



研究与开发

基于 Transformer 模型的社交网络影响力最大化算法

于树科, 姚瑶, 严晨雪

(江苏商贸职业学院电子与信息学院, 江苏 南通 226011)

摘要: 基于网络拓扑结构的社交网络影响力最大化算法受网络结构影响大, 导致在不同规模、不同拓扑结构的社交网络上的性能不稳定。针对此问题, 提出一种基于改进 Transformer 模型的社交网络影响力最大化算法。首先, 基于 K-shell 分解法筛选社交网络中影响力高的节点; 然后, 运用随机游走策略发现候选节点的拓扑结构信息和连接框架信息; 最终, 对 Transformer 模型进行改进, 使其支持可扩展的节点特征序列, 利用改进 Transformer 模型预测社交网络中的种子节点。在 6 个不同规模的真实社交网络上完成了验证实验。结果表明, 所提算法在不同规模、不同拓扑结构的社交网络上均实现了较好的影响力最大化性能, 且大幅提高了种子节点识别的时间效率。

关键词: 社交网络; 影响力节点; 影响力最大化; 信息传播; 神经网络

中图分类号: TP391

文献标志码: A

doi: 10.11959/j.issn.1000-0801.2024256

Influence maximization algorithm of social networks based on Transformer model

YU Shuke, YAO Yao, YAN Chenxue

School of Electronics and Information, Jiangsu Vocational College of Business, Nantong 226011, China

Abstract: The network topology structure based influence maximization algorithms are greatly influenced by the network structure, which leads to unstable performance of social networks of different scales and different topology structures. In view of this problem, a improved Transformer model based social network influence maximization algorithm was proposed. Firstly, the high influential nodes of the society network were selected based on the k-shell decomposition method. Seconclly, the topology structure information and connection framework information of the candidate nodes were discovered by use of the random walk strategy. Finally, the Transformer model was improved, in order to support scalable node feature sequences, and the improved Transformer model was taken advantage to predict the seed nodes of the social network. Validation experiments were carried on six real social networks of different scales. The results show that the proposed algorithm realizes a good influence maximization performance on social networks of different scales and topology structures, and the time efficiency of the seed node recognition has been increased significantly.

Key words: social network, influence node, influence maximization, information propagation, neural network

0 引言

影响力最大化^[1]通过发现社交网络的种子节点,使信息在社交网络中实现最大范围的传播效果,此技术在社交推荐、产品营销和谣言阻断等领域得到广泛应用^[2]。近年来,基于贪心思想的影响力最大化算法展现出良好的性能,文献[3]将节点集影响力评估转化为局部概率解计算,以提高爬山贪心算法的运行效率。文献[4]通过社区选择和社区内影响力传播两阶段策略来改善影响力传播性能,并加快种子节点选取速度。文献[5]通过增加边界约束策略以快速确定估计最优采样次数,提出基于节点度筛选策略以提升种子集质量,提出基于影响力增量剪枝策略以提高算法运行效率,通过这3个策略加快了基于贪心思想的影响力最大化算法速度。从文献[3-5]来看,当前基于贪心思想的影响力最大化算法主要面临时间复杂度高的问题,导致其无法应用于大规模社交网络。相较于基于贪心思想的影响力最大化算法,基于网络拓扑结构的影响力最大化算法具有时间效率高的优点,可有效处理大规模社交网络。文献[6]利用社区结构特性提出一种正向连接和反向连接的影响力最大化算法,该算法通过最小化种子节点数来加快运算效率。文献[7]采用 Louvain 算法将大规模社交网络划分成小规模社交网络,利用一跳扩散值和两跳扩散值计算节点影响力,该算法成功降低了大规模社交网络种子节点的发现时间。文献[8]使用 DyHATR 算法提取网络节点的传播特征向量,再通过 k -means 聚类算法得到粗粒度网络,大幅提高了种子节点的发现速度。上述基于网络拓扑结构的影响力最大化算法重点关注于提高算法运算效率,然而这些算法仅考虑某单一的网络中心性,其性能受网络拓扑结构的影响较大,导致算法稳定性不佳。

Transformer 模型^[9]是一种基于自注意力机制的神经网络模型,它将输入序列的所有 token 作比

较,可捕捉 token 间的长距离依赖关系。文献[10]提出一种基于节点度数、邻居信息和 Transformer 模型的节点影响力识别方法,采用目标节点与其邻居节点的度数作为节点的基本特征,将目标节点与其邻居节点转换成序列形式。虽然该方法在邻居信息聚合过程中取得了较好的效果,但依然存在两点不足:(1)仅聚合了邻居节点的信息,未考虑更大范围的网络拓扑分布情况,得到的影响力值存在不可忽略的失真;(2)以度数作为节点特征可能导致节点表示的相似性较高,进而出现过平滑现象。

针对上述方法的局限性,本文提出一种基于 Transformer 模型的社交网络影响力最大化算法。该算法通过随机游走策略扩大聚合的邻居节点范围,采用节点索引、节点度数和节点 shell 数作为节点特征,以期提高节点的特征表达力。具体而言,采用 K-shell 分解法^[11]选择网络中的传播影响力高的候选节点,通过缩小搜索空间来降低计算成本。然后,提出一种基于随机游走的节点嵌入方法获得候选节点的嵌入表示,将高维度候选节点映射到二维空间,并尽可能地保留节点在网络中的拓扑结构信息和连接框架信息。最终,提出一种改进的 Transformer 模型,简称 TranR 模型,利用 TranR 模型捕获社交网络包含的复杂模式,不仅可提取社交节点在拓扑结构上的依赖性,也可提取节点在连接框架中的依赖性。

