在影响力最大化问题中寻找种子节点的替补节点

马 茜 马 军

(山东大学计算机科学与技术学院 济南 250101)

摘 要 在线社会网络的发展为市场营销提供了新的机遇和挑战.对于广告投放者来说,面临的问题是如何从一个有 n 个用户的社会网络中,选取 $k(0<k\ll n)$ 个有影响力的用户,称作种子节点,通过提供报酬、试用品等方式激活他们,让他们为产品做宣传,通过口口相传的方式使尽可能多的用户了解或者购买该产品.这个问题也被称作影响力最大化(Influence Maximization),简称 IM 问题.IM 问题的相关工作往往会默认所选出的 k 个种子节点均可被激活.而在实际应用中,受各种因素的影响,$t(0<t\leq k)$ 个种子节点很有可能无法激活.因此该文的研究问题是如何选取替补节点来代替不能被激活的种子节点,该文称该问题为在影响力最大化中寻找替补种子节点(Substitutes Discovery in Influence Maximization),简称 SDIM 问题.SDIM 问题的提出有利于解决营销中面临的实际问题,帮助广告投放者更顺利地完成营销目标.为此,该文首先给出了 SDIM 问题的形式化定义,并提出对该问题求解的优化函数.在证明了该问题属于 NP 难的基础上,说明了基于该文提出的优化函数得到的贪心算法具有精度保证.该文首先利用社会网络的无尺度特性,给出了保留网络中度较大的节点作为初始候选节点集的策略,在此基础上,分别提出了 3 个求解 SDIM 问题的算法:(1)找出恰好 t 个替补节点的全局静态贪心算法 GSG;(2)在选择种子节点的同时选取 $t'(t'\geq t)$ 个替补节点的预选式贪心算法 GIA,可防止新选的 t 个替补节点中仍存在不能被激活的节点;(3)可以改善 GSG 算法执行时间且不影响精度的全静态算法 AS.由于 GSG 运行时间过长,我们对其进行了 CELF 优化,在实验中我们称其为 GSG-CELF.实验结果表明:根据节点度减少候选节点数量的方法不会影响各算法的效果,却可以有效地减少运行时间;GSG-CELF 选出的替补节点的影响力很接近原始种子节点集的效果;GIA 具有更好的鲁棒性,同时传播效果也十分接近 GSG-CELF;AS 与 GSG-CELF 这类有 CELF 优化的贪心算法相比,运行时间是 GSG-CELF 的 10%~50%,且传播效果不受影响.

关键词 影响力最大化;社会网络;独立级联模型;信息传播;社会计算;社会媒体;社交网络

中图法分类号 TP399 **DOI 号** 10.11897/SP.J.1016.2017.00674

Discovering the Substitutes for the Seeds in Influence Maximization Problem

MA Qian MA Jun

(*School of Computer Science and Technology, Shandong University, Jinan* 250101)

Abstract The development of online social networks provides new opportunities and challenges for viral marketing. For a social network with n user-nodes, the advertisers face an important problem, i.e., how to select $k(0<k\ll n)$ influential nodes called seeds, once they become active because of the rewards or free samples, probably the number of the users who know or buy their products is the maximum at the end of diffusion through the word-of-mouth effects. This problem is also called Influence Maximization, or IM for short. Currently in the study of IM problem it is usually assumed that the k selected seeds can be activated. However, maybe there are $t(0<t\leq k)$ seeds that cannot be activated for various reasons in practice. In this situation a new research issue is how to reselect t substitutes to replace the inactive seeds. In this paper, we name this problem as Substitutes Discovery problem in Influence Maximization, or SDIM for short. The

收稿日期:2015-07-20;在线出版日期:2016-03-26.本课题得到国家自然科学基金(61272240,61103151)资助.马 茜,女,1989 年生,博士研究生,中国计算机学会(CCF)会员,主要研究方向为社会网络分析、社会影响力分析、信息传播.E-mail:maqiansdu@126.com.
马 军,男,1956 年生,博士,教授,博士生导师,主要研究领域为信息检索、数据挖掘、并行计算和自然语言处理.

