计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200200117

基于分层注意力的信息级联预测模型

张志扬　张凤荔　陈学勤　王瑞锦
电子科技大学信息与软件工程学院　成都 610054
(13980044734@163.com)

摘　要　信息级联预测(Information Cascade Prediction)是社交网络分析领域的一个研究热点,其通过信息级联的扩散序列与拓扑图来学习在线社交媒体中信息的传播模式。当前的信息级联预测模型大多以循环神经网络为基础,仅考虑信息级联的时序结构信息或者序列内部的空间结构信息,无法学习序列之间的拓扑关系。而现有的级联图结构学习方法无法为节点的邻居分配不同的权重,导致节点之间的关联性学习较差。针对上述问题,文中提出了基于节点表示的信息级联采样方法,将信息级联建模为节点表示而非序列表示。随后提出一种基于分层注意力的信息级联预测(Information Cascade Prediction with Hierarchical Attention,ICPHA)模型。该模型首先通过结合了自注意力机制的循环神经网络来学习节点序列的时序结构信息;然后通过多头注意力机制学习节点表示之间的空间结构信息;最后通过分层的注意力网络对信息级联的结构信息进行联合建模。所提模型在 Twitter,Memes,Digg 这 3 种数据集上达到了领先的预测效果,并且具有良好的泛化能力。

关键词:在线社交媒体;深度学习;循环神经网络;图表示学习;信息级联预测;多头注意力机制
中图法分类号　TP183

Information Cascade Prediction Model Based on Hierarchical Attention

ZHANG Zhi-yang,ZHANG Feng-li,CHEN Xue-qin and WANG Rui-jin
School of Information and Software Engineering,University of Electronic Science and Technology of China,Chengdu 610054,China

Abstract　Information cascade prediction is a research hotspot in the field of social network analysis. It learns the propagation mode of information in online social media through the diffusion sequence and topology map of the information cascade. Most current models for solving this task are based on recurrent neural networks and only consider information cascading time series structure information or spatial structure information inside sequences, and cannot learn topological relationships between sequences. And the existing cascade graph structure learning methods cannot assign different weights to the neighbors of the nodes, resulting in poor association learning between the nodes. In response to the above problems, this paper proposes an information cascade sampling method based on node representation, which models the information cascade as a node representation rather than a sequence representation. This paper also proposes an information cascade prediction model based on hierarchical attention network (ICPHA), which learns the time series structure information of the node sequence through a recurrent neural network layer with self-attention mechanism, and learns the spatial structure information between node representations through a multi-head attention mechanism. By this way, ICPHA jointly models the structural information of the information cascade through a hierarchical attention network. ICPHA has achieved leading prediction results on Twitter, Memes, and Digg, and has good generalization ability.

Keywords　Online social media, Deep learning, Recurrent neural network, Graph representation learning, Information cascade prediction, Multi-head attention mechanism

到稿日期:2020-02-26　返修日期:2020-04-08　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(61802033,61472064,61602096);四川省科技计划(2018GZ0087,2019YJ0543);博士后基金项目(2018M643453);广东省国家重点实验室项目(2017B030314131);网络与数据安全四川省重点实验室开放课题(NDSMS201606)
This work was supported by the National Natural Science Foundation of China(61802033,61472064,61602096),Sichuan Science and Technology Program (2018GZ0087,2019YJ0543),Chinese Postdoctoral Science Foundation(2018M643453),Guangdong Provincial Key Laboratory Project (2017B030314131) and Network and Data Security Key Laboratory of Sichuan Province Open Issue(NDSMS201606).
通信作者:张凤荔(fzhang@uestc.edu.cn)

1 引言

当前,在线社交媒体的迅猛发展极大地促进了海量信息之间的传播与交互,并突出了预测信息级联的重要性。在这些社交媒体中,信息不仅可以通过人群之间的社交关系进行共享与传播,还会在此过程中衍生出一系列新的共享关系与模式[1]。通过研究信息传播规律与结构特性,信息级联预测使社交媒体、公共卫生、网络安全和关键基础架构等领域的各种应用受益,具体的下游应用包括虚假新闻检测和病毒式营销识别、众包和舆情调控等[1]。

信息级联预测问题源于对博客[3]、电子邮件[4]、产品推荐[5]以及社交网站(如 Facebook 和 Twitter[6-7])中识别到的信息级联进行分析,一般的数据样本包括若干信息级联序列与信息级联拓扑图。深度学习模型可以自动学习信息级联的某些表示,这些表示在模型训练过程中将会被嵌入到新的向量空间,然后通过复杂的神经网络进行学习。这类方法已经在社交网络分析中的节点推荐[8]与影响力预测[9]等领域得到了广泛应用。在信息级联预测中,深度学习模型主要学习信息级联的时序结构信息与空间结构信息,分别体现在数据样本的节点序列与节点构成的级联图中。

随着图表示学习的发展,许多与信息级联相关的研究致力于挖掘级联图的拓扑结构信息。基于图表示学习的信息级联预测方法不学习每个级联序列的表示,而是尝试学习级联拓扑图结构的表示,这类方法的灵感大部分来自应用于文本[9]和图像[11]等各个领域的表示学习。受图像表示学习的启发,这类方法通过图卷积神经网络等方法进行特征聚合与学习,可以较好地学习拓扑图的空间结构。一些最新的研究[12]尝试将图表示学习方法与循环神经网络结合,以同时学习信息级联的时序结构信息与空间结构信息。然而,基于卷积网络的图表示方法采用图的拉普拉斯矩阵进行节点特征的学习,使得模型与图的结构紧密相关,且无法为邻居节点分配不同的权重,局限了训练所得模型在其他图结构上的泛化能力。另一类方法将注意力机制应用于图表示学习[13],使用注意力操作对级联图中节点的邻近节点特征进行加权求和。由于邻近节点特征的权重完全取决于当前节点,独立于图结构,因此保证了模型的泛化能力。但由于级联规模一般远大于序列长度,因此计算代价较大,注意力权重的获取也十分困难。另外,这类方法忽略了级联图的时序结构信息,其信息捕捉的完整性也存在不足。

针对以上问题,本文提出基于分层注意力的信息级联预测模型,主要贡献有以下 3 方面。

(1)提出融合序列注意力机制与图注意力机制的信息级联预测模型,实现了对级联图的时序结构信息与空间结构信息的同时捕捉,并在信息级联预测领域首次提出利用多头图注意力机制进行级联拓扑图结构学习,使模型可应用于不同的图结构,提高了模型的泛化能力。

(2)针对分层注意力信息级联预测模型设计了基于节点表示的序列采样方法,使信息级联样本可以通过节点来同时表征信息级联序列内部的时序结构信息与序列之间的空间结构信息,能够更全面、更完整地捕捉信息级联节点包含的信息。

