一种新型的社会网络影响最大化算法

田家堂 王轶彤 冯小军
(复旦大学计算机科学技术学院 上海 201203)

摘 要 社会网络中影响最大化问题是对于给定 k 值,寻找 k 个具有最大影响范围的节点集.这是一个优化问题并且是 NP-完全的.Kemple 和 Kleinberg 提出具有较好影响范围的贪心算法,但其时间复杂度很高,不能适用在大型社会网络中,并且不能保证最好的影响范围.文中利用线性阈值模型的"影响力积累"特性,提出了一个该模型下影响最大化算法的框架,并在此框架基础上给出一个新的算法 HPG.HPG 综合考虑网络的结构特性和传播特性,首先启发式选择 PI 值最大的节点,然后寻找最具影响力的节点.实验结果显示 HPG 在最终影响范围和运行时间上都获得比贪心算法更好的效果.

关键词 社会网络;贪心算法;影响最大化;带符号网络;信息传播
中图法分类号 TP311 **DOI 号**: 10.3724/SP.J.1016.2011.01956

A New Hybrid Algorithm for Influence Maximization in Social Networks

TIAN Jia-Tang WANG Yi-Tong FENG Xiao-Jun
(*School of Computer Science, Fudan University, Shanghai* 201203)

Abstract Influence maximization is a problem of finding a small subset of nodes (target set) in a social network that could maximize the spread of influence. This optimization problem of influence maximization is NP-hard under several most widely studied diffusion models and is even challenging for current online social networks which contain both positive and negative relations. Kemple and Kleinberg proposed a natural climbing-hill greedy algorithm that chooses the nodes which could provide a good marginal influence. This greedy algorithm has large spread of influence, but is very costly and cannot be applied to large social networks. Also, greedy algorithm could not guarantee the best influence spread. In this paper, we propose a framework on the linear threshold model and a hybrid potential-influence greedy algorithm (HPG) which can make good use of the "influence accumulation" property of the linear threshold model. Considering the network structure and propagation characteristics, HPG algorithm first heuristically choose half of the initial seeds with the biggest potential influence (PI) and then greedily choose the other half initial seeds with the most influence. Experiments are conducted comprehensively on different real datasets (including weighted social networks, directed social networks and signed social networks). Experimental results demonstrate that HPG algorithm significantly outperforms the local-optimal greedy algorithm and could achieve reduced running time.

Keywords social network; greedy algorithm; influence maximization; signed social network; information diffusion

收稿日期:2011-07-18;最终修改稿收到日期:2011-08-19.本课题得到国家自然科学基金(61033010)资助.**田家堂**,男,1987 年生,硕士研究生,主要研究方向为社会网络、Web 挖掘、Web 信息检索. E-mail: 09210240036@fudan.edu.cn. **王轶彤**(通信作者),女,1973 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为数据库、Web 挖掘及数据挖掘、Web 信息检索等. E-mail: yitongw@fudan.edu.cn. **冯小军**,男,1984 年生,硕士研究生,主要研究方向为 Web 挖掘、Web 信息检索、社会网络.

1 引　言

社会网络是指由个体及个体之间的关系所组成的一个复杂网络，这种复杂的社会结构对信息的传播和扩散起着至关重要的作用. 当一个人采纳一个新的思想或接受一种产品时，他会向他的朋友或同事推荐，某些人可能会接受或采纳他的推荐，并进一步向他们自己的朋友或同事推荐，这个过程称为传播或扩散(Propagation or Spreading). 一个人的行为在很大程度上取决于周边的朋友或同事的决定.

社会网络的传播和扩散过程在社会科学中已有很长的研究历史. Richardson 和 Domingos 等人[1]将影响最大化问题归纳为一个算法问题，即如何定位网络中某些最有影响力的成员，提供给他们免费的样品，希望通过他们向网络中其他成员推荐，从而达到营销的目的，那么该如何选择这 k 个初始成员使得最终购买人数最多？影响最大化问题的研究有着十分重要的现实意义，在市场营销、广告发布、舆情预警以及社会安定等方面有十分重要的应用. 随着 WEB2.0 的出现及流行，目前出现了很多大型在线社交网站，如 Facebook、Flickr 等. 这些大型在线社会网络的成员数目庞大，它们的出现对传统社会网络中的影响最大化算法，包括传播模型均提出了巨大的挑战. 近年来，社会网络中影响最大化算法再次成为研究热点. 目前研究的目标主要集中在如何扩大影响范围同时降低算法的时间复杂度.

社会网络中影响最大化问题(即如何选择 k 个种子节点，使其在传播过程结束之后，传播的范围达到最大)已被证明是一个 NP-hard 问题[2]. Kemple 和 Kleinberg 提出了一种自然的爬山贪心算法. 它在每一步都选择当前“最具影响力”的节点作为初始传播对象进行传播. 所谓“最具影响力”的节点，即是当前能够激活最多节点的节点. 然而，选择“最具影响力”的节点是一个非常耗时的过程，并且这种局部最优并不能保证最终的传播结果最优. 对于大型社会网络，这种贪心算法由于高耗时更加不适用. 除了贪心算法之外，还有一些常见的启发式节点选择策略，包括基于点的度数或中心度等[2]. 然而，就传播范围而言，完全基于度数的启发式规则的效果并不理想. 因为，该方法显然没有考虑到社会网络的传播特性.

当然信息扩散有其本身的规则，或者叫做模型. 当前所有社会网络影响最大化问题的研究都是基于以下两个基本传播模型：线性阈值模型(Linear Threshold Model，LT 模型)和独立级联模型(Independent Cascade Model，IC 模型)，我们将在第 2 节进行介绍.

目前影响最大化问题的主要研究工作集中在 IC 模型下利用次模特性(sub-modularity)来提高贪心算法的运行效率[3]. 本文考虑 LT 模型(介绍见第 2 节)下的影响最大化问题. 通过考察，我们发现 LT 模型具有“影响力积累”的特性. 利用这个特性，在本文我们提出了一种混合算法，通过综合考虑网络的特性和传播特性来提高最终的影响范围和降低算法的运行时间. 本文的主要贡献有：(1) 基于线性阈值模型提出一种算法框架和一种新型的混合式影响最大化算法(HPG)；(2) 将 HPG 算法推广到带符号的社会网络中，并给出合理的节点之间的影响力 b_{uv} 估计公式；(3) 在各种不同特性的数据集上进行了实验，并分析和验证了 HPG 算法的有效性.