problem of SDIM is beneficial to solve the practical problems in marketing, and help advertisers to complete the marketing target more smoothly. For this purpose, we first give a formal definition on SDIM, and propose an optimization function for the problem solving. We prove that SDIM is NP-Hard, and show that our proposed function is of approximation guarantee in developing greedy algorithms for SDIM. In the approach on SDIM, we first filter the nodes according to the scale-free property of social networks and reserve nodes with high degree as candidates. Then we propose three algorithms for SDIM respectively, i. e., (1) the Global Static Greedy (GSG) algorithm which just selects t substitutes; (2) the Greedy In Advance (GIA) algorithm which selects $t'(t' \geq t)$ substitutes in case of there are still inactivated seeds in the substitutes; and (3) All Static (AS) algorithm which can improve the efficiency of GSG. As the running time of GSG is unacceptable, we optimize it with the CELF strategy and we call it GSG-CELF in the experiments. The experimental results show that reducing the number of candidates according to the degree will not affect the accuracy of the algorithms but rather reduce the running time. The effect of the substitutes selected by GSG-CELF is close to the original seeds' effect. GIA is of the clear robustness with the effect close to the effect of GSG-CELF. Compared with the static greedy algorithm with CELF optimization—GSG-CELF, the running time of AS can still be reduced to 10%—50% of GSG-CELF's time, and the effect is not affected.

Keywords　influence maximization; social network; independent cascade model; information diffusion; social computing; social media; social networks

1　引　言

近几年来，Facebook、Twitter、微博、微信等在线社会网络的兴起和快速发展，使得信息能够以前所未有的速度扩散，同时用户的身份发生了转变，由信息的被动接收者变为主动参与者. 在社会网络中，每个用户都可以成为信息源，通过信息的快速传播来影响更多的人. 社会网络中这种影响力传播有着广泛的应用价值，比如推荐系统、市场营销、广告投放、谣言控制、专家发现等. 其中，影响力传播在广告营销方面的应用越来越受到重视，例如微博营销、网络营销等. 在社会网络中选择一些有影响力的用户，通过提供试用品或者折扣等方式，让这些用户向其他用户推荐该产品，通过口碑效应使得网络中了解或者购买该产品的用户数量达到最多. 同传统的广告方式相比，这种方式投入更少，广告投放更精准. 选择哪些用户投放广告才能影响最多的用户，这个问题被称作影响力最大化问题.

影响力最大化问题可定义为：在给定的社会网络中寻找大小为 k（k 为大于等于 1 的整数）的种子节点集合 S，激活这些种子节点并通过这些种子节点的传播使网络中的其他节点被激活，最终使整个网络中被激活节点的期望数目 $R(S)$ 达到最大. $R(S)$ 也被称作 S 的影响力.

在影响力最大化问题中，种子节点的规模是固定的，一般与企业的广告预算有关. 尽管在影响力最大化方面已存在大量的相关工作，但这些工作大都只考虑了从网络的所有节点中选定哪些节点组成种子节点集合可以影响最多的节点，并未考虑从全部网络节点中选择种子节点的现实性，也未考虑这些种子节点能否被激活.

通信技术的发展使得社会网络用户规模庞大，根据 2015 年 1 月《中国互联网络发展状况统计报告》[①]，截止 2014 年 12 月，我国微博用户规模为 2.488 亿，微信用户数量突破 5 亿. 社会网络规模如此庞大，加上影响力传播的模拟本身复杂度高，直接从网络中的所有节点中选取种子节点或本文提出的替补节点计算量都十分巨大；且企业在对产品进行营销时，考虑到产品受众、效益等因素，也要对种子节点进行一些限制. 一般来说，企业会找名人、明星等有一定知名度的用户进行推广. 另一方面，企业在实际营销中，受预算、用户喜好等一些因素的限制，有些种子节点可能无法激活. 例如，企业在有限的预算内，通过影响力最大化算法在一个社会网络中选出了 k 个有影响力的用户作为种子节点，但通过提

① http://www.cnnic.net.cn

供报酬、试用品、折扣等方式只能说服其中 $k-t$ ($0<t\leq k$)个用户,还有 t 个用户不愿为该产品做宣传.在这种情况下,一种做法是从网络中重新选取 k 个种子节点.这种做法没有考虑到原来已经同意为产品做宣传的 $k-t$ 个种子节点可能不会全部出现在新的种子集合中,企业面临着违约的风险;且新选出的种子节点集与原来的种子节点集会有大量重合,选择过程中有很多重复性计算.因此,本文要做的工作是在保留原来能激活的 $k-t$ 个种子节点的基础上,从未被选择的节点中重新寻找 t 个替补节点,使得传播效果与原来的种子节点的传播效果尽可能接近.

