

# 基于加权 PageRank 的异质网络影响力最大化

韩 婷<sup>1</sup>, 周丽华<sup>1\*</sup>, 黄亚群<sup>1</sup>, 姜懿庭<sup>2</sup>

(1. 云南大学 信息学院, 云南 昆明 650504;

2. 云南师范大学 信息学院, 云南 昆明 650500)

**摘 要:**影响力最大化问题是信息网络挖掘中的热门研究问题之一,大多数信息网络包含了多种不同类型的节点和连接边,其本质属于异质信息网络,然而以前关于影响力最大化问题的研究大多停留在同质信息网络,它们考虑的节点和连接边类型单一,这与现实的信息网络有所差别。异质信息网络的影响力最大化问题其关键在于如何识别异质信息网络中最有影响力的节点。为了能融合网络中的异质信息并衡量节点影响力,提出了一种基于加权 PageRank 的异质信息网络影响力最大化算法。该算法保留了网络中所有类型节点和连接边的信息,通过考虑异质信息网络中不同类型节点之间的影响关系来得到节点的最终影响力,从而实现异质信息网络的影响力最大化。该算法能更好地描述节点和连接边的异质性,并在两个真实的数据集上验证了算法的有效性。

**关键词:**异质信息网络;信息网络挖掘;信息扩散;影响力最大化;加权 PageRank

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2022)03-0046-08

doi:10.3969/j.issn.1673-629X.2022.03.008

## Influence Maximization of Heterogeneous Networks Based on Weighted PageRank

HAN Ting<sup>1</sup>, ZHOU Li-hua<sup>1\*</sup>, HUANG Ya-qun<sup>1</sup>, JIANG Yi-ting<sup>2</sup>

(1. School of Information, Yunnan University, Kunming 650504, China;

2. School of Information, Yunnan Normal University, Kunming 650500, China)

**Abstract:** The influence maximization problem is one of the hot research problems in information network mining. Most information networks contain a variety of different types of nodes and connection edges, which belong to heterogeneous information networks in nature. However, previous studies on influence maximization mostly stay in homogeneous information networks, which consider a single type of nodes and connection edges and is different from the reality of information networks. The key to influence maximization of heterogeneous information network lies in how to identify the most influential nodes in heterogeneous information network. In order to fuse heterogeneous information in the network and measure the influence of nodes, we propose a weighted PageRank based influence maximization algorithm for heterogeneous information networks, which calculates the influence of nodes through the influence relationship between different types of nodes. The proposed algorithm can better describe the heterogeneity of nodes and edges, of which the validity is verified on two real data sets.

**Key words:** heterogeneous information network; information network mining; information diffusion; influence maximization; weighted PageRank

## 0 引 言

随着各种各样社交网络的出现,人与人之间的联系越来越紧密,人们的学习、工作和生活正在不断地被改变。社交网络中信息的传播和影响力无处不在,通过社交网络,具有高影响力的名人可以影响他人的看

法和行为。准确度量不同对象之间的影响力,有助于识别社交网络中最具影响力的对象并促进信息的快速传播,对谣言传播、流行病传播、产品营销以及推荐系统等工作起着至关重要的作用<sup>[1-3]</sup>,因此影响力最大化研究受到了研究人员的极大关注。

收稿日期:2021-04-02

修回日期:2021-08-04

**基金项目:**国家自然科学基金(61762090,62062066,61966036);云南省自然科学基金(2016FA026);国家社会科学基金(18XZZ005);云南省高等学校科技创新团队项目(IRTSTYN)

**作者简介:**韩 婷(1996-),女,硕士研究生,研究方向为社会网络分析;通信作者:周丽华(1968-),女,博士,教授,博导,CCF会员(42099M),研究方向为数据挖掘、社会网络分析等。

影响力最大化问题被认为是病毒营销的一个直接数学刻画。其目的就是希望利用病毒式营销手段,在社交网络找到少数重要的节点作为种子集,利用这些种子集进行信息的传播从而达到在社交网络中影响力的最大化<sup>[4]</sup>。目前,传统的影响力最大化方法有基于中心度、PageRank、特征向量和启发式算法等,其中 PageRank<sup>[5]</sup>是一种重要的算法,该算法最初是 Google 公司为了衡量网页等级和重要性而提出的,它从网页数量和质量综合考虑了页面的重要性,能较好地刻画页面的性质,并描述对象之间的关系。这些传统的算法在同质信息网络中取得了较好的结果,但是同质信息网络中的节点和链接关系类型单一,没有区分对象及其关系的异质性<sup>[6]</sup>,这与实际的现实网络不符。现实中的网络大多是异质信息网络,包含了多种类型的对象及多种关联类型的链接关系<sup>[7-8]</sup>,网络中的一个实体对象的影响力不仅受到同种类型对象的影响,还与其他类型对象有关。由此关于影响力最大化问题的研究正逐步从同质信息网络转向异质信息网络。

异质信息网络包含了多种类型的对象及链接关系,相对于同质信息网络,节点和链接类型、语义关系更为丰富<sup>[8]</sup>,这些丰富的信息可以更全面地评价节点的影响力。如图 1 所示的文献信息网络 DBLP 就是典型的异质信息网络,它包含了四种类型的节点:A(作者)、P(文章)、C(会议)和 T(主题),六种关系:A-P(编写/被编写),P-C(发表/被发表),P-T(包含/被包含)。评价一个作者的影响力不仅要从作者发表的文章数量和质量来衡量,还要从他撰写的文章的内容主题、所属会议以及与他合作的作者等方面考虑,通过融合这些丰富的信息能更好地刻画现实网络中不同节点对象之间的影响力情况。