本文第 2 节介绍两个最常用的传播模型和相关工作；第 3 节介绍提出的算法框架和混合型算法并给出不同特性网络上 b_{uv} 的估计公式；第 4 节介绍在 6 个不同特性数据集上进行的实验及结果分析；最后一节进行总结并探讨未来的工作.

2 背景知识

2.1 两个基本传播模型

一般将社会网络抽象为一张有向(无向)图 $G(V,E)$，V 代表节点的集合，每个点表示个人或组织；E 代表边的集合，每条边表示个体之间的关系(合作、朋友、敌对等). 每个节点有两种状态，激活状态(购买了某产品或接受了某观念等)和未激活状态(还未购买或未接受). 处于激活状态的节点对处于未激活状态的节点存在影响，如果这个影响导致了某个节点从未激活状态变为激活状态则这个过程称为激活. 某节点的邻居节点被激活的越多，则该节点被激活的可能性就越大. 新激活的节点又会影响其处于未激活状态的邻居节点. 随着时间的推移，越来越多的节点从未激活状态转变为激活状态. 整个传播过程是不可逆的，即：一个节点可以从未激活状态变为激活状态，反之则不可.

2.1.1 线性阈值模型[4]

线性阈值模型是所有基于节点特异性阈值模型的核心. 给定一个社会网络 $G(V,E)$，定义 $N(v)$ 为节点 v 的邻居节点集合(有向图中，“邻居”的定义为

“入边邻居”).被激活的节点 u 对邻居节点 v 存在影响 b_{uv},一个节点 v 的所有邻居节点对 v 的影响力总和小于等于 1.定义 $A(v)$为节点 v 的邻居节点中已激活的节点集合.每个节点 v 有一个特异性阈值 θ_v,如果满足 $\sum_{u\in A(v)} b_{uv}\geq\theta_v$,则 v 被激活.在 LT 模型中,当一个激活节点 u 尝试去激活它的未激活邻居 v 而没有成功时,节点 u 对节点 v 的影响力 b_{uv} 被“积累”下来,而不是被抛弃.这种积累对后面节点 v 的其它邻居对 v 的激活是有贡献的,直到节点 v 被激活或传播过程结束.这就是 LT 模型的“影响积累”特性,这和 IC 模型是不同的.LT 模型的扩散过程如下:给定初始传播节点集合 S_0、所有节点的特异性阈值 θ_v 以及节点之间的影响力 b_{uv}.在第 t 步扩散时,基于集合 S_{t-1} 激活满足阈值的节点,被激活的节点加入到集合 S_{t-1} 形成 S_t.重复这一过程,直至不再有新的节点被激活.

2.1.2 独立级联模型[5-6]

独立级联模型是基于相互粒子系统(Interacting Particle Systems)设计的一个信息扩散的模型,这是一个概率模型.给定初始传播节点集合 S_0 以及所有节点之间相互激活成功的概率 p_{uv}.当传播至第 t 步时,利用在 $t-1$ 步中被激活的节点,根据成功概率 p_{uv} 试图去激活它们的邻居节点,并将在这一步中被激活的节点加入到 S_{t-1} 形成 S_t.重复这一过程,直至不再有新的节点被激活.值得注意的是成功概率 p_{uv} 是一个系统变量,与其它尝试激活节点 v 而未成功的节点无关,这也是该模型命名的来历.显然,在 IC 模型下,当节点 u 尝试激活其邻居节点 v 而失败时,这种激活行为就被抛弃了.Kemple 和 Kleinberg 认为 p_{uv} 并不是独立的,随着时间的推移会变得越来越小.也就是说 v 已经被其它很多节点尝试激活过很多次都没有成功,新激活的邻居节点 u 对 v 的影响就会被削弱,由此提出了一个新的模型叫做递减级联模型[2](Decreasing Cascade Model).

2.2 相关工作

2.2.1 爬山贪心算法

为了找到模型中要求的初始扩散集合 S_k,一个简单有效的策略是每一步根据算法的标准确定初始集合中的一个节点,直到找到 k 个(预定义)节点.为了便于介绍算法,我们定义:

(1) $S_0=\varnothing$;

(2) $I(S_i)$:集合 S_i 扩散后已激活节点的集合;

(3) $m(u|S_i)=|I(S_i\cup\{u\})|-|I(S_i)|$:节点 u 相对于集合 S_i 的边际影响范围.

Kemple 和 Kleinberg 在文献[2]中提出一种自然的爬山贪心算法(我们用作者的姓名缩写 KK 来表示):每一步都选择当前最具影响力的节点.从 S_0 开始,在第 i 步,根据局部最优策略选择节点 u,$u=\underset{v}{argmax}\,m(v|S_{i-1})$,并令 $S_i=S_{i-1}\cup\{u\}$.尽管 KK 算法能够在 $1-1/e$ 的因子内近似最优值,但是其缺点非常明显,每一步都要计算所有未激活节点 u 的边际影响 $m(u|S_i)$,这导致了 KK 算法运行十分耗时.文献[3]中,利用 IC 模型的次模特性给出了 KK 算法的改进方法 CELF 算法.次模特性指当我们添加一个节点 v 到种子集合 S 时,如果 S 集合越小,v 对影响范围的增量影响就越大.CELF 利用次模特性,在每一步选择初始种子节点时,大量节点的增量影响不需要被重新计算,这是因为它们在之前步骤中的值已经小于其它节点在当前步骤中的值.它缩短了 KK 算法的时间,但在影响范围上没有提高.然而,次模特性并不适合 LT 模型,我们希望利用 LT 模型的“影响力积累”特性来提高算法的运行效率.