本文的主要贡献是:(1)对 SDIM 问题给出了形式化的定义,并且给出了具有子模性质的组合优化函数帮助设计贪心近似算法;(2)提出可以利用社会网络的无尺度特性对节点进行预处理,减少候选节点数量以提高算法的执行效率;(3)分别提出了 3 个 SDIM 问题的求解算法:(a)通过对 IM 问题中静态贪心算法的扩展,提出了寻找 t 个替补节点的 GSG 算法;(b)提出了在选择种子节点的同时选取 t' ($t'\geq t$)个替补种子节点的预选式贪心算法 GIA,可解决选出的替补节点中仍存在无法激活的节点的问题;(c) GSG、GIA 都存在着效率低的问题,为了提高选择效率,本文利用模拟传播过程中的静态优势,提出了 GSG 的优化算法——全静态算法 AS.实验结果表明:减少候选节点数量的方法不会影响各算法的精度,且可以有效地减少运行时间;带 CELF 优化的 GSG 算法选出的替补节点效果最好;GIA 算法结果比预期好,效果很接近 GSG-CELF,且具有很好的实用性和鲁棒性;与 GSG-CELF 相比,AS 运行时间可降低数倍且精度无损失.

本文第 2 节介绍相关工作;第 3 节定义 SDIM 问题并进行分析;第 4 节提出有效的求解算法并分析算法复杂性;第 5 节展示实验结果;第 6 节对全文进行总结并对未来工作进行展望.

2 相关工作

IM 问题就是如何选取 k 个种子节点进行传播,从而最大程度地影响整个网络的问题.这个问题最早是由 Domingos 和 Richardsom[1] 提出来的. Kempe 等人[2] 将该问题定义为离散优化问题,证明了这个问题是 NP 难的,并提出用贪心算法来近似求解,即每一步都选择当前最有影响力的节点加入种子节点集.这种方法具有(1−1/e)的求解精度,但该算法复杂度高,无法适用于大规模网络.针对这一缺陷,很多优化策略被提出以优化该算法使其能在大规模网络中运行. Leskovec 等人[3] 根据问题的子模性(submodular)提出了 CELF 优化策略,可使算法的执行效率提升 700 多倍. Goyal 等人[4] 对 CELF 进一步优化提出了 CELF++. Chen 等人[5] 从原图中移去不参与传播的边,在生成的较小的子图上计算影响力的传播. Cheng 等人[6] 指出文献[5]中的方法由于子模性无法得到保证,所以要进行大量的模拟.他们针对这个问题进行改进提出了 StaticGreedy 算法.该算法精度不受影响,效率可提升两个数量级.另一方面,还有很多工作从减少贪心算法中参与种子节点选择的节点数量的角度来提高选择效率. Luo 等人[7] 提出首先用 PageRank 算法对网络中的节点排序,再从排序较高,即较有影响力的节点中选择种子节点. Chen 等人[8] 和 Wang 等人[9] 分别提出用划分社区的方法减少参与种子节点选择的节点数量,从而获得更好的运行时间.

此外,也有很多工作使用启发式方法来提高问题的求解效率.如 Chen 等人[5] 提出了 DegreeDiscount 算法,根据节点的度对每个节点加入种子集合中能增加的影响力做简单估算. Kimura 和 Saito[10] 提出了基于最短路径的传播模型,并给出了有效的算法来估算影响力的传播. Chen 等人[11] 用最大影响力路径来估算影响力的传播,并据此提出了 PMIA 算法,具有良好的性能和精确度. Jung 等人[12] 提出的 IRIE 算法对所有节点的影响力进行排序,选择排名最靠前的节点作为种子节点;在选择一个种子节点后,估算网络中剩下节点的影响力重新排序选择,该方法比 PMIA 可扩展性更强. Cheng 等人[13] 分析发现贪心算法获得的解集是一种自洽排序,因此他们提出了一个迭代的框架 IMRank,将任意给定的初始排序通过迭代调整的方式得到自洽排序,这种方法能大大降低计算复杂度,且精度和贪心算法相当.

