags on Twitter in a certain period of time in combination with content feature, topology, and time factor. Xiao *et al*. [2] analyzed the causes of influence from two aspects, personal memory and user interaction, and proposed a method to measure the social influence of users based on the multiple linear regression model. Li *et al*. [3] proposed an information communication model based on heterogeneous mean field and evolutionary game, taking into account the real topological relationship between the participants and the psychological characteristics of users. The above research proves that it is reasonable to measure node influence from three dimensions: topology structure, user interaction behavior, and information content; however, the studies have failed to construct a node influence measurement model embodying the aforementioned three dimensions.

Given the few studies on state transition probability in the information communication model and the incomplete consideration of the influencing factors in the measurement of node influence, we present a method for measuring the influence of multi-dimensional nodes based on the multiple linear regression model. Taking node influence as the cause of state transition in the epidemic model, we analyzed the factors affecting information dissemination and described the trend of information dissemination in social networks. Lastly, we verified the feasibility of this method by experiments on a real data set. The main contributions of this paper can be summarized as follows:

(i) analysis of the influencing factors of information dissemination in social networks; construction of a measurement model that integrates three dimensions (topology, user interaction behavior, and information content) based on multiple linear regression;
(ii) proposal of a new SIR information dissemination model based on node influence to model the information dissemination process and predict its future propagation trend;
(iii) our method is more in line with the actual situation of information dissemination than existing baseline methods.

2. Problem description

2.1. Description of social network

Social networks are often represented by graph $G=\{V,E\}$; $V=\{v_1,v_2,\cdots,v_n\}$ is the node set, that is, the users set, where $v_i(1\le i\le n)$ is a user (node), and $|V|=n$ represents the size of the node set, that is, the total number of users; $E\subseteq V\times V$ is the edge set, that is, the following relationship between users, where e_{ij} represents an edge, that is, user v_i follows v_j.

2.2. Description of information communication

Given a social network, $G=\{V,E\}$, information travels along its edges – the process of information communication in the epidemic model (SIR). At different times, users in the social network are in one of the three states in the epidemic model, namely the unknown (S), who may participate in information dissemination; the already known (I), who conducts information dissemination behavior; and the immune (R), who loses interest in information and no longer produces communication. Among them, the transition between states is determined by the node influence.

3. Model framework

As shown in **Figure 1**, we proposed an information dissemination model of multi-dimensional node influence by combining social user relationships and historical data. The features that influence information dissemination in social networks were extracted multi-dimensionally, and the node influence features were constructed based on the multi-dimension linear regression model. The quantization value of node influence features was taken as the state transition probability of the epidemic disease model to establish the dynamic differential equation and describe the trend of information dissemination as well as the change of group status in social networks at different times.

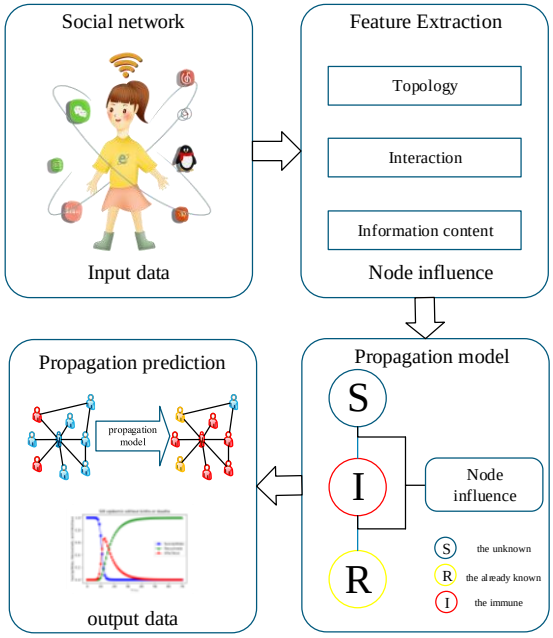

Figure 1. Information dissemination model framework

3.1. Feature extraction

Information dissemination in social networks is a complex and dynamic process, which is affected by multi-dimensional factors, including topology structure, user interaction behavior, information content, *etc*. We proposed the quantification features of node influence from the aforementioned three dimensions, as shown in **Figure 2**.

