计算机科学
COMPUTER SCIENCE

用户行为驱动的时序影响力最大化问题研究

魏鹏，马玉亮，袁野，吴安彪

引用本文

魏鹏，马玉亮，袁野，吴安彪. 用户行为驱动的时序影响力最大化问题研究[J]. 计算机科学, 2022, 49(6): 119-126.

WEI Peng, MA Yu-liang, YUAN Ye, WU An-biao. Study on Temporal Influence Maximization Driven by User Behavior[J]. Computer Science, 2022, 49(6): 119-126.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210700145

用户行为驱动的时序影响力最大化问题研究

魏 鹏[1] 马玉亮[2] 袁 野[3] 吴安彪[1]
1 东北大学计算机科学与工程学院 沈阳 110000
2 东北大学工商管理学院 沈阳 110000
3 北京理工大学计算机学院 北京 100081
(vpeng1009@163.com)

摘 要 影响力最大化IM问题旨在查找社交网络中的一组用户,通过这些用户,使信息在网络中传播的范围最大化。现有研究主要关注静态网络中的IM问题,然而在现实生活中,社交网络是不断演化的,基于静态网络的传播模型(如独立级联模型、线性阈值模型)无法适用于演化网络中的信息传播过程。同时,现有研究忽略了用户行为对信息传播的影响。因此,针对该问题,提出了一种用户行为驱动的独立级联BDIC传播模型,该模型主要根据用户行为对信息的传播过程进行建模,可有效刻画演化社交网络中的信息传播过程。在该模型的基础上,提出了用户行为驱动的影响力最大化算法,主要包括3个步骤:首先,建模消息传播过程,计算演化社交网络中的信息传播概率;然后,提出一种用户行为驱动的反向影响力采样方法,有效查询单个时间点下的种子用户;最后,设计一种不同时间节点(时间序列)下的种子节点查询方法,有效反映演化社交网络中种子节点动态变化的特性。为了评估所提算法的有效性,设计了种子节点与受影响节点的相似度对比方法。通过大量真实数据集上的实验,验证了信息传播概率算法的高效性和扩展性,证明了相比普通的独立级联模型,BDIC模型能更好地建模演化社交网络中的信息传播过程。

关键词:演化社交网络;行为驱动模型;影响力最大化;传播概率矩阵;反向可达集
中图法分类号 TP399

Study on Temporal Influence Maximization Driven by User Behavior

WEI Peng[1],MA Yu-liang[2],YUAN Ye[3] and WU An-biao[1]
1 School of Computer Science and Engineering,Northeastern University,Shenyang 110000,China
2 School of Business Administration,Northeastern University,Shenyang 110000,China
3 School of Computer Science,Beijing Institute of Technology,Beijing 100081,China

Abstract Influence maximization(IM) aims to find a group of users in a social network,through whom information can spread most widely in the network. Existing studies mainly focus on the IM problem in static networks. However,social networks are constantly evolving in real life,and propagation models(such as independent cascading model and linear threshold model) based on static networks are not suitable for the information propagation process in evolving networks. Meanwhile,the existing researches ignore the influence of user behavior on information propagation. Therefore,to tackle this problem,this paper proposes a behavior driven independent cascade(BDIC) propagation model,which can effectively describe the information propagation process in the evolving social networks. Based on this model,a user behavior-driven IM algorithm is proposed. It mainly includes three steps. Firstly,the process of message transmission is modeled to calculate the probability of information transmission in evolving social networks. Then,a user behavior-driven reverse influence sampling algorithm is proposed,which can effectively query the most influential user with a specific time. Finally,a seed query algorithm under different time(time series) is designed,which can effectively reflect the dynamic change characteristics of seed nodes in evolving social networks. To evaluate the effectiveness of the proposed algorithm,a similarity comparison method between seed nodes and the affected nodes is designed. Experiments on real datasets verify the efficiency and scalability of the proposed approaches. The results also demonstrate that the BDIC model can effectively reflect the information propagation process in evolving social networks.

到稿日期:2021-07-14 返修日期:2021-10-20
基金项目:国家自然科学基金(61932004,62002054);中国博士后科学基金(2020M670780);东北大学博士后科研基金
This work was supported by the National Natural Science Foundation of China(61932004,62002054),China Postdoctoral Science Foundation(2020M670780) and Postdoctoral Research Fund of Northeastern University.
通信作者:马玉亮(mayuliang@mail.neu.edu.cn)

Keywords Evolving social networks, User behavior driven model, Influence maximization, Propagation probability matrix, Reverse reachable set

1 引言

随着互联网技术的发展，越来越多的虚拟社交网络相继出现，如大型社交网站 Facebook 以及手机通信形成的人际关系网络等。人们在社交网络中分享并传播自己的想法、新闻和其他信息，从而影响网络中的其他用户。现有研究表明[1-2]，人与人之间的影响力可以被量化。为了研究信息传播过程，Domingos 等[3]首次提出了影响力最大化问题，该问题旨在网络中寻找 k 个用户作为种子节点，使得信息在特定的传播模型(如 IC 传播模型)下，通过 k 个用户在网络中尽可能多地影响到其他用户。影响力最大化具有广泛的应用场景，如市场营销、个性化推荐、专家用户发现等。因此，分析社交网络数据，剖析社会现象，度量用户影响力成为了当前研究的一大热点。

尽管上述影响力最大化问题已受到广泛关注，但是现有的工作主要在静态图中展开，即在传播过程中，社交网络的拓扑结构和节点关系保持不变。然而在现实生活中，信息传播的社交网络均是动态变化的。根据 CNNIC 在 2020 年 9 月给出的中国互联网络发展统计报告[4]，截至 2020 年 6 月，我国即时通信用户规模达 9.31 亿，占网民整体人数的 99.0%，网络视频用户规模达 8.88 亿，占网民整体人数的 94.5%，社交网络中的用户规模不断扩大，社交网络的结构也在不断演化。

