Linking Transformer to Hawkes Process for Information Cascade Prediction (Student Abstract)

Liu Yu[1], **Xovee Xu**[1], **Ting Zhong**[1], **Goce Trajcevski**[2], **Fan Zhou**[1*]

[1] University of Electronic Science and Technology of China
[2] Iowa State University
liu.yu@std.uestc.edu.cn, xovee@ieee.org, zhongting@uestc.edu.cn, gocet25@iastate.edu, fan.zhou@uestc.edu.cn

Abstract

Information cascade is typically formalized as a process of (simplified) discrete sequence of events, and recent approaches have tackled its prediction via variants of recurrent neural networks. However, the information diffusion process is essentially an evolving directed acyclic graph (DAG) in the continuous-time domain. In this paper, we propose a transformer enhanced Hawkes process (**Hawkesformer**), which links the hierarchical attention mechanism with Hawkes process to model the arrival stream of discrete events continuously. A two-level attention architecture is used to parameterize the intensity function of Hawkesformer, which captures the long-term dependencies between nodes in graph and better embeds the cascade evolution rate for modeling short-term outbreaks. Experimental results demonstrate the significant improvements of Hawkesformer over the state-of-the-art.

Introduction

Sharing content through social media platforms such as Twitter and Weibo has become a main channel to express individual opinions. The initial tweet, along with the subsequent retweets forms an *information cascade* (Zhou et al. 2021). Predicting the *size* of an information cascade after a certain time-period is one of the typical tasks. The core challenge is how to model the underlying diffusion process governing the popularity dynamics of information cascades.

Prior *Temporal Point Processes* (TPPs) based methods and *Deep Learning*-based approaches especially recurrent neural network (RNN) based sequential models exhibit three notable drawbacks: **(i)** they use simple intensity functions and make strong assumptions on the diffusion mechanism; **(ii)** the irregular time intervals between events and their order are important to describe the diffusion dynamics, however, they are difficult to capture in a way from discrete-time domain; **(iii)** they cannot fully exploit the diffusion process of DAGs and capture the long-term dependency due to the intrinsic limitations of recurrent models; and **(iv)** previous works (Tang et al. 2021) have shown that information cascades with short-term outbreaks are more likely to be popular in the future. They cannot capture such trends intrinsically and RNN-based models are prone to face error accumulation issue, especially when the cascade size is large.

*Corresponding author

Methodology

Let C denote an information cascade with a retweet history $\mathcal{H} = \{(t_j, u_j)\}_{j=1}^{L}$ with $t_j \in [0, t_o)$, where each pair (t_j, u_j) corresponds to infected user u_j occurs at time t_j. L denotes the number of retweets that have occurred up to the observation time t_o. Generally, we focus on the information cascade prediction as a complex and nonlinear regression task, predicting the popularity size $P(t_p|\mathcal{H})$ of a cascade over a period of time, where t_p is the time of prediction.

Hawkesformer provides an elegant mathematical tool for modeling event occurrence in continuous-time domain and takes the underlying DAG into consideration, which links a two-level attention (Vaswani et al. 2017) framework to parameterize the intensity function of Hawkes process (Hawkes 1971). We characterize the long-term dependence as $\mathbf{h}(t)$ and short-term outbreak as $\boldsymbol{\vartheta}(t)$. Given an information cascade C – e.g., a tweet, and its retweet history $\mathcal{H} = \{(t_j, u_j)\}_{j=1}^{L}$ – we model the continuous dynamics of temporal point processes, with the following conditional intensity function:

$$\lambda(t|\theta, \mathcal{H}) = f\Big(\underbrace{\alpha \frac{t - t_j}{t_j}}_{\textit{current}} + \underbrace{\boldsymbol{w_1}^\top \mathbf{h}(t_j)}_{\textit{long-term}} + \underbrace{\boldsymbol{w_2}^\top \boldsymbol{\vartheta}(t_j)}_{\textit{short-term}} \Big) \quad (1)$$

where θ denotes the model parameters; retweet time t is defined on interval $[t_j, t_{j+1})$; and $f(\cdot)$ is a softplus function to constrain the intensity function to be positive. The first term *current* indicates the evolutionary process in the continuous-time domain; $\boldsymbol{w_1}, \boldsymbol{w_2}$ are the learned weights for long-term dependency and shot-term outbreak at time t respectively; and $\lambda(t)$ denotes the arrival rate of a retweet at time t.

Specifically, at the first level, we design a long-term dependency module to dynamically capture the diffusion process, where we make a *primary/non-primary* path assumption to adaptively integrate the diffusion process on the underlying DAG, see Figure 1. This also enables each node attending the cascade at any position to update its current hidden states. At the second level, we design a short-term outbreak module which considers the evolution rate of a cascade and learns the local patterns in a time-slice window.

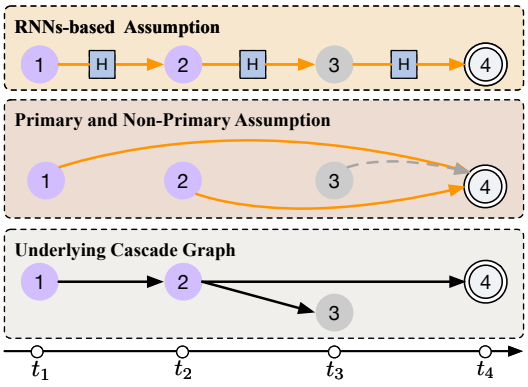

Figure 1: Our primary and non-primary assumption: an example for current node u_4. Purple (gray) dots denote nodes on the primary (non-primary) path.