2.2.2 基于度数的节点选择策略

中心度是分析社会网络的一个最重要的和常用的概念工具之一.它是关于行动者在社会网络中的中心性位置的测量概念,反映的是行动者在社会网络结构中的位置或优势的差异.在一个社会网络中,某节点度数最高,该点就居于中心位置,这表明该点所对应的行动者为此网络中的中心人物即最具影响力的人物[7].在社会网络和其它网络中,以度数递减的顺序选择 k 个最大度数节点的启发式节点选择策略,是长期以来的一个标准方法,在社会科学中被称为“度中心性”.此方法的一个缺点就是静态选择初始节点,没有考虑影响的扩散过程,不能保证最终影响范围最优.

2.2.3 其它相关算法

集合覆盖贪心算法[8]每次选择最高“uncovered”度数的节点,一旦一个节点被选中,它的所有邻居节点被标记为“covered”,这个过程一直持续 k 步.它选择覆盖范围最大的节点,但是在影响最大化的约束下,覆盖并不等于激活,所以其实验结果并不好.

文献[9]中讨论了如何在作者合作网络中找到 k 个研究员,使得与他们合作的其他研究员人数最多.这是影响最大化问题的一个特例.该文中提出的计算 Shapley 值的方法并不适合本文提出的问题.

3　一种新型的混合式影响最大化算法

3.1　框架的提出及 HPG 算法

由于 KK 算法时间复杂度高，且局部最优并不能保证具有最好的最终影响范围，这里的影响范围是指能够激活的节点个数. 我们利用 LT 模型的“积累特性”并综合网络的结构特性和传播特性，提出了解决问题的框架，并在此框架基础上提出了新的初始节点选取策略（Hybrid Potential-influence Greedy Algorithm，HPG），这是一种混合式的影响最大化算法. 根据 LT 模型的“积累特性”，我们知道尽管一个节点 u 没能成功激活节点 v，但影响力 b_{uv} 却被积累了. 因此，我们的出发点是与其花费大量的时间寻找“最具影响力”的节点，不如迅速地找到一些具有潜在影响力的节点：这些节点虽然当前并不能激活最多的节点，但却能积累大量的“潜在影响力”，使得在后面的阶段能够激活更多的节点. 这样可以大大地节省运行时间，因为寻找“最具影响力”的节点是非常耗时的，我们需要针对当前所有未激活节点去计算它的扩散影响范围，在算法初始阶段，即大部分节点均未被激活时，尤其耗时. 基于这个出发点，我们提出了一个解决问题的框架：框架的核心是节点的选取过程分为两个部分.（1）启发阶段. 选取最具潜在影响力的节点. 这部分节点虽然在当前不能带来最大的影响范围，但却蕴含着巨大的潜在影响力.（2）贪心阶段. 选取最具影响力的节点.

文献[8]中提到度数大的节点往往都处在社会网络的中心位置. 而 KK 算法考虑影响的传播过程，能够达到最优值的 63%[2]. 综合度中心性和贪心算法的优势并结合 LT 模型的特性，首先静态选择最具潜在影响力的节点来激活，尽管当前不能激活最多的节点，在接下来贪心阶段时，很大部分没有被激活但是积累了很多潜在影响力的节点一触即发. 尽管这一框架不是局部最优，但其最终影响范围更大.

为了找到潜在影响力最大的节点，通过网络结构分析和实验，我们发现有两个因素影响最具潜在影响力节点的选择：节点的度数（$outdegree$）和一个激活节点对其所有未激活邻居影响力之和（inf），并且这两个因素的效果也不相同. 根据实验结果，节点的度数对最终影响范围的贡献更加明显. 因此，综合上述原因，我们给出下列公式，使得最具潜在影响力节点的选择以节点的度数为主，并综合考虑 inf 因素. 潜在影响力 PI（Potential Influence）定义如下：

$$inf(u)=\sum_{v\in \overline{N}(u),v\notin \overline{A}(u)} b_{uv} \tag{1}$$

$$PI(u)=outDegree(u)+(1-e^{-inf(u)}) \tag{2}$$

其中 $\overline{N}(u)$ 表示 u 的出边邻居节点集合，$\overline{A}(u)$ 表示 $\overline{N}(u)$ 中已激活的节点，b_{uv} 表示节点 u 对 v 的影响力. 因此 $inf(u)$ 表示节点 u 施加于所有未激活邻居节点的影响力之和，我们称之为节点 u 的影响力，它由节点 u 未激活邻居的个数以及 b_{uv} 大小这两个因素决定. $outDegree(u)$ 表示节点 u 的出度. 当节点的出度相同时，选择影响力较大的节点，而不是盲目随机地选择一个度数最大的节点作为当前潜在影响力最大的节点. 因此，“最具潜在影响力”的节点就是 PI 值最大的节点. 显然我们将计算每个节点的边际影响范围简化为计算每个节点的 PI 值. 这里 PI 值在常数时间内可以计算获得.

需要特别注意的是，当处理带符号社会网络时，式(2)中的 $outDegree(u)$ 的取值需要进行特别处理. 理论上我们是选择当前节点中度数最大的节点，但是由于负边的存在，我们重新定义

$$outDegree(u)=outDegree_{+}(u)-outDegree_{-}(u) \tag{3}$$

其中 $outDegree_{+}(u)$ 表示 u 出边中正边的个数，$outDegree_{-}(u)$ 则表示 u 出边中负边的个数. 这是由于 u 的出度中包含正边和负边，正边对影响的传播产生正影响，负边对影响的传播产生负影响，正影响减去负影响才能表示节点 u 的度中心性.

我们给出的框架中将 k 个初始节点的选择分为两个阶段：启发阶段和贪心阶段. 启发阶段选择 PI 值最大的节点，贪心阶段选择最具影响力的节点. 框架中引入了启发因子 c（$c\in[0,1]$），$\lceil ck\rceil$ 表示贪心阶段的步数，$k-\lceil ck\rceil$ 表示启发阶段的步数. 显然，当 $c=1$ 时框架中的方法为 KK 算法. 框架中启发因子 c 是一个经验值，我们需要确定 c 的取值，给出一个确定的影响最大化算法即提出的 HPG 算法. 在后文实验部分，我们得出 HPG 算法是 $c=0.5$ 时的一种影响最大化算法.

