A social commerce information propagation prediction model based on transformer

Huibing Zhang, *Jingrui* Dai, *Junfei* He, and *Huacheng* Zhang[*]

Guangxi Key Lab of Trusted Software, Guilin University of Electronic Technology, 541004 Guilin Guangxi, China

Abstract. Precise forecasts of the propagation patterns of social commerce information play a crucial part in precision marketing. Traditional forecast relies on machine learning diffusion models, in which the forecast accuracy is dependent on the quality of the designed features. Researchers using these models are required to have experience in this regard, but due to the complexity and variations of real-world social commerce information propagation, design of features for the prediction model turns out difficult and is likely to cause local or universal errors in the model. To address these problems, this study proposed an information propagation prediction model based on Transformer. First, the fully-connected neural network was employed to code the user nodes to low-dimension vectors; then, Transformer was employed to perform information of the user-node vectors; last, the output of the Transformer was uploaded to the output layer to forecast the next user node in information propagation. The model was tested on data sets obtained from Sina Weibo, and the test result shows that the proposed model outperformed baseline models in terms of the indicators of Acc@k and MRR.

1 Introduction

Rapid development of Internet technology has greatly facilitated information propagation and utilization. To probe into and forecast information propagation on social media, a major channel for information propagation, is of great importance for product promotion and marketing. Information cascade forecast emerges because of the initiatives to promote and market products on social media. A previous study [1] proposed a model parameter learning algorithm based on real-world data, and found that the nodes with the highest propagation strength are not necessarily the nodes of strongest connectivity or the highest centrality, but nodes at the "core" of the network. The "core" nodes in the network were obtained based on specified parameters introduced to the training model. When applied to personal information propagation in blogs, the model showed high efficiency in detecting trending topics on the blog. Jackson et al. [2] proposed a method that merged information diffusion selection bias and the Galton–Watson process, which solved the information diffusion problem that the information reaches the target node through multiple

---

[*] Corresponding author: glietzhc@guet.edu.cn

intermediate nodes. Qiu et al. [3] designed DEEPINF, a graphics-based deep learning model, which takes the user's local network as the input, collects samples via random walk, and realizes learning and forecast by graphics convolutional network and graphics attention network. Ma compared the performance of five methods, i.e., Naive Bayes, K-nearest neighbor, decision tree, SVM and logic regression, in forecasting the popularity degree of new tags on Twitter. The result shows that classifiers based on feature extraction performed better, and the context features had large impact than content features on the classification result. Gabor et al. [5] used the earlier views of content of users provided by social websites to forecast the long-term popularity of the content, and found that their proposed model performed better in forecasting outdated content than in forecasting constantly updated content. The aforementioned studies relied on feature extraction to perform forecasts, but had poor performance in extracting complete cascade information.

There are three types of conventional information propagation prediction models — feature-based models, generative models, and diffusion-based models [6]. Feature-based models classify the manually collected datasets into different features like information content and attribute content, and upload these features into a machine learning algorithm to perform forecast. Weng et al. [7] compared information propagation to complex spread of infectious diseases, and employed the random forecast classifier to forecast the propagation effect in the early stage of information propagation in communities. Generative models are principally used to forecast the forward quantities or popularity of information on social media. Bao et al. [8] proposed the SEHP model to forecast the popularity of a single Weibo article, and found that the model performed better than enhanced Poisson process models. Diffusion-based prediction models simplify the complex networks based on hypotheses to perform forecast tasks. Jure et al. [9] adopted the survival theory and proposed additive and multiplicative models to realize effective reasoning of networks. These models apply not only to regular reasoning models, but also to conventional models.

Wide adoption of deep learning in different fields in these years have led to many achievements in deep learning-based information propagation forecast. Manavoglue et al. [10] proposed a customized behavior model to recognize complex behavior patterns of users, and proved that user behavior models based on the largest entropy and Markov had the strongest performance. Mikolov et al. [11] put forward a RNNLM prediction model based on recurrent neural networks, and found that the model could achieve higher accuracy than other models. Wang et al. [12] proposed the IARNN-GATE model, an RNN model with the attention mechanism; the model added the attention information to each function, optimized the IARNN-WORD model for word input, and showed good performance in information forecast. It is fair to conclude that deep learning models have good performance in information propagation forecast. The reason is that compared with machine learning models that rely on manual extraction of features, the deep learning models can abstract the complex network information propagation into sequence modelling, which thus preclude the errors of manual extraction of features and ensure integrity of information cascade sequences.