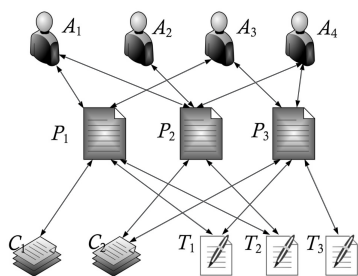


图 1 DBLP 网络

由于异质信息网络包含了多种类型的对象及链接关系,并且网络结构复杂,各节点之间不是相互独立的,他们通过各种关系相互影响。如何有效地利用不同类型对象间的关系成为异质信息网络分析的一个难点。Zhao 等人<sup>[9]</sup>对异质信息网络中两种不同类型的节点使用 PageRank,保留了节点间的连接关系,更好地考虑到不同类型节点彼此间的影响。但是他们只考

虑了直接相连的两种类型节点,忽略了异质网络中同种类型节点和其他非直接相连的不同类型节点的影响,同时他们还将所有节点之间的初始连接关系的权重视为相同。然而,现实中并非所有节点间的连接关系都是同等重要的,为了能进一步区分连接关系的重要性并考虑到所有类型节点间的影响,该文针对异质信息网络提出了一种基于加权 PageRank 的影响力最大化算法 (Comprehensive Weighted PageRank, CWPR)。CWPR 根据不同节点之间的连接关系赋予对应的权重,这样可以更全面地考虑节点的重要性。

主要工作如下:

(1)将异质信息网络分解为若干个只含一种连接类型的网络,再根据各节点之间连接关系的次数分配对应的权重。网络的分解简化了复杂的网络结构,权重的分配区分了节点间连接关系的重要性,有助于准确度量不同节点之间的影响力。

(2)提出了一种基于加权 PageRank 的影响力最大化算法 CWPR,其中影响力的度量考虑了不同类型节点的直接影响和间接影响,从而更好地描述了节点影响力的复杂性和异质性,全面保留了异质信息网络中的信息,使找到的种子节点具有较高的影响力。

(3)在 DBLP 和 Yelp 两个数据集上进行了实验,通过与其他同质和异质影响力最大化算法的对比,验证了 CWPR 的合理性和准确性。同时讨论了参数和边权重对于算法性能的影响。

## 1 相关工作

Kempe 等人<sup>[10]</sup>首次将影响力最大化问题表示成离散的优化问题,证明了该问题是一个 NP-hard 问题,并基于单调次模性提出了有效的贪心算法。该算法能得到最优解,但是不能改进算法的时间复杂度。后来 Leskovec 等人<sup>[11]</sup>提出了 CELF 算法,CELF 算法在实验中的效率得到了很大的提升。随着对影响力最大化问题研究的进一步深入,相关工作也越来越多,Goyal 等人<sup>[12]</sup>又进一步改进了 CELF 算法,提出 CELF++算法。当问题规模较大时,CELF++算法并不适用,于是 Chen 等人<sup>[13-15]</sup>又提出了 DegreeDiscount、PMIA、LDAG 等算法,大大提高了运算速度。周明洋等人<sup>[16]</sup>从多节点的综合影响力角度出发,基于 Rayleigh 熵机制,提出了一种指标刻画多节点的综合影响力算法。曹玖新等人<sup>[17]</sup>基于用户交互的主题偏好计算不同类别信息下节点间的影响概率,并结合扩展的传播模型和信息扩散的特点,提出基于节点子图的影响力计算算法。杨书新等人<sup>[18]</sup>基于三度影响力原则,综合考虑局部度量的适宜层次及大规模网络的可扩展性,提出一种基于 3 级邻居的节点影响力度量算法。Oriedi 等

人<sup>[19]</sup>提出选择性广度优先遍历算法,对来自社交网络成员之间实际社交行为进行影响力建模,有效地生成影响最大化的最佳种子集。目前从运算效率、网络结构等方面对影响力最大化问题的研究工作越来越多,影响力最大化问题也正逐步从同质信息网络转向异质信息网络。

在异质信息网络中,Deng 等人<sup>[20]</sup>设计了一个基于互动记录、社交友谊、标签和话题的 MIF 模型来衡量用户之间的社交影响力,还有基于同元路径考虑信息熵<sup>[21-22]</sup>来定位有影响力的节点等。Keikhar 和 Rahgoza 等人<sup>[23]</sup>利用深度学习技术获得异质信息网络节点的特征,根据节点的本地和全局结构特性得到最具影响力的节点。然而,由于异质信息网络相对于同质信息网络的网络结构、性质更为复杂,目前对于异质信息网络影响力最大化的研究还未足够成熟,因此在异质信息网络中对于影响力最大化的研究还存在很大的进步空间。

## 2 相关概念及问题定义

该文的主要目的是利用加权 PageRank 综合考虑各种类型节点之间的影响关系,从而挖掘出异质信息网络中影响力最大的节点。本节主要介绍所涉及的一些相关概念及问题定义。