框架的伪代码详见算法 1.

算法 1.　HPG 算法框架.

输入：图 $G(V,E)$，b_{uv}，θ_v，初始集合大小 k，启发因子 c

输出：初始传播集合 S_k

1. 集合 $S_0=\varnothing$，$k1=k-\lceil ck\rceil$，$k2=\lceil ck\rceil$
2. FOR $i=1$ TO $k1$
3.　　从未激活的节点中选择 PI 最大的节点 u
4.　　令 $S_i=S_{i-1}\cup\{u\}$
5.　　基于集合 S_i 激活尚未激活的节点
6.　　更新所有还未被激活节点的 PI 值

7. ENDFOR
8. FOR $i=1$ TO $k2$
9. 计算每个尚未被激活节点的边际影响值
10. 选择边际影响值最大的节点 u
11. 令 $S_{i+k1}=S_{i-1+k1}\cup\{u\}$
12. 更新节点 u 所影响的节点状态
13. ENDFOR

3.2 b_{uv} 估计公式的改进

线性阈值模型中 b_{uv} 表示激活的节点 u 对邻接点 v 存在的影响，通常用 $1/d(v)$（$d(v)$ 表示 v 的度数）估计，意味着 v 的所有的邻居对它的影响力相同. 显然，这一假设忽视了节点之间的差异性，并不符合现实. 这里我们针对不同特性的社会网络给出了不同的 b_{uv} 估计公式.

3.2.1 无权图上 b_{uv} 的估计公式

b_{uv} 的大小是节点 v 自身的一个感受，这个感受是它对于指向它的节点的权威性的一种体现，与其它节点没有关系，因此只需要考虑节点 v 邻居节点的结构关系即可[10]. 邻接图 NG（Neighbor Graph）由 v 和指向它的邻居节点以及这些节点之间的关系构成. b_{uv} 估计公式中节点的度数是根据邻接图中节点的度数得来的. 式(4)给出了它的定义.

$$NG(v)=G'(V',E'),V'=\{v\}\cup N(v),$$
$$E'=\{(x,y)\mid x,y\in V',(x,y)\in E\},$$
$$b_{uv}=\frac{outDeg(u)}{\sum_{w\in N(v)}outDeg(w)} \tag{4}$$

其中 $outDeg(u)$ 表示节点 u 在邻接图中的出度. $N(v)$ 表示节点 v 在邻接图中的入边邻居节点集合. 节点 u 对 v 的影响力主要由 v 的邻接结构来决定.

图 1 给出一个简单的邻接图示例，深色节点 v，以及节点 u_1，u_2 和 u_3，组成一张邻接子图. u_1，u_2 和 u_3 在 $NG(v)$ 中的出度分别为 1,1,2，因此 b_{u1v}，b_{u2v} 和 b_{u3v} 分别为 0.25，0.25 和 0.5.

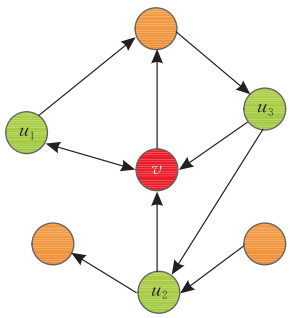

图 1 节点 v 的邻接图

3.2.2 带权图上 b_{uv} 的估计公式

在考虑边上带权重的情况下，节点 u 对 v 的影响力主要由边上的权重来决定. 我们规定 b_{uv} 的定义如下：

$$b_{uv}=\frac{W(u,v)}{\sum_{w\in N(v)}W(w,v)} \tag{5}$$

其中 $W(u,v)$ 表示边 (u,v) 上的权重，$N(v)$ 表示节点 v 在邻接图中的入边邻居节点集合.

图 2 给出了一个简单的示例. 边 (u_1,v)，(u_2,v) 和 (u_3,v) 的权重分别为 2,5,3，根据式(3)计算得到 b_{u1v}，b_{u2v} 和 b_{u3v} 分别为 0.2，0.5 和 0.3.

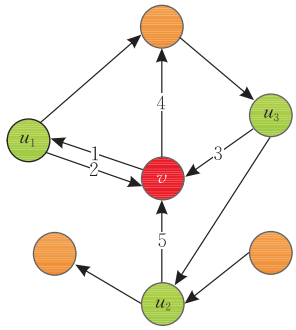

图 2 节点 v 的带权邻接图

3.2.3 带符号图上 b_{uv} 的估计公式

之前大量社会网络的研究只关注传统社会网络，而关于带符号社会网络的研究相对很少. 带符号的社会网络是指个体与个体之间的关系不仅存在正关系还存在负关系，其中正关系对影响传播产生正影响，负关系产生负影响. 在真实的社会网络中，考虑正关系和负关系之间的相互作用是非常重要的[11]. 例如资讯科技网站 Slashdot 的用户之间可以相互标注为“朋友”或者“敌人”，评论网站 Epinions 的用户之间可以相互表达“信任”或者“不信任”等.

我们将一个带符号的社会网络抽象为一个由正边和负边组成的有向图 $G(V,E)$. 其中，边的符号代表影响的正负. 定义 $Triangle(a,b,c)$ 表示由边 (a,b)，(a,c) 和边 (b,c) 三条有向边所组成的三角形，如图 3 所示，其中 x,y,z 代表边的符号即影响的正负. 在带符号图中，当考虑 a 对 c 的影响时，a 对 c 产生直接影响，a 通过 b 对 c 产生间接影响，但需要注意的是并不是所有的间接影响都有效. 根据带符号图上的 Balance 理论[11]，给出乘法规则：若要 a 通过 b 对 c 产生的间接影响有效，则必须满足 $z=x\times y$. 有了乘法规则我们就可以通过邻接图来进行带符号图上 b_{uv} 的计算.