(3)在 Twitter, Memes, Digg 这 3 个数据集上进行的实验表明,相比传统的基于循环神经网络或基于图神经网络的信息级联预测模型,基于分层注意力的信息级联预测模型可以更好地结合级联的时序结构信息与空间结构信息,取得了更好的预测效果,其有效性得到进一步验证。

2 相关工作

2.1 信息级联预测

传统的信息级联预测方法主要分为 3 类。

(1)基于特征建模的方法[14]。从原始数据中提取各种手工特征,然后将它们输入判别式机器学习算法中进行预测。其缺点是过于依赖人工设计与先验知识。

(2)基于生成建模的方法[15]。将级联图的变化模式视为转推的到达过程,并针对每个消息分别对到达过程的强度函数进行建模。这类方法无法充分利用所有级联的流行度动态中隐含的信息进行预测,模型的可解释性和预测能力较差。

(3)基于扩散模型的方法[16]。这类方法通常做出各种强有力的假设,简化了现实中的一些复杂网络关系,在很大程度上取决于基础的扩散模型,并不适合信息级联预测。

近年来,深度学习被广泛应用于社交网络分析领域,以循环神经网络为核心的序列算法是信息级联预测的重要工具。文献[8]通过随机游走算法对级联序列进行采样,通过将注意力机制嵌入循环神经网络模型框架中来学习级联图的表示。Cao 等[17]针对常见生成建模方法解释性不足的问题,在信息级联预测模型中加入了霍克斯过程的可解释因素,即端到端监督框架下的用户影响力,但该模型忽略了复杂的时间依赖性。这些方法的共同点是将信息级联建模为序列表示,通过循环神经网络的记忆特性捕获级联序列中的顺序依赖性,致力于表征这些级联的结构特性和内容信息;但由于抛弃了信息级联的拓扑图结构,因此其在空间结构信息的捕捉方面表现不足。

融合图表示学习算法的信息级联预测模型是这一领域的最新成果,这类方法通常借助图谱的理论来实现拓扑图上的特征聚合操作,包括卷积操作与注意力计算。文献[18]提出了基于子图序列的信息级联预测模型,该模型将图谱理论桥接到深度网络中,提高了信息级联预测对空间局部结构信息的捕捉能力。文献[12]设计了一种端到端的方法来自动发现社会影响中的隐藏和预测信号,该方法将网络嵌入和图形卷积[19]构建到一个统一的框架中,通过重启随机游走[20]对其本地邻居进行采样,随后利用图形卷积和注意技术来学习潜在的级联预测信号。这类方法解决了拓扑图空间结构的学习问题,证明了图表示学习方法在信息级联预测领域的可行性,为今后的研究提供了新的思路。但基于图表示学习的信息级联预测模型仍然受限于基于序列表示的采样方法与图卷积网络无法处理动态有向图的瓶颈,模型特征学习与泛化的效果并不理想。

2.2 注意力机制

注意力机制最早由 Mnih 等[21]提出,其作用是使模型在训练中可以高度关注指定区域或者个体,从而实现特征的加

权变化，并将计算资源分配给更重要的目标。文献[22]提出了一种具有结构关注度的新型顺序神经信息扩散模型，在RNN的框架基础上加入了结构注意模块(Structure Attention Module，SAM)，用于捕获信息级联空间结构中的扩散上下文。Islam等[23]提出一次分析级联的一部分，并根据其评估移至级联的下一个位置，通过计算序列中所有隐藏状态的加权总和来分析较长的级联。这些方法将注意力机制应用于时序结构特征的学习，相比基于循环神经网络的模型，可以更好地捕捉级联序列中节点之间的关联特性，但仍无法充分学习信息级联的空间结构信息。为了学习拓扑图的空间结构，文献[13]将注意力机制应用于图表示学习中，提出了图注意力网络，通过注意力机制对邻近节点特征进行加权求和，其中邻近节点特征的权重完全取决于节点特征，独立于拓扑图结构。基于注意力机制的模型通常在节点间的特征关联性学习上有更好的学习效果，因此图注意力网络在诸多图表示任务中取得了最好的效果，这为信息级联预测中如何捕捉级联图的空间结构信息以及如何提高模型的泛化能力提供了新的思路。

3 问题定义与方法概论

3.1 问题定义

本文信息级联预测任务的目标是预测级联序列下一个或下一组被感染的节点。现有 n 条转发记录，记为 $P=\{p_i, i\in[1,n]\}$，每条转发记录对应一条级联序列 $C_i=\{u_1,u_2,\cdots,u_n\}$，其中，元素 u_i 为在此级联中被感染的用户 u_i 的节点id，被感染的用户按时间排序。所有级联序列中的节点都在级联图 $G=(V,E)$ 中，其中 V 表示节点集，$E\subseteq V\times V$ 表示边集。通常，信息级联预测问题可以表示为：给定级联图 G 中的级联序列 $C_i=\{u_1,u_2\cdots,u_n\}$，预测下一个将被感染的用户 $u_{n+1}\in V$。

3.2 方法概述

为了同时捕捉信息级联的时序结构信息与空间结构信息，本文提出了基于节点的级联采样方法，将序列表示建模为其末尾节点的表示，并分层采用两种注意力机制来分别学习信息级联的两种结构信息。

(1)信息级联的时序结构信息：首先使用双向自注意力循环神经网络学习序列末尾节点的状态表示，通过注意力机制为级联序列内所有顶点的状态表示分配注意力权重，再进行加权求和，得到最终的节点表示。

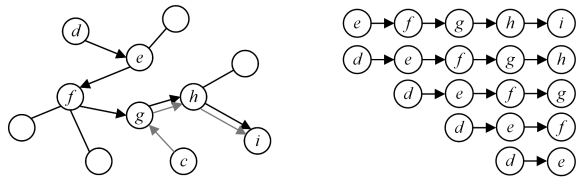
(2)信息级联的空间结构信息：使用多头图注意力网络学习级联图的空间结构，将注意力权重分配至节点的邻居节点集，然后聚合邻居节点的特征，以学习节点的空间结构特征。

