

基于多标签传播的社交网络用户影响力评估

许 为 林柏钢 林思娟 杨 旻

(福州大学数学与计算机科学学院 福州 350108)

(网络系统信息安全福建省高校重点实验室 福州 350108)

摘 要 为了识别出社交网络中的关键人物,需要对用户影响力进行评估。由于影响力是借助信息在网络中的扩散而逐步形成的,因此需首先对影响力传播过程进行建模;然后以该模型为基础,用标签表示影响力的所有者,以隶属度表示用户被影响的程度,利用多标签传播来模拟影响力传播的过程,实现了一种新的用户影响力评估算法 MLPIA (Multi-label Propagation User Influence Assessment Algorithm);最后,在真实数据集上测验排名靠前的用户的影响力覆盖范围和紧密中心性,结果证明了该算法的合理性和有效性。

关键词 用户影响力,多标签传播,影响力传播,社交网络,效果评判

中图法分类号 TP309 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.10.025

Assessment of User Influence in Social Networks Based on Multi-label Propagation

XU Wei LIN Bo-gang LIN Si-juan YANG Yang

(College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China)

(Key Lab of Information Security of Network System in Fujian Province, Fuzhou 350108, China)

Abstract In order to identify the key figures in social networks, we need to assess the user influence. Influence is gradually formed on the base of the information propagation. This paper built a model to describe the process of influence propagation. Then, based on the model, we used label to indicate the original owner of influence and used degree of membership to indicate the effect degree of which users are affected. We utilized multi-label propagation to simulate the process of users' influence propagation to achieve a new user influence assessment algorithm MLPIA (Multi-label Propagation User Influence Assessment Algorithm). Lastly, the coverage area and closeness centrality of top users' influence were tested on real data set. The results demonstrate the rationality and validity of the proposed algorithm.

Keywords User influence, Multi-label propagation, Influence propagation, Social networks, Effect judgment

随着互联网技术的蓬勃发展,新型社交媒体形成了强劲的发展势态。这类社交媒体潜移默化地影响着人们的生活方式,已逐渐成为人们日常社交的重要组成部分。社交媒体平台上庞大的用户群体之间相互作用建立起社会关系,连接形成了社交网络(Social Network)^[1]。网络中的用户可以将信息通过推送、转发等方式影响其他成员。影响力大的用户往往是信息传播中的关键节点,其言论会引起广泛的追随和效仿,甚至发挥着引导社会舆论的作用。由于新型社交媒体的高度开放性,使得用户获得了更加自由的话语权利,网络舆论失控现象频频发生。通过用户影响力评估,这在挖掘出社交网络中意见领袖在舆论引导和降低事件危害等方面发挥着不可替代的积极作用,对构建和谐网络具有重要意义。

现有的社交网络用户影响力评估主要采用以下两种方法。一是通过节点中心性指标^[2],如节点度中心性^[3]、K-壳^[4]等,从网络结构的角度分析节点在网络中的中心地位,但是这类方法没有考虑信息传播,并且缺乏与用户属性等因素

的融合。另一种主流的方法是借鉴网页排名中的经典算法PageRank^[5]的原理,如在PageRank的基础上考虑主题相似度的TwitterRank算法^[6]、综合考虑用户领袖特征的Weibo-Rank算法^[7]和OpinionLeaderRank算法^[8]等。该类算法的核心思想是将用户影响力看作在信息流动网络中随机游走到该节点的概率,越容易被访问到的节点的影响力越大。

本文将对基于PageRank思想以及点度中心性的评估方法存在的问题进行分析,针对它们的不足,考虑节点之间的信息传递,借鉴标签传播的思想设计并实现一种新的影响力评估算法。

1 预备知识

1.1 好友关系网络

在社交平台中,两个用户互相关注时被认定为具有好友关系。定义一个无向图 $G_f(V_f, E_f)$ 来抽象表示好友关系网络,当两个用户互相关注时,用一条边来连接,如图1所示。

到稿日期:2015-10-24 返修日期:2015-12-13 本文受国家自然科学基金(61402112),福建省安全课题(828398)资助。

许 为(1990—),女,硕士生,主要研究方向为信息安全、社交网络分析,E-mail:happyxivi@163.com;林柏钢(1953—),男,教授,博士生导师,主要研究方向为网络与信息安全、编码与密码;林思娟(1990—),女,硕士生,主要研究方向为信息安全、社交网络分析;杨 旻(1984—),女,博士,讲师,主要研究方向为密码学与信息安全。

其中 V_f 表示用户节点集合, E_f 表示节点间边的集合。由于社交网络具有高聚集系数,即朋友的朋友很可能也是朋友,关系越为紧密的用户之间越可能保持相对一致的思想观念;而没有直接相连的用户也可能间接通过自己信任的好友接收到其他用户发表的观点。因此在研究用户影响力的评估中,有必要构建用户的好友关系网络图。