现有研究大多忽略了用户行为对信息传播的影响，Matsubara 等[5]于 2012 年发现信息扩散过程往往会持续数天甚至数月，在此期间，用户行为的变化，如新用户的加入与离开、信息的发布或删除，都会对信息的传播产生影响。

因此，本文研究的演化网络中用户行为驱动的影响力最大化问题主要面临两个挑战：

(1)当前影响力最大化问题的解决方案没有考虑用户行为对信息传播的影响，但是用户行为的改变会使信息传播概率发生改变，从而对网络中信息的传播产生影响；

(2)网络演化过程中用户的规模和社交关系不断变化，这种不断变化的特性在静态图中难以观察到，阻碍了目前影响力最大化问题解决方案对最具影响力用户的识别。

为解决上述问题，本文提出演化社交网络中的 BDIC(Behavior Driven Independent Cascade)模型，该模型根据用户行为对信息的传播过程进行建模，使其可以应用于演化社交网络中，从而求得演化社交网络中各个节点的影响力。然后在该模型的基础上，设计演化网络中影响力最大化算法，主要分为 3 个步骤：首先，针对演化社交网络中信息传播概率设计 PBA(Probability Based Actions)算法，该算法根据影响时间窗口策略计算不同时间下的信息传播概率；其次，提出用户行为驱动的反向影响力采样算法 UD-RIS(User Driven Reverse Influence Sampling)，以查询特定时间节点下的种子用户及其影响力；最后，查询演化社交网络中不同时间下的种子用户，该过程有效反映了演化社交网络中种子节点的动态变化特性，设计 SISC(Seeds and Influenced Nodes Similarity Comparison)算法以验证种子集合的有效性。

本文的主要贡献如下：

(1)首次研究了用户行为驱动的影响力最大化问题，并对用户行为如何对演化社交网络中的消息传播产生影响进行了探讨。

(2)提出了用户行为驱动的独立级联 BDIC 传播模型，用于解决演化社交网络中的影响力最大化问题。

(3)首先提出了基于用户行为的消息传播概率计算算法 PBA，然后在 PBA 算法结果的基础上，利用 UD-RIS 算法查询种子用户及其影响力，在不同的时间点执行该过程，可以解决网络演化给影响力最大化问题带来的困难。

(4)在 4 个数据集上的实验验证了 BDIC 模型可以解决演化社交网络中的 IM(Influence Maximization)问题。

本文第 1、第 2 节调研静态社交网络和动态社交网络中的 IM 问题，研究已有的成果和存在问题；第 3 节介绍演化社交网络模型，并给出演化社交网络中影响力最大化问题的定义及符号说明；第 4 节介绍独立级联(Independent Cascade, IC)传播模型，并在 IC 模型的基础上提出用户行为驱动的独立级联模型；第 5 节介绍用户行为驱动的影响力最大化算法的 3 个步骤，即 PBA, UD-RIS 以及 SISC 算法；第 6 节介绍实验设置和实验结果分析；最后总结全文并展望未来。

2 相关工作

如前文所述，影响力分析已经成为了当前社交网络研究的热点方向，研究者们针对社交网络中影响力最大化问题提出了众多的解决方案。本节将从静态社交网络、演化社交网络两个方面就影响力最大化问题的研究现状进行阐述。

2.1 静态社交网络 IM 问题研究现状

Domingos 等[3]先从数据挖掘的角度考虑了影响力的传播和有影响力用户的识别问题。此问题通过一个概率交互模型来解决，并给出了一种启发式方法来选择种子用户。Kempe 等[6]将影响最大化问题建模为离散优化问题(已知为 NP-hard)，并获得了可证明的近似保证，特别是他们提出了两个基本的传播模型，用于描述信息的传播过程，即线性阈值(Linear Threshold, LT)模型和 IC 模型。

Goyal 等[2]利用上述影响力扩散模型，从数据集 Flickr 中的用户行为记录推测用户之间的传播概率，利用最大似然估计分析了用户之间在固定时间的影响力大小，但是该研究在动作失效后将用户之间的传播概率直接设置为 0，这并不符合实际情况，因为用户之间可能会继续出现新的互动。

Yoshida 等[7]提出了一个衡量网站吸引力的新框架，同时考虑了用户利益的分布，并且定义了吸引力因素作为索引，评估用户对网站的关注程度，用户对网站的关注趋势可以反映出网站的影响力变化。Tian 等[8]提出了一个广义启发式框架来解决话题感知的影响力最大化问题，并提出了基于独立级联模型和线性阈值模型的两种主题感知社会影响传播模型。

Yang 等[9]于 2020 年研究了在有限预算下给用户提供怎样的折扣才能使产品使用率最大化的问题，并改进了坐标下降法，以解决该问题。Huang 等[10]于 2021 年研究了传播实体之间的竞争与互补关系，提出了针对实体间不同关系的传播扩散模型，用于解决影响力最大化的问题。

2.2　演化社交网络中 IM 问题研究现状

在演化社交网络中影响力最大化问题中，也涌现了很多先进的模型，Ohsaka 等[11]于 2016 年提出了一个实时的全动态索引数据结构设计的影响分析演进网络，并且设计了网络更新算法，利用这个索引提出了影响力估计和影响力最大化的查询算法。Wang 等[12]于 2017 年从降低谣言影响力着手，提出了一个基于用户体验的动态谣言影响最小化模型，该模型基于一个真实场景，旨在降低谣言的影响力，是一个考虑谣言的全球流行度和个体吸引力的动态传播模型。Wang 等[13]为处理网络发展时种子质量差和处理时间长的问题，定义了一个名为流影响最大化的社交流 IM 方法，维持一组对最近的社交活动具有最大影响价值的用户，并提出了一种 SIC 方法，用于对 IM 问题进行连续查询处理。

Xie 等[14]于 2019 年建立了动态扩散模型，该方法引入了动态扩散网络中连续时间约束影响力最大化问题，利用理论推导证明了模型内影响力扩散函数是单调和子模的，并且该方法能快速逼近最优解。Wu 等[15]研究了时序图影响力最大化问题，并对 IC 模型进行了改进，使其可以应用在时序图上，并提出了 BIMT 算法及其优化算法 AIMT 和 IMIT，以更快速地解决大规模时序图影响力最大化问题。