3.3 基于节点的信息级联采样方法

现有的基于循环神经网络的信息级联预测模型一般是在拓扑图上进行随机游走采样[20]，在每一步采样中，从当前节点 v 的邻居节点中选取一个节点 w 作为下一个待采样节点，v 以一定概率进行跳跃、游走或进入终止状态。随机游走采样有统计误差小、执行效率高等优点，但级联预测模型中的随机游走算法存在两个缺点。1)在随机游走过程中，是否执行另一个随机跳转或进入终止状态的概率 P_a 决定了采样序列的预期数量，而是否执行随机跳转或过渡到邻居的概率 P_j 对应于采样序列的长度，这两个因素在确定级联图的表示中起着关键作用。通常情况下，这些转移概率是固定的超参数，因此基于随机游走的采样方法无法针对不同的级联图结构进行合适的采样，导致模型泛化能力较差。2)目前大部分数据集中的级联序列并不能与级联图完全吻合，为了保证采样样本所对应的级联图中所有可能的路径必须真实存在，一些现有的采样算法将级联图按序列分解为若干子图进行图表示学习，这样虽然保证了采样路径与级联图的吻合度，但丢失了级联拓扑图中一些序列之外的空间结构信息，比如两条交叉序列之间的空间关系、序列内节点的邻居节点特征等。

针对上述问题，本文设计了一种基于节点表示的信息级联采样方法，以一条序列的末尾节点作为序列样本的唯一索引，将每个样本采样为一条以末尾节点为索引节点的子序列。对于信息级联序列 $C_i=\{u_1,u_2,\cdots,u_n\}$，该采样方法首先生成 n 条以 C_i 中每一个节点为末尾节点的子序列，然后对每条子序列分别进行等长的逆序一阶游走采样。由于信息级联序列的传播模式是单向的链式传播，因此对于初始信息级联中的每个节点都可以采样出一条与之对应的唯一子序列，实现了采样序列与节点的一一映射。

如图1(a)所示，对级联序列 $\{d,e,f,g,h,i\}$ 进行基于节点的信息级联采样，假设序列的给定采样长度为5，则分别以级联序列中的每个节点为末尾节点生成6条子序列，之后对每一条子序列进行长度为5的逆序一阶游走采样(若无前一个节点，则直接停止采样)。如图1(b)所示，基于节点的信息级联采样的结果可表示为序列集 $\{\{d\},\{d,e\},\{d,e,f\},\{d,e,f,g\},\{d,e,f,g,h\},\{e,f,g,h,i\}\}$，其中每一条序列作为其末尾节点的时序表示。注意到，基于分层注意力的信息级联预测模型在特征学习过程中以节点为单位进行学习，因此每一个节点的序列表示必须是唯一确定的，当级联图中有末尾节点同为 i 的级联序列 $\{c,g,h,i\}$(在图中用灰色标明其轨迹)时，则选择包含更多未采样节点的序列 $\{e,f,g,h,i\}$。



(a)采样路径示意图 (b)采样过程示意图

图1 基于节点表示的信息级联采样方法示意图

Fig. 1 Diagram of information cascading sampling method based on node representation

相较于随机游走采样方法，基于节点表示的信息级联采样方法有以下优点：1)采样数量等于样本所构成的级联图的节点总数，因此信息级联序列样本的采样数量会随着级联图结构的变化而改变，解决了模型在不同数据集上的泛化问题；2)该方法通过样本与级联图节点一一对应，使信息级联序列与级联图形成了映射关系，保持了级联拓扑图的完整性。通过在特征学习时纳入其他节点的表示，基于此采样方法的模型可以学习序列之间的空间结构信息。

4 基于分层注意力的信息级联预测模型

基于分层注意力的信息级联预测模型如图 2 所示。以信息级联序列$\{e,f,g,h,i\}$为例，首先通过嵌入层将序列中的每个节点以及节点 i 的邻居节点映射为一个多维向量，之后将$\{e,f,g,h,i\}$的向量序列表示传入双向自注意力 LSTM，得到节点 i 的隐藏状态表示$\tilde{\boldsymbol{h}}_i$，再通过图注意力网络将$\tilde{\boldsymbol{h}}_i$与其邻居节点的状态向量集$\{\tilde{\boldsymbol{h}}_j, j\in\mathcal{N}_i\}$进行注意力操作，得到新的节点表示$\boldsymbol{h}_i'(K)$，最后经过输出层得到最终的预测结果。

(1)节点嵌入层：输入为节点序列，本层将节点序列中的每个节点映射为一个多维的向量，得到输入序列的向量矩阵。

(2)自注意力循环神经网络层：将嵌入层的输出向量矩阵传入循环神经网络，学习信息级联序列的时序依赖性，得到并输出序列中每个节点的隐藏状态的向量表示。随后，自注意力机制将各个节点的隐藏状态表示进行加权求和后作为信息级联序列末尾节点的隐藏状态表示输出，以代表整条序列的时序结构信息。

(3)图注意力网络层：将循环神经网络层输出的节点状态表示传入图注意力网络，通过计算节点与其邻居节点之间的注意力权重，捕捉节点之间的依赖关系，然后将邻居节点的特征聚合于该节点，得到新的节点向量表示，以代表该节点所捕捉到的空间结构信息。

(4)输出层：得到图中每个节点的特征表示后，训练一个多层感知器进行特征学习，最后通过 softmax 函数输出最终的预测结果。

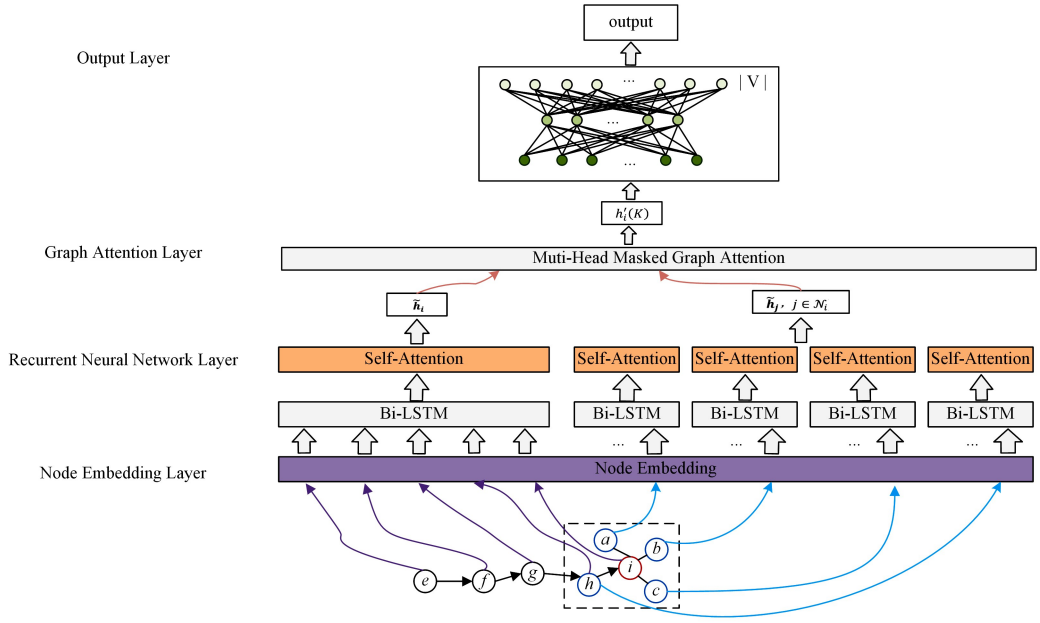


图 2 ICPHA 模型框架示意图

Fig. 2 Framework of ICPHA

4.1 模型构建

4.1.1 节点嵌入层

本文将拓扑图中的每个节点表示为向量而非索引，首先将所有节点的 id 嵌入向量空间$\boldsymbol{X}_v\in\mathbb{R}^{k\times|V|}$中，其中 k 是节点的嵌入维度。模型使用独热编码向量 $\boldsymbol{q}$ 来提取节点v_q的嵌入，形式为：

