基于LT模型的个性化关键传播用户挖掘

郭　静[1),2)]　张　鹏[2)]　方滨兴[1)]　周　川[2)]　曹亚男[2)]　郭　莉[2)]

[1)](北京邮电大学计算机学院　北京　100876)

[2)](中国科学院信息工程研究所　北京　100093)

摘　要　文中针对社交网络中特定用户展开个性化关键传播用户挖掘研究，目标是在线性阈值传播模型的基础上，挖掘出能够最大程度影响网络中特定用户的节点集合.尽管在社交网络影响最大化问题方面已存在相关工作，但该文工作偏重于针对网络中的特定用户展开，该问题的解决将有助于企业有效的进行个性化产品营销.为此，文中提出一种基于LT模型的个性化关键传播用户挖掘问题的解决框架.首先，在线性阈值模型的基本传播机制下，提出一个随机函数来模拟基于LT模型的个性化关键传播用户挖掘问题的目标函数，该随机函数具有较小方差的理论保证；然后，提出一个有效的求解算法从网络中挖掘针对特定用户的关键传播节点集合，理论证明该算法具有$(1-1/e)$的近似精度保证.实验使用真实的社交网络数据验证了算法的有效性.

关键词　社交网络；线性阈值模型；特定用户；关键传播用户；社会计算

中图法分类号　TP399　　**DOI号**　10.3724/SP.J.1016.2014.00809

Personalized Key Propagating Users Mining Based on LT Model

GUO Jing[1),2)]　ZHANG Peng[2)]　FANG Bin-Xing[1)]　ZHOU Chuan[2)]　CAO Ya-Nan[2)]　GUO Li[2)]

[1)](*School of Computer Science, Beijing University of Posts and Telecommunications, Beijing*　100876)

[2)](*Institute of Information Engineering, Chinese Academy of Sciences, Beijing*　100093)

Abstract　In this paper, we study a new problem of personalized key propagating users mining based on LT model. This problem aims to mine out the most influential nodes under the basic linear threshold model for a target user in the social network, which will in favor of better product marketing. While different from existing research on influence maximization area focuses on identifying a seed set to maximize the influence spread over the entire network, this problem focuses on identifying a seed set which can maximize the influence spread to a given target user. For this purpose, we present a solution framework for the proposed target-based influence maximization problem. Specially, we first provide a random function to randomly simulate the objective function of our problem with low variance guarantee. Then, we present an efficient algorithm to identify influential nodes for a given target user with approximation guarantee $(1-1/e)$. Experimental results on several real-world social networks validate the performance of the proposed algorithms.

Keywords　social networks; linear threshold model; target user; key propagating users; social computing

收稿日期：2013-06-19；最终修改稿收到日期：2014-01-24.本课题得到中国科学院战略性先导科技专项“面向感知中国的新一代信息技术研究”(XDA06030200)、中国科学院信息工程研究所科研项目“网络信息源发现和信息传播关键技术研究”(Y3Z0062101)、青年科学基金项目“多数据流关联挖掘的模型研究”(61003167)资助.**郭　静**，女，1982年生，博士研究生，主要研究方向为社会计算、数据挖掘.E-mail：guojing@nelmail.iie.ac.cn.**张　鹏**，男，1982年生，博士，副研究员，主要研究方向为社会计算、数据挖掘.**方滨兴**，男，1960年生，博士，教授，博士生导师，中国工程院院士，主要研究领域为网络分析、信息安全和网络安全等.**周　川**，男，1984年生，博士，主要研究方向为社会计算、概率统计.**曹亚男**，女，1985年生，博士，助理研究员，主要研究方向为Web语义抽取、文本挖掘.**郭　莉**，女，1969年生，教授，主要研究领域为网络信息安全.

1 引 言

随着互联网技术的迅速发展,越来越多的用户参与到内容丰富的社交网络应用中,使社交网络发展成一种普及、有效的日常交流平台.在社交网络中个体成员的行为和选择往往能够影响和带动其周边节点,例如,用户在接受一个新服务或新产品后往往能影响自己的朋友或者同事对该事物的观点.这种在社交网络中口口相传的口碑传播作用,对商品推广、市场营销、意见传播等应用具有重要的现实意义,吸引了大量研究者展开研究.

社交网络影响最大化问题虽然源自于市场营销领域,但它却是影响力传播领域的基本问题.其主要研究内容就是在社交网络中挖掘能够最大程度影响整个网络的种子节点集合.早在 2001 年 Domingos 和 Richardson[1-2]就针对信息在网络中的传播情况,提出了一个基础性的问题,即如何选择网络中最有影响力的 k 个营销节点,使得该产品的最终购买人数达到最多?随后 Kempe 等人[3]进一步将此问题定义为一个离散最优的数学问题,并利用目标函数的子模特性,设计一种具有近似精度理论保证的贪婪求解算法;即每一步都选择当前最有影响力的节点加入初始节点集合.

尽管在社交网络影响最大化问题方面已存在大量相关工作,但是这些工作将社交网络看成一个整体,关注于挖掘能最大程度影响整个网络的节点集合,以及在提高影响范围的同时,尽量降低算法的时间复杂度,鲜有针对网络中特定的用户,挖掘能最大程度影响该用户的节点集合,或者说,挖掘特定用户的关键传播节点.