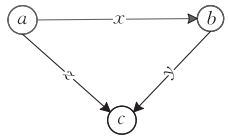

图 3 带符号图中(a,b,c)组成的三角形

在考虑存在负边的情况下，我们规定 b_{uv} 的定义如下：

$$b_{uv}=\frac{sign(u,v)\times W(u,v)}{\sum_{w\in N(v)}W(w,v)} \tag{6}$$

其中 $W(u,v)$ 不是指边上的权重，而是指邻接图中 u 对 v 的影响权重，由边 (u,v) 产生的直接影响和 $Triangle(u,u',v)$ 产生的间接影响来决定. 首先考察边 (u,v) 产生的直接影响，将 $W(u,v)$ 初始化为 1，再考察间接影响，若存在 $Triangle(u,u',v)$ 满足乘法规则，则 $W(u,v)$ 加 1，否则什么都不做.

图 4 给出了一个简单示例. $Triangle(u_1,u_3,v)$ 满足乘法规则，$Triangle(u_1,u_2,v)$ 不满足乘法规则，所以在邻接图中 u_1 对 v 的影响权重为 $2=1+1+0$. 图中 $W(u_1,v)=2$，$W(u_2,v)=2$，$W(u_3,v)=1$，$W(u_4,v)=2$，$W(u_5,v)=1$. 因此根据公式 3 计算得到 b_{u1v}，b_{u2v}，b_{u3v}，b_{u4v} 和 b_{u5v} 分别为 0.25，−0.25，−0.125，0.25 和 0.125.

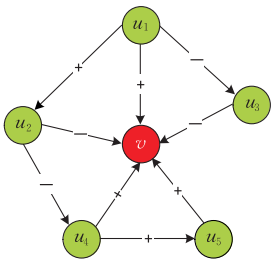

图 4　节点 v 的带符号邻接图

3.3　KK 算法和 HPG 算法的算法时间复杂度分析

下面，我们对两个算法的运行时间进行简要的说明和分析. KK 算法考虑了影响的整个传播过程，每一步都需要计算所有未激活节点的边际影响. 刚开始时，图中所有的节点均处于未激活状态，计算每个未激活节点的边际影响均要遍历整张图来进行影响的传播，非常耗时.

HPG 算法是在所提框架下启发因子 $c=0.5$ 时的一种影响最大化算法. 我们从两个阶段来分析 HPG 算法的时间复杂度.

启发阶段，每一步都是从未激活节点中选择 PI 值最大的节点. PI 值的计算基本不消耗时间（时间复杂度为常量，用 $O(C)$ 表示），这是因为公式（1）中的 $outDegree(u)$ 是不变的，$inf(u)$ 在确定上一个初始种子节点之后进行整个图的更新时也已计算完毕. 启发阶段结束之后，图中积累了很大的潜在影响力，同时激活了大量的节点.

贪心阶段，虽然在贪心阶段每一步选择的都是边际影响最大的节点. 但是由于已经经历过了启发阶段，图中已经有大量的节点被激活，此时未激活的节点比原始数据集会少很多，相应的会比 KK 算法少遍历很多遍图. 可以将之看作 KK 算法在小规模数据集上的运行. 因此 HPG 算法的时间复杂度要比 KK 小很多.

表 1　HPG 和 KK 算法两个阶段的时间复杂度

	HPG 算法		KK 算法	
	启发阶段	贪心阶段	相应阶段 1	相应阶段 2
节点情况	大量未激活节点	少量未激活节点	大量未激活节点	少量未激活节点
时间复杂度	$T(1)=k/2\times O(C)$	$T(2)=k/2\times O(\ll NM)$	$T'(1)=k/2\times O(NM)$	$T'(2)=k/2\times O(\ll NM)$

表 1 中给出了 HPG 算法和 KK 算法两个阶段的时间复杂度对比. 我们将 KK 算法也相应地分成两个阶段，每个阶段均选择 $k/2$ 的初始节点. 需要注意的一点是，随着图中未激活节点个数的减少即随着初始节点选择的进行，贪心算法的运行时间也在单调减少. 假设要选择 k 个初始节点，图中有 N 个节点，M 条边.

表 1 中，$\ll$ 表示远小于的含义. HPG 算法的时间复杂度为 $T(1)+T(2)=k/2\times O(\ll NM)$. KK 算法的时间复杂度为 $T'(1)+T'(2)=k/2\times O(NM)$. 因此，HPG 算法时间复杂度比 KK 算法时间复杂度低很多.

4　实验和评估

4.1　实验数据集

我们在 6 个真实数据集上进行了实验，它们的统计信息详见表 2.

数据集 1 是酵母菌蛋白质之间的相互作用①，有关它的社会网络性质详见文献[12]，无向无权图.

数据集 2 是个作者合作社会网络②，点代表作者，边代表两个作者有合作关系，无向无权图.

① http://vlado.fmf.uni-lj.si/pub/networks/data/bio/Yeast/Yeast.htm

② http://www-personal.umich.edu/~mejn/netdata/

表 2 数据集的统计信息

序号	数据集	节点数	边数（正边/负边）	平均度数	社会语义
1	酵母菌	2361	13292	11.6	蛋白质相互作用
2	航天物理合作网络	16706	121251	14.5	作者合作
3	计算几何合作网络	7343	11898	3.2	作者合作
4	Wiki_vote	7115	103689	26.6	Wiki 投票
5	Slashdot	77357	396378/120197	13.4	资讯科技网站（敌/友）
6	Epinions	131828	717667/123705	12.8	评论网站（信任/不）

注：在这些统计数据中，有向图都当作无向图来处理.

数据集 3 是个带权的作者合作社会网络①，点代表作者，边上的权值表示作者之间合作的次数，是带权社会网络的代表，无向带权图.

数据集 4 是个 Wikipedia 的投票历史网络②，其中点代表 Wikipedia 的用户，u 到 v 的有向边意味着 u 投票给了 v，有向无权图.

数据集 5 是一个来自 Slashdot 网站的朋友敌人网络③，点代表网站的会员，u 到 v 的有向边意味着 u 把 v 标注为朋友或者敌人（由边上的符号决定）. 这是个带符号的有向图.

数据集 6 是一个来自 Epinions 的信任网络④，节点代表网站的会员，u 到 v 的有向边意味着 u 信任 v 或者不信任 v（由边上的符号决定）. 这是个带符号的有向图.