$$\boldsymbol{x}_q=\boldsymbol{X}_v\cdot\boldsymbol{q} \tag{1}$$

其中，$|\boldsymbol{q}|=|V|$，级联序列$C_i=\{u_1,u_2,\cdots,u_n\}$所对应的嵌入矩阵$\boldsymbol{X}_i\in\mathbb{R}^{k\times n}$在模型的训练过程中也会同步学习。

4.1.2 自注意力循环神经网络层

本文将自注意力机制与循环神经网络结合，以更好地学习信息级联序列的内部时序依赖性与级联固有的时间衰减效应[24]。其中，循环神经网络的目的是学习级联序列中每个节点的隐藏时序状态，并将其汇聚至末尾节点的状态向量中；自注意力机制通过计算序列中所有节点的隐藏状态的加权总和来学习节点的预测比重。给定级联序列C_i的嵌入矩阵$\boldsymbol{X}_i\in\mathbb{R}^{k\times n}$，对于输入序列的第 j 个元素u_j，假设u_j的嵌入是$\boldsymbol{x}_{u_j}$，则正向序列的 LSTM 中的门控单元公式为：

$$\boldsymbol{i}_j=\sigma(\boldsymbol{W}_i\cdot[\boldsymbol{h}_{j-1},\boldsymbol{x}_{u_j}]+\boldsymbol{b}_i) \tag{2}$$

$$\boldsymbol{f}_j=\sigma(\boldsymbol{W}_f\cdot[\boldsymbol{h}_{j-1},\boldsymbol{x}_{u_j}]+\boldsymbol{b}_f) \tag{3}$$

$$\widetilde{\boldsymbol{C}}_j=\tanh(\boldsymbol{W}_c\cdot[\boldsymbol{h}_{j-1},\boldsymbol{x}_{u_j}]+\boldsymbol{b}_c) \tag{4}$$

$$\boldsymbol{o}_j=\sigma(\boldsymbol{W}_o\cdot[\boldsymbol{h}_{j-1},\boldsymbol{x}_{u_j}]+\boldsymbol{b}_o) \tag{5}$$

$$\boldsymbol{C}_j=\widetilde{\boldsymbol{C}}_j\odot\boldsymbol{i}_j+\boldsymbol{f}_j\odot\boldsymbol{C}_{j-1} \tag{6}$$

$$\boldsymbol{h}_j=\boldsymbol{o}_j\odot\tanh(\boldsymbol{C}_j) \tag{7}$$

其中，$\boldsymbol{W}_i,\boldsymbol{W}_f,\boldsymbol{W}_c,\boldsymbol{W}_o$为门控单元的权重，$\tanh(\cdot)$为双曲正切函数，$\odot$表示逐项乘积操作。隐藏状态向量$\boldsymbol{h}_j$捕获级联序列$C_i$，直到节点 u_j 的正向时序信息。在信息级联预测中，考虑到当前节点的时序状态信息不仅与节点未来的状态有关，可能还与之前的状态有关，本文使用双向 LSTM 模型来学习信息级联的演变模式：

$$\boldsymbol{h}_j=\boldsymbol{h}_{fj}\parallel\boldsymbol{h}_{bj} \tag{8}$$

其中，$\boldsymbol{h}_{fj}$与$\boldsymbol{h}_{bj}$分别表示末尾节点为u_j的信息级联序列的正向时序信息与逆向时序信息，$\boldsymbol{h}_{bj}$由序列$\{u_1,u_2,\cdots,u_j\}$逆序得到，再通过式(2)－式(7)进行更新，$\parallel$表示两个状态向量的拼接操作。

如图 3 所示，在得到信息级联C_i中所有节点的隐藏状态

之后，我们通过序列内部的自注意力机制来学习一条序列内节点的注意力权重，并以非参数的形式建模信息级联的时间衰减效应。注意力计算的形式为：

$$\boldsymbol{u}_k=\tanh(\boldsymbol{W}_a^{\mathrm{T}}\boldsymbol{h}_k+\boldsymbol{b}_a) \tag{9}$$

$$\alpha_k^a=\frac{\exp(\boldsymbol{u}_k^{\mathrm{T}}\boldsymbol{u}_a)}{\sum_z\exp(\boldsymbol{u}_z^{\mathrm{T}}\boldsymbol{u}_a)} \tag{10}$$

其中，$\boldsymbol{h}_k$是级联序列的第 k 个节点的隐藏状态表示，注意力权重矩阵$\boldsymbol{W}_a\in\mathbb{R}^{K\times K_a}$，$\boldsymbol{u}_a,\boldsymbol{b}_a\in\mathbb{R}^{K_a}$。最终节点的时序结构信息表示为注意力分数与 LSTM 输出表示的加权求和，形式为：

$$\tilde{\boldsymbol{u}}_i=\sum_k\alpha_k^a\boldsymbol{h}_k \tag{11}$$

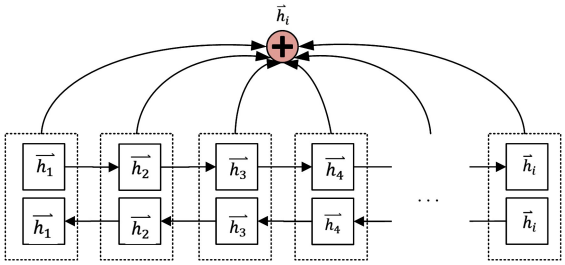


图 3 自注意力循环神经网络层

Fig. 3 Recurrent neural network layer based on self-attention mechanism

4.1.3 图注意力网络层

图注意力网络的目的是将邻居顶点的特征聚合到中心顶点上，学习新的顶点特征表达。级联图的有向无环图(Directed Acyclic Graph，DAG)特性，导致无法直接定义级联图的拉普拉斯矩阵，因此信息级联预测无法直接使用图卷积网络进行邻居节点特征的聚合。图注意力网络利用注意力机制替代了图卷积中固定的标准化操作，逐节点地为邻居节点分配不同的注意力权重以对其特征进行聚合，解决了信息级联预测模型的拓扑图结构学习问题。给定级联图 $G=(V,E)$中节点集的特征表示矩阵 $\boldsymbol{H}=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_{N'}\}$，$h_i\in\mathbb{R}^F$，其中$N'$表示节点的个数，$F$ 表示节点的特征维度。图注意力网络首先使用一个共享参数的线性映射对节点的特征进行增维，具体通过执行 Mask Graph Attention 操作[13]将拓扑图结构与注意力机制结合，其中 Mask Graph Attention 指注意力机制的运算只在邻居顶点上进行，而非对图中所有顶点进行。对于节点 u_i与其邻居节点集合 $J=\{j,j\in\mathcal{N}_i\}$，首先逐个计算节点$u_i$与其邻居节点之间的相似系数：