Therefore, the present study proposes an information propagation prediction model based on Transformer. First, the input user nodes in the propagation route were coded by a fully-connected neural network to obtain the vectors of user nodes; then, the user-node vectors were input to the Transformer for information representation; at last, the Softmax layer was employed to forecast the next user nodes in the information propagation route. Experiments on Sina Weibo data showed that the proposed method had good performance in forecasting information propagation.

2 Problem definition and methods

2.1 Problem definition

Given a user node set that consists of n users $U = \{u_i\}\,(i = 1, 2, \cdots, n)$, and the propagation route $L = \{p_1, p_2, \cdots, p_k\}$, where $p_i \in U$. Forecasting the social commerce information propagation trend is to identify a propagation route L based on the propagation patterns of the former k users to forecast the probability P_i that the next user node is u_i.

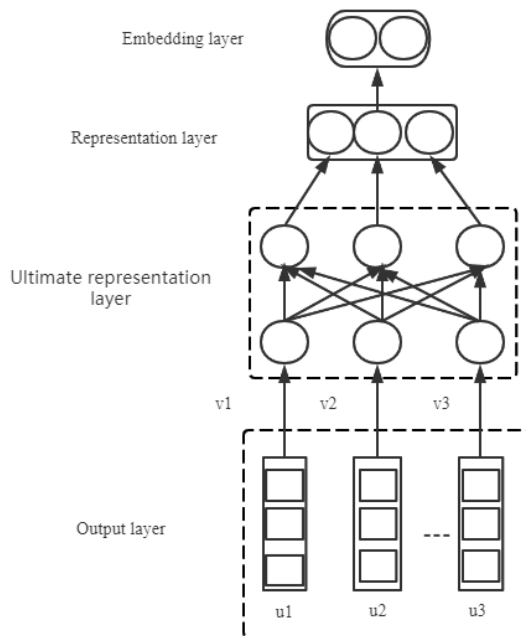### 2.2 Transformer-based information propagation model

Fig. 1. Structure of the Transformer-based information propagation model.

Figure 1 shows the structure of a transformer-based information propagation model. It consists of an input layer that consists of vectors obtained by coded user nodes in the user set, an information representation layer that extracts features of the propagation process, and an output layer that classifies and forecasts the next user node.

2.2.1 Input layer

The one-hot coding method was employed to code the user nodes, which were then uploaded to the fully-connected neural network. Each user node in the propagation route was projected into a low-dimensional vector, $v_i \in R^{d_u}$, where d_u is the dimension of the user node. The vector of the set of user nodes in the propagation route is $\{v_1^s, v_2^s, \cdots, v_n^s\} \in R^{n \times d_u}$.

2.2.2 Information representation layer

The information representation layer is mainly the Transformer coder, as shown in Figure 2. The Transformer consists of three modules: the location coder, the multi-head attention module, and the feedforward network. The multi-head attention module and the feedforward network are connected by the residual layer.
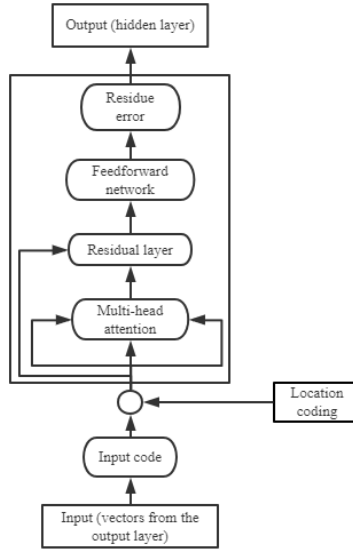

Fig. 2. Structure of Transformer.

1) Location coding

For the vectors obtained by the fully-connected neural network, the location coding is required to market the input sequence. As the Transformer uses the universal information, the sequence information in the route cannot be utilized. For the propagation route, the route from u_2 to u_1 is different from that from u_1 to u_2, so the location of the user node in the route is of great importance. Thus, we introduced the location code to mark the input sequence of the user nodes. The trigonometric function was used to mark the location, as shown in Eq. 1. The trigonometric function can accurately represent the location of the input nodes and the maximum length of the input sequence, as well as the changes in the sequence of the user nodes as the time proceeded.

$$\begin{cases} P_{(p,2i+1)} = \sin(p/1000^{2i/d}) \\ P_{(p,2i+1)} = \cos(p/1000^{2i/d}) \end{cases} \tag{1}$$

where P represents the location of the user node, and i is the dimension of the code.