这 6 个数据集都是社会网络领域公开的用于各种测试的数据集，它们具有不同的特性（带权\不带权，有向\无向，带符号\不带符号）. 我们注意到对于数据集 Wiki_vote、Slashdot 和 Epinions，需要对原始图进行反向处理，因为我们研究的是影响的最大化，我们把 v 投票给 u，信任 u 或者把 u 标注为朋友看成 u 对 v 产生影响.

4.2 实验设计

基于线性阈值模型进行实验，b_{uv} 按照 3.2 节中给出的估计公式计算. 文献[2]中给出一个经典的阈值 θ_v 取值方法是固定每个节点的阈值为 0.5. 启发因子 c 意味着贪心阶段拥有 $\lceil ck \rceil$ 步，启发阶段拥有 $k-\lceil ck \rceil$ 步. 文献[3]给出的算法虽然降低了贪心算法的时间复杂度，但最终影响范围没有提升，所以这里我们不予比较.

4.3 实验结果

4.3.1 算法框架在无向网络上的效果

首先考察启发因子 c 和目标集合大小 k 的联合效果，即当 k 相同时，不同 c 值对影响范围的影响. 数据集 1 上的实验结果见图 5. 由图可知对于不同的 k 值，除了 $c=0$ 的情况，其它的 c 值大部分情况都比 $c=1$ 的影响范围大. 当 c 取 0.5，k 取 60 时算法框架下的算法的影响范围高出贪心算法 10% 左右. 当 $c=0$ 时，所有初始节点都是静态选取的 PI 值最大的节点，未考虑影响的传播过程，所以其影响范围最差. 在接下来的实验中，我们直接忽略 $c=0$ 的情况.

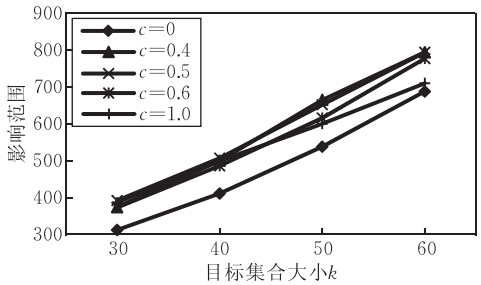

图 5 数据集 1 上不同 k 和 c 影响范围曲线

数据集 2 的实验结果见图 6，同样算法框架在取到合适的 c 值时绝大部分都要比 KK 算法效果好，仅有个别的情况效果差于 KK 算法.

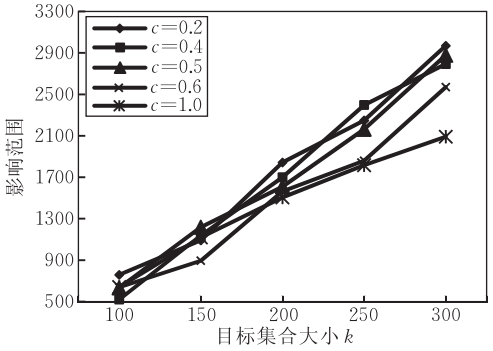

图 6 数据集 2 上不同 k 和 c 影响范围曲线

4.3.2 算法框架在带权网络上的效果

为了验证框架在带权网络上的有效性，我们在计算几何作者合作网络数据集上进行了同样的实验. 实验中 b_{uv} 采用 3.2.2 节中的 b_{uv} 的估计公式，θ_v 取 0.5. 结果见图 7.

从图中可以看到，算法框架下算法大部分影响范围都优于 KK 算法，特别是 $c=0.5$ 的情况，最好情况下得到 10% 左右的提升. 图 5、图 6 和图 7 都显示出随着 k 增大算法框架下算法的影响范围曲线的斜率越大，表明这种提升随着 k 值的增大而愈加的

① Jones B. Computational Geometry Database, February 2002; FTP/HTTP
② http://snap.stanford.edu/data/wiki-Vote.html
③ http://snap.stanford.edu/data/soc-sign-Slashdot081106.html
④ http://snap.stanford.edu/data/soc-sign-epinions.html

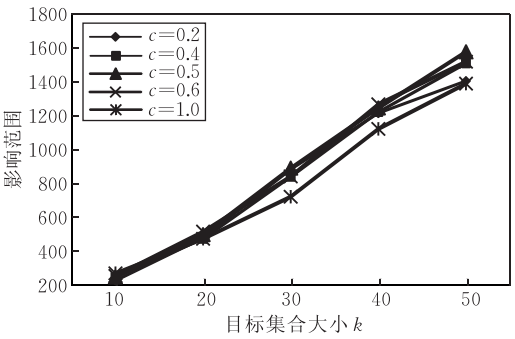

图 7　数据集 3 上不同 k 和 c 影响范围曲线

明显. 然而，当前的大型社会网络均包含上百万个节点，过小的 k 值没有实际的意义. 因此，本文提出的算法框架在大型社会网络中的优势愈加明显，因为该框架需要足够的节点来积累潜在的影响力.

4.3.3　算法框架在有向网络上的效果

为了验证框架在有向网络上的有效性，我们在 Wiki_vote 数据集上进行了同样的实验. 实验中 b_{uv} 采用 3.2.1 节中的 b_{uv} 估计公式，θ_v 取 0.5. 结果见图 8.

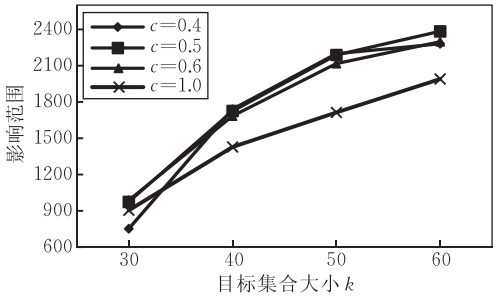

图 8　数据集 4 上不同 k 和 c 影响范围曲线

从图中可以看到，实验结果跟无向网络和带权网络上的结果类似，算法框架下的算法大部分情况下影响范围都优于 KK 算法，特别是 $c=0.5$ 的情况. 这说明了算法框架在有向网络上也能够取得很好的效果.