然而,现实中,企业在进行个性化产品营销时,往往需要对网络中特定用户展开分析,所以更关注于挖掘影响特定用户的关键传播节点.例如:某大型网上商城借助社交网络对其重要的用户 A 开展产品营销时,希望在有限预算的前提下,从网络中选择若干节点最大程度地影响用户 A,从而将产品推销出去.因此,解决社交网络中特定用户的影响最大化问题有助于更加有效的进行产品营销,具有一定实际意义.然而,一方面,现有的影响最大化方法仅仅针对整个网络而设计,所以无法满足此类应用的需求;另一方面,由于推销预算限制,所以通常无法发动特定用户的所有邻居来实现影响最大化.虽然从邻居中选择若干用户能够对目标用户有所影响,但是这种方法无法保证精度.从信息传播的角度来看,社交网络中的特定用户主要受其邻居影响,所以在有限预算的前提下,如果能够挖掘出一个或者较少几个节点,使其能够尽可能多地去影响目标用户的邻居集合,那么就可以在满足预算的要求下,达到所需的影响效果.

为此,本文提出一种挖掘特定用户关键传播节点的解决框架.首先,本文在线性阈值模型的传播机制下,提出一个随机函数来模拟本文问题的目标函数;该随机函数将网络中其他节点对目标节点的影响力计算分为两段,一段计算其他节点对目标节点邻居的激活情况,另一段计算目标节点邻居节点对目标节点的影响力传播权重;理论证明,该随机函数具有较小方差.然后,本文提出了一个有效的算法来挖掘能够最大程度影响网络中特定用户的节点集合;该算法采用贪婪策略,依次从网络中选择影响目标节点的 k 个关键传播节点;理论证明,该算法具有 $(1-1/e)$ 的近似求解精度.实验使用真实的社交网络数据分析了算法的性能.

本文第 2 节介绍相关工作;第 3 节定义个性化关键传播节点挖掘问题;第 4 节提出有效的求解算法;第 5 节实验分析所提算法的性能;在第 6 节对全文进行总结.

2 相关工作

社交网络影响最大化问题就是如何选取 k 个初始节点进行传播,从而最大程度地影响整个网络的问题.早期的研究工作是由 Domingos 和 Richardson 开展的,他们[1-2]针对信息在网络中的传播情况,提出了一个基础性的算法问题.随后 Kempe 等人[3]进一步将此问题定义为一个离散最优的数学问题,并根据目标函数的子模特性,设计了一种具有精度保证的近似求解算法.

由于社交网络影响最大化问题在市场营销、信息传播等领域的应用背景,使其得到了广泛的关注.目前,已有相关研究工作主要解决社交网络影响最大化的效率和精度保证问题.其中,研究[4-10]关注于具有精度保证的近似求解.研究者主要利用贪婪算法或者它们的扩展形式来对问题进行近似求解,以便在保证计算精度的前提下,尽可能地提高问题求解的效率.例如,Leskovec 等人[4-5]以及 Goyal 等人[5]根据问题目标函数的子模特性来减少计算影响力的评估次数;Kimura 等人[7]利用键渗透理论以及

图论知识来评估所有量值；Chen 等人[8]首先从原始社交网络图中移除所有不参与传播的边，然后在较小的生成子图上进行影响力传播的计算；Wang 等人[9]首先对社交网络进行社区划分，然后在社区内挖掘出最有影响力的节点；Barbieri 等人[10]将主题模型引入传播计算，提出主题驱动的传播模型，并在此模型基础上展开影响最大化研究.

随着社交网络规模的不断增大，研究[8,11-14]开始关注于大规模社交网络影响最大化的在线计算问题. 这些工作主要使用启发式算法来保证问题求解的效率和可扩展性，但它们在理论上不能保证计算精度. 例如，Chen 等人[8]提出了基于度折扣的启发式算法，该算法将计算速度提升了百万倍；Kimura 和 Saito[11]提出了基于最短路径的传播模型，并以此来快速估算影响力的传播情况；随后，Chen 等人[12-13]分别在独立级联模型和线性阈值模型的基础上，研究了构建有效局部传播区域的方法和快速估算的影响力传播的方法；Goyal 等人[14]利用邻居节点间的简单路径对影响力传播进行快速估算.

虽然上述工作能够帮助企业快速挖掘出最大程度影响整个网络的节点集合，但是企业在进行个性化产品营销时，往往需要对网络中特定用户展开分析，所以更关注于挖掘影响特定用户的关键传播节点. 即，针对一个给定的目标用户，从网络中挖掘能够最大程度影响该用户的节点. 虽然 Guo 等人[15]研究了个性化影响最大化问题，但该工作假设用户间相互影响力是独立的，没有考虑影响力所具有的累积效应，其适用范围有限. 为此，本文在线性阈值传播模型的基础上，挖掘能够最大程度影响网络中特定用户的节点集合. 另外有一些研究工作对社交网络的影响力传播开展研究，例如 Anagnostopoulos 等人[16]研究如何区分影响力和相关性；Cui 等人[17]研究如何根据重要的节点去预测网络突现的行为；Goyal 等人[18]研究如何计算用户间影响力的传播概率；Cui 等人[19]针对给定的物项，研究如何计算用户邻居点击该物项的概率；但这些工作不涉及特定用户的影响力最大化问题.

3　基于 LT 模型的个性化关键传播节点挖掘问题

本节在线性阈值模型（Linear Threshold Model，简称为 LT 模型）的传播框架下，对个性化关键传播节点挖掘问题（即个性化影响最大化问题）进行定义和介绍. 其常用数学符号表示如表 1 所示.