2) Multi-head attention

The in-built multi-head attention mechanism of Transformer was employed to represent the interactions between user nodes in the propagation route. The multi-head attention mechanism consists of multiple self-attention. The self-attention mechanism, on the basis of the current nodes, observes all nodes in the input route, and obtains the contribution of other nodes to the code of the current node. The self-attention mechanism is used to obtain the importance degree of each node to the propagation route. Eq. 2 is obtained by linear transformation of the input sequence.

$$Q = W_q \cdot x$$
$$K = W_k \cdot x \tag{2}$$
$$V = W_v \cdot x$$

where W_q, W_k, W_v is the weighted matrix obtained through training, x is the input matrix, Q is the query matrix, K and V are the key values. Q was then multiplied by K to obtain the score of each node. At last, the score was multiplied by the value of V so that the model could focus on the important nodes in the propagation sequence. Details of calculation are shown by Eq. 3.

$$Attention(Q, K, V) = soft \max(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

where d_k is the column number of Q and K vectors, i.e., the dimension of the vector.

The multi-head attention mechanism is to obtain different matrices through multiple self-attention calculation. Compared with the self-attention mechanism, the multi-head attention mechanism pays more attention other user nodes in the propagation sequence and thus improve the model's performance. Multi-head attention calculation is to connect the matrices obtained through multiple self-attention operations.

$$MultiHead(Q, K, V) = Concat(head_1, head_2, ..., head_n) \tag{4}$$

3) Feedforward Network

The attention module and the feedforward network are connected by the residual network. The residual error is usually used to solve the degradation problem caused by multiple layers. The feedforward network usually consists of two fully-connected layers. The first fully-connected layer contains a Relu activation function, as shown in Eq. 5.

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{5}$$

where W_1 and W_2 are the weighted matrices, and b_1 and b_2 are the errors.

2.3 Output layer and model training

On the output layer, the representation obtained by the Transformer model is uploaded to the Softmax layer to forecast the next user node in the information propagation route, as shown by Eq. 6.

$$p(u_i \mid L_k) = soft \max(W_c L_k + b_c) \tag{6}$$

where p is the probability that the (k+1)-th node in the propagation route L is u_i, u_i, represents the (k+1)-th user in the route L, W_c is the weighted matrix, b_c is the variance. The minus-log-likelihood function is used to train the model, as shown by Eq. 7. The backpropagation algorithm is used to optimize the model.

$$\varsigma(U) = -\sum_{m=1}^{M} \sum_{i=1}^{n_m - 1} \log p(u_i \mid L_k) \tag{7}$$

3 Experiment

3.1 Dataset

The dataset used in the present study is the Sina Weibo data from [13]. The dataset consists of 221,473 users' posts from July 2014 to April 2016. The dataset involves five categories, i.e., sports, science, politics, business and music. The user posts under the category of

business were labelled manually. The forwarding of each post was considered a round of propagation, and each forwarding operation is represented by the website link of corresponding timestamp. Table 1 shows the details of the dataset.

Table 1. Details of the dataset used in the present study.

Dataset	User node	User connection	Propagation route	Average length of the propagation route
Sina Weibo	12568	42586	425	40.23

3.2 Evaluation indicators

The accuracy of the first k-th ranking (Acc@k) and the MRR in Eq. 8 were used as the two evaluation indicators to assess the model's performance [12, 14]. Acc@k and MRR are positively correlated to the performance of the model; the larger the values of these two indicators, the better the performance of the model.

$$MRR = \frac{1}{N} \sum_{i=1}^{|N|} \frac{1}{r_i} \tag{8}$$

where $|N|$ is the number of user nodes, and r_i is the user node of the *i*-th user.