那么对于不同的数据集类型，该如何确定 c 值？我们不可能对每一个 c 值做实验，然后选择最优的. 但从实验结果中可以看到 $c=0.5$ 时大部分情况下都要优于 KK 算法. 因此我们选择中间值 0.5 作为算法的参数，不管 c 值偏大还是偏小，$c=0.5$ 在统计意义上离中心最近. 因此我们把 $c=0.5$ 时算法框架下的影响最大化算法确定为 HPG 算法.

4.3.4　HPG 算法在带符号社会网络上的效果

我们已经确定 HPG 算法是 $c=0.5$ 时的影响最大化算法，下面直接验证 HPG 算法和 KK 算法在带符号网络上的效果. 我们选取两个大规模的带符号社会网络进行实验. 分别是 Slashdot 和 Epinions 数据集. 实验结果分别见图 9 和图 10.

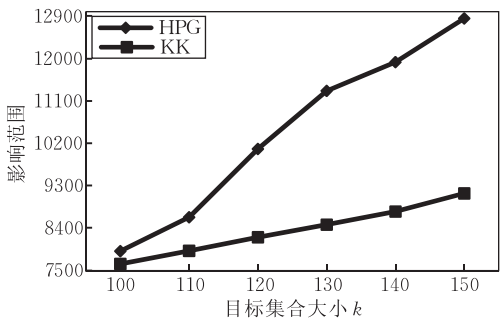

图 9　数据集 5 上 HPG 和 KK 影响范围曲线

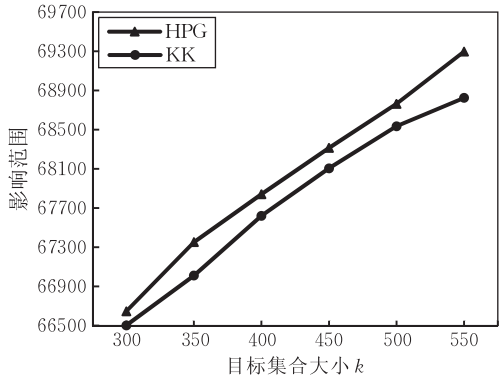

图 10　数据集 6 上 HPG 和 KK 影响范围曲线

从图中可以看出，在带符号网络上 HPG 算法相对于 KK 算法在影响范围上同样有很大的提升，同时也验证了参数 c 赋值的正确性.

4.3.5　KK 算法和算法框架之间的详细比较

下面考察算法框架两个阶段的区别，说明为什么 HPG 算法要比 KK 算法的影响范围广.

表 3 和表 4 中列出了算法框架两个阶段的节点平均影响数 AI(Average Influence). 为了便于比较，我们将 KK 算法也对应地分为两个阶段(相应阶段 1 和相应阶段 2). 这里我们用 AI 作为比较的工具. 例如在贪心阶段选择了 10 个节点，而它们共激活了 60 个节点，那么贪心阶段的 $AI=6.0(60/10)$.

表 3　数据集 1 上 $k=50$ 各阶段的 AI

平均影响数	算法框架		贪心算法	
	启发阶段	贪心阶段	相应阶段 1	相应阶段 2
$c=0.2$	10.27	20.4	14.6	11.35
$c=0.4$	10.4	17.65	13.75	10.83
$c=0.6$	10.65	15.36	13.24	10.76
$c=0.8$	11.9	12.15	12.47	10.1

表 4　数据集 3 上 $k=40$ 各阶段的 AI

平均影响数	算法框架		贪心算法	
	启发阶段	贪心阶段	相应阶段 1	相应阶段 2
$c=0.2$	26.15	47.25	28.13	28.03
$c=0.4$	19.42	47.21	24.19	30.63
$c=0.6$	18.31	40.58	25.08	32.5
$c=0.8$	18.63	36.5	24.06	34.3

从表 3 和表 4 中可以看出，启发阶段要比相应阶段 1 的平均影响数低，但是贪心阶段要比相应阶段 2 的平均影响数高出很多，因此只要选择合适的参数 c，综合两个阶段，我们所提框架下算法的效果要优于 KK 算法. 启发阶段比相应阶段 1 低的原因容易解释，因为启发阶段寻找到的节点都是静态选择 PI 值最大的节点并不是最具影响力的节点. 贪心阶段比相应阶段 2 高出很多主要是启发阶段中所选择节点的特性和 LT 模型的“影响力积累”特性导致的，虽然它们没有激活很多节点，但是广泛地施加影响给网络中的其它节点，这些被影响但还没有被激活的节点在贪心阶段一触而发，这就是为什么贪心阶段能够激活更多节点的原因.

4.3.6 时间复杂度的比较

前面对算法影响范围进行了详细的对比分析，这里进行时间复杂度的比较. 为了更好地区别算法在时间复杂度上的差别，我们选择规模较大的 Slashdot 数据集，给出了算法框架在不同参数下所消耗的时间，结果见图 11.

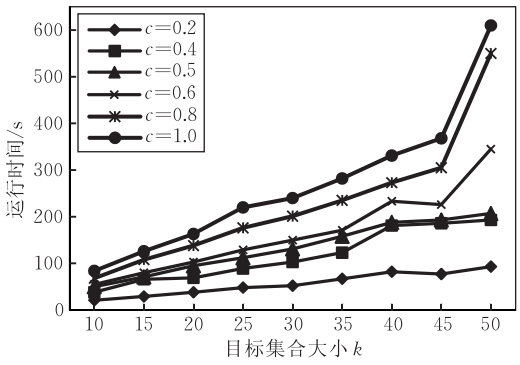

图 11 数据集 5 上算法时间图

从图中可以看到当 k 值相同时，c 越大算法花费的时间越多，这是因为 c 越大算法在它贪心阶段花费的步骤越多. 值得注意的是，当 c 越来越大时，时间曲线的斜率也越来越大. 我们还可以看出 HPG 算法(即 $c=0.5$)比 KK 算法(即 $c=1$)高效很多. 随着 k 值的增加，这种时间上的优势更加明显，这对大型社会网络来说将是一种非常重要的贡献.