$$e_{ij}=a([\boldsymbol{W}\boldsymbol{h}_i\,\|\,\boldsymbol{W}\boldsymbol{h}_j]),j\in\mathcal{N}_i \tag{12}$$

其中，$\mathcal{N}_i$表示节点 i 在图中的一些邻居节点(本文实验中采用节点的一阶邻居进行注意力操作)；$a(\cdot)$将高维节点特征映射为实数，然后使用 softmax 函数对e_{ij}进行归一化，从而使系数可以在不同节点之间进行比较。节点u_i与u_j之间的注意力系数可表示为：

$$\alpha_{ij}=softmax(e_{ij})=\frac{\exp(LeakReLU(e_{ij}))}{\sum_{k\in\mathcal{N}_i}\exp(LeakReLU(e_{ik}))} \tag{13}$$

其中，$LeakReLU(\cdot)$为泄露修正线性单元。如图 4 所示，对注意力系数进行加权求和，得到节点u_i的新特征$\boldsymbol{h}_i'$为：

$$\boldsymbol{h}_i'=\sigma(\sum_{j\in\mathcal{N}_i}\alpha_{ij}\boldsymbol{W}\boldsymbol{h}_j) \tag{14}$$

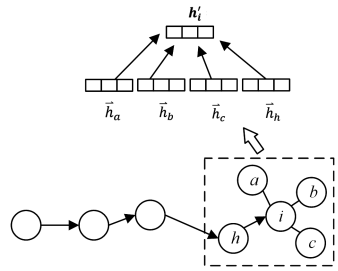


图 4 图注意力机制示意图

Fig. 4 Graph attention mechanism

本文的信息级联预测模型采用多头图注意力网络进行节点的特征学习。设置 K 个独立的注意力机制执行上述共享权重$\boldsymbol{W}$的线性映射，最后将 K 个特征向量进行拼接，得到节点u_i的多头特征表示$\boldsymbol{h}_i'(K)$：

$$\boldsymbol{h}_i'(K)=\Big\|_{k=1}^{K}\sigma(\sum_{j\in\mathcal{N}_i}\alpha_{ij}^k\boldsymbol{W}^k\boldsymbol{h}_j) \tag{15}$$

其中，$\|$ 表示拼接操作，α_{ij}^k是通过第 K 个注意力机制计算得到的归一化注意系数，$\boldsymbol{W}^k$是对应的输入线性变换的权重矩阵。

4.1.4 输出层

在输出层，$\boldsymbol{h}_i'(K)$通过一个多层感知器，得到最终的节点表示$\boldsymbol{h}_i^{\mathrm{fin}}$：

$$\boldsymbol{h}_i^{\mathrm{fin}}=MLP(\boldsymbol{h}_i'(K)) \tag{16}$$

节点表示$\boldsymbol{h}_i^{\mathrm{fin}}$作为全连接层的输入。节点被激活的分数可以通过以下公式进行计算：

$$\boldsymbol{w}_{u_{i+1}}=\boldsymbol{V}_u\boldsymbol{h}_i^{\mathrm{fin}}+\boldsymbol{b}_u \tag{17}$$

其中，$\boldsymbol{V}_u\in\mathbb{R}^{|V|\times k'}$为全连接层的权重矩阵，$k'$为节点表示$\boldsymbol{h}_i^{\mathrm{fin}}$的维度大小；$\boldsymbol{b}_u\in\mathbb{R}^{|V|}$为全连接层的偏置项矩阵。模型最后通过一个 softmax 层输出节点集中每个节点被激活的概率为：

$$P\,u_{i+1}\,|\,\boldsymbol{h}_i^{\mathrm{fin}}=\frac{\exp(\boldsymbol{w}_{u_{i+1}})}{\sum_{z\in V}\exp(w_{z_{i+1}})} \tag{18}$$

4.2 模型训练

本文使用反向传播算法来优化模型，损失函数为交叉熵。对于数据集 $P=\{p_i,i\in[1,N]\}$，采样为序列集 $C=\{C_{u_i},u_i\in V\}$，其中C_{u_i}表示末尾节点为u_i的级联序列，则交叉熵损失函数的形式为：

$$\ell(C)=-\sum_{v\in V}\sum_{j=1}^{|V|}\log P(u_{i+1}\,|\,\boldsymbol{h}_i^{\mathrm{fin}}) \tag{19}$$

5 实验

5.1 实验数据集

表 1 列出了本文所使用的 3 个数据集的数据信息统计，其中的节点数与边数对应于每个数据集的全局拓扑图。

(1)Twitter[23]：每个样本对应一个网站 URL 的转发记录，其中包含转发用户 id，以及这些转发发布的时间戳。

(2)Memes[25]：收集并跟踪网站内最频繁的引用和短语。每个 meme 都被视为信息项，网站或博客的每个 URL 被视为用户。

(3)Digg[26]：收集人们对网页内容进行的投票。数据样

本包括投票的时间戳以及匿名用户的 id。

表 1 实验数据集统计

Table 1 Statistic of datasets

数据集	节点数	边数	训练集	验证集	测试集
Twitter	137093	3589811	26871	1574	6663
Memes	5000	313669	N/A	N/A	N/A
Digg	279632	2617993	N/A	N/A	N/A

5.2 超参数设定

在本文实验中，嵌入层单元数为 128，双向 LSTM 单元数为 256，图注意力网络单元数为 8 * 8，多层感知器单元数为 32 * 64，dropout 率为 0.2，Adam 优化器的初始学习率为 0.0001，序列采样长度为 20，迭代 5000 轮。