影响力最大化是应用性很强的问题,因此,还有很多工作致力于改善该问题模型的实用性.如 Goyal 等人[14] 研究如何利用已知的用户行为历史信息计算用户间影响力的传播概率,以便更好地对传播效果进行估算; Carnes 等人[15]、Shirazipourazad 等人[16] 考虑了有多条信息同时传播的情况,研究了存在竞争关系的影响力最大化问题; Bhagat 等人[17] 将投入、收益等经济因素加入模型中希望取得收益最大化; Guo 等人[18] 从对用户进行推荐的角度提出

了个性化影响力最大化问题等.

虽然上述工作能够帮助企业更快速更准确地挖掘出种子节点集合,但产品推广要达到理想效果的前提是要激活所有的种子节点.以前的工作都默认选出的所有种子节点都是可以激活的,然而在实际应用中,受许多不可控因素的影响,很有可能无法激活某些种子节点,如激活某个种子节点用户的费用过高超出企业的预算、种子节点用户对企业的产品不感兴趣或者不愿意做推广等. Lappas 等人[19]提出在社会网络中寻找那些能够达到当前传播效果的节点——效应者(effector),即要求效应者能激活特定的节点集合,这对营销问题来说过于精确.在营销中,对于一般的商品,企业通常只要求最终卖出的产品数量,而不要求一定要卖给某些特定的用户. Li 等人[20]提出在社会网络中寻找种子节点继任者这个问题,与本文的出发点相似.但这篇文章设定从未激活种子节点的邻居节点中寻找继任者,这种方法难以保证全局最优,且社会网络中的营销模式和传统的推销模式有所不同,文中提到的"一个推销员退休则从他的同事中寻找继任者"的应用场景在社会网络营销中不适用.因此,本文提出从全局网络中寻找种子节点替补者的方法,这种方法更适用于社会网络,效果也更好.

3　SDIM 问题

本节在独立级联模型的传播框架下,对 SDIM 问题进行了定义和介绍.本文采用的数学符号及含义如表 1 所示.

表 1　本文采用的数学符号及含义

符号	含义	符号	含义
G	社会网络结构图	T	替补节点集合
S	G 中的种子节点集合	S'	含替补节点的新种子节点集合
S_1	能激活的种子节点集合	C	候选节点集合
S_2	不能激活的种子节点集合	c	候选节点个数
k	种子节点个数	$p_{u,v}$	节点 u 激活节点 v 的概率
t	不能激活的种子节点个数	λ	全局传播概率

3.1　独立级联模型及影响力最大化问题

在解决 IM 问题的过程中,要用传播模型来模拟现实中影响力的传播,常用的模型有独立级联模型(Independent Cascade Model)和线性阈值模型(Linear Threshold Model)等.本文采用的是独立级联模型.独立级联模型将社会网络抽象为一个有向图 $G=(V,E,P)$,其中 V 代表节点集合,E 代表节点之间的有向边集,每条有向边 $(u,v)\in E$ 都关联一个传播概率 $p_{u,v}$,表示 u 通过边 (u,v) 激活 v 的概率.在独立级联模型中,节点有两种状态:活跃(active)或非活跃(inactive).每个节点同一时刻只能处于一种状态中,且节点可从非活跃状态变为活跃状态,反之则不可.S_t 表示在时刻 t $(t\geq 0)$ 被激活的节点集合.初始时,种子节点集 S 被激活,$S_0=S$.在 $t+1$ 时刻,每个在 t 时刻被激活的节点 $u\in S_t$,有且仅有一次机会去尝试激活它仍处于非活跃状态的邻居节点 v,激活成功的概率为 $p_{u,v}$,这一时刻被激活的节点集合为 S_{t+1}.这个过程重复直到某一时刻,整个网络都没有激活的情况发生,即 $S_t=\{\}$.