表 1　对应数学符号表示

符号	含义
G	社交网络的结构图
X	G 中基于 LT 模型所得的随机激活结果
w	G 中特定用户
$G_{\overline{w}}$	去除 G 中 w 后所得图
Y	基于 $G_{\overline{w}}$ 的随机激活结果
C_w	w 的邻居节点集合
$F(w,x)$	在激活结果 x 中去除 w 节点
$1_{\{w\in X\}}$	用于指示 $w\in X$ 的指示函数
Ω	所有可能激活结果的样本空间
$\mathbb{P}(X)$	X 在样本空间 Ω 中的概率
D	蒙特卡洛随机模拟的模拟次数
U	初始节点集合
δ_U	被 U 激活的节点个数
$R_w(U)$	U 到特定用户 w 的影响力
k	初始节点个数
p_{uv}	基于 LT 模型，沿有向边 $\boldsymbol{uv}$ 的激活权重

3.1　预备知识

本节主要介绍个性化影响最大化问题的相关基础知识. 我们用 $G=(V,E)$ 来表示社交网络，其中，点集 V 和边集 E 分别对应社交网络中的用户及其关系. 根据用户行为的发生情况，每个节点通常存在两种状态：激活状态（active）或者未激活状态（inactive），且同一时刻只能处于这两种状态中的一种. 在社交网络中，处于激活状态的节点能够影响处于未激活状态的节点. 如果一个节点周围有越来越多的邻居节点变成激活状态，则该节点被激活的可能性也就越来越大.

为了刻画这种传播行为，线性阈值模型作为社交网络的基本传播模型之一，主要关注于影响力传播过程中的阈值行为，即影响力在传递过程中所具有的累积效应. 当一个激活节点尝试去激活它的非激活邻居节点时，尝试失败节点在该次激活过程的影响力会被累积，并对后面其他节点对该非激活节点的激活行为产生贡献. 例如，社交网络中的个体成员 C 的朋友向 C 推荐新产品时，朋友推荐的影响力将被累积，当达到一定程度（阈值）时，个体成员 C 将被成功激活购买该新产品.

在本文中，种子节点集合 U 对特定用户 w 的影响力记为 $R_w(U)$；在给定初始种子节点集合 U 的前提下，$R_w(U)$ 的实质就是特定用户 w 被成功激活的程度，其形式化表达如式(1)所示.

$$R_w(U):=\mathbb{P}^U(w\in X) \tag{1}$$

其中，$\mathbb{P}^U$ 表示节点集合 U 的传播概率，X 表示一个随机激活结果，包括激活路径以及被种子节点激活的节点集合，$w\in X$ 则表示特定用户 w 属于该激活结果 X.

3.2 目标函数模拟

与社交网络影响最大化问题的目标不同，本研究的目标就是要在社交网络中挖掘能够最大程度影响特定用户的节点集合. 在社交网络影响最大化问题中，问题的目标函数可由式(2)来表示.

$$U^{*}=\underset{|U|=k,U\subseteq V}{\arg\max}R(U) \tag{2}$$

$$\text{s.t. } U^{*}=\{u_1,u_2,\cdots,u_k\}\subseteq V$$

其中，U 是社交网络中任意包含 k 个节点的集合，U^{*} 是所有可能集合中，能够满足 $R(U)$ 最大的节点集合，$R(U)$ 是整个网络被初始种子节点集合 U 激活的程度.

本问题目标的形式化表达如式(3)所示.

$$U^{*}=\underset{|U|=k,U\subseteq V\backslash\{w\}}{\arg\max}R_w(U) \tag{3}$$

$$\text{s.t. } U^{*}=\{u_1,u_2,\cdots,u_k\}\subseteq V\backslash\{w\}$$

其中，U 是社交网络中任意包含 k 个节点(除 w 节点以外)的集合；U^{*} 是所有可能集合中，能够满足 $R_w(U)$ 最大的节点集合；$V\backslash\{w\}$ 表示节点集 V 中去除 w 点；$R_w(U)$ 表示特定用户 w 被初始种子节点集合 U 激活的程度.

从式(3)可以看出，对社交网络图 G 中任何一个节点集合 $U(U\subseteq V\backslash\{w\},|U|=k)$ 来说，都满足 $R_w(U^{*})\geq R_w(U)$，所以目标函数关键在于 $R_w(U)$ 的计算.

由于影响力在传播路径及时间上具有不确定性，所以精确计算 $R_w(U)$ 属于 #P-hard 难题. 目前，通用的方法是利用蒙特卡洛(Monte Carlo)模拟法去近似估算 $R_w(U)$. 于是，结合式(1)可以获得 $R_w(U)$ 的随机统计量，具体如式(4)所示.

$$R_w(U)=\mathbb{E}^{U}(1_{\{w\in X\}}) \tag{4}$$

其中，$1_{\{w\in X\}}$ 用于指示特定用户 w 是否属于该激活结果 X. 如果 $1_{\{w\in X\}}=1$，则特定用户 w 属于该激活结果；如果 $1_{\{w\in X\}}=0$，则特定用户 w 不属于该激活结果.

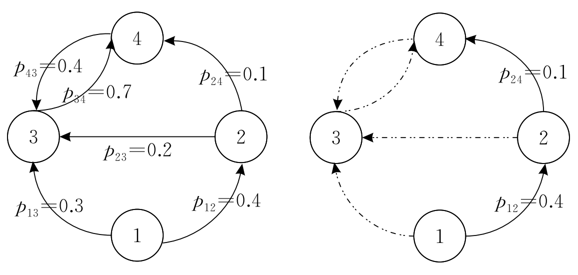
在基于线性阈值模型的传播机制下，本文用 Ω^{U} 表示节点集合 U 在图 G 上所有激活结果的样本空间，那么，可将式(4)进一步展开获得式(5).

$$R_w(U)=\sum_{x\in\Omega^{U}}\mathbb{P}(X=x)\times 1_{\{w\in X\}} \tag{5}$$

其中，$\mathbb{P}(X=x)$ 表示 $X=x$ 在 Ω^{U} 中的概率，同时满足 $\sum_{x\in\Omega^{U}}\mathbb{P}(X=x)=1$.