### 2.1 相关概念

定义 1 异质信息网络<sup>[7,24]</sup>:信息网络由一个带有对象类型的映射函数  $\tau: V \rightarrow A$  和关系类型映射函数  $\varphi: E \rightarrow R$  的有向图  $G = (V, E, \tau, \varphi)$  组成,其中  $V = \{v_1, v_2, \dots, v_n\}$  是对象集合,它属于对象类型集合  $A$  的某一个特定对象类型集合,  $E = \{e_1, e_2, \dots, e_n\}$  是对象之间的链接集合,属于关系类型集合  $R$  的某一个特定关系类型集合,当信息网络中的对象类型数  $|A|$  或者关系类型数  $|R|$  大于 1 时,称这个信息网络为异质信息网络。

定义 2 网络模式<sup>[7,24]</sup>:网络模式是定义在对象类型  $A$  上的有向图,它的边为  $R$  中的关系,记为  $T_G = (A, R)$ ,表示信息网络的元模式。

定义 3 元路径<sup>[24]</sup>:元路径  $P$  是网络模式  $T_G = (A, R)$  的图上形式为  $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_l} A_{l+1}$  的一条路径,该路径定义了类型  $A_1$  和  $A_{l+1}$  之间的复合关系  $R = R_1 \circ R_2 \circ \dots \circ R_l$ ,其中“ $\circ$ ”表示关系上的复合运算。以图 1 所示的 DBLP 网络为例,作者和会议可以通过元路径  $A \xrightarrow{\text{write}} P \xrightarrow{\text{publish}} C(APC)$  相连,作者和主题可以通过元路径  $A \xrightarrow{\text{write}} P \xrightarrow{\text{include}} T(APT)$  相连。其中作者和会议以及作者和主题均只通过一个中介  $P$

相连,称为单中介元路径。

定义 4 加权 PageRank<sup>[25]</sup>:根据信息网络中对象的连接结构及连接频次对每个对象的质量进行排名,进而利用链接和对象质量排名来衡量整个网络对象的重要性。其重要性的表示如下。

$$PR_i = \frac{1 - \alpha}{n} + \alpha \sum_{j \in N(i)} PR(j) * W_{(j,i)}^{\text{out}} \quad (1)$$

其中,  $n$  为总的对象数量,  $\alpha$  为标度常数。初始时,为所有的对象赋一个初始的 PageRank 值  $PR_i(0)$ ,表示该对象被连接的概率,且满足  $\sum_{i=1}^N PR_i(0) = 1$ ,然后将每个对象的 PR 值进行迭代分配给它所指向的对象直至所有 PR 值收敛,最后的 PR 值表示对象的重要性。

$W_{(j,i)}^{\text{out}}$  是  $j$  到  $i$  的链接权重比,为  $W_{(j,i)}^{\text{out}} = \frac{w(j,i)}{\sum_k w(j,k)}$ ,

$w(j,i)$  为  $j$  到  $i$  的链接权重,  $\sum_k w(j,k)$  是  $j$  所有出度链接权重之和。

### 2.2 问题定义

在异质信息网络  $G = (V, E, \tau, \varphi)$  中,影响力是在不同实体之间的交互关系中产生的,实体类型和交互关系复杂多样,但是每一个实体的影响力都能通过直接相连或间接相连去影响其他实体,从而都能通过直接影响和间接影响获得目标实体的综合影响力。如何在结构复杂的异质信息网络中找到一组节点作为种子集合  $S^*$ ,使得该种子集合的影响力扩展度  $\sigma(S_0)$  在给定的扩散模型下达到最大值,即被影响的节点数量达到最多是该文的研究目标,表示如下:

$$S^* = \arg \max_{S_0} \sigma(S_0) \quad (2)$$

## 3 CWPR 算法

该文提出了一种基于加权 PageRank 的影响力最大化算法(CWPR),用于解决异质信息网络中的影响力最大化问题。该算法包含两个步骤。第一,首先将原始异质信息网络分解成若干个只含一种连接类型的网络,并根据节点间的连接关系分配对应的边权重。第二,利用加权 PageRank 来衡量节点的直接和间接影响力,最终融合所有影响力得到节点的最终影响力,并筛选出影响力最大的前  $k$  个节点。

### 3.1 边权重的分配

由于异质信息网络结构复杂,每个节点都能通过不同的元路径与其他类型节点相连得到不同类型的连接边并产生不同的影响关系,其中每条连接边的权重都不尽相同,这与连接边的两个节点间的交互程度密切相关,若两节点间交互次数过多,则对应的边权重也大。为了能减少不同类型边的差异性,简化整个复杂的异质信息网络,同时保留网络的异质性和不同类型



节点之间的影响关系,并为影响关系分配相应的权重,该文将包含多种连接关系的异质信息网络分解成若干个只含一种连接类型的网络。如图 1 中的 DBLP 网络,APTC 四种类型节点,直接相连关系  $A-P/P-A$ ,

$P-T/T-P$ ,  $P-C/C-P$ , 故可以分解成只含有  $AP$ ,  $PT$ ,  $PC$  类型的三个异质信息网络,使每条边的权重分别为对应的连接次数,如图 2 所示。

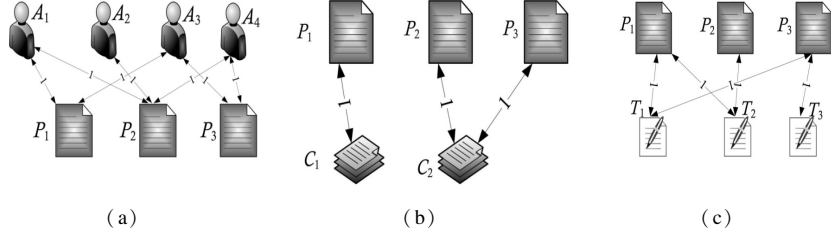


图 2 分解后的网络

一个节点的影响力除了与它直接相连的邻居有关,还与它邻居的邻居有关,因此可以基于单中介元路径获得节点的间接相连关系,若一个节点到达另一个间接相连的节点的路径数越多,则说明它们之间的联系越紧密,因此根据路径数为节点的间接相连关系分配对应的权重。对图 1 中的 DBLP 网络而言,  $A$  能通