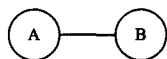


图1 好友关系网络示例

1.2 信息转发网络

社交平台的转发机制为信息的迅速流动创造了条件。当用户A发布了一条信息之后,受到这条信息感染的粉丝B就会对其进行转发,若B的粉丝C在接收到此信息之后继续转发,那么C就对B和A同时进行了一次转发,当上述过程不断被循环以后,A用户的影响力将以该条信息为载体,沿着信息转发网络以一种裂变^[9]的传播方式扩散到网络中去。图2中,以用户A,B为例,用一条有向带权边从B指向A,表示用户B对A的信息进行了转发,边的权重表示转发的频繁程度。由此给出信息转发网络的抽象表示 $G_r(V_r, E_r)$, 其中 V_r 表示用户节点集合, E_r 表示节点间有向边的集合。

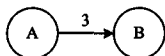


图2 转发关系网络示例

2 影响力传播过程建模

2.1 问题分析

传统的影响力分析方法通常借鉴 PageRank 的思想^[5]或是网络节点中心性的思想对信息转发网络中的节点重要性进行分析,用重要性来衡量用户在网络中的影响力。PageRank 的提出最初是应用于网页等级排名,它模拟了上网者随机点击链接访问网站的过程,若网页B存在一个指向网页A的超链接,则访问B的上网者会以一定的概率继续跳转到A,从而B网页会以自身的重要性对A网页进行打分,若分别以 $PR(B)$ 和 d_{out}^B 表示网页B的重要性和出链数,那么B网页对A赋予的重要性分值就为 $PR(B)/d_{out}^B$ 。若将信息转发网络中的转发和被转发关系比作Web页面的链出和链入,那么就可以应用PageRank算法对社交网络的用户重要性进行评估。同理可得,一个用户的重要性由所有对其信息进行了转发的用户的重要性来决定。然而这种方法忽略了用户转发行为和上网者随机点击过程的区别。以图3中的网络为例,虽然B的入度高于A,但是根据PageRank算法,A却与B具有同等的重要性,这是因为节点B的出度全指向A。若将图3看作网页之间的相互链接,可以用上网者随机点击过程对这一结果进行解释:即使网页D,E,F,C没有链向网页A,但是通过网页D,E,F,C到达网页B的上网者会以很大的概率继续浏览A网页,因此网页A也具有较高的重要性。然而若将图3看作是信息转发网络,那么其表现的涵义就与上网者的随机浏览行为不同。从图中可知,用户B可以直接影响到D,E,F,C 4个用户,而用户A仅直接影响到用户B,虽然A可以通过B间接影响到D,E,F,C,但是从上述4人未转发A的信息可知,他们对B的转发仅来源于B的原创信息,说明他们对B发表的信息比对A发表的更感兴趣。所以在这个

网络中,用户B比A的影响范围更大,应该具有更高的影响力。而PageRank算法中,即使用户A没有很高的入度,只要指向A的用户B具有很高的权重,就能获得较高的重要性,本文认为这种基于PageRank的评估得到的重要性更多地反映出用户的权威度,在社交网络中对用户影响力的评估也应考虑影响力的覆盖范围,仅以权威度作为评估影响力的标准有失公允。

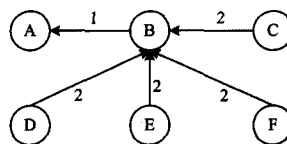


图3 PageRank分析示例

从上述分析可知,在用户影响力评估中,节点的入度也起着重要的作用。在用户影响力评估中也常以节点中心性分析中的入度中心性来衡量节点的影响力的大小。然而入度仅能体现出信息转发网络中用户影响力对相邻节点的覆盖度,单纯地用入度中心性来评估用户影响力无法体现出信息传播。以图4为例,用户A与B的入度都为2,但是用户A的信息也会经由B间接影响到E和F,所以A应该比B具有更大的影响力。

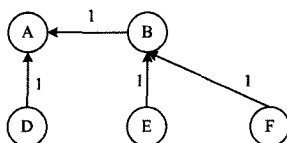


图4 入度中心性分析示例

2.2 建模

针对上述两种传统的影响力评估方法存在的问题,本文从影响力覆盖范围的角度来考虑用户影响力的评估。因此,需要通过模拟影响力传播的过程,使影响力能够更多地流动到可达用户上。首先需要对影响力传播的过程进行建模。图5描述了用户不断将自身携带的影响力向外传递的过程。用户以自身发布的信息为介质,在网络中注入影响力;接着以转发网络为传播路径,影响潜在用户,潜在用户在多种因素的作用下一定程度上接受影响,成为影响力的携带者,接着继续循环上述过程。网络中任意用户都可以将自身携带的影响力传递给其在信息转发网络中的可达用户。这是因为信息转发网络中边的权重可以反映出源节点对目的节点的认可程度。所以即使用户A没有直接转发B的信息,但是由于A到B的可达路径上的每一条边的源节点都一定程度上认可目的节点的思想观念,那么A也很有可能在潜移默化中受到B的影响。