5.3 评价指标

信息级联预测一般将下一个感染节点的预测任务视为排名问题，并将用户的转移概率作为其得分进行排序。本文采用两种常用的基于排名的性能评价指标：

(1)Map@k，即平均 AP(Average Precision)值；

(2)Hits@k，即前 k 个节点中包含下一个激活节点的比率。

文中遵循 Topo-LSTM[27] 与 DeepDiffuse[23] 中的实验设置，将 k 分别设置为 10，50，100。

5.4 对比实验

本节在上述数据集上将所提模型与以下 4 种方法进行对比实验。

(1)DeepDiffuse[23]：基于 LSTM 与注意力机制的模型，利用来自表示学习的嵌入和注意力机制进行级联序列的变化推理与预测。

(2)Topo-LSTM[27]：基于 LSTM 的模型，在给出级联及其网络结构的情况下预测下一个节点。

(3)DeepCas[8]：基于双向 GRU 的模型，采用自学习的随机游走进行采样，用于预测级联的增量。本文修改了模型的输出层，通过 softmax 层来预测下一个节点。

(4)Bi-LSTM：基础的双向 LSTM 模型。通过去除 ICPHA 的图注意力网络层，保留其余全部层，得到 Bi-LSTM 模型。

5.4.1 模型预测指标对比

不同模型在 3 种数据集上迭代至拟合(即损失在 100 轮迭代之内没有继续下降)后，统计每种模型的预测指标，如表 2 和表 3 所列。

表 2 不同模型的 Map@k 指标统计

Table 2 Map@k of different models

(单位：%)

Model	Twitter			Memes			Digg		
	@10	@50	@100	@10	@50	@100	@10	@50	@100
Bi-LSTM	9.77	10.53	10.98	10.46	11.00	12.59	2.25	2.81	3.04
Topo-LSTM	12.08	12.63	13.95	13.78	14.05	14.54	2.60	3.04	3.17
DeepCas	12.27	15.80	16.87	17.54	17.89	18.93	2.74	3.39	3.51
DeepDiffuse	19.46	20.92	21.51	**20.10**	20.76	21.63	2.87	3.53	3.68
ICPHA	**21.09**	**22.45**	**24.26**	19.59	**21.08**	**22.64**	**3.11**	**3.84**	**4.01**

表 3 不同模型的 Hits@k 指标统计

Table 3 Hits@k of different models

(单位：%)

Model	Twitter			Memes			Digg		
	@10	@50	@100	@10	@50	@100	@10	@50	@100
Bi-LSTM	22.58	24.45	28.36	39.26	57.62	60.69	6.88	20.12	30.58
Topo-LSTM	20.44	23.83	24.80	34.94	50.72	58.23	5.66	19.34	32.52
DeepCas	21.61	23.28	25.72	41.85	58.15	66.58	6.19	21.29	32.76
DeepDiffuse	24.46	27.26	30.91	43.42	61.27	67.65	7.64	23.78	34.27
ICPHA	**31.74**	**34.55**	**38.14**	**45.73**	**70.29**	**76.40**	**11.47**	**26.85**	**39.52**

(1)首先将 ICPHA 与 Bi-LSTM 进行比较，结果显示，相较于加入注意力机制或者图结构学习的模型，Bi-LSTM 的信息级联预测指标较低。这是因为 Bi-LSTM 只建模了序列内部节点之间的时序依赖关系，并未建模各序列之间的关系，而这种关系则代表由这些序列所构成的级联拓扑图的空间结构。这类仅包含循环神经网络的方法的另一个缺点是对节点特征的要求较高，本文实验并未进行额外的特征工程，级联序列的特征只包含节点 id 与拓扑图结构，模型抛弃节点的拓扑图信息会导致节点所持的信息量过少，从而无法学到有效的特征表示。

(2)在与 Topo-LSTM 的对比实验中，ICPHA 成功体现了基于节点表示的信息级联采样方法的优势，即可以学习信息级联样本之间的空间结构信息。Topo-LSTM 设计了新的扩散拓扑模型，将扩散拓扑的动态上下文学习纳入循环神经网络，这样做虽然可以有效利用级联序列的空间结构信息，但其采样方法将级联图拆分为若干子图，破坏了拓扑图的完整性，因此其预测指标低于 ICPHA，从而证明了基于节点表示的信息级联采样方法的有效性。同时，注意到 Topo-LSTM 未采用注意力机制，因此其预测指标低于 DeepDiffuse 和 DeepCas，证明了注意力机制在信息级联预测任务上的有效性。

(3)在另外两种加入了注意力机制的信息级联预测模型中，我们观察到其预测指标均高于未使用注意力机制的模型，这再一次验证了注意力机制的有效性。其中，DeepCas 使用具有注意力机制的 GRU 网络将序列转换为单个向量来表示扩散过程，其预测结果优于普通的循环神经网络模型，但相较于文本或者语句，其级联图的序列长度一般较长。因此，将注意力机制直接应用于整条序列的扩散预测，会使注意力机制

的计算代价大幅提高，导致模型训练困难，因此 DeepCas 的预测指标相较于 Bi-LSTM 并没有出现较大的提升。针对这一问题，DeepDiffuse 通过一次分析级联的一部分并根据其评估移至级联的下一个位置的方法降低了注意力机制的计算代价，该模型的 Hits@100 指标相较于 DeepCas 在 3 个数据集上分别提高了 5.19%，1.07%，1.51%。

(4)本文提出的 ICPHA 在对比实验中取得了最好的预测效果。在将注意力机制与神经网络结合的模型中，一般都只针对信息级联序列的时序信息进行特征学习，ICPHA 在使用循环神经网络与自注意力机制学习信息级联序列的时序信息的同时，用多头注意力机制来捕捉级联拓扑图的结构信息，成功将级联的时序结构信息与空间结构信息进行了有效结合。另外，一些最新的信息级联预测模型尝试直接将级联样本视为级联子图进行建模，ICPHA 通过学习信息级联序列之间的空间结构信息与节点的邻居节点特征，更加完整地捕捉了级联图的空间结构信息，因此得到了更好的预测效果。同时，ICPHA 支持在一次迭代中对所有的节点进行表示学习，不仅实现了级联图的全局结构捕捉，也加快了模型的训练速度。

5.4.2 模型整体复杂度对比

本文对 3 个预测指标取值较高的模型 DeepCas，DeepDiffuse 与 ICPHA 的参数数量进行限制，以对 3 种模型的整体复杂度进行分析。将所有模型的嵌入层单元数设为 128，多层感知器单元数设为 32 * 64，序列采样长度设为 20；此外，由于 3 种模型均在循环神经网络的基础上使用了注意力机制，因此实验忽略序列注意力参数对模型复杂度的影响，仅考虑循环神经网络层与图神经网络层上的隐藏层神经元数量对模型复杂度的影响。

本实验将循环神经网络层与图神经网络层上的神经元数量总和分别设为 128，192，256，320，384，其中前 4 组的图注意力层单元数设为 64，最后一组设为 128。各模型在 Twitter 数据集上的 Hits@100 对比结果如表 4 所列。