过  $P$  与  $A$ 、 $C$ 、 $T$  间接相连,则对应的间接相连关系有  $A-A$ ,  $A-C/C-A$ ,  $A-T/T-A$ , 将图 1 中的间接相连关系提取出来,例如  $A_1$  和  $C_2$  只能通过元路径  $A_1P_2C_2$  相连,路径数为 1,而  $A_4$  和  $C_2$  可以通过元路径  $A_4P_2C_2$ ,  $A_4P_3C_2$  相连,路径数为 2,则  $A_1$  和  $C_2$  的边权重为 1,  $A_4$  和  $C_2$  的边权重为 2,如图 3 所示。

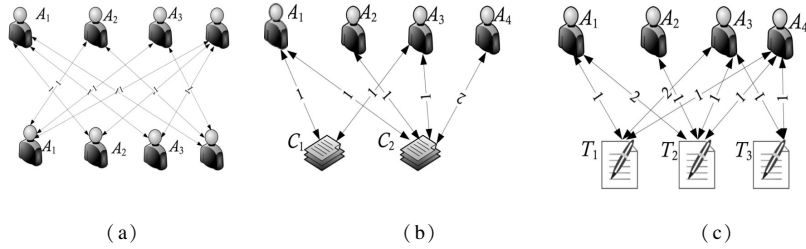


图 3 基于元路径的间接相连关系

### 3.2 影响力度量

加权 PageRank 延续了 PageRank 的优点,能够通过节点间的连接数量和质量来综合描述节点的重要性,同时又根据节点之间的交互程度分配对应的权重。该文利用加权 PageRank 来度量节点对不同类型节点的影响力,其综合影响力主要由直接影响力和间接影响力组成。

#### 3.2.1 直接影响力

在异质信息网络中,不同类型的实体  $u, v$  直接相连,  $u$  和  $v$  两者之间对彼此都有一定的影响,用  $DI_u^v$  表示  $u$  对  $v$  的直接影响。

为给定的若干个直接相连两种不同类型节点,只包含一种类型边的异质信息网络  $G$  构建一个加权有向图,  $i, j$  是两种不同类型  $A, B$  的节点,若  $j$  指向  $i$ , 则  $j$  到  $i$  有边,边的权重等于  $j$  到  $i$  边的个数  $k$ , 即  $w_{j,i} = k$ , 否则  $w_{j,i} = 0$ , 使用加权 PageRank 计算得到  $j$  对  $i$  贡献的 PR 值,即  $i$  对  $j$  的重要性为:

$$PR_{i,j}^B = \frac{1-\alpha}{n} + \alpha PR(j) * W_{(j,i)}^{\text{out}} \quad (3)$$

其中,  $PR(j)$  为  $j$  节点的初始值,在异质信息网络  $G$  中,每个节点的初始值都为  $\frac{1}{n}$ ,  $W_{(j,i)}^{\text{out}}$  是  $j$  到  $i$  的链接

权重比,为  $W_{(j,i)}^{\text{out}} = \frac{w(j,i)}{\sum_k w(j,k)}$ ,  $w(j,i)$  为  $j$  到  $i$  的链接

权重,  $\sum_k w(j,k)$  是  $j$  的所有出度链接权重之和。那么  $i$  对所有与它直接相连  $B$  类型的节点的重要性总和为:

$$PR_i^B = \frac{1-\alpha}{n} + \alpha \sum_{j \in N(i)} PR(j) * W_{(j,i)}^{\text{out}} \quad (4)$$

其中,  $N(i)$  为与  $i$  直接相连的不同类型节点的集合,故得  $i$  对所有与它相连的  $B$  类型节点的直接影响力为:

$$DI_i^B = PR_i^B \quad (5)$$

#### 3.2.2 间接影响力

在异质信息网络中,不同类型的实体  $u, w$  通过元路径  $uvw$  间接相连,  $u$  和  $w$  两者之间对彼此都有一定的影响,用  $\Pi_u^w$  表示  $u$  对  $w$  的间接影响。

若某两种不同类型的节点是间接相连的,如图 4 所示,  $i, j, t$  是三种不同类型  $A$ 、 $B$ 、 $C$  节点,其中  $i_1$  和  $t_1$  是间接相连,  $i_1$  对  $t_1$  的影响力是通过  $j$  间接传播的,使用加权 PageRank 计算出在只含  $BC$  类型的异质信息网络中  $t_1$  的重要性值  $PR_{t_1}^B$ 。由加权 PageRank 算法的定义知,  $t_1$  的重要性与指向它的节点数量和质量有关,

那么基于元路径  $ABC$  可以间接得到  $t_1$  对  $i_1$  的贡献值, 即  $i_1$  对  $t_1$  的间接影响力为:

$$\Pi_{i_1, t_1} = P_{i_1, t_1} \text{PR}_{t_1} \quad (6)$$

其中,  $P_{i_1, t_1}$  为  $i_1$  到达  $t_1$  的概率,  $P_{i_1, t_1} = \frac{\text{Path}(i_1, t_1)}{\text{inPath}(t_1)}$ ,  $\text{Path}(i_1, t_1)$  为  $i_1$  经过某个类型的中介到达  $t_1$  的路径数,  $\text{inPath}(t_1)$  为所有 A 类型的节点经过中介到达  $t_1$  路径数总和。图 4 中的  $\text{Path}(i_1, t_1) = 2$ ,  $\text{inPath}(t_1) = 5$ , 故  $P_{i_1, t_1} = \frac{2}{5}$ 。