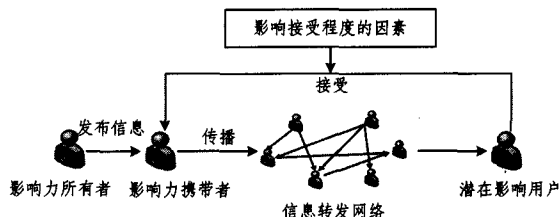


图5 影响力传播过程

在上述影响力传播过程中,用情感强度和衰减系数两个概念来度量影响潜在用户所能接受影响程度的因素,其中情

感强度表示用户对某个影响力拥有者的情感倾向,代表主观因素;而衰减系数则从客观的角度刻画了随着路径长度的增长用户所携带影响力的衰减程度。若先排除主观因素的影响,设定影响力所有者在发布信息后,影响力沿着传播路径 x 每一时刻感染一位潜在用户,并用 $\kappa_{cur \rightarrow pre}$ 表示该潜在用户从路径中的上一携带者那获取影响的衰减系数,那么在 t 时刻影响力所有者在路径 x 上新产生的影响力可以表示如下:

$$I_x = I_x^{-1} * \kappa_{cur \rightarrow pre} \quad (1)$$

在 t 时刻影响力所有者若有 n 个可达用户,那么 t 时刻影响力所有者新产生的影响力总和为:

$$I^t = \sum_{x=1, \dots, n} I_x^t \quad (2)$$

由式(1)、式(2)可知,影响力所有者在刚刚发布信息后的短时间内,其影响力会得到大量传播,并且传播范围逐渐扩大。当传播范围达到一定程度之后,即使存在多个潜在用户,其传播效率也将进入下降状态,这是由于随着传播路径的加深,路径上的用户受到越来越多来自其他人的影响,对该影响力的携带能力逐渐被削弱,导致该影响力的传播受到抑制。

在影响力传播途中,用户对影响力的接受程度受多个因素的共同作用。本文认为,情感强度代表的主观因素包括影响力所有者的权威度、活跃度、信息质量以及潜在用户与其的紧密程度等。若影响力所有者具有越高的权威度、活跃度以及信息质量,则其对潜在用户形成的吸引就越大;潜在用户与影响力所有者的关系越紧密,就越有可能对其产生认可。在信息转发网络中节点不会对其出边邻接节点携带的影响力全盘接收,应结合节点的出入度对衰减系数进行量化。根据上述分析,本文以 Twitter 平台上的用户社交网络为例,做如下定义。

定义 1(权威度) 根据 2.1 节所述,在有向图中节点的 PageRank 值越高,级别越高。本文以这种级别的高低来衡量用户在信息转发网络中的权威度,权威度越高,发布的信息越容易得到他人的信赖。对用户的权威度定义如下:

$$UR(U_i) = (1-c) + c \sum_{j \in R(U_i)} \frac{UR(U_j) * W_{ji}}{d_{in}^i} \quad (3)$$

其中, $R(U_i)$ 表示对 i 的信息进行转发的用户集合, W_{ji} 为用户 j 对 i 的推文转发的频率。为了保证即使用户没有入度,也至少能有一个最小的非零值作为权威度,在式(3)中加入阻尼系数 c , 通常取值为 0.85。

定义 2(活跃度) 用户的活跃度取决于用户发布信息的频率,用户活跃度越高,其他用户看到他的信息的概率就越大,越容易受到影响。Twitter 网络中,用户发布的文本信息称为推文(tweet),本文将活跃度定义为一个时间窗口内用户发布推文的数量,记为:

$$Active(U_i) = n_{tw} \quad (4)$$

定义 3(信息质量) 优质的推文一般会引起大量的转发和收藏,因此可以用平均每条推文被转发和收藏的次数来判断这个用户发布信息的质量的好坏,用 $retweet_i$ 和 $favorite_i$ 分别表示用户 i 的推文在单位时间窗口内被转发的次数和被收藏的次数,那么信息质量计算公式如下:

$$Q(U_i) = \frac{retweet_i + favorite_i}{n_{tw}} \quad (5)$$

定义 4(吸引力) 用权威度、活跃度、信息质量来衡量用户的吸引力,如式(6)所示。在实际应用中,上述 3 种属性均需先进行归一化处理。其中, α, β, γ 分别表示对应属性值的