从式(5)可以看出，$1_{\{w\in X\}}$ 是 $R_w(U)$ 的一个无偏差统计量，能够对目标函数进行模拟. 然而，使用 $1_{\{w\in X\}}$ 对目标函数进行模拟是否足够好？能不能找到更好的随机函数？

为解答上述问题，下面将以图 1(a)所示的网络结构为例来展开分析和说明. 其中，节点与节点之间的激活权重直接标记在图 1(a)中的边上. 在线性阈值模型的传播框架下，对于每一个节点，其邻居节点集合对其激活权重和小于等于 1.



(a) 图G (b) 生成图$G_{\overline{w}}$

图 1 网络结构举例

例 1. 假定图 1(a)中的特定用户为 $w=$③，初始选定的节点集合为 $U=\{$①$\}$，则计算图 1(a)中的影响力 $R_w(U)$，可通过下列步骤获得.

首先，针对激活结果 $x=$(①→②→④)，给出 $\mathbb{P}(X=x)$ 的具体计算. 由于线性阈值模型中历史激活权重对激活过程的累积作用，所以在样本空间 Ω^{U} 中，激活结果 $x=$(①→②→④)发生的概率计算如下.

$$\begin{aligned}\mathbb{P}(X=x)&=\mathbb{P}(X=(①\rightarrow②\rightarrow④))\\&=p_{1,2}\times p_{2,4}\times\Big(1-\sum_{i=1,2,4}p_{i,3}\Big)\\&=0.4\times0.1\times(1-0.3-0.2-0.4)\\&=0.004.\end{aligned}$$

于是，按照同样的计算方法，基于节点集合 $U=\{$①$\}$，计算出所有可能激活结果的概率统计情况如表 2 所示.

表 2 图 G 每个激活结果的 $R_w(U)$

序号	x	$\mathbb{P}(X=x)$	$F(w,x)$
1	①	0.420	①
2	①→②	0.180	①→②
3	①→②→④	0.004	①→②→④
4	①→②→④→③	0.016	①→②→④
5	①→②→③	0.016	①→②
6	①→②→③→④	0.056	①→②
7	①→②→④ ↓ ③	0.008	①→②→④
8	①→③	0.054	①
9	①→③→④	0.126	①
10	①→② ↓ ③	0.024	①→②
11	①→②→④ ↘ ↗ ③	0.096	①→②→④
合计	—	1	—

根据表 2 中所有包含 $w\in x$ 的概率 $\mathbb{P}(X=x)$，按照式(5)可计算获得 $R_w(U)$ 的取值.

$$R_w(U)=\sum_{w\in X}\mathbb{P}(X=x)=\sum_{x=x_4,\cdots,x_{13}}\mathbb{P}(X=x)$$
$$=0.396 \tag{6}$$

那么，如果我们将 w 从激活结果 x 中移除，记为 $F(w,x)$，则对表 2 中第 4 列的相同项进行合并计算可得下列等式.

$$\mathbb{P}(F(w,x)=①)=0.6,$$
$$\mathbb{P}(F(w,x)=①\rightarrow②)=0.36,$$
$$\mathbb{P}(F(w,x)=①\rightarrow②\rightarrow④)=0.04.$$

从上述等式可以看出，$F(w,x)$ 与基于生成图 $G_{\overline{w}}$（如图 1(b)）的随机激活结果 Y 同概率分布.

由于任何节点对特定用户 w 的激活作用都绕不开其邻居节点，且每一个邻居对特定用户 w 都存在一定的激活权重，所以本文将图 G 中节点集合 U 到特定用户 w 的影响力计算过程分为两步：$U\rightarrow C_w$ 和 $C_w\rightarrow w$. 于是在线性阈值模型的传播机制下，我们引入生成图 $G_{\overline{w}}$ 的随机激活结果 Y，获得下列计算结果.

$$\sum_y\mathbb{P}(Y=y)\times\sum_{v\in Y}p_{vw}=$$
$$0.6\times0.3+0.36\times(0.2+0.3)+$$
$$0.04\times(0.2+0.3+0.4)=0.396 \tag{7}$$

对比式(6)和式(7)，发现二者计算结果相同. 实际上，这个结果的获得并非偶然，它是有理论依据的，下面将给出理论证明.

定理 1. $\sum_{v\in Y}p_{vw}$ 是 $R_w(U)$ 估计的无偏差统计量，且其方差小于 $1_{\{w\in X\}}$. 即满足下列公式.

$$\mathbb{E}^U\left[\sum_{v\in Y}p_{vw}\right]=R_w(U) \tag{8}$$

$$\mathbb{D}^U\left[\sum_{v\in Y}p_{vw}\right]<\mathbb{D}^U\left[1_{\{w\in X\}}\right] \tag{9}$$

证明. 首先，证明 $\sum_{v\in Y}p_{vw}$ 是 $R_w(U)$ 的一个无偏差统计量. 根据条件概率的 tower property[20]，有

$$\begin{aligned}R_w(U)&=\mathbb{P}^U(w\in X)\\&=\mathbb{E}^U\left[\mathbb{P}^U(w\in X\mid F(w,X))\right]\\&=\mathbb{E}^U\left[\sum_{v\in F(w,X)}p_{vw}\right]\\&=\mathbb{E}^U\left[\sum_{v\in Y}p_{vw}\right],\end{aligned}$$

所以 $\mathbb{E}^U\left[\sum_{v\in Y}p_{vw}\right]=R_w(U)$.