那么  $i_1$  对所有与它相连的类型 C 的节点的间接影响力为:

$$\Pi_{i_1}^C = \sum_{t \in \text{ii}(i)} \Pi_{i_1, t} \quad (7)$$

其中,  $\text{ii}(i)$  为与  $i_1$  间接相连的节点类型的集合。

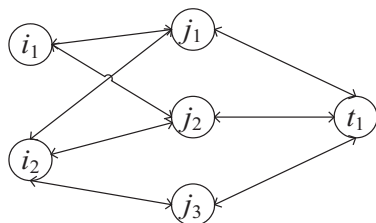


图 4 间接相连关系图

### 3.2.3 综合影响力

异质信息网络中, 节点类型丰富, 它们之间的影响力是通过直接或间接关系相连去传播的。该文将融合节点的直接影响力和间接影响力, 作为节点的最终综合影响力  $I_i$ ,  $I_i$  的表示如下:

$$I_i = \sum_{T \in \text{di}(i)} \lambda_T \text{DI}_i^T + \sum_{T \in \text{ii}(i)} \eta_T \Pi_i^T \quad (8)$$

其中,  $\text{di}(i)$  为与  $i$  直接相连的节点类型的集合,  $\lambda_T$  为  $i$  与  $T$  类型节点相连时直接影响力的一个可变参数, 取值  $(0, 1)$ ,  $\text{ii}(i)$  为与  $i$  间接相连的节点类型的集合,  $\eta_T$  为  $i$  与  $T$  类型节点相连时间接影响力的一个可变参数, 取值  $(0, 1)$ , 有  $\sum_{T \in \text{di}(i)} \lambda_T + \sum_{T \in \text{ii}(i)} \eta_T = 1$ 。

### 3.2.4 筛选种子节点

通过融合节点的直接影响力和间接影响力得到节点的最终影响力, 因此可以对所有节点的最终影响力进行排序。为了避免种子节点影响力重合, 该文采用边际增益策略筛选种子节点。先选择一个影响力最大的节点作为种子节点, 然后去除其余节点与其影响重叠的部分, 再选择剩余节点中影响力最大作为种子节点, 不断重复此过程, 直到筛选出给定数量的节点作为种子集为止。

CWPR 是基于加权 PageRank 迭代计算获得节点的影响力, 因此使用邻接矩阵对节点间的关系进行表示存储, 需要的空间复杂度为  $O(n^2)$ , 其中  $n$  为节点数, 而 PR 值的计算是一个迭代过程, 向量与矩阵相乘

所需要的时间复杂度为  $O(n^2)$ , 经过若干次迭代达到收敛所需的时间复杂度为  $O(cn^2)$ , 其中  $c$  为迭代次数。

## 4 实验评估

### 4.1 实验准备

数据集: 使用了两个最常见的文献网络数据集 DBLP 和 Yelp, 数据集的情况分别如表 1 和表 2 所示。

表 1 DBLP 数据集详细信息

对象	数量	关系	数量
Author( A )	14 475	A - P	41 794
Paper( P )	14 376	-	
Conference( C )	20	P - C	14 376
Type( T )	8 920	P - T	114 624

表 2 Yelp 数据集详细信息

对象	数量	关系	数量
User( U )	5 019	U - B	62 146
Business( B )	10 283	-	
City( C )	47	B - C	10 283
Category( Cat )	496	B - Cat	29 307

对比算法: 为了验证 CWPR 方法的有效性, 该文将与已有的同质信息网络影响力度量方法 Degree (DC)、PageRank (PR) 以及异质的方法 APR 和 CWPR 的变种方法 CWPR-II 进行实验对比。由于目前对于网络中关键节点的度量方法大多都是针对同质信息网络的, 而本实验是利用异质信息网络的关键节点进行影响力最大化研究, 为了进行对比实验, 故将常用的同质信息网络关键节点的度量方法直接运用于异质信息网络, 在使用这些方法时, 忽略不同类型节点之间关系的差异, 根据度量方法计算得到每个节点对应的度量值。在选取种子集时, 由于不同类型节点在信息扩散中扮演的角色不同, 为了减少实验的差异性, 种子集类型固定, 本实验以人作为目标类型, 选取度量值最大的目标类型节点作为种子集。对比算法描述如下。

Degree centrality (DC): 一个节点  $v$  与它直接相连的邻居节点的个数, 称为度, 一个节点度越大, 就意味着这个节点越重要。

PageRank (PR): 网页重要性度量方法, 如果一个网页被很多网页链接, 或者被知名度很高的网页链接, 则这个网页的重要性就越大, 也可以用于社交网络节点分析。

APR: 一种在异质的文献网络中的节点重要性度量方法, 利用 PageRank 度量的异质信息网络中作者和文章两种类型节点之间的影响力, 对于 DBLP、Yelp 则分别考虑了作者和文章, 用户和商业之间的影响力。