5 总 结

本文中，基于 LT 模型影响力积累的特性，综合度中心性和贪心算法的优点，提出了一个影响最大化算法的框架并给出新的混合式影响最大化算法 HPG. 在 6 个真实的具有不同特性的数据集上进行了详尽的实验. 实验结果表明，我们提出的 HPG 算法相比 KK 算法在影响范围上有很好的提升. 时间复杂度上 HPG 算法也很直观地优于 KK 算法.

尽管 HPG 算法影响范围和时间复杂度都比较好，但依然还有许多值得改进与进一步研究的地方. 例如如何将已有的社团信息整合到影响最大化问题中；现有的模型都没有涉及时间的概念，但是现实的信息扩散都是需要时间的，如何扩展已有的模型使之包括时间因素，或者创建新的模型等等.

致 谢 感谢本文审稿专家和编辑所提出的宝贵意见和建议！

参 考 文 献

[1] Richardson M, Domingos P, Glance N. Mining knowledge-sharing sites for Viral Marketing//Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Canada, 2002: 61-70

[2] Kempe D, Kleinberg J, Tardos E. Maximizing the spread of influence in a social network//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2003: 137-146

[3] Chen Wei, Wang Ya-Jun, Yang Si-Yu. Efficient influence maximization in social networks//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 199-208

[4] Young H P. The diffusion of innovations in social networks//Blume L, Durlauf S. The Economy as a Complex System III. USA: Oxford University Press, 2003: 1-19

[5] Watts D J. A simple model of global cascades on random networks. National Academy of Sciences, 2002, 99(9): 5766-5571

[6] Goldenberg J, Libai B, Muller E. Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing Letters, 2001, 12(3): 211-223

[7] Lin Ju-Ren. Social Network Analysis: Theory, Methods and Applications. Beijing: Beijing Normal University Press, 2009 (in Chinese)
(林聚任. 社会网络分析：理论、方法与应用. 北京：北京师范大学出版社，2009)

[8] Estevez Pablo A, Vera Pablo, Saito Kazumi. Selecting the most influential nodes in social networks//Proceedings of the International Joint Conference on Neural Networks. Orlando, Florida, USA, 2007: 2397-2402

[9] Suri N Rama, Narahari Y. Determining the top-k nodes in social networks using the shapely value (Short Paper)//Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems. Estoril, Portugal, 2008: 1509-1512

[10] Wang Yi-Tong, Feng Xiao-Jun. A potential-based node

selection strategy for influence maximization in a social network//Proceedings of the 5th International Conference on Advanced Data Mining and Applications. Beijing, China, 2009: 350-361

[11] Leskovec J, Huttenlocher D, Kleinberg J. Signed networks in social media//Proceedings of the 28th International Conference on Human Factors in Computing Systems. Atlanta, USA, 2010: 1361-1370

[12] Sun Shi-Wei, Ling Lun-Jiang, Zhang Nan, Li Guo-Jie, Chen Run-Sheng. Topological structure analysis of the protein-protein interaction network in budding yeast. Nucleic Acids Research, 2003, 31(9): 2443-2450

TIAN Jia-Tang, born in 1987, M. S. candidate. His research interests include social networks, Web mining and Web information retrieval.

WANG Yi-Tong, born in 1973, Ph. D., associate professor. Her research interests include database, Web mining, data mining and Web information retrieval.

FENG Xiao-Jun, born in 1984, M. S. candidate. His research interests include social networks, Web mining.

Background

Social network analysis (SNA) has been developed as an independent research domain for many years. Social networking sites such as MySpace, Facebook and Flickr enable people and organizations to contact each other. Today, hundreds of millions of internet users are using online social networking sites to discover new "friends", share user-created content (photos, videos, blogs etc.) as well as diffuse news or ideas. So, the Web is now considered more like a giant "Social Network" or "Social Web". Social networks also serve as an important medium for the diffusion of information, ideas or innovations. With the development of Web technology and the advent of Web 2.0, the research of SNA is facing great challenges. The NFS project (61033010) was proposed under such situation and our main purpose is to study "information fusion and its corresponding knowledge service under the WEB environment". In this project, we will focus on various information diffusions such as integrating different media data, identifying relationships between different data objects, extracting social communities and investigating how information diffuses among people, organizations and communities in large online social networks.

The problem we study in this paper is about information diffusion; especially about how to maximize diffusion range under certain diffusion model in large online social networks. This problem is very important and meaningful and it has many real applications such as social security, pre-warning of public opinion and marketing by identifying "most influential rioter, opinion leaders, potential buyers" etc. Traditional marketing is being replaced by new strategies in which the product (ideas, innovations) is turned into "epidemics". Many products promote rather easily in a social system through a domino effect. Comparing with traditional social network, network structure and diffusion behavior become even more complex in online social networks, e.g. relationships could be positive (friends) and negative (foes), diffusion model might be different. Influence maximization is the problem of finding a small subset of nodes (seed nodes) in a social network that could maximize the spread of influence and it has attracted much attention and some works have been proposed in recent years. The optimization problem of selecting the most influential nodes is proved to be NP-hard and the running time is the main bottleneck for large social networks. In this paper, we mainly attempt to tackle the problem of influence maximization in large online social networks in terms of both efficiency and effectiveness. Most prior work on influence maximization is mainly focused on IC (Independent cascading) diffusion model due to its sub-modularity property, and few works are on LT (Linear Threshold) model. However, LT model is the basis of "word-of-mouth" effect for online marketing. In this paper, we investigate LT model and observe "accumulation" feature of LT model. According to this observation, we propose new algorithms for influence maximization under LT model. We conduct thorough experiments on different real datasets (including sighed social networks) and experimental results show that the proposed algorithm obtains a much better performance than the greedy algorithm in terms of both influence spread and running time while few works so far has done this. Moreover it is extended to the signed social networks and has a good performance too. We have been working on Web mining and social network for a long time and our work is mainly focused on: community extraction, spam detection, influential nodes identification, blog experts identification etc. and has published in international conferences DASFAA11, ADMA09, ADMA10, HT09.

This work is supported by National Natural Science Foundation of China under grant No. 61033010. The main result of this paper is to solve information diffusion and influence maximization in the online social network environment.