然后，证明 $\sum_{v\in Y}p_{vw}$ 要比 $1_{\{w\in X\}}$ 的方差更小. 根据方差计算的表达式，有

$$\begin{aligned}\mathbb{D}^U\left[\sum_{v\in Y}p_{vw}\right]&=\mathbb{E}^U\left[\left(\sum_{v\in Y}p_{vw}\right)^2\right]-[R_w(U)]^2\\&<\mathbb{E}^U\left[\sum_{v\in Y}p_{vw}\right]-[R_w(U)]^2\\&=R_w(U)-[R_w(U)]^2\\&=\mathbb{E}^U\left[1^2_{\{w\in X\}}\right]-[R_w(U)]^2\\&=\mathbb{D}^U\left[1_{\{w\in X\}}\right],\end{aligned}$$

所以 $\mathbb{D}^U\left[\sum_{v\in Y}p_{vw}\right]<\mathbb{D}^U\left[1_{\{w\in X\}}\right]$. 于是定理 1 得证.

证毕.

由定理 1 可知，$\sum_{v\in Y}p_{vw}$ 不仅是 $R_w(U)$ 估计的无偏差统计量，且其具有方差较小的理论保证. 为此，本文以式(10)来计算集合 U 对特定用户 w 的影响力大小.

$$R_w(U)=\mathbb{E}^U\left[\sum_{v\in Y}p_{vw}\right] \tag{10}$$

4　近似算法

本节将根据个性化影响最大化问题目标函数的特点，给出求解问题的近似算法.

由于个性化影响最大化问题的特例属于经典的顶点覆盖问题，所以该问题属于 NP 难问题. 因为本问题具有 NP 难问题的特点，所以不太可能采用蛮力算法去解决，特别是网络规模很大的时候. 为此，本文首先对目标函数的特点进行分析，以获得具有精度保证的结果. 通过分析我们获得 $R_w(U)$ 具有子模特性(sub-modularity)[21-22]（即定理 2).

定理 2. 目标函数式(3)具有子模特性.

证明. 首先证明 $R_w(U,Y)$ 具有子模特性. 假设 $A,B(A\subseteq B\subseteq V')$ 是生成图 $G_{\overline{w}}$ 中的节点集合；同时 $\nabla R_w(A_v,Y)=R_w(A\cup\{v\},Y)-R_w(A,Y)$ 代表作用在目标节点 w 上影响力的边际值，即节点集合 $A\cup\{v\}$ 对 w 的影响力与节点 v 对 w 的影响力差值. 显然，由于 $A\subseteq B$ 所以 $\nabla R_w(A_v,Y)$ 不少于 $\nabla R_w(B_v,Y)$，则 $R_w(A\cup\{v\},Y)-R_w(A,Y)\geq R_w(B\cup\{v\},Y)-R_w(B,Y)$ 该表达式满足子模特性的定义. 因此，$R_w(U,Y)$ 具有子模特性. 同时，由于子模特性在非负线性组合条件下具有封闭性，所以目标函数同样具有子模特性.　　证毕.

同时，根据文献[1]的描述，对一个非负单调且满足子模特性的函数 $r(\cdot)$ 来说，如果 S^* 是所有包含 k 个元素集合中能够使 $r(\cdot)$ 最大化的集合，S 是一个包含 k 个元素的集合，且满足集合中每个元素是通过最大边际效益依次选择出来的. 则二者满足

不等式 $r(S)\geq(1-1/e)r(S^*)$，即集合 S 能够提供 $(1-1/e)\approx63\%$ 的计算精度保证.

为此，本文利用目标函数的子模特性，采用贪婪策略来对问题进行求解，并提出具有理论精度保证为 $(1-1/e)\approx63\%$ 的近似求解算法，记为目标贪婪算法(Target-based Greedy Algorithm，简称为 TGA).

算法 1. 目标贪婪算法.

```
输入：(1) 正整数k；(2) 目标节点w；(3) 社交网络图G；
     (4) 蒙特卡洛模拟次数 D
输出：对 w 最有影响力的 k 个种子节点集合 U
U←∅；
Create G_w̄(G,w)；
WHILE |U|<k DO
    FOR each node v∈V_w̄\U DO
        R_w(U∪{v})=0；
        FOR i=1 to D DO
            RanCas Y(i,v,G_w̄)；
            R_w(U∪{v})+=R_w(U∪{v},Y)；
        END
        R_w(U∪{v})=R_w(U∪{v})/D；
    END
    U=U∪ argmax_{v∈V_w̄\U} R_w(U∪{v})；
END
Output：U
```

TGA 算法的输入由正整数 k、目标节点 w、社交网络图 G 和蒙特卡洛模拟次数 D 所组成. 算法的执行主要由下列两步组成：首先，根据特定用户产生影响力社区；然后，计算对特定用户最有影响力的节点，具体如上述算法描述所示.

在算法执行的第 1 步中，影响力社区 $G_{\overline{w}}$ 主要由图 G 和输入的目标节点 w 生成，其生成过程就是在图 G 中去除目标节点 w，并删除与目标节点 w 相关的有向边. 在算法执行的第 2 步中，TGA 算法将通过贪婪策略，每次选择一个具有最大边际效应的节点加入种子节点集合中. 即从非种子节点的集合中，依次选择一个节点使其对目标用户的影响力最大，具体如式(11)所示.

$$U^*=\bigcup_k \mathop{\arg\max}_{v\in V_{\overline{w}}\setminus U} R_w(U\cup\{v\})-R_w(U) \quad (11)$$

于是通过上述步骤就能获得具有(1−1/e)近似精度保证的节点集合.

5 实 验

本节以真实的社交网络数据集为仿真数据，使用 TGA 算法和 4 种基准算法在不同参数设置下进行对比实验，并对单个实验结果和统计实验结果进行分析，其详细描述如下.