权重比例,可根据评估的侧重点对参数进行灵活设置。

$$Attract(U_i) = \alpha UR(U_i) + \beta Active(U_i) + \gamma Q(U_i) \quad (6)$$

定义 5(相似度) 相似度可以衡量用户在好友关系网中的紧密程度。对于直接相连的用户,他们的相似度与彼此之间的边权以及与共同好友的边权相关;对于不直接相连的用户,相似度为最短路径上所有节点对的相似度的乘积,路径越长,相似度越低。当存在多条最短路径时,取乘积较大者作为他们的相似度。因此,将相似度定义如下:

$$Sim(U_i, U_j) = \begin{cases} \frac{w_{ij} + 2 \sum_{k \in N(i) \cap N(j)} \min(w_{ik}, w_{jk})}{\sum_{k \in N(i)} w_{ik} + \sum_{k \in N(j)} w_{jk} - w_{ij}}, & U_i, U_j \text{ 相邻} \\ \max_{(U_n, U_m) \in path_{ij}} \prod sim(U_n, U_m), & 1 < |path_{ij}| \end{cases} \quad (7)$$

其中, $N(i)$ 和 $N(j)$ 分别表示用户 i 和用户 j 在好友关系网络中的邻居节点集, $path_{ij}$ 表示用户 i 与用户 j 之间的最短路径。

定义 6(衰减系数) 衰减系数 $\kappa_{i \rightarrow j}$ 用来度量潜在用户 i 接收用户 j 携带的影响力的程度,由于用户 i 在信息转发网络中会接收到所有其他转发过的用户(以下称为邻接用户或邻接节点)传递来的影响力,所以其中一个邻接用户 j 对 i 的影响与 w_{ij} 和 i 出度的比值有关;另外,由于 i 的信息同时也可能被其他用户转发,被转发的频率越高, i 自身携带的影响力就越大,就越不容易受到他人的影响。因此,根据上述分析给出如下衰减系数的计算公式:

$$\kappa_{i \rightarrow j} = \frac{w_{ij}}{d_{out}^i + d_{in}^i} \quad (8)$$

3 基于多标签传播的用户影响力评估

3.1 基本思想

由于用户的影响力会以信息为传播介质在网络中不断积累,因此在用户影响力评估中需要考虑信息传播的作用。受标签传播思想的启发,本文将把信息网络中抽象的影响力具体化为标签,影响力的大小具体化为标签隶属度,应用多标签传播算法,兼顾网络结构和用户属性特征模拟影响力在信息转发网络中的传递过程,使影响力能够更充分地扩散到网络中,最后将网络中影响力的平均值作为用户影响力的度量,如图 6 所示。

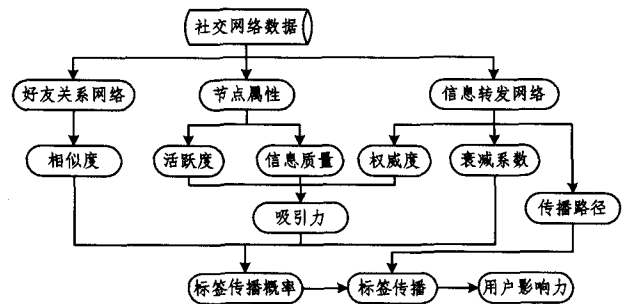


图 6 用户影响力评估解决框架

3.2 COPRA 算法

标签传播算法是一种复杂网络中的社区发现算法。Gregory 等人^[10]首次应用多标签传播提出了基于标签传播的重叠社区发现算法(Community Overlap Propagation Algorithm, COPRA),让具有相同标签的节点归属于同一社区。

由于 COPRA 算法允许一个节点携带多个标签,因此可以挖掘出重叠社区。

COPRA 算法的基本步骤如下。

(1)初始化节点标签。对于网络中的每一个节点 v , 对其赋予一个唯一的标签 k , 表示其从属的社区。初始时指定所有标签的隶属度 $P_v(k)=1$ 。

(2)迭代地进行标签传播。网络中的任意节点 v 在第 t 轮迭代中会更接受邻接节点的影响来更新自己的标签集, v 的某个标签 k 在 t 轮迭代中的隶属度计算公式如下:

$$P_v^t(k) = \frac{\sum_{i \in N(v)} P_i^{t-1}(k)}{|N(v)|} \quad (9)$$

算法设置参数 n 来控制每个节点可以携带标签的数量。保留隶属度从大到小的前 n 个标签后, 进一步对标签的隶属度进行标准化, 使其满足以下条件:

$$\sum_{k \in \text{label}(v)} P_v^t(k) = 1 \quad (10)$$

(3)上述标签传播过程满足终止条件后停止。

(4)根据最终节点的标签集合确定其从属的社区。

3.3 用户影响力评估算法

受 COPRA 算法的启发, 本文拟将原本表示社区的标签用来表示影响力的所有者, 用标签隶属度来表示用户被影响的程度, 模拟影响力的传播。然而 COPRA 算法并不完全适用于信息转发网络中的用户影响力传播。本文对 COPRA 算法进行修正以满足影响力评估的需要, 提出基于多标签传播的用户影响力评估算法 (Multi-label Propagation User Influence Assessment Algorithm, MLPIA)。

将 COPRA 算法应用于信息转发网络后, 由于节点的标签更新完全依赖于相邻节点的标签情况, 那么一轮迭代以后, 节点将不携带标注为自身的影响力标签。以图 7 为例, 经过一轮迭代以后, 由于节点 a 只有入度没有出度, 因此 a 无法携带任何影响力标签; 用户 b 虽然携带标签 a , 但是却丢失了自身的标签, 这会导致在之后的传播中 b 的影响力会无法延续而从网络中消失。为了保证在标签传播的迭代过程中节点的影响力能够持续沿着网络扩散, 节点每一轮更新标签时都要重新携带以自己命名的标签, 如图 8 所示, 经过修正的更新策略使得一轮迭代后 b 节点同时携带了 a, b 标签, 在接下来的迭代中, b 就可以顺利将 a 的影响力继续向 c 和 d 传播。

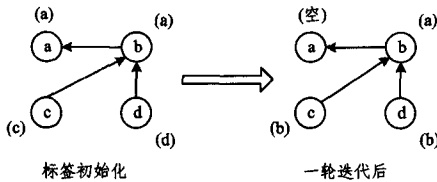


图 7 COPRA 算法标签更新示例

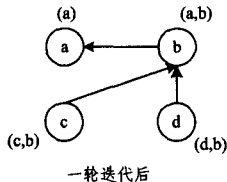


图 8 修正后的标签更新示例

定义 7(标签更新规则) 第 t 轮中, 节点 i 的标签更新规则如式(11)所示, 其中 N_i 表示节点 i 的邻接节点, L_i 表示 i 携带的标签集合。

$$L_i^t = \bigcup_{j \in N(i)} L_j^{t-1} \cup \{l_i\} \quad (11)$$

定义 8(标签隶属度更新规则) 考虑到节点对影响力标签的接受程度与相似度、吸引力以及衰减系数有关, 节点 i 在 t 轮迭代时标签隶属度 $P_i^t(l_k)$ 更新的公式为:

$$P_i^t(l_k) = \begin{cases} \sum_{j \in N(i)} P_j^{t-1}(l_k) * \text{Attract}(U_i) * \text{Sim}(U_i, U_j) * \kappa_{i \rightarrow j}, & k \neq i \\ \frac{\text{Attract}(U_i) * d_m^i}{d_m^i + d_{out}^i}, & k = i \end{cases} \quad (12)$$

$$P_i^t(l_k) = \frac{P_i^t(l_k)}{\sum_{k \in L_i^t} P_i^t(l_k)} \quad (13)$$

其中, 当 $k=i$, 即影响力标签为自己时, 式(12)的含义为: 节点 i 被转发的次数占被转发与转发次数之和越大, 自己的影响力标签占有所有标签的比重越大。节点 i 携带的标签 l_i 的隶属度与邻接节点对 i 的传播无关, 这是因为用户不可能通过其他人受到自己的影响。另外, 由于社区发现中为了避免所有节点被标上一样多的社区标签, 对标签个数给出了限制。然而在考虑影响力传播时, 限制了标签数量反而会阻碍影响力向更多的节点扩散, 所以 MLPIA 中暂不考虑对标签个数设限。

定义 9(迭代终止条件) 影响力标签传播的终止条件为所有节点的标签集合不再发生变化, 且标签的隶属度前后差异小于一定的精度范围。

定义 10(用户影响力) 让用户携带自己的标签是为了让影响力得以延续, 并非指用户对自己产生影响。所以标签传播结束以后, 要将每个用户携带的标签中属于自己的标签删除, 重新标准化标签隶属度。最后将网络中每种标签隶属度总和求平均, 结果即为用户的影响力。