CWPR-II: 该文提出的 CWPR 的变体, 在异质网络中只考虑人与人之间的影响力。

CWPR: 该文提出的异质信息网络影响力的度量算法, 基于异质信息的连接结构, 考虑了不同类型节点之间的影响力。

扩散模型: 采用线性阈值模型 LT 作为传播模型, 将每一节点的入度边的度数归一化, 作为每个节点被自己入邻居节点激活的概率, 使它们和为 1, 每个非激活节点都有一个  $[0, 1]$  的激活阈值, 当非激活节点的已激活邻居节点对其影响总和超过该阈值, 则此节点被激活。该文的扩散指标分别为在  $k$  个有影响力的作者和用户作为种子集时被影响的作者和用户的个数, 影响的人越多说明实验效果越好。为了减小实验的偶然性, 进行了 10 000 次蒙特卡洛仿真来估计影响扩散结果

## 4.2 实验结果

### 4.2.1 算法参数的影响

在异质信息网络中, 包含了多种类型的节点, 每种类型节点在信息扩散中扮演的角色也是不尽相同, 故在 CWPR 算法中假设  $\sum_{T \in \text{di}(i)} \lambda_T + \sum_{T \in \text{ui}(i)} \eta_T = 1$ , 其中  $\lambda_T, \eta_T$  的取值决定了对不同类型节点影响力所占比重。对于数据集 DBLP,  $\lambda_{AP} + \eta_{AA} + \eta_{AC} + \eta_{AT} = 1$ , 数据集 Yelp 则是  $\lambda_{UB} + \eta_{UU} + \eta_{UC} + \eta_{UCat} = 1$ 。本节将设置多组不同的权重, 并通过选出  $k = 50$  个种子得到的影响范围大小进行实验对比, 从而得到一组合理的权重。实验结果如图 5 所示。

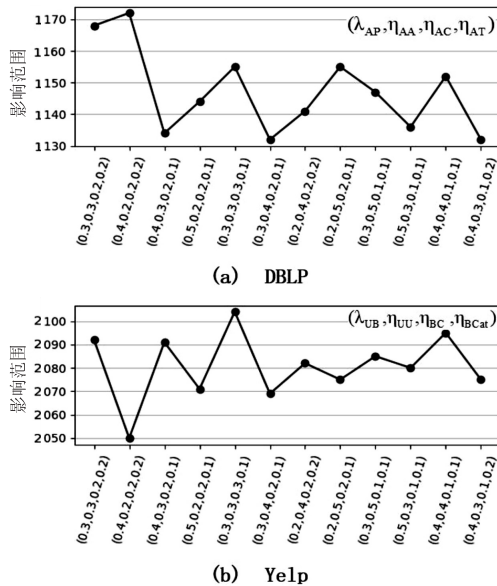


图5 算法参数的影响

对以上实验结果分析可知, 在数据集 DBLP 中, 当  $\lambda_{AP} = 0.4, \eta_{AA} = 0.2, \eta_{AC} = 0.2, \eta_{AT} = 0.2$  时, 影响范围的值达到最大, 这表明在信息扩散过程中作者对论文的影响力是作者的综合影响力的重要组成部分。而对

于数据集 Yelp, 当  $\lambda_{UB} = 0.3, \eta_{UU} = 0.3, \eta_{BC} = 0.3, \eta_{BCat} = 0.1$  时, 影响范围的值达到最大, 此时在信息扩散过程中, 用户和领域的之间的影响力在用户的综合影响力所占的比重最小。用户和用户、商业、城市之间的影响力则是用户综合影响力的重要组成部分。

### 4.2.2 边权重参数的影响

在异质信息网络中, 包含了多种类型的边, 每种类型的边在信息扩散中同不同类型的节点一样也是扮演着不同的角色。同不同类型节点一样, 该文也假设异质信息网络中不同类型边的权重等于 1, 则数据集 DBLP 中有  $W_{AP} + W_{PC} + W_{PT} = 1$ , 数据集 Yelp 中有  $W_{UB} + W_{BC} + W_{BCat} = 1$ 。通过设置多种不同的权重并选出  $k = 50$  个种子得到的影响范围大小进行结果对比, 从而获得一组合理的边权值。实验结果如图 6 所示。

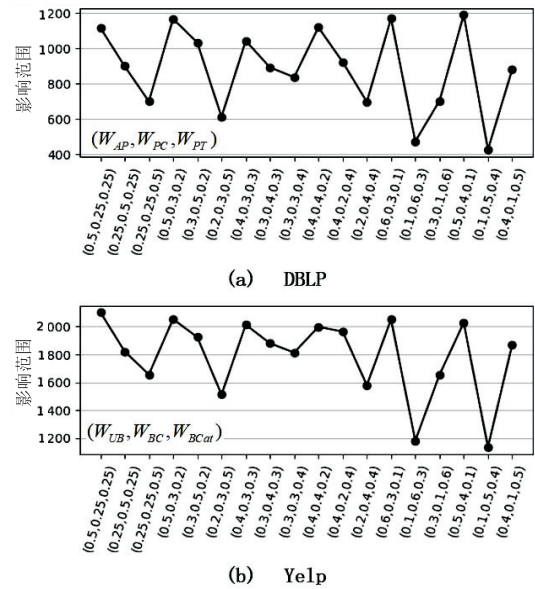


图6 边权重的影响

由实验结果可知, 在数据集 DBLP 中, 当  $W_{AP} = 0.5, W_{PC} = 0.4, W_{PT} = 0.1$  时, 影响范围的值达到最大, 此时作者与论文之间的边权值是三个中间最大的, 这说明在信息扩散过程中作者与论文之间的关系起着重要作用, 同时发现对于每一组权重, 若是论文与主题之间的边权重是三者中最大的一个, 则影响范围的值将会下降, 则可以认为在信息扩散中, 论文与主题之间的关系影响作用较小。在数据集 Yelp 中, 当  $W_{UB} = 0.5, W_{BC} = 0.25, W_{BCat} = 0.25$ , 影响范围的值达到最大, 此时用户和商业之间的关系在信息扩散过程中起着重要的作用。通过对这两个数据集的边权重分析发现, 均是人和与人直接相连的类型节点的边权重在所有边权重所占的比重是最大的, 这也表明了直接的影响会比间接影响更有力。