5.1 实验设置

为保证实验结果的有效性，本文选取两组具有不同统计特性的真实社交网络数据来进行仿真. 其中，数据集 1 取自新浪微博，数据集 2 从 Leskovec Jure 的主页①获得，二者的统计特性描述如表 3 所示.

表 3 真实数据集描述

	数据集 1	数据集 2
数据来源	Weibo.com	Wikipedia.org
数据描述	用户间关注关系	用户间选举关系
节点信息	602	7115
边信息	17595	103689
平均度	29.2	14.6
平均聚类系数	0.3867	0.1400

为分析本算法的性能，实验以如下基准算法(baseline methods)为对比对象，利用文献[3]中的蒙特卡洛模拟方法去估计所有算法挖掘的结果(即选出的种子节点)对目标用户的影响力. 其中基准算法描述如下：

(1) Greedy 算法[3]. 该算法是经典的社交网络全局影响力最大化算法. 该算法以全局影响最大为目标函数，每次选取网络中最具影响力的一个节点，直到取满 k 个为止.

(2) PageRank 算法[23]. 该算法通过网络超链接关系来确定一个页面的等级，是一种由搜索引擎根据网页之间相互的超链接计算的技术. 一个高等级的页面可以使其他低等级页面的等级提升.

(3) LND 算法. 该算法从目标节点 w 的邻居节点集合中选择 k 个节点作为种子节点.

(4) LDAG 算法[13]. 该算法是快速的社交网络全局影响力最大化算法. 该算法将影响力计算限制在每个节点周围的有向无环图中，很大提高了算法的效率.

为全面分析算法的性能，本节将在不同参数设置下展开实验. 其中，主要的参数设置由关键传播节点个数 k 的取值和社交网络中用户间的影响力 p 所组成. k 的取值由 1～10，用户 v 到用户 u 的影响传播权重 p_{vu} 由 $p_{vu}=1/d_{in}(u)$ 计算，$d_{in}(u)$ 表示用户 u 的入度.

5.2 算法的性能分析

为分析算法的性能，实验使用上述算法分别对数据集 1 和数据集 2 中给定目标的关键传播节点进

① http://snap.stanford.edu/data/index.html

行挖掘，获得下列实验结果图. 其中，图 2～图 5 是不同方法对数据集 1 中两个目标节点挖掘的执行时间和挖掘结果对目标节点的影响力. 在此次实验中，被选目标节点的 ID 分别为 11 和 64.

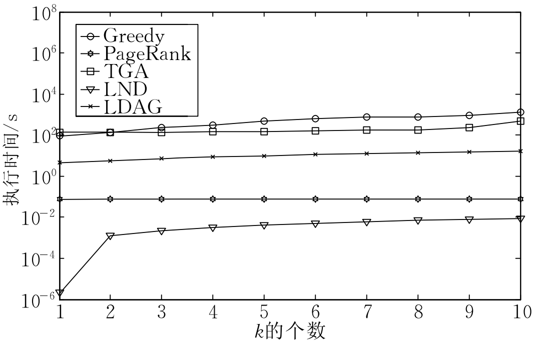


图 2　不同算法针对节点 ID=11 挖掘分析的执行时间

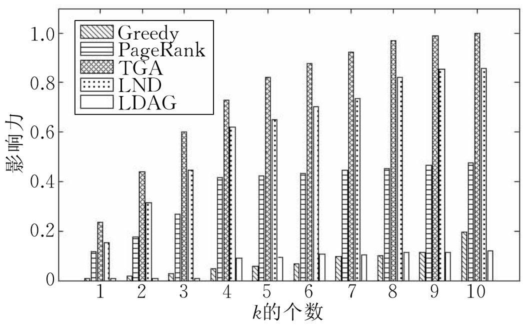


图 3　不同算法针对节点 ID=11 挖掘获得的影响力

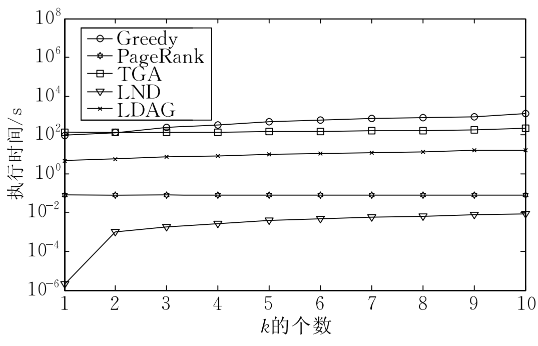


图 4　不同算法针对节点 ID=64 挖掘分析的执行时间

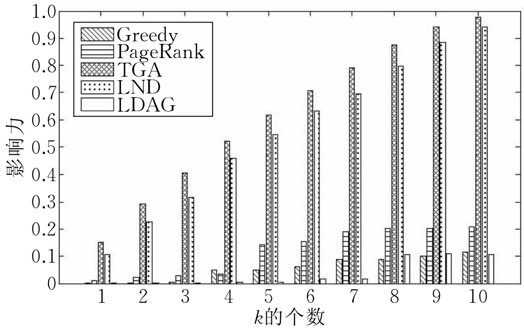


图 5　不同算法针对节点 ID=64 挖掘获得的影响力

从图 2～图 5 中可以看出：(1) TGA 算法在执行时间上的性能不如 PageRank 算法、LND 算法和 LDAG 算法，但在总体上要好于 Greedy 算法. 这是因为 TGA 算法和 Greedy 算法为获得算法精度，采用蒙特卡洛模拟进行模拟传播运算，因此在时间上损耗较大；(2) TGA 算法在针对给定节点（如 ID=11 或者 ID=64 节点）的局部影响力优于其他 4 种算法. 这是因为 Greedy 算法和 PageRank 算法、LDAG 算法都是针对网络中全局目标设置，其求解目标与本文目标有所差异，所以挖掘结果不能使特定用户的影响最大化. 但是对 LND 算法而言，由于特定用户的邻居对该节点具有影响力，所以从其的邻居节点选取，具有相对较好的效果，但是这种方法忽略了其他节点对特定用户的间接影响，其挖掘效果不如 TGA 算法. 因此，TGA 算法挖掘出的关键节点具有影响力性能上的优势.