由于无法精确求解 S 最终能激活的期望节点数目 $R(S)$,所以一般用蒙特卡洛模拟来代替.蒙特卡洛模拟分为两种,一种是按照独立级联传播过程直接模拟[3-4,10],还有一种方式是图快照(snapshot)[5-6].快照是指对图 G 中的每一条边 (u,v) 以 $p_{u,v}$ 的概率保留下来获得一个图 G 的子图.一个快照可以看作是对影响力传播图 G 取样获得的一个实例,命名为 G'.在求解 $R(S)$ 时,可提前获取 r 个 G',$R(S)$ 是在所有 G' 中从 S 出发可到达的节点数目的平均值.

定义 1.　给定一个社会网络 $G=(V,E,P)$ 及种子节点的数目 k,IM 问题要从网络中寻找一个集合 S^* 使得

$$S^*=\underset{S\subseteq V,\ |S|=k}{\arg\max}(R(S)) \tag{1}$$

3.2　SDIM 问题定义

定义 2.　给定一个社会网络 $G=(V,E,P)$,使用影响力最大化算法找到的含有 k 个节点的种子集合 $S=S_1\cup S_2$,$|S|=k$,$S_1\cap S_2=\varnothing$.其中 S_1 是能激活的种子节点集合,S_2 是不能激活的种子节点集合,$|S_2|=t$.目标是从剩下的节点中找到一个替补集合 T,满足 $T\subseteq V\backslash S$,且 $|T|=t$.T 与 S 中能被激活的种子节点组成新集合 $S'=S_1\cup T$,使得 S' 最终在社会网络中的传播结果与原来的种子节点集 S 的传播结果的差值最小,即

$$S'^*=\underset{|S'|=k}{\arg\min}(R(S)-R(S')) \tag{2}$$

由于 S 是提前给定的且认为是最优的,可认为 $R(S)$ 是固定的,所以目标函数也可记作

$$S'^*=\underset{|S'|=k}{\arg\max}R(S') \tag{3}$$

定理 1.　SDIM 问题是 NP 难的.

SDIM 问题由经典的 IM 问题衍生而来,是 IM 问题的一个特例.根据定义可知,SDIM 可看作是在 $V\backslash S$ 中寻找 t 个种子节点的 IM 问题,而 IM 问题是

NP 难的[1],所以该问题也是 NP 难的.

由于 SDIM 问题是 NP 难的,且社会网络规模很大,所以只能近似求解.文献[2]中指出对一个非负单调且具有子模性质(对任意集合 S、T 及元素 v,若 $S\subseteq T$,则有 $f(S\cup\{v\})-f(S)\geq f(T\cup\{v\})-f(T)$)的函数 f,S 为通过贪心算法选取的 k 个元素集合,则有 $f(S)\geq(1-1/e)f(S^*)$,S^* 为最优解.显然,本文中的目标函数(3)具有单调子模性,因此我们有定理 2.

定理 2. 在 SDIM 问题中使用贪心算法获得的解有 $(1-1/e)\approx 63\%$ 的精度保证.

SDIM 问题同 IM 问题一样,也存在着无法精确计算节点影响力的问题,我们同样也使用模拟的方法来获得节点影响力的近似值.无论是直接模拟还是图快照的方式,耗时都很长,而贪心算法又需要对大量的节点进行模拟.为此,本文希望在分析数据、传播模型以及问题本身特点的基础上,提出更高效的算法来解决 SDIM 问题.

4 算 法

本节提出了解决 SDIM 问题的策略.首先对节点进行过滤,生成候选节点集,减少参与替补节点选择的节点个数,提高算法效率;然后提出了 3 种算法分别从候选节点中选择替补节点.

4.1 选取候选节点

在求解 SDIM 问题之前,我们首先对参与替补节点选择的节点进行过滤,即筛选出影响力较大的节点作为候选节点.研究表明,大部分社会网络是无标度网络[21],即少量节点拥有大量连接,而大部分节点拥有少量连接.例如在微博网络中,名人、明星等往往有大量的用户关注,而普通用户一般受到的关注较少.因此,理论上不必让所有节点都参与替补节点的选取.这种做法不但可以提高算法的选择效率,在应用上也有实际意义.企业在考虑让谁替自己做宣传时,更多的是选择影响力大、粉丝数目多的用户,这样才能取得更好的效果.