3.3 Experiment result and analysis

To assess the accuracy of the Transformer-based information propagation prediction model, the following four models were used as the baseline models for comparison.
    RNN: the basic recurrent neural network;
    LSTM: the long short-term memory model. The set of user nodes is coded as input, and the gate function is used to control information transfer of the previous time;
    ATT-RNN: the RNN network with attention mechanism;
    ATT-LSTM: the LSTM model that combines the attention mechanism.
    The proposed Transformer model and the aforementioned four baseline models were tested on the Sina Weibo dataset, as shown in Figure 3.
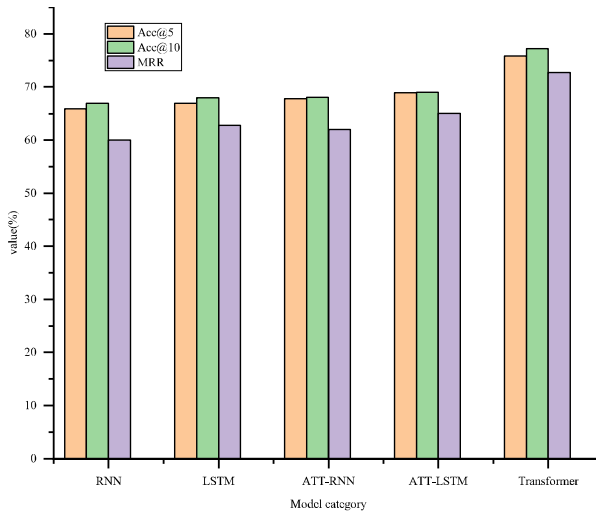

Fig. 3. Forecast results of the five models.

As Figure 3 shows, the Transformer model achieved higher values for three indicators than the baseline models, because Transformer could obtain the time series of the bi-directionally propagated information. The performance of the model that had the attention mechanism performed better than those without the attention mechanism because the attention mechanism could adjust the contribution of single propagation operation in real time in the information propagation route. Besides, the LSTM model performed better than the RNN model mainly because of the dataset selected for this experiment. The dataset used in this study are historical data on Sina Weibo, which are standard time series data. Experiments conducted in different fields have revealed that the LSTM models are more suitable for analysis and forecast of time series data than the RNN models.

4 Conclusion

Most traditional propagation prediction models rely excessively on prior knowledge and the feature quality. To address this problem, the present study proposed a Transformer-based information propagation prediction model. The multi-head attention mechanism in the model can fully learn the valid information in the propagation, and pays special attention to the important nodes to generate more accurate representation of user nodes. The experiments show that the proposed model performed well in forecasting the next user node. However, this study did not consider some attributes and relations of each user node attachment in the information propagation route. In future studies, we plan to introduce graphics neural network into the Transformer model to construct a more accurate and applicable information propagation prediction model.

References

1. D. Gruhl, R. Guha, D. Liben-Nowell, A. Tomkins. Information diffusion through blogspace. Proceedings of the 13th international conference on World Wide Web. 491-501(2004).
2. Golub, Benjamin, Jackson, O. Matthew, Using selection bias to explain the observed structure of internet diffusions. Proceedings of the National Academy of Sciences.107(24): 10833-10836(2010).
3. J. Qiu, J. Tang, H. Ma, Deepinf: Social influence prediction with deep learning. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2110-2119(2018).
4. Z. Ma, A. Sun, G. Cong. On predicting the popularity of newly emerging hashtags in T witter. Journal of the American Society for Information Science and Technology.64(7): 1399-1410(2013).
5. G. Szabo, B. A. Huberman, Predicting the popularity of online content. Communications of the ACM.53(8): 80-88(2010).
6. Z. Y. Zhang, F. Z. Zhang, Q. Tan, et al. Review of Information Cascade Prediction Methods Based on Deep Learning. Computer science, 47(7): 141-153. (in Chinese)
7. L. Weng, F. Menczer, Y. Y. Ahn. Virality prediction and community structure in social networks. Scientific reports. 3: 2522(2013).

8. Q. Zhao, M. A. Erdogdu, H. Y. He, et al. Seismic: A self-exciting point process model for predicting tweet popularity. Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. 1513-1522(2015).

9. M. Gomez-Rodriguez, J. Leskovec, B. Schölkopf. Modeling information propagation with survival theory. International Conference on Machine Learning. 666-674(2013).

10. E. Manavoglu, D. Pavlov,L. C. Giles, Probabilistic user behavior models. Third IEEE International Conference on Data Mining. IEEE.203-210(2003).

11. T. Mikolov, S. Kombrink, L.Burget, et al. Extensions of recurrent neural network language model. 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 5528-5531(2011).

12. Y. Wang, H. Shen, S. Liu, et al. Cascade Dynamics Modeling with Attention-based Recurrent Neural Network. IJCAI. 2985-2991(2017).

13. H. Chen, J. Liu, Y. Lv, et al. Semi-supervised clue fusion for spammer detection in Sina Weibo. Information Fusion.44: 22-32(2018).

14. Z. Wang, W. Li. Hierarchical Diffusion Attention Network. IJCAI.3828-3834(2019).