1 问题模型与框架设计

影响力最大化问题的目标是从社交网络中选择一组最具影响力的种子节点,使信息在社交网络中实现最大范围的传播效果^[12]。将社交网络表示为图结构 $G=(\tilde{V}, E)$,其中 \tilde{V} 为节点集, E 为边集,假设 v_i 表示第 i 个节点, e_{ij} 表示节点 v_i 和 v_j 之间的边,反映了两者的社交关系,它与邻接矩阵 A 的元素 a_{ij} 对应,网络节点数和边数分别为 n 和 m 。影响力最大化问题的目标是从 \tilde{V} 中选择种子节点集 S ,使信息传播范围最大化。



所提算法的种子节点识别流程如图1所示。该算法主要分为4个步骤。

步骤1: 从社交网络选择影响力高的节点作为候选节点。此步骤可缩小搜索空间和降低运算量。

步骤2: 对候选节点进行节点嵌入处理。通过随机游走采样各候选节点的邻域拓扑结构和连接框架信息,再结合节点属性构建候选节点的嵌入表示。

步骤3: 将节点嵌入表示输入Transformer模型进行特征提取,提取候选节点与各级邻居节点间的依赖性。

步骤4: 利用回归器将节点影响力预测问题建立为回归问题,借助SIR(susceptible infected recovered)模型^[13]计算社交网络各节点的正定传播率。

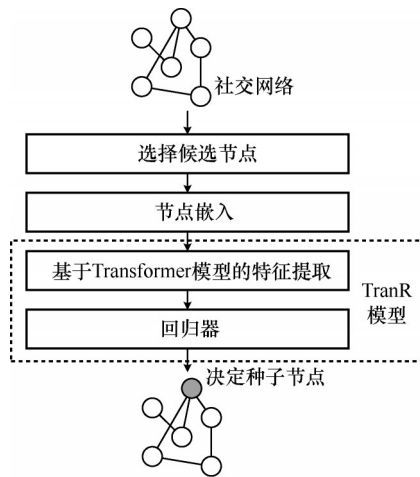


图1 种子节点识别流程

2 社交网络影响力最大化算法

2.1 选择候选节点

本文基于节点位置及其拓扑结构为各节点分配一个权值,该权值体现了节点在网络中的影响力。首先,使用K-shell分解法^[14]获得社交网络的核心节点集 O ,再计算各核心节点 $v_i \in O$ 的权重:

$$w(v_i) = \text{eig}nv(v_i) \times \left(\sum_{v_j \in \Gamma(v_i)} s(v_j) + k(v_i) \right) \times J(v_i, c) \quad (1)$$

其中, $\text{eig}nv(v_i)$ 为节点 v_i 的特征向量, $\Gamma(v_i)$ 为节点 v_i 邻居节点集, $s(v_j)$ 为节点 v_j 的shell数, $k(v_i)$ 为节点 v_i 的度。 $J(v_i, c)$ 表示shell c 的权重, $v_i \in \text{shell } c$, $J(v_i, c)$ 的计算方法为 $k_{\max}(c) \times k_{\min}(c)^{CC(c)}$, $CC(c)$ 为shell c 的平均聚类系数, $k_{\max}(c)$ 和 $k_{\min}(c)$ 分别为shell c 的最大度和最小度。

为避免“富人俱乐部”现象^[15],基于权重对核心节点集 O 的节点进行进一步筛选。将 O 的节点按权重进行降序排列,选择前 r 个节点作为候选节点。决定 r 的方法如下:

$$r = \frac{\text{lb } n \times \text{lb } m}{k_{\text{avg}} \times \text{lg } m} \quad (2)$$

其中, n 为网络节点总数, m 为网络边总数, k_{avg} 为平均网络度。式(2)根据网络规模和平均网络度决定候选节点数量。

2.2 节点嵌入

本文提出一种基于随机游走的节点嵌入方法,该方法由两个步骤构成:基于随机游走的拓扑结构探索和节点特征序列建立。

步骤1: 基于随机游走的拓扑结构探索。

随机游走结合广度优先搜索和深度优先搜索两种策略,广度优先搜索关注于搜索当前节点的一跳邻居,深度优先搜索则关注于搜索当前节点的二跳邻居。随机游走的迁移概率分布如图2所示。假设上一次游走的边为 $t \rightarrow v$,通过两个参数 p 和 q 控制 v 游走的迁移概率,可表示为:

$$\alpha(t, x) = \begin{cases} \frac{1}{p}, & d(t, x) = 0 \\ 1, & d(t, x) = 1 \\ \frac{1}{q}, & d(t, x) = 2 \end{cases} \quad (3)$$

其中, $d(t, x)$ 表示节点 t 与节点 x 之间的距离, $d(t, x) = 0$ 表示沿上一跳返回; $d(t, x) = 1$ 表示节点 t 和节点 x 相邻; $d(t, x) = 2$ 表示节点 t 和节点 x 不相邻。 p 为返回概率:如果 $p > \max(q, 1)$,则尽可能不往回走;如果 $p < \max(q, 1)$,则倾向于返回上一个节点。 q 为搜索参数:如果 $q > 1$,则倾向于在起始

点周围搜索，即广度优先搜索（breadth first search, BFS）；如果 $q < 1$ ，则倾向于往远处搜索，即深度优先搜索（depth first search, DFS）。将节点 v_i 经步骤一得到的邻居节点序列记为 $L(v_i)$ 。

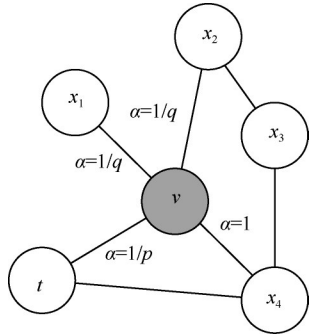


图2 随机游走的迁移概率分布

步骤2：节点特征序列建立。

将序列 $L(v_i)$ 中各节点的属性添加到节点索引之后， $L(v_i)$ 的长度由 B 扩张至 $B \times U$ ，其中 U 为节点特征长度。本文采用节点索引 v_i 、节点度 $k(v_i)$ 和节点 shell 数 $s(v_i)$ 组成每个节点的特征，因此 $U=3$ 。一个候选节点生成节点嵌入序列的流程如图3所示。