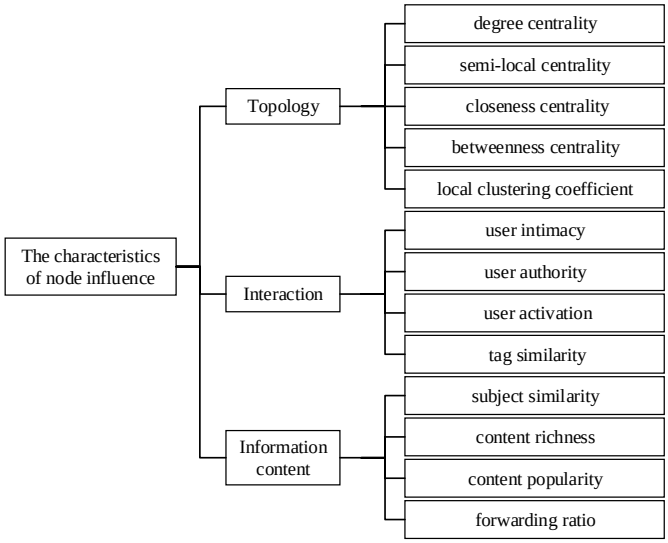

Figure 2. Node influence feature framework

3.2. Feature processing

We normalized the feature extracted in section 3.1. The data were normalized to [0,1]. The normalized feature data were represented as the feature vectors of topology structure (T), user interaction behavior (B), and information content (M), whose symbolic meanings are shown in **Table 1**, and the vector forms are shown in equation (1). The dimensions of a vector were defined by d_T=5, d_B=4, and d_M=4, *i.e.*, the number of features of topology, interaction, and information content.

Table 1. Node influence features

Dimension	Feature	Symbol
Topology (T)	degree centrality	t1
	semi-local centrality	t2
	closeness centrality	t3
	betweenness centrality	t4
	local clustering coefficient	t5
Interaction behavior (B)	user intimacy	b1
	user authority	b2
	user activation	b3
	tag similarity	b4
Information content (M)	subject similarity	m1
	content richness	m2
	content popularity	m3
	forwarding ratio	m4

$$\begin{aligned} T&=\left(\mathrm{t}_1,\mathrm{t}_2,\mathrm{t}_3,\mathrm{t}_4,\mathrm{t}_5\right)^{\mathrm{T}} \\ B&=\left(b_1,b_2,b_3,b_4,b_5\right)^{\mathrm{T}} \\ M&=\left(m_1,m_2,m_3,m_4\right)^{\mathrm{T}} \end{aligned} \tag{1}$$

3.3. Feature fusion

According to section 3.1., the node influence that affects information communication is not only related to the features of the topology in social networks, but also to the features of user interaction and information content. Multiple linear regression was used to find a linear equation from the historical data in the social network to describe the relationship between the multi-features (independent variables) of topology structure, user interaction behavior, and information content and the node influence (dependent variables). The linear equation was used to predict the new unknown node influence. The mathematical form of the model is shown in equation (2).

$$\mathrm{f}_{\mathrm{ri}}(v_i)=\omega_1 T+\omega_2 B+\omega_3 M+b \tag{2}$$

where $\omega_i=\left\{\omega_1,\omega_2,\omega_3\right\}$ is the weight of the node influence feature dimension. The greater the weight, the greater the influence of the feature dimension on the node influence. B is the bias and scalar.

3.4. SIR information dissemination model

In order to demonstrate the influence of node influence on information dissemination in social networks, the classical SIR epidemic model was used to simulate the information dissemination process [4,5].

Considering the complexity and uncertainty of information dissemination, the following assumptions were made when constructing the SIR information dissemination model:

(i) information dissemination in social networks is time-sensitive; it is assumed that the number of users in social networks is N, which remains unchanged in a period of time, that is, S + R + I = 1;

(ii) users in the social network will definitely forward when they reach the threshold of node influence after they come into contact with the information;

(iii) after the forwarding behavior generated by the user, if there is no longer forwarding after a period of time, the user will become an immune user.