MLPIA 算法实现如下。

算法 1 MLPIA 算法流程

Input: 信息转发网络 G_r , 节点相似度矩阵 sim , 吸引力数组 Attract

Output: 节点影响力数组 I

for $i \in V_r$ do // 初始化节点标签集合

$L_i \leftarrow \text{push}(l_i, 1)$

end for

while 不满足迭代终止条件

for $i \in V_r$ do

$L_i^t \leftarrow \{l_i\}$

$P_i^t(l_i) \leftarrow \text{UpdatePval}()$ // 根据式(12)更新标签隶属度

for $j \in N(i)$ do

$L_j^t \leftarrow L_j^{t-1} \cup L_i^{t-1}$

for $l_k \in L_j^t$ do

$P_j^t(l_k) \leftarrow \text{UpdatePval}()$

end for

end for

end while

for $i \in V_r$ do

$L_i \leftarrow \text{remove}(l_i)$ // 删除自己的影响力标签

$P_i \leftarrow \text{Normalize}(P_i^t)$

end for

```

for  $i \in V_r$  do
  for  $l_k \in L_i$  do
     $I_k \leftarrow P_i(l_k) + I_k$ 
  end for
   $I_i = I_i / \text{nodeCount}$ 
end for
return  $I$ 

```

本算法的时间复杂度分析:用 n 表示节点个数, m 表示有向边的条数, k 表示迭代次数。对节点赋予初始标签的复杂度为 $O(n)$, 在更新节点标签的每一轮迭代中, 由每个节点最多可以携带 n 个标签可知, 对于每一个节点 x , 遍历其标签的时间为 $O(n)$, 因此一轮迭代所需的时间为 $O(mn)$ 。迭代后删除所有节点自身标签并标准化的时间为 $O(n)$, 对每种标签的隶属度进行统计的时间为 $O(n^2)$, 所以算法总耗时为 $O(n + kmn + n^2)$ 。

4 实验

4.1 实验数据及参数设置

本文采用的数据集采自 Twitter 平台。以 Kai-Fu Lee、Ning Hu 等用户为种子节点采集用户在 Twitter 平台上的好友关系网络, 再在好友关系网络的基础上从用户发布的推文信息中提出转发关系, 构建信息转发网络, 得到的实验数据如表 1 所列。实验中根据经验, 对式(6)中参数 α, β, γ 的值分别取 0.5, 0.25 和 0.25。

表 1 真实网络数据集

数据集编号	用户数	信息转发边数	时间范围
1	148	683	2014.7—2015.6
2	534	11733	2014.7—2015.6
3	1741	41981	2014.7—2015.6

4.2 评价标准

为了验证 MLPIA 算法的合理性和有效性, 将其与节点入度评估法、PageRank 算法下的结果从影响力覆盖范围、紧密中心度两方面进行评价。

(1) 影响力覆盖范围

本文用线性阈值模型 (Linear Threshold Model, LT)^[11,12] 来测试影响力传播的覆盖范围。LT 模型中初始时选取种子节点激活, 每一个被激活的节点都将会尝试激活自己的入边邻接节点, 直到不再有新的激活节点产生。未被激活的节点 u 被出边邻接节点中已激活的节点 v 以 f_{vu} 进行影响, 当对 u 施加的 f 累积达到一定阈值时, u 转化为激活态, 激活条件如下所示:

$$\sum_{u \in N(v), \text{act}(v)=1} f_{vu} \geq \theta \quad (14)$$

其中, f_{vu} 与节点 v 的影响力和 u 到 v 的出边的权重有关。影响力越大, f_{vu} 越大; 同时 w_{uv} 占 u 出边总权重的比值越大, f_{vu} 越大。在种子用户数量相同的情况下, 被激活的用户越多, 说明种子用户的影响力具有越大的影响范围, 相应的影响力评估算法效果越好。

(2) 紧密中心度

Newman^[13] 根据网络中两点之间的最短路径将节点的紧密中心度定义为与其他所有节点最短路径和的倒数。紧密中心度从网络结构的角度刻画节点对他人产生影响的难易程度。由于相似度与距离负相关, 因此可以利用定义 5 中给出的相似度来计算紧密中心度, 计算公式如下:

$$CC_i = \frac{\sum_{j \in G_f} \text{Sim}(i, j)}{n} \quad (15)$$

实验中使用紧密中心度占比^[14] CCP_i 对 MLPIA 算法下用户影响力排序的合理性进行验证, 其计算公式为:

$$CCP_i = \frac{\sum_{i=1}^{n \times k\%} CC_i}{\sum_{i=1}^n CC_i} \quad (16)$$

4.3 实验及分析

4.3.1 影响力覆盖范围对比

在节点数最大、有向边最多的数据集 3 上用 LT 模型测试影响力覆盖范围。实验中选取影响力排名前 $k\%$ 的用户作为处于激活状态的种子节点, 代表网络中初始时携带有影响力的用户, 对所有节点设置相同的激活阈值 θ 。对 θ 和种子用户比率 k 分别取不同值下的影响力覆盖范围进行对比的结果如图 9 和图 10 所示。