选取候选节点就是对网络中节点的影响力作评估、去掉一些影响力较小的节点的过程.有很多度量方法可以对节点的影响力进行评估,如节点的度(degree)[21]、中心性(centrality)[21]、用 PageRank[22]、Hits[23]算法进行排序等.在本文中我们利用节点的度对节点进行过滤,即只从度大于等于 d 的节点中选择替补节点.中心性、排序等度量方法虽然对节点影响力的评估更准确一些,但计算复杂度高,在大规模网络中耗时很长,并且在这一步只需对节点的影响力做粗略评估,无太高的精度要求.节点的度一般代表着用户在社会网络中的粉丝数量或者朋友数量,计算简单,本身就是一种常见且有效的度量指标[24].

4.2 选取替补节点

在 SDIM 问题中,当 S 中大部分的种子节点都不能被激活时,要选取的替补节点数目较多,既要保证替补节点的传播效果和原始种子节点的效果尽可能接近,还要保证算法的可扩展性使其能在大规模网络中运行,因此本文通过对 IM 问题中的静态贪心算法(StaticGreedy)[6]进行扩展提出了 GSG 算法,并在进一步分析贪心算法选择过程及问题特性的基础上提出了 GIA 算法.GIA 直接采用了 IM 问题中的静态贪心算法来选择替补节点,但选择策略与 GSG 不同.GSG 算法存在着运行时间长、可扩展性差的问题,因此本文又对其进行了优化提出了 AS 算法.

4.2.1 全局静态贪心算法(Global Static Greedy)

在 SDIM 问题的贪心算法中,每次选择加入当前集合 S' 后能激活的节点数目与原来的种子集合 S 能激活的节点数目相差最小的那个节点作为替补节点.由于 S 是提前给定的,且本文在对 $R(S)$ 的模拟过程中使用了静态化方法,可认为 $R(S)$ 是固定的,所以可以将每次加入 S' 后能激活节点数目最多的那个节点作为替补节点.在算法 1 中,1~4 步是初始化及生成 r 个快照图的过程,5~14 步为选取替补节点的过程.当要选择一个替补节点时,需要将当前所有候选节点 $v\in C\setminus(S\cup S')$ 依次加入到当前种子节点集 S' 中,并计算它在 r 个快照图上能影响的节点个数的总和 Add_v(7~9),最后将在 r 个快照图上影响节点数目的平均值 Add_v/r 最大的节点加入 S'(12).

算法 1. 全局静态贪心算法(GSG).

输入:有向图 $G=(V,E,P)$,含有 k 个种子节点的集合 S,未激活种子节点数目 t,含有 $k-t$ 个能激活的种子节点集合 S_1,候选节点集合 C,快照图数目 r

输出:含有 k 个节点的新种子节点集合 S'

1. Initialize $S'=S_1$
2. FOR $i=1$ to r do
3. 对 G 中的每一条边 (u,v) 以 $p_{u,v}$ 的概率保留生成 $G'=(V',E')$
4. END FOR
5. FOR $i=1$ to t do

6. Set $Add_v=0$ for all $v\in C\backslash(S\cup S')$
//Add_v是在r个快照图上$\{S'\cup v\}$的影响力总和
7. FOR $j=1$ to r do
8. FOR all $v\in C\backslash(S\cup S')$ do
9. $Add_v=Add_v+|BFS(G_j',\{S'\cup v\})|$
//BFS表示宽度优先搜索
10. END FOR
11. END FOR
12. $S'=S'\cup\{\underset{v\in C\backslash(S\cup S')}{\arg\max}\{Add_v/r\}\}$
13. END FOR
14. RETURN S'

4.2.2　预选式贪心算法(Greedy In Advance)

在选取种子节点的时候，提前考虑部分种子节点不能被激活的情况，多选一部分节点作为替补. 这是在应用中较为实际的一种做法，可以防止选出的替补节点中仍存在无法激活的节点. 在本文中，假设要选取的种子节点的个数为k，则用IM问题中常用的贪心算法选取$k+t'$个节点，$t'\geq t$，并按选取的顺序排列，前k个节点为种子节点. 当t个种子节点不能激活时，则选取$k+1$到$k+t$的节点作为替补节点. 若替补节点中仍有节点无法激活，则继续依次向后选取.