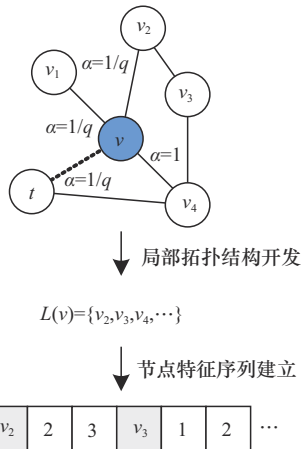


图3 候选节点生成节点嵌入序列的流程

2.3 基于 TranR 模型的节点特征学习

TranR 模型由输入层、编码器和回归器构成，输入层将候选节点特征序列转换成位置嵌入序列，位置嵌入序列输入 Transformer 编码器进行特征学习，回归器将估计影响力值与正定影响力值作比较。

2.3.1 输入层

本文采用 BERT 模型^[16]作为 Transformer 模型，TranR 模型输入层的处理流程如图4所示。首先，接收第2.2节的候选节点特征序列 $L(v_i)$ ， $L(v_i)$ 包含节点自身信息、拓扑结构信息以及连接框架信息。BERT 模型采用 [CLS] 标记序列的开始，在每个序列前增加一个 [CLS] 标记， $L(v_i)$ 长度从 $B \times U$ 变为 $(B+1) \times U$ 。然后，在 $(B+1)$ 个元素前增加位置标记 E_{POS} ，获得位置嵌入序列。最终，将位置嵌入序列变换成 $(B+1) \times U$ 的二维矩阵 E 。

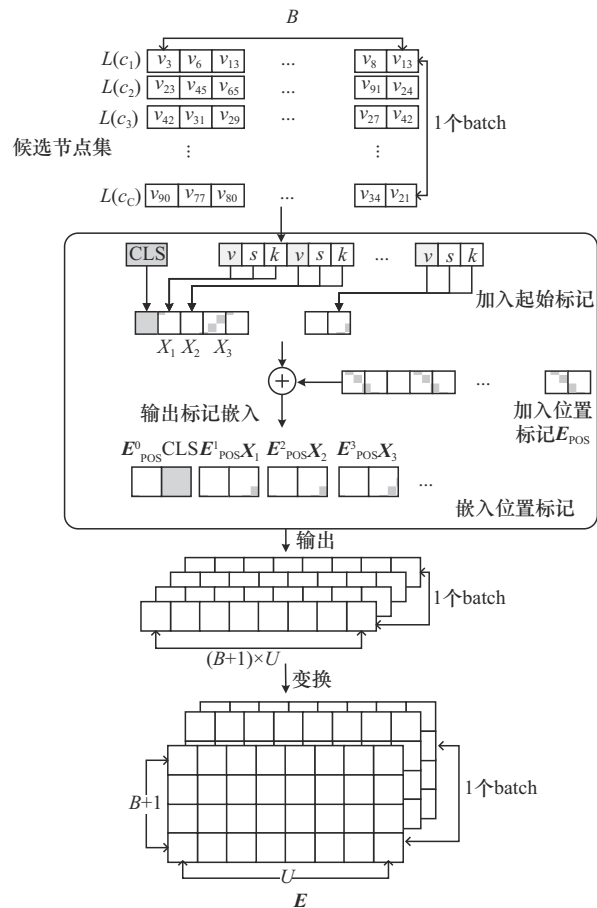


图4 TranR 模型输入层的处理流程

2.3.2 Transformer 编码器

Transformer 编码器由若干个结构相同的编码块构成，Transformer 编码块的内部结构如图5所示。Transformer 编码器的输入为矩阵 E ，输出序列记为 $\{Z_0, Z_1, \dots, Z_B\}$ ，其中第一块 Z_0 包含输入序



列的全局信息。

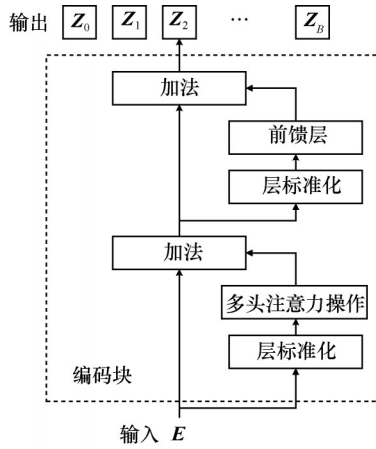


图5 Transformer编码块的内部结构

自注意力机制需学习3个矩阵 W^Q 、 W^K 和 W^V ，这3个矩阵分别与输入序列相乘得到3个新矩阵 Q 、 K 和 V ，分别称为输入序列的查询 (Query)、键 (Key) 和值 (Value)。此过程可表示为：

$$\begin{cases} Q = E \times W^Q \\ K = E \times W^K \\ V = E \times W^V \end{cases} \quad (4)$$

其中， W^Q 、 W^K 和 W^V 为3个可学习矩阵。

然后， Q 的转置与 K 相乘，得到注意力矩阵 A 。 A 经过标准化和 Softmax 计算后，其元素 A_{ij} 体现了节点 v_i 和节点 v_j 的相关性。

$$A = \text{Att}(Q, K, V) = \text{Softmax}\left(\frac{Q^T K}{\sqrt{d_K}}\right) \quad (5)$$

其中， d_K 表示 K 的向量长度。

A 与 V 做点积运算，得到包含注意力的结果：

$$o = A \cdot V \quad (6)$$

将第 i 个 head 输出的注意力向量记为 o_i ，多头自注意力的总注意力可定义为：

$$O_{\text{Multi-head}} = \text{Concat}(o_1, o_2, \dots, o_H) W^L \quad (7)$$

其中， H 为 head 总数，Concat()表示向量拼接操作，即将向量 o_1, o_2, \dots, o_H 首尾相连成一个总向量， W^L 表示线性变换层的权值向量。