Node influence probability $p(v_i)$ is expressed by sigmoid function:

$$p(v_i)=\frac{1}{1+\exp\{-f_{ri}(v_i)\}} \tag{3}$$

where, $f_{ri}(v_i)$ is the value from the output of the node influence model in section 3.3.

When node v_i has n neighbor nodes, the probability of the number of nodes with state changes conforms to the binomial distribution.

$$P_{(x=k)} = C_n^k p(v_i)^k (1-p(v_i))^{n-k} \tag{4}$$

where k = (0, 1, 2,…, n).

Therefore, the probability of any node with state change in the social network is as follows:

$$\lambda(t) = \sum_{k=0}^{n} \frac{kC_n^k p(v_i)^k (1-p(v_i))^{n-k}}{n} \tag{5}$$

where k = (0, 1, 2,…, n).

According to equation (3), the probability that the state of any node in the social network does not change is as follows:

$$\mu(t) = 1-\lambda(t) \tag{6}$$

According to the dynamics equation of the epidemic model, when the node in the social network is in the state of information unknown, it is converted into the state of information known with a probability of $\lambda(t)$ and the state of information immunity with a probability of $\mu(t)$. Therefore, the dynamic equation of information dissemination based on node influence is as follows:

$$\begin{cases} \dfrac{dS}{dt} = -\lambda(t)S(t)I(t) \\ \dfrac{dI}{dt} = \lambda(t)S(t)I(t) - \mu(t)I(t) \\ \dfrac{dR}{dt} = \mu(t)I(t) \\ S + R + I = 1 \end{cases} \tag{7}$$

4. Experiment

In order to verify the model proposed in this paper, we evaluated the effect of the proposed model from two aspects on a real data set from Sina Weibo [6]: (i) analysis of information dissemination process (research on the changes of SIR nodes in the model proposed in this paper); (ii) information dissemination fitting experiment to verify the fitting degree of the model proposed in this paper with the real communication process. The total number of nodes in the dataset is 10,000, and the number of edges is 62,475.

4.1. Analysis of information dissemination process

One node in the social network was set as the node with known information, while the other nodes were nodes with unknown information. The model parameters were as follow: $\lambda(t) = 0.65$; $\mu(t) = 0.35$. The changes in trend of the number of known nodes, unknown nodes, and immune nodes in the social network with time were analyzed. The experimental results are shown in **Figure 3**.

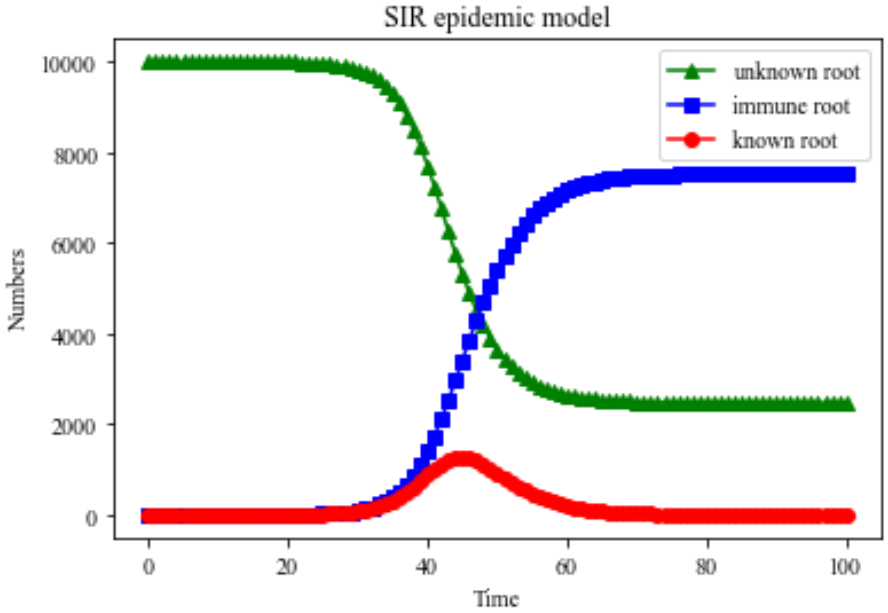

Figure 3. Multi-dimension node influence information dissemination model

It can be seen from **Figure 3** that the number of nodes with known information is 1 and the number of nodes immune to information is 0 at the initial time. With the information known nodes spreading information, the number of unknown nodes decreased sharply at about 30, and the information began to accelerate the propagation. With the passage of time and the deepening of information dissemination, information immune nodes began to appear around 40, and the number of information immune nodes increased rapidly with time. When the number of nodes with known information reached the peak at about 45, the number of immune nodes increased rapidly, while the number of unknown nodes decreased rapidly. At the time point of 100, the information immune node is 10,000, while the information unknown node and information known node are both 0, indicating that the information propagation process is over.