Wei 等[16]于 2020 年提出了一种基于社区的动态社会网络算法来解决影响力最大化问题，该算法分为 3 个阶段：1）社区检测；2）动态生成候选种子集；3）查询最终种子集，并在实验中取得了良好的效果。Dupuis 等[17]等针对实时竞价系统提出了一种影响力最大化方法，该方法在实时投标环境下进行影响力最大化决策，并且在静态和动态网络中进行了实验对比，取得了良好的结果。

Li 等[18]于 2021 年提出了一种基于内聚熵的动态影响力最大化算法，该算法利用局部聚集因子来计算最具影响力的节点。Wang 等[19]为解决动态网络中的 IM 问题引入了一种基于图嵌入和强化学习的影响力最大化算法，并在此基础上建立了一种强化学习模型来寻找种子节点。

3　问题定义

本节介绍演化社交网络模型，并说明网络中消息的传播过程，而后给出演化社交网络中影响力最大化问题的定义并对本文中使用的符号进行说明。

3.1　演化社交网络模型

演化社交网络即不断发展的网络[20]，本文以时间轮向前演化，第 $0\sim T$ 时间为第一轮，且第 m 轮为 $(m-1)T\sim mT$ 的时间间隔，使用 $G_m=(V_m,E_m)$ 表示时刻 m 的社交网络。本文假设所有的边都是有向边，这里，(u,v) 表示从用户 u 到用户 v 的边，用户之间信息只能沿着边 (u,v) 从 u 到 v。如果存在 (u,v) 则说明 u 是 v 的邻居，在每一个时刻 m，都可以识别出最具影响力的一组种子节点。图 1 为某一时刻下的消息传播过程，蓝色点和线表示消息正在该处传播，随着时间 t 的推移，收到消息的点不断通过边继续传播消息，同时传播消息的边也不断变化，直到没有边可以传播消息，过程停止。

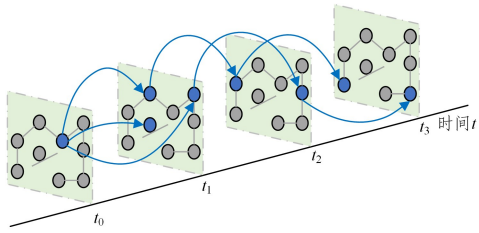


图 1　演化社交网络（电子版为彩图）

Fig. 1　Evolving social network

3.2　问题描述

给定一个 m 时刻下的社交图 $G_m=(V_m,E_m,W_m)$，其中 V 是用户节点集合，E 是边的集合，(u,v) 表示用户 u 和用户 v 之间的联系，W 是边上权重的集合。同时，本文收集各个用户的动作日志，通过分析这个日志，得到一个关系动作元组 $Actions(User,Action,Time)$，具体内容可以表示为 (u,a,t_k)，即用户 u 在 t_k 时刻执行了动作 a。令 A_u 表示用户 u 在数据集中执行的动作，$A_u\cap A_v$ 表示 u 和 v 都执行的动作，接下来定义动作传播。

定义 1（动作传播）　如果存在一个从 v_i 到 v_j 的动作 $a\in A$，当且仅当 $(v_i,v_j)\in E$ 且 $\exists(v_i,a,t_i),(v_j,a,t_j)\in Actions$，并且 $t_i<t_j$ 时，动作传播的概率为 $p_{u,v}$。需要注意的是，v_i 和 v_j 之间建立活动的同时，也建立了两个人的社会联系，这就产生了传播图的概念。

定义 2（传播图）　对于每一个行动 a，都定义一个传播图 $PG(a)=(V(a),E(a))$，并且有 $V(a)=\{v\mid \exists t:(v,a,t)\in Actions\}$，这就形成了一条在 $E(a)$ 中的有向边 $v_i\xrightarrow{\Delta t}v_j$。令 $I_C(S)$ 表示集合 S 的影响力，即在演化图 G_t 的一个传播过程 C 中 S 激活的节点数量，这里 C 是 G_t 上的一个级联传播过程。

与朴素的蒙特卡洛模拟方法相比，Borgs 等[21]提出的反向影响力采样（Reverse Influence Sampling，RIS）方法大大提高了估计影响扩散的效率，并且没有降低准确率。因此，本文改进了 RIS 以进行影响力估算，并根据引理 1 给出计算影响力的具体公式。

定义 3（反向可达集[22] RRS）　一个随机反向可达集的生成过程如下：

（1）随机选择一个节点 $v\in V$；

（2）从 v 节点开始，对指向节点 v 的节点（节点 v 的入邻居）进行随机广度优先遍历，将遍历的节点以激活概率 p 加入集合 R；

（3）直到没有节点可以加入，遍历结束，得到一个反向可达集 R。

引理 1[6]　令 $S\subset V$ 为种子节点集，R 是随机反向可达集，影响力计算式为：

$$I_C(S)=n\cdot Pr(S\cap R\neq\varnothing) \tag{1}$$

其中，n 表示网络 G 中的节点总数。如果 $S\cap R\neq 0$，则称 S 覆盖了一个 RR 集 R，假设 $\mathcal{R}$ 是生成的随机 RR 集的集合，将种子集合 S 对于 $\mathcal{R}$ 的覆盖定义为 $\mathcal{R}$ 中被 S 覆盖的 RR 集的

个数，记为$\Lambda_{\mathcal{R}}(S_k)$，通过 S 对 $\mathcal{R}$ 的覆盖即可估计集合 S 的影响力。

根据以上定义，可以给出演化社交网络影响力最大化问题的定义：给定一个演化图G_t、级联模型 C、正整数 k 以及时间节点 t，在演化社交网络中寻找一个大小为 k 的节点集合 S_k，使得集合S_k在 t 时刻的网络中影响的节点数量最多。