最终，多头自注意力机制需学习的权重参数

包括： (W_1^Q, W_1^K, W_1^V) ， (W_2^Q, W_2^K, W_2^V) ， \dots ， (W_H^Q, W_H^K, W_H^V) 和 W^L ，在神经网络训练阶段学习得到上述权重值。

2.3.3 回归器

采用全连接层对输入特征进行回归处理，回归器的网络结构如图6所示。将 Z_0 输入回归器，回归器由两个全连接层和一个输出层构成。全连接层的神经元数量为64，输出层的神经元数量为1，它输出候选节点的传播影响力值。

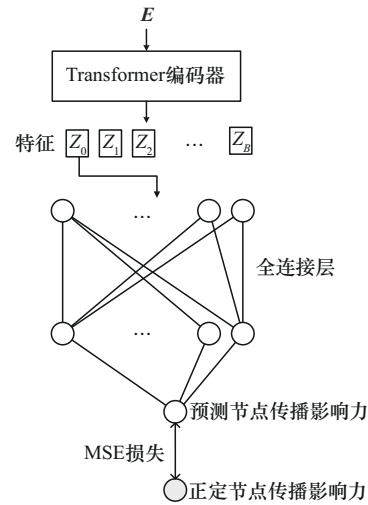


图6 回归器的网络结构

2.4 TranR模型训练

2.4.1 收集正定标签

使用 SIR 模型^[17]评估社交网络中各节点的正定传播影响力。SIR 模型将节点分成3种状态：易感者、感染者和移出者。在初始化阶段，网络的种子节点被感染，其他节点均为易感者。在之后的每个传播周期内，感染者基于一个给定传染概率感染其邻居中的易感者，然后它基于概率 λ 变为移出者，本文将 λ 设为1。当网络内的感染者消失时，传染过程结束，此时网络中各节点的感染率即对应该节点的传播影响力。

在大规模社交网络上进行 SIR 模型模拟的计算成本较高，本文在生成的小规模社交网络上进行模拟，以此收集训练集和验证集。通过编程生成3种

虚拟网络，即随机网络、小世界网络和无标度网络。随机网络生成方法为：给定网络节点数，以一定概率在任意两个节点间随机建立不重复连接。小世界网络生成方法为：生成一个环状的规则网络，再以一定概率随机重连网络中的每个边，即固定边的一个端点，随机选择另一个端点。无标度网络生成方法为：采用巴拉巴西-阿尔伯特 (Barabási-Albert, BA) 模型^[18]生成无标度网络。

在网络分析中，通过蒙特卡罗 (Monte-Carlo) 模拟器评估节点的传播影响力。生成 300 个小规模虚拟网络构成训练集，其中随机网络、小世界网络和无标度网络各有 100 个。每个社交网络的 Monte-Carlo 模拟次数设为 1 000，计算各节点 1 000 次模拟的影响力平均值作为对应的正定值。

2.4.2 神经网络训练方法

使用均方误差 (mean squared error, MSE) 作为损失函数，训练过程将最小化神经网络估计节点影响力和正定影响力 (来自 SIR 模型) 之间的 MSE 作为优化目标。损失函数定义为：

$$\ell_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^N (y_i - y'_i)^2 \quad (15)$$

其中， N 为生成网络的节点数， y_i 和 y'_i 分别为第 i 个节点的真实影响力值和预测影响力值。

TranR 模型训练与测试的流程如图 7 所示，具体如下。

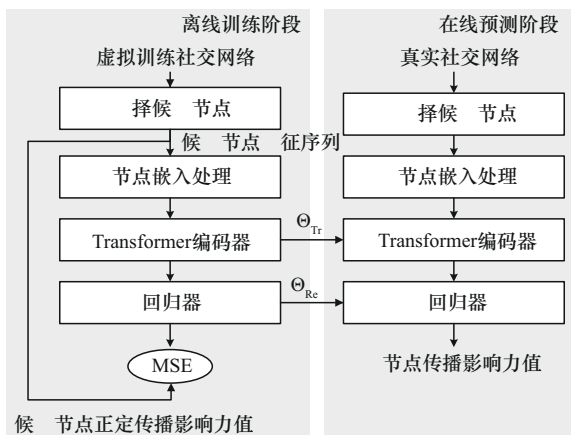


图 7 TranR 模型训练与测试的流程

(1) 离线训练阶段：在生成的虚拟网络上对 TranR 模型进行端到端训练，将节点嵌入向量作为 TranR 模型输入，SIR 模型评估的该节点平均影响力值作为 TranR 模型输出，以最小化损失函数为目标来寻找 TranR 模型的参数集 Θ_{Tr} 和 Θ_{Re} 。

(2) 在线预测阶段：基于参数集 Θ_{Tr} 和参数集 Θ_{Re} 建立 TranR 模型，估计真实社交网络各节点的传播影响力值。

3 实验结果与分析

3.1 数据集

为了评估本文算法的有效性，在真实社交网络上进行对比实验。选择 6 个真实社交网络作为测试集，其中 2 个小规模社交网络为 Dolphin 和 FootBall，2 个中等规模社交网络为 Advogato 和 Chess，2 个大规模社交网络为 Amazon 和 DBLP。6 个真实社交网络的相关属性见表 1。

表 1 6 个真实社交网络的相关属性

网络	节点数	边数	聚类系数	最大度	平均度
Dolphin	62	159	0.308	12	5.13
FootBall	115	613	0.407	12	10.66
Advogato	6 541	51 127	0.092	943	15.63
Chess	7 301	65 053	0.125	280	17.82
Amazon	262 111	1 234 877	0.236	425	9.42
DBLP	317 080	1 049 866	0.632	343	6.62