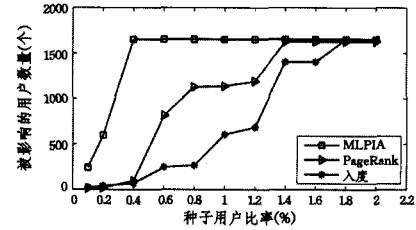


图 9 $\theta=0.8$ 时的影响力覆盖范围对比

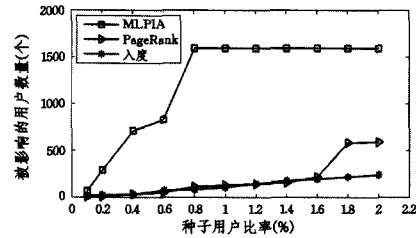


图 10 $\theta=0.9$ 时的影响力覆盖范围对比

由实验结果可知, 在相同数量种子用户的情况下, MLPIA 算法所产生的排名前 $k\%$ 的用户能够影响到的用户数量明显高于 PageRank 算法和入度中心性法, 具有更大的影响力覆盖范围。随着 k 值的提高 MLPIA 算法下的影响人数迅速增长, 当增长到一定阶段后, 开始趋于稳定; 而 PageRank 算法和入度中心性法产生的种子用户的影响人数随着 k 的提高, 呈现相对缓慢的增长, 当 $\theta=0.8, k=1.8$ 左右时, 逐渐向 MLPIA 算法的种子用户影响范围靠拢, 之后也开始保持相对稳定的状态。由于 θ 取值越大, 节点越难被激活, 因此可以从图 10 中看出, 当 $\theta=0.9$ 时, 3 种算法的影响人数的增长速度都受到了抑制, 但是 MLPIA 算法仍然能够对网络中其他用户产生较大范围的影响。

4.3.2 紧密中心度对比

接着本文分别对 3 组数据计算 $k=1, 5, 10$ 时的紧密中心度占比, 结果如表 2 所列。对比表中数据可知, MLPIA 算法的 Top $k\%$ 用户在 3 组数据上的紧密中心度占比几乎都比其他两种算法高, 表明 MLPIA 算法下的高影响力用户在好友关系网络中与其它节点的关系更近。由于传递同一条信息给其他的人所要经过的路程越短, 花费的时间就越少, 产生的效应越大, 因此具有更高紧密中心度的节点更有可能拥有较大的影响力, 所以 MLPIA 的排序效果与其他两种算法相比更为合理。

表2 Top $k\%$ 用户紧密中心度占比

紧密中心度占比	Top1%			Top5%			Top10%		
	网络1	网络2	网络3	网络1	网络2	网络3	网络1	网络2	网络3
入度法	0.014	0.018	0.015	0.057	0.087	0.061	0.125	0.159	0.135
PageRank 算法	0.014	0.010	0.015	0.064	0.082	0.072	0.133	0.146	0.137
MLPIA 算法	0.014	0.020	0.022	0.070	0.097	0.093	0.144	0.185	0.169

结束语 本文应用多标签传播算法来解决社交网络中的用户影响力评估问题,该方法更多地还原信息在网络中传递的过程,产生的高影响力用户具有较大的影响力覆盖范围,更能代表信息传播中的关键节点。在网络舆情监控中可以应用该方法来识别高影响力用户,通过控制他们的言论,引导舆情向有利的方向发展。

由于在 MLPIA 算法中未对节点所能携带的标签数量设限,在迭代过程中节点将会携带越来越多的标签,需要较高的迭代次数才能收敛。在今后的工作中,将进一步研究如何在保证影响力评估质量的前提下提高算法的效率。