4 用户行为驱动的影响力传播模型

4.1 独立级联模型

考虑一个有向社交网络 $G=(V,E,W)$，其中 V 是顶点集合，E 是边的集合，表示用户之间的关系，每一条边$(u,v)\in E$都有一个影响概率 $p_{uv}\in[0,1)$。每一个节点都有两个状态：激活与未激活。每一个节点都只能从未激活状态转变为激活状态，每个节点可以被它的相邻节点激活。任意节点 v 是否被激活由一个邻居对它的影响值p_{uv}决定。对于给定一个初始活跃节点集合，以下过程说明了级联的传播过程[6]。

(1)在传播开始，给定一个初始活跃的节点集合 S，激活的节点将在之后的级联过程中保持激活状态。

(2)在 t 时刻，新近被激活的节点 u 对它的邻居节点 v 产生影响，成功的概率为 $p(u,v)$。如果 v 有多个邻居节点都是新近被激活的节点，那么这些节点将以任意顺序尝试激活节点 v。

(3)如果节点 v 被成功激活，那么在 $t+1$ 时刻，节点 v 转为活跃状态，并尝试激活其邻居节点；否则，节点 v 在 $t+1$ 时刻状态不发生变化。

(4)不断重复上述过程，直到没有可以被激活节点时，影响传播终止。

4.2 用户关注与行为趋势的关联分析

社交网络中用户与关注者之间交互的一般过程为：用户在某一时间发起或参与了一个活动(如发布了一条原创微博，或者转发参与一项投票)，关注者浏览该用户发布的信息，然后转发或者参与一些自己感兴趣的信息。因此，从用户行为的角度出发进行用户影响力估计，首先需要回答这样一个问题，即两个有关注关系的用户之间的行为是否相似。

本文在数据集 Digg 中选择了 10000 对有直接关注关系的用户和 10000 对没有直接关注关系的用户，将它们分别作为实验组和对照组，实验组和对照组的行为相似度均值和方差如表 1 所列。

表 1 均值方差表

Table 1 Mean and variance

	Mean	Variance
Experimental group	0.2490	0.0171
Control group	0.2390	0.0180

本文选择零假设$H_0:a_1=a_2$，$H_1:a_1\neq a_2$，a_1和a_2分别为实验组和对照组的期望。由于样本数量较大，且均满足正态分布，根据假设检验原理，在方差已知的条件下零假设的拒绝域为：

$$\frac{|\bar{\xi}-\bar{\eta}|}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}>u_{1-\alpha/2} \tag{2}$$

在 $\alpha=0.001$ 的情况下，通过计算可证上式成立，因此在显著性水平为 0.001 的情况下拒绝零假设。这说明用户之间的关注关系确实会给信息的传播带来影响，而用户相似的行为动作就是信息被接受的表现。

4.3 用户行为驱动的影响力传播模型

基于对用户行为与影响力的关联分析，本文进一步探究了用户行为是如何引起用户影响力变化的，进而利用用户行为对信息传播过程进行建模，将网络的演化过程按照离散的时间展开，建立用户行为驱动的独立级联传播模型。本文用$I_C(S)$表示用户集合 S 的影响力，并作出如下假设。

(1) 给定某时刻下的网络状态，每个节点的状态有两种情况，活跃的(如实施动作)和非活跃的。每个节点对于非活跃状态的邻居只进行一次激活尝试，被激活的节点会保持激活状态。

(2)节点受到多个邻居的影响时，它们之间的影响关系相互独立。

(3) 若节点的所有邻居都未成功激活该节点，则视为信息在该节点处的传播中断。

(4)节点间的传播概率会在影响持续时间窗口内保持不变，但会随窗口的移动而改变。

从基于用户行为的微博用户社会影响力分析[23]中可以了解到，用户使用微博的时间主要分为工作时间和晚间时段，因此本文在 BDIC 模型中设计了一个时间窗口$\tau_{u,v}$作为用户对其邻居的影响持续时间，并且利用该持续时间来计算影响概率。

本文利用互动行为定义传播概率，代表用户之间的影响力，即每个用户受到某个邻居影响的操作总数与用户执行操作数之间的比率，其表达式如下：

$$p_{u,v}=\frac{|\{a\mid \exists \Delta t: prop(a,v,u,\Delta t)\wedge 0\leqslant \Delta t\leqslant \tau_{u,v}\}|}{|A_u|} \tag{3}$$

其中，$\Delta t=t_v-t_u$代表用户u_u执行操作 a 的时间与用户u_v执行动作的时间差；$prop$ 代表用户 u 和用户 v 在时间差之内执行相同动作的数量；$\tau_{u,v}$为用户动作的持续性窗口，并且有：

$$\tau_{u,v}=\frac{|\sum_{a\in A}(t_u(a)-t_v(a))|}{|A_u\bigcap A_v|} \tag{4}$$

根据已有研究工作[6]，集合对节点的影响力一般是单调的，即函数$p_u(S)$应满足：当 $S\subset T$ 时，有 $p_u(S)\leqslant p_u(T)$。此外它还应该是子模的，即当 $S\subset T$ 时，有：

$$p_u(S\bigcup\{w\})-p_u(S)\geqslant p_u(T\bigcup\{w\})-p_u(T) \tag{5}$$

由于本文的影响力计算是在离散时间中提出的，因此定义一个活跃用户 v 对其邻居 u 的影响$p_{v,u}$在时间窗口$\tau_{v,u}$之内保持不变，即用户 v 在$[t-(\tau_{(v,u)}/2),t+(\tau_{(v,u)}/2)]$时间间隔内对 u 的传播概率为$p^t_{v,u}$。然后根据算法 UD-RIS 可以返回当前时刻的种子节点集，具体算法见第 5 节。

5 用户行为驱动的影响力最大化算法

本节将介绍基于 BDIC 模型的影响力最大化算法，该算法分为 3 个步骤：针对消息传播的 PBA 方法，查询种子集合的 UD-RIS 方法和评估节点有效性(Seeds and Influenced Nodes Similarity Comparison，SISC)方法。