3.2 性能评价标准

采用影响力传播度 (influence spread) 和运行时间评估算法性能。

(1) 影响力传播度：给定一个种子集 S ，在独立级联 (independent cascade, IC) 模型^[19]上通过 Monte-Carlo 模拟器评估 S 的影响范围。IC 模型的传播概率设为 0.01，每个种子集完成 100 次蒙特卡罗模拟，统计 100 次结果的平均值作为影响范围结果。

(2) 运行时间：从真实网络输入算法到算法输出种子节点的总运行时间。



3.3 参数实验

实验硬件环境为 Intel(R) Xeon(R) Silver 4208 CPU, 内存为 64 GB, 操作系统为 Ubuntu 20.04. 显卡为 GPU NVIDIA A100. SIR 模型的传染率为 0.01, 将种子节点数 k 设为 1~30. 在 Tensorflow 2.9 和 CUDA 11.2 上建立所提 TranR 模型, 按第 4.2.4 节的方法训练 TranR 模型, TranR 模型编码块数设为 4. 使用 Adam 优化器作为 TranR 模型的优化器, 批尺寸设为 128, 学习率为 0.000 1, 最大 epoch 数为 300.

节点特征长度 U 等于式 (2) 计算的 r 值, 即根据网络规模和平均网络度决定节点特征长度. 为了扩大节点邻域拓扑的信息量, 将控制参数对 (p, q) 分别设为 (1, 0.5)、(1, 1)、(1.5, 1)、(2, 1)、(2, 2)、(3, 2) 和 (3, 3), 不同控制参数对的 TranR 模型验证损失如图 8 所示. 从图 8 来看, 当 $p=1$ 时, 随机游走的返回概率较高, 重复聚合了一些邻居节点信息, 所提取特征的判别力较弱, 导致验证损失较高. 当 $q \leq 1$ 时, 随机游走探索远距离邻居的概率较高, 此时, 验证损失较高, 说明近距离邻居节点的判别力高于远距离邻居节点. 经比较, 验证损失在 (3, 2) 处取得最低值, 因此下文将 p 和 q 分别设为 3 和 2.

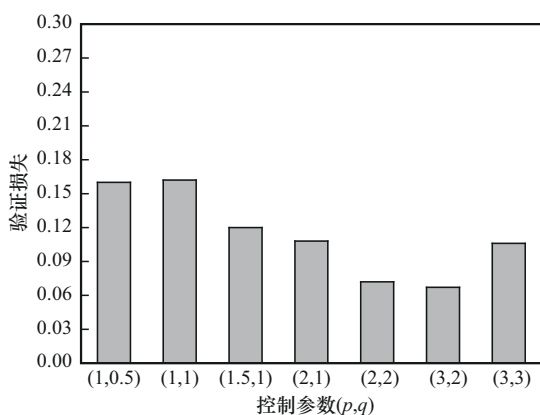


图8 不同控制参数对的 TranR 模型验证损失

3.4 对比算法

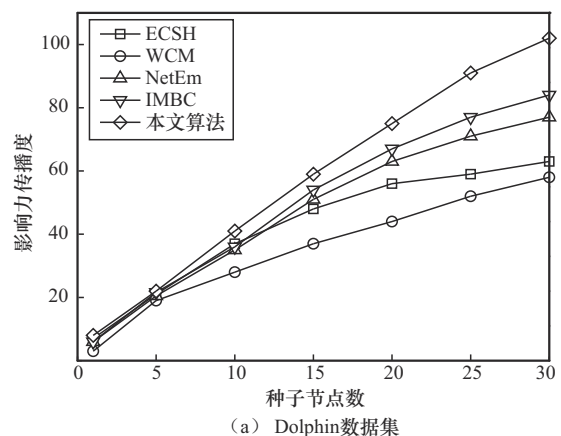
选择 4 个先进的影响力最大化算法进行对比实验, 包括 ECSH 算法^[6]、WCM 算法^[7]、NetEm

算法^[8]和 IMBC 算法^[20]. ECSH 算法^[6]、WCM 算法^[7]、NetEm 算法^[8]在引言有介绍, IMBC 算法^[20]是一种基于“富人俱乐部”现象和社区结构检测的影响力最大化算法.

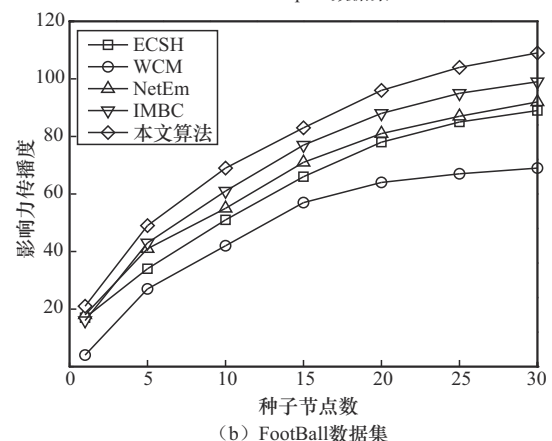
3.5 实验结果对比

(1) 影响力传播度性能

各算法在小规模社交网络上的影响力传播性能如图 9 所示, 统计了种子数量从 1 到 30 的结果. 从图 9(a) 可看出, 当种子节点数为 1 和 5 时, 各算法的影响力传播度性能较接近. 随着种子节点数的增加, WCM 算法的传播性能显著低于其他对比算法, 可知 WCM 算法识别种子节点的准确率不佳. 当种子节点数为 20、25 和 30 时, ECSH 算法的传播性能不佳, 可知 ECSH 算法无法准确识别大规模的种子集. NetEm 算法和 IMBC 算法在不同种子节点数量下的影响力传播性能均十分接