Prediction and Optimization: With the proposed intensity function, the likelihood of observing an event at time t and the popularity prediction are respectively defined as:

$$p(t \mid \mathcal{H}_t) = \lambda(t \mid \mathcal{H}_t) \exp\Big(- \int_{t_o}^{t} \lambda(\tau \mid \mathcal{H}_\tau) d\tau \Big) \quad (2)$$

$$\log_2 \hat{y} = \underbrace{\int_{t_o}^{t_p} t \cdot p\left(t \mid \mathcal{H}_t\right) dt}_{\Lambda} + FC\left(\mathbf{h}\left(t_L\right) \oplus \boldsymbol{\vartheta}(t_L)\right) \quad (3)$$

where t_p is the prediction time and t_o is the observation time.

We concatenate the last hidden states of two parts $\mathbf{h}\left(t_L\right)$ and $\boldsymbol{\vartheta}(t_L)$, plus the integral term Λ and feed it to a fully-connected layer to get predicted incremental popularity $\hat{y}$. Suppose we have N information cascades, the joint likelihood of observing retweets up to an observation window t_o and our optimization loss are:

$$\ell(C) = \underbrace{\sum_{j=1}^{L} \log \lambda(t_j \mid \mathcal{H}_j)}_{\text{retweet}} - \underbrace{\int_{t_1}^{t_o} \lambda(t \mid \mathcal{H}_t) dt}_{\text{non-retweet}}, \quad (4)$$

$$\mathcal{L}\left(y_i, \hat{y}_i\right) = \frac{1}{N} \sum_{i=1}^{N} \Big(\left(\log_2 y_i - \log_2 \hat{y}_i\right)^2 - \log \ell(C_i) \Big)$$

Experiment

We compare several strong baselines over Weibo and ASP datasets with evaluation metrics MSLE and MAPE. The observation (prediction) times of Weibo and APS datasets are set to 0.5 (24) hour and 5 (20) years, respectively. As shown in Table 1, the superior performance of Hawkesformer lies in its consideration of both long-term dependencies and short-term outbreaks in continuous-time domain. We also verify their benefit by ablating *Long Dependency* and *Short Outbreak* module respectively, denoting – `Long` and – `Short`. As visualized in Figure 2, the left shows our model pays more attention to the primary path, which confirms our assumption that primary path is more important; and the right shows our attention module is also able to model the both short-term and long-term dependencies for different nodes located in different position and/or time.

Model	Weibo		APS	
	MSLE	MAPE	MSLE	MAPE
SEISMIC	4.678	0.412	1.736	0.325
DeepHawkes	2.556	0.320	1.576	0.295
VaCas	2.032	0.246	1.337	0.279
TempCas	2.135	0.263	1.321	0.272
Ours	**1.825**	**0.223**	**1.179**	**0.206**
– `Long`	2.907	0.359	1.616	0.293
– `Short`	2.135	0.261	1.328	0.237

Table 1: Overall prediction performance comparison.

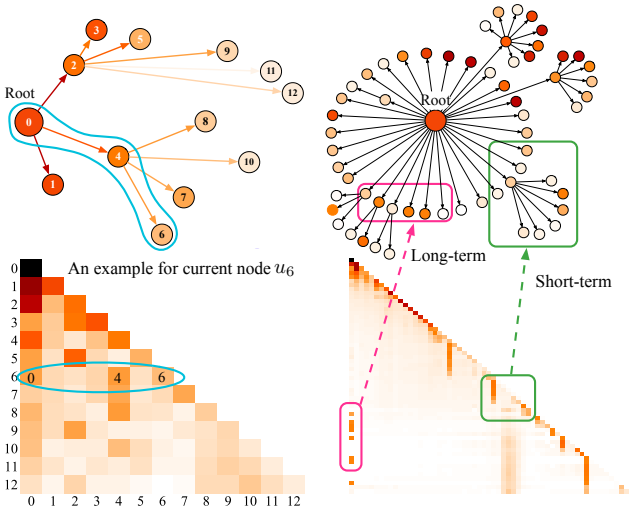

Figure 2: Attention pattern visualization. Left: short diffusion (ID: 85840); Right: popular diffusion (ID: 34484).

Acknowledgements

This work was supported by National Natural Science Foundation of China (Grant No. 62072077 and No. 62176043), and National Science Foundation SWIFT (Grant No. 2030249).

References

Hawkes, A. G. 1971. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1): 83–90.

Tang, X.; Liao, D.; Huang, W.; Xu, J.; Zhu, L.; and Shen, M. 2021. Fully Exploiting Cascade Graphs for Real-time Forwarding Prediction. In *AAAI*, volume 35, 582–590.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*, 5998–6008.

Zhou, F.; Xu, X.; Trajcevski, G.; and Zhang, K. 2021. A survey of information cascade analysis: Models, predictions, and recent advances. *ACM Computing Surveys*, 54(2): 1–36.