PBA 算法首先扫描用户动作日志，根据用户的行为计算

对邻居的影响力持续窗口。给定时间点时，可以计算时间点前后窗口内的动作数量，从而计算出信息在用户之间的传播概率。

UD-RIS 算法首先随机选择一组节点，然后从每个节点出发，反向深度遍历生成反向可达集，查找覆盖反向可达集最多的节点，将其加入种子集合，并从反向可达集中删除，循环 k 次返回 k 个种子用户。

SISC 算法首先根据节点的动作将每个节点建模成一个向量，然后根据向量计算节点与受其影响的节点的行为相似度，即可评估种子用户的有效性。

5.1　基于行动的影响概率计算算法

根据 4.2 节中提出的 BDIC 模型，本小节将给出算法 PBA 的具体过程，如算法 1 所示。

算法 1　PBA

输入：一个用户行为网络和时间 t

输出：用户影响力矩阵 **M**

1. Initialize influence matrix **M**
2. for each user u in users
3. for each user v:(u,v)∈E
4. $A_u^t \cap A_v^t = \emptyset$
5. for each action a in $A_u \cap A_v$
6. if $t-\frac{\tau_{u,v}}{2} \leqslant t_u \leqslant t+\frac{\tau_{u,v}}{2}$
7. $A_u^t \cap A_v^t ++$
8. $p_{uv}=\frac{A_u^t \cap A_v^t}{A_u^t}$
9. update p_{uv}
10. add p_{uv} in influence matrix **M**
11. return matrix **M**

首先遍历网络中用户的邻居（算法 1 第 1－3 行），并检查邻居与该用户是否有相同的行为，如果邻居有跟随行为并且在该用户的影响力持续窗口之内，则重新计算用户与该邻居之间的信息传播概率（算法 1 第 4－9 行），遍历所有行为，将最终的传播概率加入传播矩阵当中，返回传播概率矩阵 ***M***（算法 1 第 10－11 行）。

5.2　种子节点查询算法

种子节点查询算法的代码如算法 2 所示。

算法 2　UD-RIS

输入：有向图 G；用户行为影响概率 P；种子集合大小 k；反向可达集的大小 m

输出：种子节点集合 Seed

1. $S_k^* = \emptyset$
2. $\mathscr{R}=\emptyset$，R=∅
3. for i=0 to mc
4. Randomly sample a node v∈V
5. R={v}
6. QueueQ. push(v) and mark v as active
7. whileQ is not empty
8. let u=Q. top
9. Q. pop
10. for each w in-neighbors of u
11. if w is inactivated & rand()≤p(u,w) //用户 w 是 u 的邻居且 w 未被激活的情况下，如果能被 u 激活，则将 w 加入 u 的反向可达集。
12. add w to R
13. mark w is active
14. $\mathscr{R} \leftarrow \mathscr{R} \cup R$
15. R=∅
16. for i=0 to k
17. $v=\arg\max_{v' \in V}(\Lambda_{\mathscr{R}}(S_k^* \cup \{v'\})-\Lambda_{\mathscr{R}}(S_k^*)$ //$\Lambda_{\mathscr{R}}$为对集合 $\mathscr{R}$ 的覆盖，v 即对集合 $\mathscr{R}$ 边覆盖最大的节点。
18. $S_k^* \leftarrow S_k^* \cup \{v'\}$
19. return S_k^*

该算法首先初始化反向可达集为空（算法 2 第 1－2 行），随机选择一个节点 v 加入队列中，并设置该节点为激活状态（算法 2 第 3－6 行），从该节点开始进行深度优先遍历，并尝试激活路径上的节点，若节点被激活则加入由 v 生成的反向可达集当中（算法 2 第 7－13 行）。重复 mc 次，将得到的所有反向可达集加入反向可达集的集合 $\mathscr{R}$（算法 2 第 14－15 行），然后查找对反向可达集的集合 $\mathscr{R}$ 边覆盖最大的节点，将其加入种子集合，并从反向可达集中删除，循环 k 次返回 k 个种子用户（算法 2 第 16－19 行）。

图 2 为算法 PBA 的一个实例，图 2(a)为 3 个用户 P，Q，R 的动作日志，其中包含 3 个动作 a_1，a_2 和 a_3；图 2(b)－图 2(d)分别为 3 个动作的传播图 PG，边上的数字表示传播动作所需要的时间。传播图 PG 中的边都是有向边，建立模型影响力矩阵 $\boldsymbol{IM}$，$\boldsymbol{IM}[i,j]=(p_{i,j},\tau_{i,j})$，例如 $\boldsymbol{IM}[P,R]=(1/2,10)$表示用户 P 在时间为 10 的窗口内执行了 a_1 和 a_2 两个动作，用户 R 在窗口内有一个动作 a_1 与用户 P 相同，因此时间间隔$\tau_{P,R}$为 10 时，概率$P_{P,R}$为 1/2。

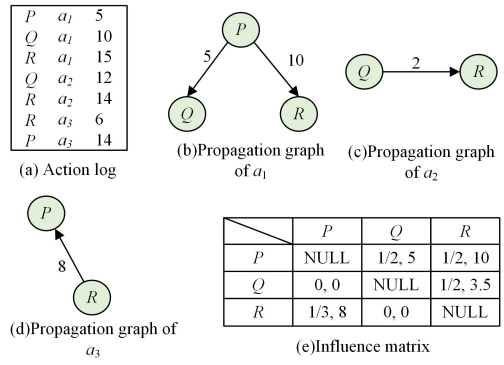


图 2　传播概率矩阵

Fig. 2　Propagation probability matrix

需注意的是，在 PBA 算法中要提前扫描用户动作日志，评估每个用户的影响力持续窗口$\tau_{u,v}$。

5.3　种子节点相似度对比算法

为了验证用户行为驱动的影响力最大化算法是否可以查询出符合真实情景的种子集合，本文设计了 SISC 算法，其包含两个过程，算法伪代码如算法 3 所示。

（1）构建节点向量 ***BuildVector***。将节点执行过的动作进行 0-1 标记，把节点抽象为动作向量（算法 3 第 1－9 行），该向量的维度等于动作的种类，若用户执行过某个动作，则动作对应位置设为 1，遍历节点邻居并将每个邻居也表示成向量