4.2. Information dissemination fitting experiment

In order to verify the performance of this model, the topic "good method of planning app promotion scheme" in Sina Weibo was chosen. The experimental model parameters were as follows: $\lambda(t) = 0.75$; $\mu(t) = 0.25$; and the time was 12 hours. The experimental results are shown in **Table 2**.

Table 2. Comparison of the number of known nodes between the model and real information propagation

Time point	1	2	3	4	5	6	7	8	9	10	11	12
Model in this paper	1	200	500	1,500	4,215	7,825	1,144	3,248	10,012	25,025	54,686	100,999
Actual propagation	1	100	375	10,258	37,465	5,412	1,002	2,749	8,899	24,986	50,326	102,689

It can be seen from **Table 2** that there is a small gap between the number of information known nodes predicted by the model and those of real information dissemination at different time points. The results showed that our model has a good fit with the information dissemination trend in the real social network.

5. Conclusion

In this paper, we introduce a model of information dissemination in social networks based on multi-dimensional node influence. We analyzed the factors that affect node influence and proposed the features of node influence based on topology, user interaction behavior, and information content. In addition, we proposed a node influence measurement model based on multiple linear regression and an information communication model based on SIR to provide reference and new methods for relevant research on extraction of influencing factors, construction of influencing features, and construction of information communication models during the process of information dissemination in social networks. The effectiveness of the proposed method was verified on a real Sina Weibo dataset. In our future work, we will propose more feature dimension and further explore the law of information dissemination in social networks.

Funding

This work was supported by the 2021 Project of the "14th Five-Year Plan" of Shaanxi Education Science "Research on the Application of Educational Data Mining in Applied Undergraduate Teaching—Taking the Course of 'Computer Application Technology' as an Example" (SGH21Y0403), the Teaching Reform and Research Projects for Practical Teaching in 2022 "Research on Practical Teaching of Applied Undergraduate Projects Based on 'Combination of Courses and Certificates"—Taking Computer Application Technology Courses as an Example" (SJJG02012), and the 11th batch of Teaching Reform Research Project of Xi'an Jiaotong University City College "Project-Driven Cultivation and Research on Information Literacy of Applied Undergraduate Students in the Information Times—Taking Computer Application Technology Course Teaching as an Example" (111001).

Disclosure statement

The authors declare no conflict of interest.

References

[1] Tsur O, Rappoport A, 2012, What's in a Hashtag?: Content Based Prediction of the Spread of Ideas in Microblogging Communities. Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, New York, 643–652.

[2] Xiao Y, Song C, Liu Y, 2019, Social Hotspot Propagation Dynamics Model Based on Multidimensional Attributes and Evolutionary Games. Communications in Nonlinear Science and Numerical Simulation, 67: 13–25.

[3] Li Q, Song C, Wu B, et al., 2018, Social Hotspot Propagation Dynamics Model Based on

Heterogeneous Mean Field and Evolutionary Games. Physica A: Statistical Mechanics and Its Applications, 508: 324–341.

[4] Liu X, He D, Yang L, et al., 2019, A Novel Negative Feedback Information Dissemination Model Based on Online Social Network. Physica A: Statistical Mechanics and Its Applications, 513: 371–389.

[5] Xiao Y, Chen D, Wei S, et al., 2019, Rumor Propagation Dynamic Model Based on Evolutionary Game and Anti-Rumor. Nonlinear Dynamics, 95: 523–539.

[6] Chen H, Liu J, Lv Y, et al., 2018, Semi-Supervised Clue Fusion for Spammer Detection in Sina Weibo. Information Fusion, 44: 22–32.