(a) Dolphin数据集



(b) Football数据集

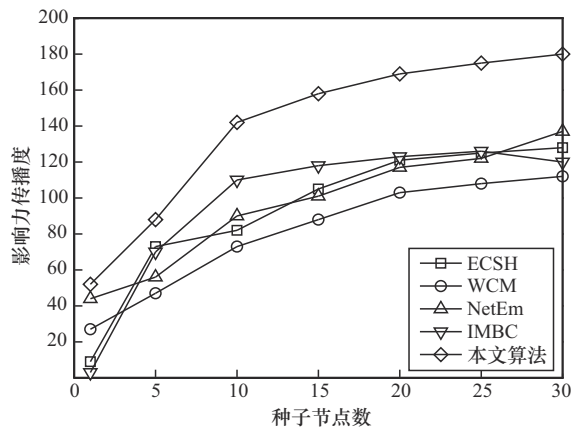
图9 各算法在小规模社交网络上的影响力传播性能

近, 本文算法在不同种子节点数量下的影响力传播性能均优于其他算法。从图9(b)可看出, WCM算法的传播性能明显低于其他对比算法, ECSH算法和NetEm算法的传播性能较接近, 而IMBC算法的传播性能在4种对比算法最佳。本文算法在Football数据集的影响力传播性能优于其他对比算法。综合图9(a)与图9(b)可看出, 尽管小规模社交网络的影响力传播度较低, 但与其他算法相比, 本文算法在小规模社交网络上依然具有一定的优势。

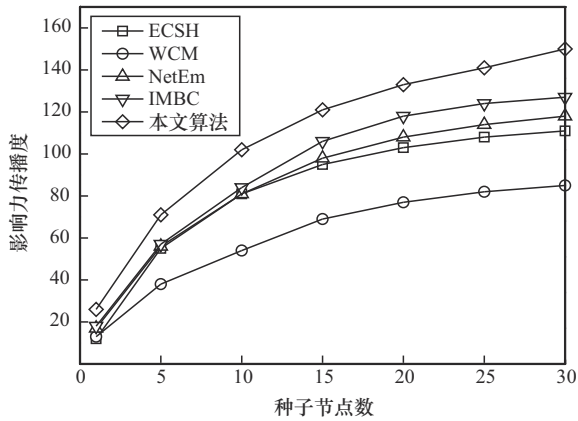
各算法在中等规模社交网络上的影响力传播性能如图10所示。从图10(a)可看出, WCM算法在中等规模社交网络上的传播性能依然低于其他对比算法。当种子节点数为1时, ECSH和IMBC的影响力传播性能较弱, 可知这两种算法无法准确识别出网络中最具影响力的节点。随着种子节

点数的增加, 这两种算法的影响力传播性能得以提升。本文算法在不同种子节点数量下的影响力传播性能均优于其他算法。从图10(b)可看出, 本文算法在Chess数据集的影响力传播性能优于其他对比算法。综合图10(a)与图10(b)可看出, 中等规模社交网络的影响力传播度显著好于小规模社交网络。相较于其他算法, 本文算法在中等规模社交网络上展现出更加明显的优势。

各算法在大规模社交网络上的影响力传播性能如图11所示。从图10(a)和图10(b)可看出, WCM算法在大规模社交网络上的传播性能依然低于其他对比算法, ECSH算法、IMBC算法和NetEm算法的传播性能则十分接近。本文算法在不同种子节点数量下的影响力传播性能均优于其他算法, 可佐证本文算法对于大规模社交网络的有效性。

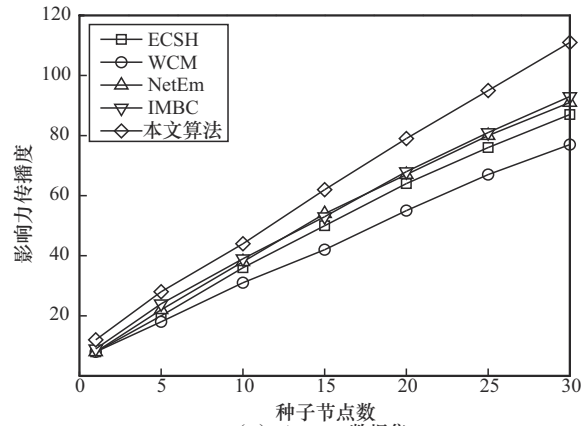


(a) Advogato数据集

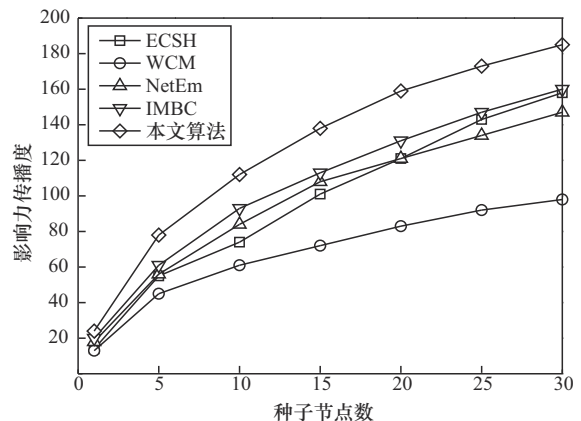


(b) Chess数据集

图10 各算法在中等规模社交网络上的影响力传播性能



(a) Amazon数据集



(b) DBLP数据集

图11 各算法在大规模社交网络上的影响力传播性能



(2) 算法效率实验

不同数据集的平均运算时间如图12所示,给出了几种影响力最大化算法选择种子节点所需的平均运行时间。ECSH算法处理Dolphin、FootBall和Advogato 3个社交网络的运行时间最长,而WCM算法的运行时间较短。对于大规模社交网络,NetEm算法和IMBC算法的运行时间最长。此外,随着社交网络规模增大,ECSH算法、WCM算法、NetEm算法和IMBC算法的平均运行时间呈现大幅升高的趋势。相较于对比算法,本文算法的平均运行时间随社交网络规模的变化较小,当处理大规模社交网络时,本文算法的平均运行时间小于20 s。