(算法 3 第 10－13 行)。

(2)计算种子节点与受影响节点行为的相似度。利用第一步中得到的动作向量,计算节点之间的余弦相似度(算法 3 第 14 行)。为了实验更加合理,对相似度取平均值,并将其作为最终结果(算法 3 第 15－17 行)。

算法 3 SISC

输入:影响力最大的种子集 S

输出:种子节点与其邻居的相似度

1. similarity=0
2. for u in S
3. u_neigh=∅
4. fo rv in User
5. if v is u's neighbor
6. u_neigh←u_neigh∪{v}
7. $\mathbf{T}_u$=Vector(|Actions|,0)
8. for a in A_u
9. $\mathbf{T}_u[a]=1$
10. for v in u_neigh
11. $\mathbf{T}_v$=Vector(|Actions|,0)
12. for a in A_v
13. $\mathbf{T}_v[a]=1$
14. u_v_simil=cosine($\mathbf{T}_u$,$\mathbf{T}_v$)
15. $u\_simil=\frac{\Sigma u\_v\_simil}{|u\_neigh|}$
16. $similarity=\frac{\Sigma u\_simil}{|S|}$
17. return similarity

6 实验和评估

为了验证传播概率算法 PBA 及测试算法 SISC 的有效性,本文选取了两种不同规模的真实数据集和两种合成数据集作为输入数据,用于在演化社交网络中用户行为驱动的节点影响力评估和种子节点的选取。

6.1 实验数据和参数设置

本文的实验在 4 个数据集上进行:Digg 社交网络数据集[24];新浪微博社交网络数据集[25-26];在 Digg 数据集的基础上基于随机分布生成的数据集 data3;在 Weibo 数据集的基础上利用正态分布生成的数据集 data4。具体如表 2 所列。

表 2 数据集

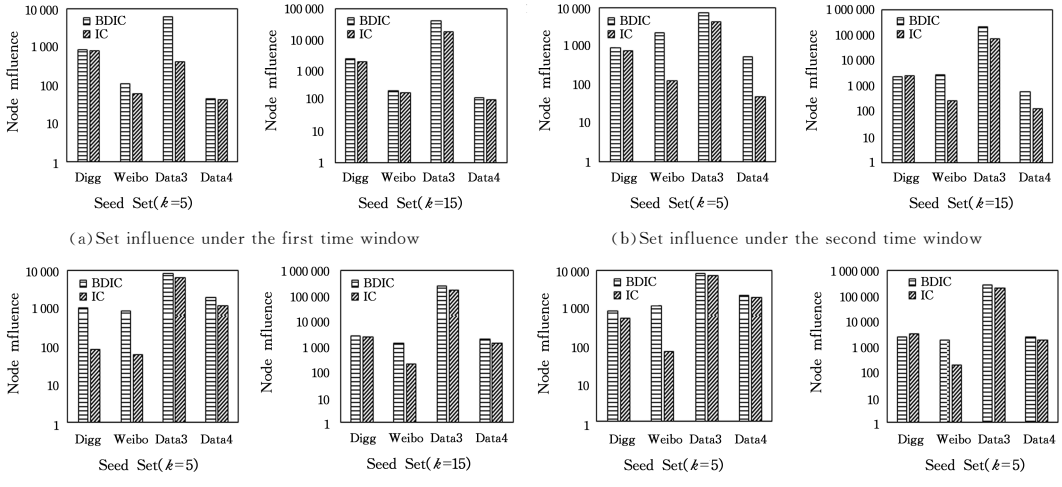
Table 2 Datasets

Datasets	Nodes	Edge	Action
Digg	139 000	100 000	301 000
Weibo	1 776 000	308 000	300 000
Synthetic data3	329	15 000	20 000
Synthetic data4	4 620	38 000	492 000

本文主要研究用户行为对影响力最大化算法的影响。为了使实验结果更具有效性,仅采用固定窗口 $\tau_{u,v}$ 并不合理,因此本文还考虑了其他的时间窗口 $\tau_{u,v}$,在实验过程中对不同数据集的时间窗口作相应的调整。对于用户活动密集的数据集,采用较小的时间窗口可以更加准确地查询种子节点;而对于活动稀疏的数据集,为了保证有效性应选择较大的时间窗口,详细的窗口设置见 6.3 节。

6.2 种子节点影响力测试

本文通过 UD-RIS 算法查询影响力最大的用户集合并计算该集合的影响力。图 3 给出了 4 个不同时间窗口下种子集合影响力的对比图,其中种子集合规模分别选取 5 和 15,横线柱形是利用 BDIC 模型得出的种子集影响力大小,斜线柱形是利用普通 IC 模型得出的种子集合影响力大小。



(a)Set influence under the first time window (b)Set influence under the second time window

(c)Set influence under thethird time window (d)Set influence under the fourth time window

图 3 不同时间窗口下种子集合影响力对比

Fig. 3 Seed sets influence comparison in different time windows

图 3(a)－图 3(d)分别为实验划分的 4 个窗口内种子集合的影响力大小对比。可以看出,种子集合的大小选择 5 或者 15 时,虽然在数据集 Digg,Weibo 以及 Data4 中两种模型得到的种子集合影响力都较为接近,但 BDIC 模型要比普通 IC 模型高 100 至 1 000,而在数据集 Data3 中的效果更加明显,两种模型种子集合的影响力差距都在 10 000 以上。