图 6 和图 7 是不同方法对数据集 2 中两个目标节点挖掘所获结果对目标节点的影响力. 在此次实验中，被选目标节点的 ID 分别为 84 和 122.

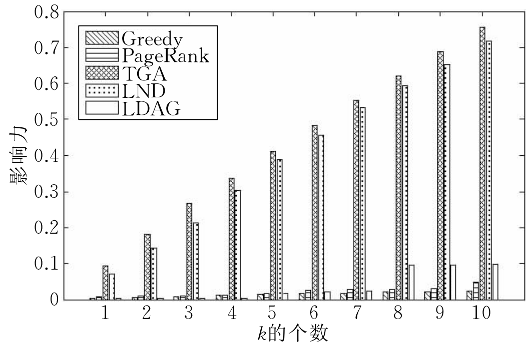


图 6　不同算法针对节点 ID=84 挖掘获得的影响力

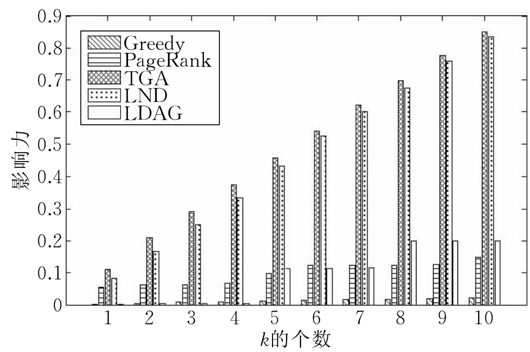


图 7　不同算法针对节点 ID=122 挖掘获得的影响力

从图 6 和图 7 中可以看出：TGA 算法在数据集 2 上挖掘的效果同样优于其他 4 种算法. 同时，由于数据集 1 和数据集 2 拥有不同的统计特性（即数据集 2 比数据集 1 拥有更大的网络规模，且拥有不同的平均度和平均聚类系数），所以可以说 TGA 算法能够对不同的社交网络挖掘获得良好效果.

为进一步研究 TGA 算法的性能，实验以数据集 1 为仿真对象，利用挖掘结果的统计信息来展开分析. 在实验中，我们随机选取网络中 5% 的节点为目标集合，利用不同算法对其关键传播节点进行挖掘. 如表 3 所示，数据集 1 共有 602 个节点，所以此次实验总共针对 31(602×5%≈31) 个目标进行挖掘，实验结果将按照平均值进行统计，其结果如图 8 和表 4 所示.

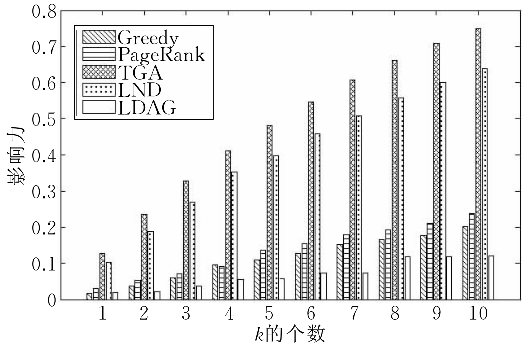


图 8 不同方法对网络中 5% 节点挖掘所获得的平均影响力

表 4 不同方法获得的平均影响力取值

k	PageRank 算法	Greedy 算法	TGA 算法	LND 算法	LDAG 算法
1	0.0176	0.0307	0.1270	0.1031	0.0185
2	0.0363	0.0531	0.2358	0.1883	0.0225
3	0.0601	0.0704	0.3298	0.2704	0.0363
4	0.0956	0.0922	0.4106	0.3531	0.0547
5	0.1097	0.1363	0.4813	0.3971	0.0572
6	0.1282	0.1542	0.5467	0.4565	0.0730
7	0.1519	0.1795	0.6066	0.5080	0.0730
8	0.1657	0.1931	0.6610	0.5569	0.1184
9	0.1783	0.2107	0.7082	0.6011	0.1184
10	0.2023	0.2378	0.7489	0.6376	0.1213

从图 8 和表 4 可以看出，TGA 算法在不同 k 值下均获得了良好的挖掘结果，其挖掘所获种子节点对目标用户的影响力均高于其他算法. 虽然与其他基准算法相比，LND 算法取得了不错的挖掘效果，但是其统计结果仍不如 TGA 算法. TGA 算法挖掘所获结果对目标用户的影响力比 LND 算法的影响力平均高 20%. 因此，上述分析表明 TGA 算法不仅能够对不同目标用户的关键传播节点进行挖掘获得良好效果，而且说明 TGA 算法具有一定的稳健性.

综上所述，TGA 算法能够有效解决个性化关键传播节点的挖掘问题.

6 结 论

本文基于 LT 模型研究了社交网络中个性化关键传播节点的挖掘问题，提出用一个具有较低方差保证的随机函数来模拟该问题的目标函数. 在保证求解精度的条件下，根据问题目标函数的子模特性，设计一个有效的求解算法来挖掘个性化关键传播节点. 实验以真实的社交网络数据验证了算法的有效性.