参考文献

- [1] Porzycki J, Was J. Novel algorithms of sensors detection in social network[OL]. 2015-10-23. <http://www.researchgate.net/publication/262560387>
- [2] Yang Kai, Zhang Ning, Su Shu-qing. Node Centrality on Individual Microblog User Network[J]. University of Shanghai for Science and Technology, 2015, 37(1): 43-48 (in Chinese)
杨凯, 张宁, 苏树清. 个人微博用户网络的节点中心性研究[J]. 上海理工大学学报, 2015, 37(1): 43-48
- [3] Wang Xiao-fang, Li Xiang, Chen Guang-rong. Network Science: An Introduction[M]. Beijing: Higher Education Press, 2012: 158-159 (in Chinese)
汪小帆, 李翔, 陈关荣. 网络科学导论[M]. 北京: 高等教育出版社, 2012: 158-159
- [4] Kitsak M, Gallos L K, Havlin S, et al. Identification of influential spreaders in complex networks[J]. Nature Physics, 2010, 6: 888-893
- [5] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine[J]. Computer Networks, 1998, 30: 107-117
- [6] Weng J, Lim E P, Jiang J, et al. TwitterRank: finding topic-sensitive influential twitterers[C] // International Conference on Web Search and Data Mining. 2010: 261-270
- [7] Xu Yang. Researches on the center nodes estimate and community discovery on microblog network[D]. Nanning: Guangxi University, 2013 (in Chinese)
徐杨. 微博网络的中心节点评估与社区发现方法研究[D]. 南宁: 广西大学, 2013
- [8] Wu Xian-hui, Zhang Hui, Yang Chun-ming, et al. An Algorithm of Topic-related Microblogging Opinion Leaders Mining[J]. Journal of Chinese Computer Systems, 2014, 35(10): 2296-2301 (in Chinese)
吴岷辉, 张晖, 杨春明, 等. 一种话题相关的微博意见领袖挖掘算法[J]. 小型微型计算机系统, 2014, 35(10): 2296-2301
- [9] Xiao Yin-tao. Research on Rumor Propagation and Public Opinion Mining in Microblogging Community[D]. Zhenjiang: Jiangsu University, 2011 (in Chinese)
萧银涛. 微博社区谣言传播和舆情挖掘研究[D]. 镇江: 江苏大学, 2011
- [10] Gregory S. Finding Overlapping Communities in Networks By Label Propagation[J]. New Journal of Physics, 2010, 12(10): 2011-2024
- [11] Kempe D, Kleinberg J, Tardos É. Influential nodes in a diffusion model for social networks[M] // Automata, Languages and Programming. Berlin Heidelberg: Springer Berlin Heidelberg, 2005: 1127-1138
- [12] Chen Hao, Wang Yi-tong. Threshold-Based Heuristic Algorithm for Influence Maximization[J]. Journal of Computer Research and Development, 2015, 49(10): 2181-2188 (in Chinese)
陈浩, 王铁彤. 基于阈值的社交网络影响力最大化算法[J]. 计算机研究与发展, 2015, 49(10): 2181-2188
- [13] Newman M E J. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality[J]. Physical Review E, 2001, 64(1): 132-158
- [14] Ma Jun, Zhou Gang, Xu Bin, et al. User Influence Analysis in Microblog Based on Topic Diffusion[J]. Journal of Information Engineering University, 2013, 14(6): 735-742 (in Chinese)
马俊, 周刚, 许斌, 等. 一种基于话题传播的微博用户影响力分析方法[J]. 信息工程大学学报, 2013, 14(6): 735-742
- [15] Sugihara R, Gupta R K. Improving the data delivery latency in sensor networks with controlled mobility[M] // Distributed Computing in Sensor Systems. Springer Berlin Heidelberg, 2008: 386-399
- [16] Sugihara R, Gupta R K. Optimizing energy-latency trade-off in sensor networks with controlled mobility[M]. IEEE, 2009: 125-130
- [17] He L, Xu J, Yu Y. Optimize Multiple Mobile Elements Touring in Wireless Sensor Networks[C] // 2009 IEEE International Symposium Parallel and Distributed Processing with Applications. IEEE, 2009: 317-323
- [18] Alhasanat A, Al-Khassawneh M, Twaisi A. A new partitioning-based data gathering algorithm for path-constrained mobile elements in Wireless Sensor Networks[C] // 2013 IEEE Jordan Conference Applied Electrical Engineering and Computing Technologies (AEECT). IEEE, 2013: 1-5
- [19] Somasundara A, Kasal A, Jead, et al. Controllably mobile infrastructure for low energy embedded networks[J]. IEEE Transactions on Mobile Computing, 2006, 5(8): 958-973
- [20] Tang Z, Luo H Q. Application of mobile sink to power saving in sensor network[J]. Journal of Chengdu University of Information Technology, 2005, 20(2): 147-150 (in Chinese)
唐轶, 罗惠琼. 可移动中继节点在传感器网络中的节能应用[J]. 成都信息工程学院学报, 2005, 20(2): 147-150
- [21] Cao H, Zhu X P, Wen H. A survey on the topology of wireless sensor networks based on data MULEs[J]. Information Technology, 2014, 38(3): 13-16 (in Chinese)
曹海, 朱西平, 文红. 基于数据骡子的无线传感器网络拓扑控制综述[J]. 信息技术, 2014, 38(3): 13-16
- [22] El-Moukaddem F, Torng E, Xing G, et al. Mobile relay configuration in data-intensive wireless sensor networks[J]. IEEE Transactions on Mobile Computing, 2013, 12(2): 261-273

(上接第 102 页)