表 4 模型复杂度分析

Table 4 Model complexity analysis

Number of Units	Hits@100/%		
	DeepCas	DeepDiffuse	ICPHA
128	24.98	29.74	20.52
192	26.11	31.80	31.46
256	25.72	30.91	36.65
320 (256+64)	22.18	30.15	38.14
384 (256+128)	17.03	24.93	28.69

在神经网络层单元数较少的情况下，ICPHA 的预测精度低于 DeepCas 与 DeepDiffuse，这是因为在相同的参数数量下，ICPHA 的单层神经元数量过少导致模型出现了一定的欠拟合现象；随着单元数的增加，ICPHA 表现出较好的性能，且在其余两种模型出现明显过拟合时，仍能进行有效的模型学习。可见，ICPHA 的模型架构可以容纳更为复杂的级联结构与内容信息。本文推荐将 ICPHA 中的图注意网络层固定为 64，而将循环神经网络层设为 128 或 256，这是因为图注意力网络层本质上属于邻域聚合运算，过深的层数或过多的单元数易导致模型出现过度平滑问题。

5.5 模型分析

为了对模型的各层进行详细分析，本文引入 ICPHA 的两种变体进行对比。

(1)ICPHA-GRU/LSTM：将自注意力循环神经网络层的 Bi-LSTM 替换为单向 GRU/LSTM。循环神经网络层参数设置为 ICPHA 的 2 倍。

(2)ICP-GCN：将图注意力网络层替换为图卷积网络层，网络层参数与 ICPHA 保持一致。

5.5.1 自注意力循环神经网络层

对 ICPHA-GRU，ICPHA-LSTM 与 ICPHA 在 Twitter 数据集上的损失进行对比，实验结果如图 5 所示。

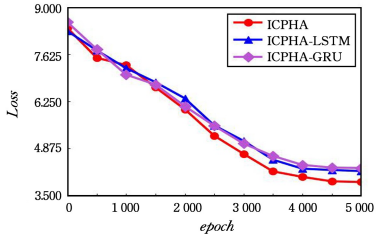


图 5 不同循环神经网络层损失的对比

Fig. 5 Comparison of losses in different recurrent neural network layers

可以看出，经过 4 500 轮迭代之后，ICPHA-GRU 与 ICPHA-LSTM 基本拟合；5 000 轮迭代之后，ICPHA 达到拟合，ICPHA 所采用的 Bi-LSTM 的训练时间稍长，但模型的预测指标更高。综合模型的运行时间与预测效果，本文选择能够更好捕捉级联前后时序依赖关系的 Bi-LSTM 作为循环神经网络层。

5.5.2 图注意力网络层

对 ICP-GCN 与 ICPHA 的预测指标进行对比分析，结果如图 6 所示。可以看出，ICPHA 所使用的图注意力网络的预测指标优于图卷积网络，图卷积网络不容易为不同的邻居节点分配不同的学习权重，而 ICPHA 利用 Mask Graph Attention 系数进行邻居节点的权重分配，顶点特征之间的相关性被更好地融入到模型中。对于有向级联图，图卷积网络还需要重新定义图的邻接关系，以保持对称性，而且模型无法针对不同的拓扑图结构进行泛化，而 ICPHA 利用注意力机制的结构无关性成功地规避了这一问题。

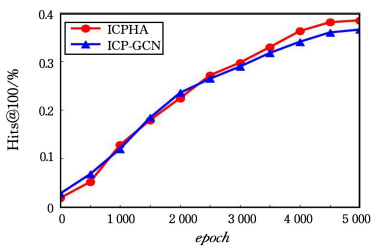


图 6 不同图表示学习方法的预测指标对比

Fig. 6 Prediction indicators comparison of different graphs representing learning methods

在图注意力网络层，本文测试了图注意力机制中的两种注意力机制运算方式，分别为 Global Graph Attention 与 Mask Graph Attention，其中 Global 指对图上的所有节点都

进行注意力计算，Mask 指仅对节点的邻居节点（一般为一阶邻居）进行注意力计算。对于一个 K 头的图注意力网络，Global 形式的图注意力网络层输出的特征维度是 Mask 形式的 K 倍，因此训练的时间成本要远高于 Mask 形式的注意力机制，但我们观察到两种注意力运算方法在信息级联的特征学习上并没有明显差异。本文综合考虑模型的性能，选择基于 Mask Graph Attention 的图注意力网络来构建模型的图注意力网络层。

最后，我们分析了序列采样长度对模型预测指标的影响。将基于节点表示的信息级联采样方法的采样窗口分别设为 5，10，15，20，25，30，监测模型的 Hits@100 指标，当精度不再提高时，模型再训练 100 轮后终止训练，记录迭代次数。实验结果如表 5 所列。

表 5 采样窗口大小分析

Table 5 Sampling window size analysis

采样窗口大小	每轮迭代所需时间/s	Hits@100/%	训练轮数
5	0.6542	35.70	4727
10	0.7008	37.66	4784
15	0.7244	38.07	4810
20	0.7322	38.14	5012
25	0.7471	38.11	6348
30	0.7686	38.18	8504

可以看出，单次迭代所需的时间与所需的训练轮次随采样序列窗口的增大而增加，而模型的预测指标在采样窗口为 20 以上时基本保持不变。综合模型的预测指标与训练时间，本文将对比实验中的采样序列长度设为 20。

5.5.3 注意力可视化

本文在 Digg 数据集上选取图 7(a)所示的子图进行注意力可视化说明，其中黑色标注了以 i 为末尾节点的信息级联序列，v 是所预测的下一个扩散节点。图 7(b)－图 7(d)表示在 3 个不同的图注意力头上的注意力权重分布，通过在一次迭代中传入所有子图节点来获得所有节点的注意力权重，图中节点的颜色越深，代表其权重越大。