通过对算法参数和边权重设置不同值, 分别选取了各自最好的结果, 作为该算法有效性验证的参数。



### 4.2.3 有效性验证

对于数据集 DBLP 和 Yelp, 本实验的种子集的类型分别设为作者和用户, 由于本实验基于不同元路径考虑了不同类型的节点直接的影响, 在数据集 DBLP 中, 对不同类型的边权重设为  $W_{AP} = 0.5$ ,  $W_{PC} = 0.4$ ,  $W_{PT} = 0.1$ 。在数据集 Yelp 中, 设各类型的边权重为  $W_{UB} = 0.5$ ,  $W_{BC} = 0.25$ ,  $W_{BCat} = 0.25$ , 实验对比方法中的同质方法 DC 和 PageRank 不区分边的类型, 权重都为 0.5。实验效果如图 7 所示。

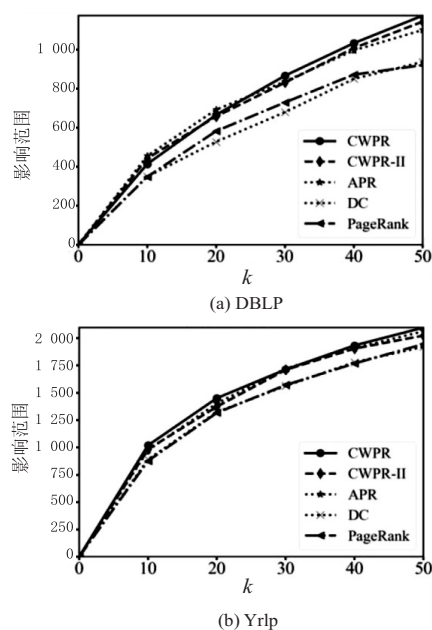


图 7 影响范围

由这些实验对比结果可知, 保留各种类型节点信息的三种异质方法要明显优于其他两种同质方法, 在 DBLP 中该文所提出的 CWPR 方法明显优于其他两种异质方法 CWPR-II、APR, 而在 Yelp 中 CWPR 也同样优于其他两种异质方法, 但是差距并不如 DBLP 明显。该文给出的三种异质方法都区分了不同类型的边的权重, 但 CWPR 考虑了不同类型节点之间的影响, 而 APR、CWPR-II 只考虑了部分的类型节点的影响。通过以上实验结果可以表明, 在异质信息网络中, 保留节点与其他类型节点之间的语义信息比只保留部分信息能更全面地评价节点的特征, 得到更好的实验效果, 从而可以借助这种方法得到最有影响力的节点。

## 5 结束语

该文提出了一种基于加权 PageRank 的异质信息网络影响力最大化算法 CWPR, 该算法将包含多种类型节点的异质信息网络分解成若干个只含一种连接类型的网络, 然后通过节点之间的连接方式考虑了所有不同类型节点之间的影响关系, 去获得影响力最大的节点作为信息扩散的种子节点, 从而实现异质信息网

络影响力的最大化。通过在两个真实数据集的实验结果表明, 在异质信息网络中, 保留节点与其他节点之间的信息越多, 筛选出的种子节点得到的影响效果越好。但是该算法的不足在于对异质信息网络中不同类型的边权重的设置是基于先验知识设定的, 在未来的研究中, 可以通过机器学习去自主获得不同类型的边权重, 使得边权重结果更加真实可靠。

### 参考文献:

- [1] 陈卫. 社交网络影响力传播研究[J]. 大数据, 2015, 1(3): 82-98.
- [2] BRACH P, EPASTO A, PANCONESI A, et al. Spreading rumours without the network[C]//Proceedings of the second ACM conference on online social networks. Dublin, Ireland: Association for Computing Machinery, 2014: 107-118.
- [3] 陈晋音, 张敦杰, 林翔, 等. 基于影响力最大化策略的抑制虚假消息传播的方法[J]. 计算机科学, 2020, 47(S1): 17-23.
- [4] 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.
- [5] BRIN S, PAGE L. The anatomy of a large-scale hypertextual web search engine[J]. Computer Networks ISDN Systems, 1998, 30(1-7): 107-117.
- [6] SOHEILA M, SAMA B, MOSTAFA S, et al. Information spread and topic diffusion in heterogeneous information networks[J]. Scientific Reports, 2018, 8(1): 9549.
- [7] SUN Y, YU Y, HAN J. Ranking-based clustering of heterogeneous information networks with star network schema[C]//Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. Paris, France: Association for Computing Machinery, 2009: 797-806.
- [8] 王锐, 张志强, 石川. 异质信息网络分析及其语义探索[J]. 电信科学, 2015, 31(7): 43-51.
- [9] ZHAO F, ZHANG Y, LU J, et al. Measuring academic influence using heterogeneous author-citation networks[J]. Scientometrics, 2019, 118(3): 1119-1140.
- [10] KEMPE D, KLEINBERG J, TARDOS É. Maximizing the spread of influence through a social network[C]//Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. Washington, D. C: Association for Computing Machinery, 2003: 137-146.
- [11] LESKOVEC J, KRAUSE A, GUESTRIN C, et al. Cost-effective outbreak detection in networks[C]//Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining. California, USA: Association for Computing Machinery, 2007: 420-429.
- [12] GOYAL A, WEI L, LAKSHMANAN L. CELF++: optimizing the greedy algorithm for influence maximization in social networks[C]//Proceedings of the 20th international