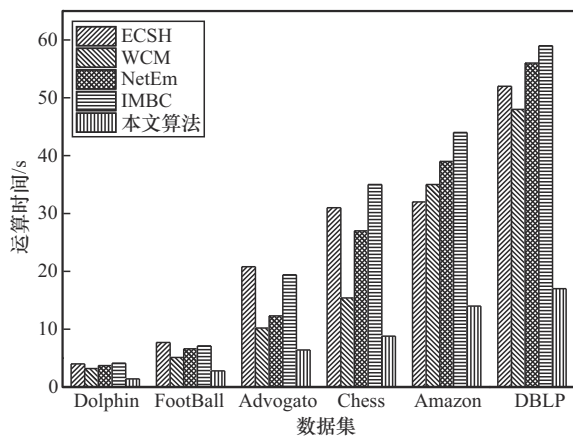


图12 不同数据集的平均运算时间

(3) 消融实验

TranR模型与经典Transformer模型主要存在两个差异之处:经典模型采用节点度数作为节点基础特征,而TranR模型采用节点索引、节点度数和节点shell数组成节点特征向量;经典模型聚合目标节点的1-hop邻居节点作为邻域序列信息,而TranR模型提出随机游走策略发现邻域序列信息。

综上,消融实验包含Transformer模型、多特征Transformer(MFTransformer)模型、随机游走Transformer(RWTransformer)模型和TranR模型。其中Transformer模型采用节点度数作为节

点基础特征,采用1-hop邻居节点作为邻域序列信息;MFTransformer模型采用节点索引、节点度数和节点shell数组成节点特征向量,采用1-hop邻居节点作为邻域序列信息;RWTransformer模型采用节点度数作为节点基础特征,采用随机游走策略发现邻域序列信息。

不同模型的影响力传播性能如图13所示,比较了4种模型在20个种子节点下的影响力传播度。在两个小规模网络上,MFTransformer模型的影响力传播度优于Transformer模型和RWTransformer模型,证明采用节点索引、节点度数和节点shell数组成节点特征向量能有效增强小规模社交网络中节点的特征表达力。随着网络规模扩大,RWTransformer模型的影响力传播度优于Transformer模型和MFTransformer模型,证明通过随机游走发现更大范围的邻域拓扑信息,能有效增强大规模社交网络中节点的特征表达力。总体而言,所提两个改进点能一定程度提高Transformer模型的影响力节点识别性能,但在不同数据集上的效果存在差异。TranR模型将两个改进点结合,在不同规模数据集上均能取得较好的性能。

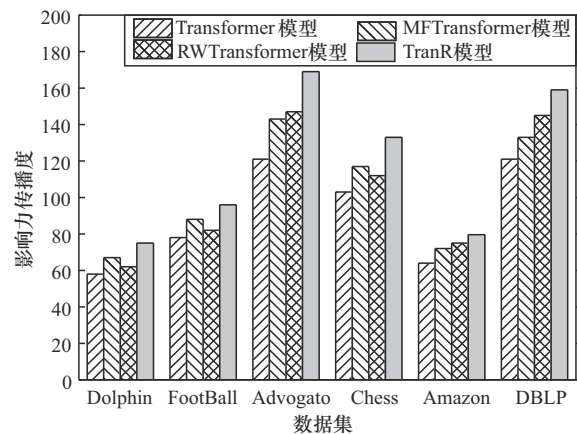


图13 不同模型的影响力传播性能

不同数据集的平均运算时间如图14所示,比较了4种模型选择种子节点所需的平均运行时间。从图14中可知,4种模型在小规模社交网络上的运行时间差异较小。随着网络规模扩大,TranR模型的

运行时间有所增长。由于 TranR 模型通过随机游走策略搜索网络拓扑结构信息，聚合的节点数据量有限，在大规模数据集上的运行时间依然处于合理范围内。

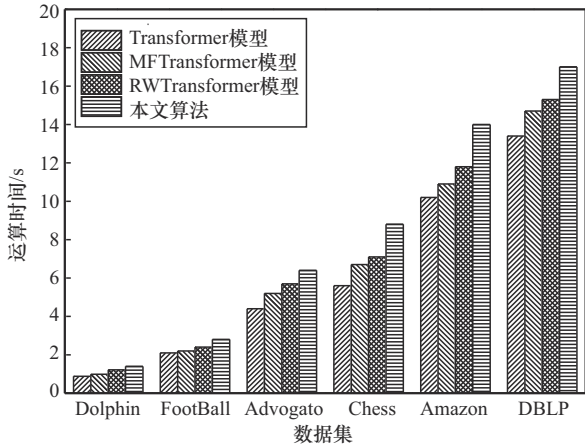


图 14 不同数据集的平均运算时间

4 结束语

为提高基于网络拓扑结构的社交网络影响力最大化算法的性能，本文提出一种基于改进 Transformer 模型的社交网络影响力最大化算法。采用 K-shell 选择网络中的候选影响力节点，运用随机游走策略发现候选影响力节点的邻居关系信息，对 Transformer 模型进行改进，使其支持多属性的节点序列，利用改进 Transformer 模型预测社交网络的种子节点。实验结果表明，本文算法在不同规模和不同结构的社交网络上的影响力传播性能均具有明显优势，且运算效率也较高，所提算法在不同结构社交网络上的性能较为稳定。

参考文献:

[1] 孔芳, 李奇之, 李帅. 在线影响力最大化研究综述[J]. 计算机科学, 2020, 47(5): 7-13.
KONG F, LI Q Z, LI S. Survey on online influence maximization[J]. Computer Science, 2020, 47(5): 7-13.

[2] CHEN W N, TAN D Z, YANG Q, et al. Ant colony optimization for the control of pollutant spreading on social networks[J]. IEEE Transactions on Cybernetics, 2020, 50(9): 4053-4065.