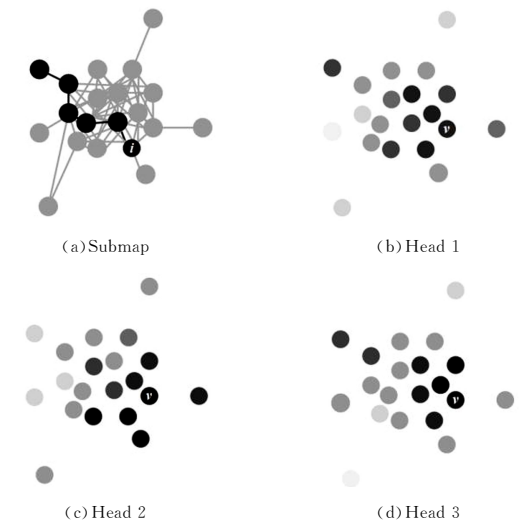


图 7 多头图注意力分布可视化

Fig. 7 Visualization of attention distribution in multi-head graph

结束语 在信息级联预测领域中，大部分方法都只针对信息级联的时序结构信息与空间结构信息中的一种进行特征建模，无法捕捉级联所包含的完整结构特征。本文基于循环神经网络与图注意力机制，提出了基于分层注意力的信息级联预测模型 ICPHA。该模型将信息级联预测任务抽象为信息级联节点的特征建模，能够对信息级联的时序结构信息与空间结构信息进行联合建模，并通过分层注意力机制聚合邻居节点的特征，更好地捕捉了级联节点之间的相关性，提高了模型的泛化能力。此外，本文提出了基于节点表示的信息级联采样方法，解决了随机游走采样无法捕捉级联序列之间结构信息的缺点，完善了拓扑图结构的特征建模方式。实验表明，本文提出的基于分层注意力的信息级联预测模型的性能相比传统循环神经网络方法有了较大的提升，同时，能取得比常见图表示学习方法更好的预测结果。

ICPHA 中使用的图注意力机制要求每一个级联图节点有唯一的序列表示，这使得模型在处理一些较为复杂的路径时可能会丢弃一些特征；此外，模型的复杂度较高也导致 ICPHA 在模型参数较少情况下的预测效果略低于最先进的算法。在未来的研究中，我们会针对这些问题对模型的图注意力网络层进行改进，并尝试简化模型现有框架，使模型能够纳入更全面的级联结构信息，并在任意复杂度下都能达到较好的预测效果。

参 考 文 献

[1] ZHU X, JIA Y, NIE Y P, et al. Event Propagation Analysis on Microblog[J]. Journal of Computer Research and Development, 2015, 52(2): 437-444.

[2] CHENG J, ADAMIC L, DOW P A, et al. Can cascades be predicted? [C]// Proceedings of the 23rd International Conference on World Wide Web. ACM, 2014: 925-936.

[3] JIANG Y, COUNTS S. Predicting the speed, scale, and range of information diffusion in twitter [C] // Fourth International AAAI Conference on Weblogs and Social Media. 2010.

[4] GOLUB B, JACKSON M O. Using selection bias to explain the observed structure of internet diffusions[J]. Proceedings of the National Academy of Sciences, 2010, 107(24): 10833-10836.

[5] LESKOVEC J. The Dynamics of Viral Marketing [J]. Acm Transactions on the Web, 2005, 1(1): 228-237.

[6] DOW A P, ADAMIC L A, FRIGGERI A. The Anatomy of Large Facebook Cascades[C]// ICWSM. 2013.

[7] KUMAR R, MAHDIAN M, MCGLOHON M. Dynamics of conversations[C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2010: 553-562.

[8] CHENG L, MA J Q, GUO X X, et al. Deepcas: An end-to-end predictor of information cascades[C]// Proceedings of the 26th international conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017: 577-586.

[9] WANG X S, MA S Z. Method of Weibo User Influence Calculation Integrating Users' Own Factors and Interaction Behavior [J]. Computer Science, 2020, 47(1): 96-101.

[10] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural

probabilistic language model[J]. Journal of Machine Learning Research,2003,3(Feb):1137-1155.

[11] KRIZHEVSKY A,SUTSKEVER I,HINTON G E. Imagenet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing Systems. 2012:1097-1105.

[12] QIU J Z,TANG J,MA H,et al. Deepinf:Social influence prediction with deep learning[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM,2018:2110-2119.

[13] VELIKOVI P,CUCURULL G,CASANOVA A,et al. Graph attention networks[J]. arXiv:1710. 10903,2017.

[14] GUO R C,SHAKARIAN P. A comparison of methods for cascade prediction[C]// Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE,2016:591-598.

[15] BAO P,SHEN H W,JIN X L,et al. Modeling and predicting popularity dynamics of microblogs using self-excited hawkes processes[C]// Proceedings of the 24th International Conference on World Wide Web. ACM,2015:9-10.

[16] KEMPE D,KLEINBERG J,TARDOS É. Maximizing the spread of influence through a social network[C]// Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM,2003:137-146.

[17] CAO Q,SHEN H W,CEN K T,et al. Deephawkes:Bridging the gap between prediction and understanding of information cascades[C]// Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. ACM,2017:1149-1158.

[18] CHEN X Q,ZHOU F,ZHANG K P,et al. Information Diffusion Prediction via Recurrent Cascades Convolution[C]// 2019 IEEE 35th International Conference on Data Engineering (ICDE). IEEE,2019:770-781.

[19] KIPF T N,WELLING M. Semi-supervised classification with graph convolutional networks[J]. arXiv:1609. 02907,2016.

[20] TONG H H,FALOUTSOS C,PAN J Y. Fast random walk with restart and its applications[C]// Sixth International Conference on Data Mining (ICDM'06). IEEE,2006:613-622.

[21] MNIH V,HEESS N,GRAVES A. Recurrent models of visual attention[C]// Advances in Neural Information Processing Systems. 2014:2204-2212.

[22] WANG Z T,CHEN C Y,LI W J. A Sequential Neural Information Diffusion Model with Structure Attention[C]// Proceedings of the 27th ACM International Conference on Information and Knowledge Management. ACM,2018:1795-1798.

[23] ISLAM M R,MUTHIAH S,ADHIKARI B,et al. DeepDiffuse:Predicting the 'Who' and 'When' in Cascades[C]// 2018 IEEE International Conference on Data Mining (ICDM). IEEE,2018:1055-1060.

[24] GAO S,MA J,CHEN Z M. Modeling and predicting retweeting dynamics on microblogging platforms[C]// Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. ACM,2015:107-116.

[25] LESKOVEC J,BACKSTROM L,KLEINBERG J. Meme-tracking and the dynamics of the news cycle[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM,2009:497-506.

[26] TAD H,KRISTINA L. Social Dynamics of Digg [C]// Proceedings of the Fourth International Conference on Weblogs and Social Media(ICWSM 2010). Washington,DC,USA,2010:23-26.

[27] WANG J,ZHENG V,LIU Z M,et al. Topological recurrent neural network for diffusion prediction[C]// 2017 IEEE International Conference on Data Mining (ICDM). IEEE,2017:475-484.

ZHANG Zhi-yang, born in 1997, postgraduate, is a member of China Computer Federation. His main research interests include machine learning, data mining and cascade prediction.

ZHANG Feng-li, born in 1963, Ph. D, professor, Ph.D supervisor, is a member of China Computer Federation. Her main research interests include network security and network engineering, cloud computing and big data and machine learning.