- conference on world wide web. Hyderabad, India; Association for Computing Machinery, 2011; 47–48.
- [13] CHEN W, WANG Y, YANG S. Efficient influence maximization in social networks[C]//Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. Paris, France; Association for Computing Machinery, 2009; 199–208.
- [14] WANG C, CHEN W, WANG Y. Scalable influence maximization for independent cascade model in large-scale social networks[J]. Data Mining and Knowledge Discovery, 2012, 25(3): 545–576.
- [15] CHEN W, YUAN Y, ZHANG L. Scalable influence maximization in social networks under the linear threshold model [C]//IEEE international conference on data mining. Sydney, Australia; IEEE, 2010; 88–97.
- [16] 周明洋, 吴向阳, 曹 扬, 等. 基于群体影响力的网络传播关键节点选择策略[J]. 中国科学: 信息科学, 2019, 49(10): 1333–1342.
- [17] 曹玖新, 闵绘宇, 王浩然, 等. 竞争环境中基于主题偏好的利己信息影响力最大化算法[J]. 计算机学报, 2019, 42(7): 1495–1510.
- [18] 杨书新, 梁 文, 朱凯丽. 基于三级邻居的复杂网络节点影响力度量方法[J]. 电子与信息学报, 2020, 42(5): 1140–1148.
- [19] ORIEDI D, RUNZ C D, GUESSOUM Z, et al. Influence maximization through user interaction modeling [C]//Proceedings of the 35th annual ACM symposium on applied computing. Brno, Czech Republic; Association for Computing Machinery, 2020; 1888–1890.
- [20] DENG X, LONG F, LI B, et al. An influence model based on heterogeneous online social network for influence maximization[J]. IEEE Transactions on Network Science Engineering, 2020, 7(2): 737–749.
- [21] YANG Y, ZHOU L, JIN Z, et al. Meta path-based information entropy for modeling social influence in heterogeneous information networks [C]//2019 20th IEEE international conference on mobile data management. Hong Kong, China; IEEE, 2019; 557–562.
- [22] MOLAEI S, FARAHBAKHS R, SALEHI M, et al. Identifying influential nodes in heterogeneous networks[J]. Expert Systems with Applications, 2020, 160(12): 113580. 1–113580. 2.
- [23] KEIKHA M M, RAHGOZAR M, ASADPOUR M, et al. Influence maximization across heterogeneous interconnected networks based on deep learning [J]. Expert Systems with Applications, 2020, 140(5): 112905. 1–112905. 11.
- [24] SUN Y, HAN J, YAN X, et al. PathSim: meta path-based top-k similarity search in heterogeneous information networks [J]. Proceedings of the VLDB Endowment, 2011, 4(11): 992–1003.
- [25] BOLLEN J, DE SOMPEL H, SMITH J, et al. Toward alternative metrics of journal impact: a comparison of download and citation data [J]. Information Processing Management, 2005, 41(6): 1419–1440.
- +++++
- (上接第45页)
- al. Text classification algorithms; a survey [J]. Information, 2019, 10(4): 150.
- [7] MENARD S. Logistic regression [J]. American Statistician, 2004, 58(4): 1–12.
- [8] RISH I. An empirical study of the naive Bayes classifier [C]//Proceedings of the 2001 IJCAI workshop on empirical methods in artificial intelligence. [s. l.]: [s. n.], 2001; 41–46.
- [9] 宋胜利, 鲍 亮, 陈 平. 多层文本分类性能评价方法[J]. 系统工程与电子技术, 2010, 32(5): 1088–1093.
- [10] RAMOS J. Using TF-IDF to determine word relevance in document queries [C]//Proceedings of the first instructional conference on machine learning. Washington, DC, USA; AAAI, 2003; 29–48.
- [11] HAVRLANT L, KREINOVICH V. A simple probabilistic explanation of term frequency-inverse document frequency (TF-IDF) heuristic (and variations motivated by this explanation) [J]. International Journal of General Systems, 2017, 46(1): 27–36.
- [12] QAISER S, ALI R. Text mining: use of TF-IDF to examine the relevance of words to documents [J]. International Journal of Computer Applications, 2018, 181(1): 25–29.
- [13] LI H, TECHNOLOGIES N A L H. Deep learning for natural language processing: advantages and challenges [J]. National Science Review, 2018, 5(1): 24–26.
- [14] CHEN P H, ZAFAR H, GALPERIN-AIZENBERG M, et al. Integrating natural language processing and machine learning algorithms to categorize oncologic response in radiology reports [J]. Journal of Digital Imaging, 2018, 31: 178–184.
- [15] SOUCY P, MINEAU G W. A simple KNN algorithm for text categorization [C]//Proceedings of the 2001 IEEE international conference on data mining. San Jose, CA, USA; IEEE, 2001; 647–648.