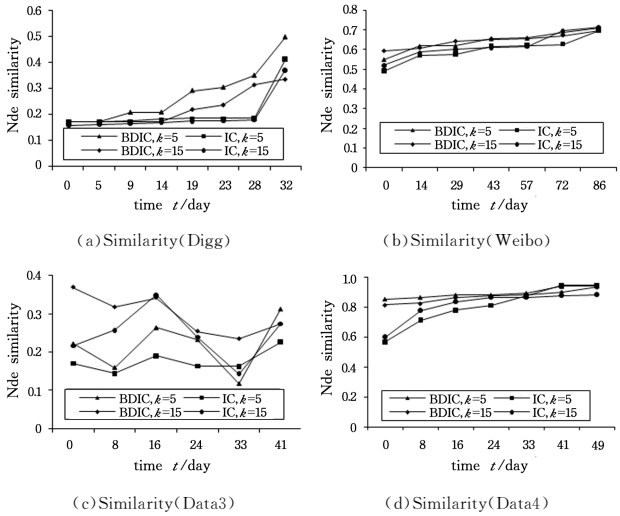
由图 3 可以看出,不论是在真实的数据集还是合成数据集上,BDIC 模型得出的种子节点影响力均大于 IC 下的种子集合。这说明在寻找最具影响力的节点方面,BDIC 模型比 IC 更有效,得出的节点集合的影响也更加广泛。为了在关注查询演化社交网络中影响力最大的节点的同时保证找到的节点集合符合真实情景,本文设计了有效性实验,用以测试两种

模型得到的种子集合的真实性。

6.3　种子节点有效性测试

为了进一步判断 BDIC 模型得出的种子集合是否符合实际情景，本文从用户行为的角度出发设计了 SISC 算法。通过比较节点与受其影响节点的行为相似度来判断该节点是否对其他节点产生过影响，从而验证 BDIC 模型得出的种子集合的有效性。

图 4 给出了在 BDIC 模型和 IC 模型下不同大小的种子集合与受其影响节点的行为相似度随时间变化的情况。通过对每个数据集中用户行为的分析，本文将数据集 1，2，3，4 的窗口大小分别设置为 4.6 天、14.3 天、8.1 天和 8.1 天。同时为了方便观察，将图 4 中的横坐标时间设置为整数。



(a)Similarity(Digg)　(b)Similarity(Weibo)

(c)Similarity(Data3)　(d)Similarity(Data4)

图 4　$k=5$，$k=15$ 时两种模型有效性测试

Fig. 4　Validity test of two models when $k=5$，$k=15$

6.3.1　Digg 数据集实验分析

图 4(a)表示在 Digg 数据集中的实验，在刚开始的 0～5 天内，节点之间的影响不明显，两种模型之下的种子节点与其邻居的行为相似度并不高，在 k 为 5 时，只有 15.50% 和 17.33%。但是在 9～28 天时，网络中的用户开始相互影响，BDIC 模型中用户的影响概率开始变化，BDIC 得出的种子节点与受其影响的节点之间相似度开始上升并达到 35.04%，而通过 IC 模型计算的种子集与受其影响的节点相似度为 18.53%，在 28～32 天时，两种模型下的种子集合与邻居相似度都呈上升趋势，在 32 天时 BDIC 模型下的相似度最高为 49.79%，IC 模型下为 37.03%。

6.3.2　Weibo 数据集实验分析

图 4(b)为在 Weibo 数据集中展开的实验，由于微博数据集记录的时间跨度很大，因此选取的实验窗口也较大，涵盖的用户动作较多。在 k 为 5 时，BDIC 模型下用户之间的行为相似度较高，能达到 54.82%，IC 模型下能达到 48.81%。随着网络的演化，0～58 天内 BDIC 和 IC 下用户的相似度都呈上升趋势，BDIC 下的节点相似度从 54.82% 上升至 65.73%，而 IC 下的节点相似度从 48.81% 仅上升至 62.15%。

6.3.3　Data3 数据集实验分析

图 4(c)为在数据集 Data3 上的实验，该数据集是 Digg 数据集仿真出的正态分布动作数据，整个数据集动作时间分布服从正态分布。在网络演化时，由于正态分布的特性，动作的分布会出现中间部分较密集、两边部分较稀疏的现象，因此用户相似度随动作数量变化呈现出在 20%～40% 之间高低起伏的趋势。在 k 为 15 时，BDIC 模型下的相似度为 22.11%，IC 下为 16.97%。在开始的 0～8 天内，BDIC 下的节点相似度下降到 15.88%，IC 下的相似度下降到 14.37%，但是在 8～16 天内动作变得密集，BDIC 下的相似度上升到 26.32%，IC 下的相似度上升到 18.98%，16 天之后动作变得稀疏，两种概率下的相似度又呈下降趋势。

6.3.4　Data4 数据集实验分析

图 4(d)为在数据集 Data4 上的实验，该数据集为 Weibo 数据集仿真出的随机数据，仿真时将数据集的时间范围压缩至与数据集 Data3 一致。一开始用户动作就较为密集，因此在 BDIC 模型下得出的相似度相对较高，为 84.90%，在 IC 模型下得出的相似度只有 56.62%。随着时间演化，BDIC 下的相似度会上升至 94.02%，IC 下的相似度会逐步上升至 94.40%，这是因为将时间跨度压缩后，用户的动作变得非常密集。

因此，综合图 4 的结果可以得出：随着时间的不断演化，不论是种子集合大小取 5 还是 15，由 BDIC 得出的种子节点集合相似度比 IC 模型下的种子节点相似度平均要高 10%～20%，这说明基于用户行为的 BDIC 模型下找到的种子节点集合比 IC 模型更加真实。这一结果符合现实中的消息传播行为：当用户接受其他节点的影响后，该用户行为会与影响他的节点趋于相似，证明了基于 BDIC 模型方法找到的种子节点更加贴合实际情景。

结束语　本文从用户行为角度出发，对演化社交网络中的影响力最大化问题进行了研究，该问题的目的是动态地查找网络中影响力最大的 k 个用户。针对该问题提出了用户行为驱动的独立级联模型，以查询最具影响力的用户。在该模型的基础上，提出了用户行为驱动的影响力最大化算法，主要包括 3 个步骤：首先，提出了 PBA 算法计算消息传播概率，并将其作为用户的影响概率；其次，针对查询种子用户问题，设计了 UD-RIS 算法查询演化社交网络下影响力最大的 k 个用户，并估计他们的影响力；最后，设计了 SISC 算法来测试种子节点的有效性。