[3] 李美玲, 钱付兰, 徐涛, 等. 基于种子候选的贪心策略影响力最大化算法[J]. 模式识别与人工智能, 2020, 33(11): 1033-1042.
LI M L, QIAN F L, XU T, et al. Greedy strategy influence maximization algorithm based on seed candidates[J]. Pattern Recognition and Artificial Intelligence, 2020, 33(11): 1033-1042.

[4] VENUNATH M, SUJATHA P, KOTI P, et al. Efficient community-based influence maximization in large-scale social networks[J]. Multimedia Tools and Applications, 2024, 83(15): 44397-44424.

[5] 王璿, 张瑜, 周军锋, 等. 基于社交网络的影响力最大化算法[J]. 通信学报, 2022, 43(8): 151-163.
WANG X, ZHANG Y, ZHOU J F, et al. Influence maximization algorithm based on social network[J]. Journal on Communications, 2022, 43(8): 151-163.

[6] BAGHERI E, MIRTALAEI R S. Community-based influence maximization in social networks under a competitive linear threshold model considering positive and negative user views[J]. International Journal of Modern Physics C, 2024, 35(1): 89-96.

[7] QIU L Q, YANG Z Q, ZHU S W, et al. ComIM: a community-based algorithm for influence maximization under the weighted cascade model on social networks[J]. Intell Data Anal, 2022, 26: 205-220.

[8] QIN X, ZHONG C, LIN H X. Community-based influence maximization using network embedding in dynamic heterogeneous social networks[J]. ACM Transactions on Knowledge Discovery from Data, 2023, 17(8): 1-21.

[9] 邓帆, 曾渊, 刘博文, 等. 基于 Transformer 时间特征聚合的步态识别模型[J]. 计算机应用, 2023, 43(S1): 15-18.
DENG F, ZENG Y, LIU B W, et al. Gait recognition model based on temporal feature aggregation with Transformer[J]. Journal of Computer Applications, 2023, 43(S1): 15-18.

[10] 席颖, 鄒学猛, 崔晓晖. 基于 Transformer 的节点影响力排序模型[J]. 计算机科学, 2024, 51(4): 106-116.
XI Y, WU X M, CUI X H. Node influence ranking model based on transformer[J]. Computer Science, 2024, 51(4): 106-116.

[11] 熊才权, 古小惠, 吴歆韵. 基于 K-shell 位置和两阶邻居的复杂网络节点重要性评估方法[J]. 计算机应用研究, 2023, 40(3): 738-742.
XIONG C Q, GU X H, WU X Y. Evaluation method of node importance in complex networks based on K-shell position and neighborhood within two steps[J]. Application Research of Computers, 2023, 40(3): 738-742.

[12] 邹晓红, 许成伟, 陈晶, 等. 大规模时序图中种子节点挖掘算法研究[J]. 通信学报, 2022, 43(9): 157-168.



- ZOU X H, XU C W, CHEN J, et al. Research on seed node mining algorithm in large-scale temporal graph[J]. Journal on Communications, 2022, 43(9): 157-168.
- [13] 唐建荣, 鲍佳彤. 基于改进 SIR 模型的反转事件舆情传播控制研究[J]. 系统仿真学报, 2022, 34(11): 2406-2415.
- TANG J R, BAO J T. Research on network public opinion transmission mechanism of inversion event based on integrating improved SIR model[J]. Journal of System Simulation, 2022, 34(11): 2406-2415.
- [14] 吴亚丽, 任远光, 董昂, 等. 基于邻域 K-shell 分布的关键节点识别方法[J]. 计算机工程与应用, 2024, 60(2): 87-95.
- WU Y L, REN Y G, DONG A, et al. Key nodes identification method based on neighborhood K-shell distribution[J]. Computer Engineering and Applications, 2024, 60(2): 87-95.
- [15] 原慧琳, 冯宠. 基于 K-shell 熵的影响力节点的排序与识别[J]. 计算机科学, 2022, 49(S2): 226-230.
- YUAN H L, FENG C. Ranking and recognition of influential nodes based on K-shell entropy[J]. Computer Science, 2022, 49(S2): 226-230.
- [16] 崔朝阳, 江爱文, 陈思航, 等. 基于 BERT 模型的多层语义粒度视觉对话算法[J]. 中文信息学报, 2023, 37(11): 120-130.
- CUI Z Y, JIANG A W, CHEN S H, et al. Bert based visual dialogue algorithm with multi-level semantic context[J]. Journal of Chinese Information Processing, 2023, 37(11): 120-130.
- [17] WU X, MA Z H. Traveling waves for a nonlocal dispersal susceptible - infected - recovered epidemic model with the mass action infection mechanism[J]. Mathematical Methods in the Applied Sciences, 2021(46): 18837-18860.
- [18] BARABÁSI A, BONABEAU E. Scale-free networks[J]. Scientific American, 2003, 288(5):60-70.
- [19] 邵玉, 陈峻, 刘维. 独立级联模型下基于最大似然的负影响力源定位方法[J]. 计算机科学, 2022, 49(2): 204-215.
- SHAO Y, CHEN L, LIU W. Maximum likelihood-based method for locating source of negative influence spreading under independent cascade model[J]. Computer Science, 2022, 49(2): 204-215.
- [20] KAZEMZADEH F, ASGHAR S A, MIRZAREZAEI M, et al. Determination of influential nodes based on the Communities' structure to maximize influence in social networks[J]. Neurocomputing, 2023, 534: 18-28.

[作者简介]



于树科 (1980-), 男, 江苏商贸职业学院副教授, 主要研究方向为物联网技术和计算机网络。

姚瑶 (1990-), 女, 江苏商贸职业学院实验师, 主要研究方向为计算机应用和图形图像处理。

严晨雪 (1994-), 女, 江苏商贸职业学院讲师, 主要研究方向为计算机应用和网络安全。