参 考 文 献

[1] Domingos P, Richardson M. Mining the network value of customers//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2001: 57-66

[2] Richardson M, Domingos P. Mining knowledge-sharing sites for viral marketing//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Edmonton, Canada, 2002: 61-70

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

[4] Leskovec J, Krause A, Guestrin C, et al. Cost-effective outbreak detection in networks//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, USA, 2007: 420-429

[5] Goyal A, Lu Wei, Lakshmanan L V S. CELF++: Optimizing the greedy algorithm for influence maximization in social networks//Proceedings of the 20th International Conference Companion on World Wide Web. Hyderabad, India, 2011: 47-48

[6] Leskovec J. Dynamics of Large Networks. Pittsburgh, USA: Addison-Wesley Publishing Company, 2008

[7] Kimura M, Saito K, Nakano R. Extracting influential nodes for information diffusion on a social network//Proceedings of the 22nd AAAI Conference on Artificial Intelligence and the 19th Innovative Applications of Artificial Intelligence Conference. Vancouver, Canada, 2007: 1371-1376

[8] Chen Wei, Wang Yajun, Yang Siyu. Efficient influence maximization in social networks//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 199-208

[9] Wang Yu, Cong Gao, Song Guojie, Xie Kunqing. Community-based greedy algorithm for mining top-k influential nodes in mobile social networks//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 1039-1048

[10] Barbieri N, Bonchi F, Manco G. Topic-aware social influence propagation models//Proceedings of the 12th IEEE International Conference on Data Mining. Brussels, Belgium, 2012: 81-90

[11] Kimura M, Saito K. Tractable models for information diffusion in social networks//Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases. Berlin, Germany, 2006: 259-271

[12] Chen Wei, Wang Chi, Wang Yajun. Scalable influence maximization for prevalent viral marketing in large-scale social networks//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 1029-1038

[13] Chen Wei, Yuan Yifei, Zhang Li. Scalable influence maximization in social networks under the linear threshold model//Proceedings of the 10th IEEE International Conference on Data Mining. Sydney, Australia, 2010: 88-97

[14] Goyal A, Lu Wei, Lakshmanan L V S. SIMPATH: An efficient algorithm for influence maximization under the linear threshold model//Proceedings of the 11th IEEE International Conference on Data Mining. Vancouver, Canada, 2011: 211-220

[15] Guo J, Zhang P, Zhou C, et al. Personalized influence maximization on social networks//Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management. San Francisco, USA, 2013: 199-208

[16] Anagnostopoulos A, Kumar R, Mahdian M. Influence and correlation in social networks//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA, 2008: 7-15

[17] Cui Peng, Yu Linyun, Wang Fei, et al. Cascading outbreak prediction in networks: A data-driven approach//Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago, USA, 2013: 901-909

[18] Goyal A, Bonchi F, Lakshmanan L V S. Learning influence probabilities in social networks//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York, USA, 2010: 241-250

[19] Cui Peng, Wang Fei, Liu Shaowei, et al. Who should share what? Item-level social influence prediction for users and posts ranking//Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China, 2011: 185-194

[20] Williams D. Probability with Martingales. Cambridge, UK: Cambridge University Press, 1991

[21] Nemhauser G L, Wolsey L A, Fisher M L. An analysis of the approximations for maximizing sub-modular set functions. Mathematical Programming, 1978, 14(1): 265-294

[22] Mossel E, Roch S. On the submodularity of influence in social networks//Proceedings of the 39th Annual ACM Symposium on Theory of Computing. San Diego, USA, 2007: 128-134

[23] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab, Technologies Project: 422, 1998

GUO Jing, born in 1982, Ph. D. candidate. Her main research interests include social computing and data mining.

ZHANG Peng, born in 1982, Ph.D., associate researcher. His main research interests include social computing and data mining.

FANG Bin-Xing, born in 1960, Ph.D., professor, Ph. D. supervisor, member of Chinese Academy of Engineering. His main research interests include network analysis, information security and network security.

ZHOU Chuan, born in 1984, postdoctoral. His main research interests include social computing and probability statistic.

CAO Ya-Nan, born in 1985, Ph.D., assistant researcher. Her main research interests include Web semantic extraction and text mining.

GUO Li, born in 1969, professor. Her main research interests include network and information security.

Background

The research problem in our paper belongs to users' influence propagation analysis problem in social computing area. For now, the users' influence propagation analysis problem has caused wide academic attention, and most of them focus on how to quantify the influence between users in the social network, or how to influence the whole network to the most degree. In this paper, we mainly concern about mining out the key propagation nodes with the novel perspective of target node. This personalized influence propagation analysis problem has practical values in many applications, such as network security, recommendation and so on.

This research is supported by the Strategic Leading Science and Technology Projects of Chinese Academy of Sciences, "A New Generation of Information Technology

Research with Perception-China Oriented (XDA06030200, 2012.1-2016.12)".

With the strategic needs on "perception-China", the new opportunity bringing by the "ternary human-cyber-physical universe", driven by the construction of "Sea-Cloud Innovative and Experimental Environment", the information technology project aims to transform the research model of information technology, promote a batch of important innovations on information technology and science research, lay the technology foundations for the deeper integration and comprehensive utilization of information resources, physical resources, and social resources, guide the leap frog development on a new generation of information technology of strategic new industries, provide key technical support on the construction and development of a generous, intelligent, safety, servicing information society.

So far, this research group has launched a widely studied in related areas, such as social network analysis, data mining theory, and Web information collection, Web semantics analysis and so on. The number of published and accepted papers by this research group cumulatively reaches to more than sixty, and eight of them are SCI retrieval, over fifty of them are EI retrieval, which lays a significant foundation for the projects. The research results of this paper will provide basic technological and theoretical support.