通过在 4 个不同规模的数据集上进行影响力的对比实验，一方面，证明了本文方法不仅能够找到网络中影响力最大的节点集合；另一方面，在追求影响力最大的同时，本文方法兼顾了最具影响力节点的有效性，证明了基于 BDIC 模型的影响力最大化算法能查询出更加真实有效的用户集合。未来的工作不仅可以探究用户行为方面的交互，还可以研究情感方面的交互，在合适的情境之下，将用户情感融合到影响力最大化算法中，以探究更加真实可靠的种子节点用户。

参 考 文 献

[1]　LI Y，FAN J，WANG Y，et al. Influence Maximization on Social Graphs：A Survey[C]// IEEE Transactions on Knowledge and Data Engineering. IEEE，2018：1852-1872.

[2]　GOYAL A，BONCHI F，LAKSHMANAN L. Learning influence

probabilities in social networks[C]// Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York: WSDM, 2010: 241-250.
[3] DOMINGOS P, RICHARDSON M. Mining the network value of customers[C]// Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2001: 57-66.
[4] CNNIC. The 46th china Statistical Report on Internet Development[R/OL]. 2020. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/202009/t2020092971257.html.
[5] MATSUBARA Y, SAKURAI Y, PRAKASH B A, et al. Rise and fall patterns of information diffusion: model and implications [C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2012: 6-14.
[6] KEMPE D, KLEINBERG J, TARDOS E. Maximizing the spread of influence through a social network[C]// Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2003: 137-146.
[7] YOSHIDA A, HIGURASHI T, MARUISHI M, et al. New Performance Index 'Attractiveness Factor' for Evaluating Websites via Obtaining Transition of Users' Interests[J]. Data Science and Engineering, 2020, 5(3): 48-64.
[8] TIAN S, MO S, PENG Z. Deep Reinforcement Learning-Based Approach to Tackle Topic-Aware Influence Maximization [J]. Data Science and Engineering, 2020, 5(3): 1-11.
[9] YANG Y, MAO X, PEI J, et al. Continuous Influence Maximization[J]. Association for Computing Transactions on Knowledge Discovery from Data, 2020, 14(3): 1-38.
[10] HUANG H, MENG Z, SHEN H. Competitive and complementary influence maximization in social network: A follower's perspective[J]. Knowledge-Based Systems, 2021, 213(3): 106600.
[11] OHSAKA N, AKIBA T, YOSHIDA Y, et al. Dynamic influence analysis in evolving networks[C]// Proceedings of the Very Large Data Bases Endowment. 2016: 1077-1088.
[12] WANG B, CHEN G, FU L, et al. DRIMUX: Dynamic Rumor Influence Minimization with User Experience in Social Networks [C]// IEEE Transactions on Knowledge and Data Engineering. 2017: 2168 -2181.
[13] WANG Y, FAN Q, LI Y, et al. Real-Time Influence Maximization on Dynamic Social Streams[C]// Proceedings of the VLDB Endowment. 2017: 805-816.
[14] XIE M, YANG Q, WANG Q, et al. DynaDiffuse: a dynamic diffusion model for continuous time constrained influence maximization[C]// Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015: 346-352.
[15] WU A B, YUAN Y, QIAO B Y, et al. The Influence Maximization Problem Based on Large-Scale Temporal Graph[J]. Chinese Journal of Computers, 2019, 42(12): 2647-2664.
[16] WEI J, CUI Z, QIU L, et al. A community-based algorithm for influence maximization on dynamic social networks[J]. Intelligent Data Analysis, 2020, 24(1): 959-971.
[17] DUPUIS D, MOUZA D, TRAVERS N, et al. Real-Time Influence Maximization in a RTB Setting[J]. Data Science and Engineering, 2020, 5(9): 224-239.
[18] LI W, ZHONG K, WANG J, et al. A dynamic algorithm based on cohesive entropy for influence maximization in social networks[J]. Expert Systems with Applications, 2021, 169(5): 114207.
[19] WANG C, LIU Y, GAO X, et al. A Reinforcement Learning Model for Influence Maximization in Social Networks[C]// Database Systems for Advanced Applications. Cham, 2021: 701-709.
[20] WU X, FU L, MENG J, et al. Maximizing Influence Diffusion over Evolving Social Networks[C]// Proceedings of the Fourth International Workshop on Social Sensing. New York, USA, 2019: 6-11.
[21] BORGS C, BRAUTBAR M, CHAYES J, et al. Maximizing Social Influence in Nearly Optimal Time[C]// Proceedings of the 2014 Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, 2013: 946-957.
[22] GUO Q, WANG S, WEI Z, et al. Influence Maximization Revisited: Efficient Reverse Reachable Set Generation with Bound Tightened[C]// Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. New York, USA, 2020: 2167-2181.
[23] MAO J X, LIU Y Q, ZANG M, et al. Social Influence Analysis for Micro-Blog User Based on User Behavior[J]. Chinese Journal of Computers, 2014, 37(4): 791-800.
[24] HOGG T, LERMAN K. Social dynamics of Digg[J]. EPJ Data Science, 2012, 1(1): 1-5.
[25] ZHANG J, LIU B, TANG J, et al. Social influence locality for modeling retweeting behaviors[C]// Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. Beijing, China, 2013: 2761-2767.
[26] ZHANG M, LI W H. Influence maximization algorithm based on user interactive representation[J/OL]. Journal of Computer Applications. http://kns.cnki.net/kcms/dtail/51.1307.TP.20201229.1645.010.html.

WEI Peng, born in 1996, postgraduate, is a member of China Computer Federation. His main research interests include temporal graph and graph data management.

MA Yu-liang, born in 1990, postdoctor, is a member of China Computer Federation. His main research interests include graph databases, location-based social networks and social networks analysis.

(责任编辑:李亚辉)