在IM问题中，用贪心算法选择种子节点集合S时，每次选择将增加影响力最多的节点加入S. 假设图1中IM贪心算法的选择序列D为1、2、3、4、5. 若要选择两个种子节点，则$S=\{1,2\}$. 此时在剩下的节点中，节点3、4、5加入S能增加的影响力分别为7、5、4，节点3加入S中增加影响力的能力要优于网络中的其他节点. 当S中有节点未能激活时，网络中的其他节点加入S能增加的影响力可能会有不同程度的增加. 例如，当种子节点2不能激活时，节点3加入S能增加的影响力由7变为8. 此时，除非有其他节点与不能激活的种子节点有大量的影响力重叠使得该节点的影响范围大幅度增加，使其加入S能增加的影响力超过了节点3加入S能增加的影响力，否则节点3仍是增加影响力最多的节点.

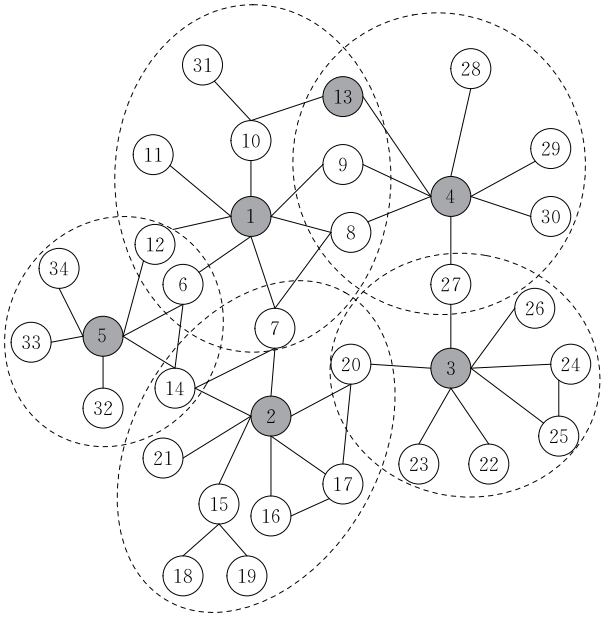

图1　节点影响力的重叠

与GSG相比GIA的优势是，当新选出的替补节点，例如节点3，仍无法激活时，GIA可以根据序列D中的顺序直接将4选为替补节点尝试激活，而GSG却要重新运行选取替补节点.

算法2.　预选式贪心算法(GIA).

输入：按IM贪心算法选择顺序排列的、含有$k+t'$个节点的序列D，$t'\geq t$，其中前k个节点组成种子节点集合S，未激活种子节点数目t，含有$k-t$个能激活的种子节点集合S_1

输出：含有k个节点的新种子节点集合S'

1. Initialize $S'=S_1$
2. FOR $i=1$ to t do
3. $S'=S'\cup D_{k+i}$
//D_{k+i}是序列D中第$k+i$个节点
4. END FOR
5. RETURN S'

4.2.3　全静态算法(All Static)

贪心算法效率低的原因是，每选择一个节点，都要进行大量的模拟以估算将该节点加入种子节点集后能增加的影响力，影响力指从该节点出发能到达的节点个数. 进行大量模拟的原因是新加入的节点可能与已选出的种子节点的影响范围有重叠. 例如在图1中，假设$\{1,2\}$是已选出的能激活的种子节点. 当考虑将3加入时，3与$\{1,2\}$的影响范围有重叠部分——节点20，所以要重新计算将3加入后$\{1,2,3\}$的影响力，进而判定节点3是否是增加影响力最多的节点.

在文献[2-3,6]等以前的贪心算法以及GSG算法中，每选一个种子节点或替补节点都要将所有候选节点v依次加入S'中求影响力值，选择加入后影响力最大的那个节点作为种子节点或替补节点. 因此选择t个节点就需要在所有快照图上对$|S'|+c$个节点搜索t次. 为了避免这种重复搜索，提高效率，在算法3中，我们利用快照图的静态优势，提前记录S'和各个候选节点v在每个子图G'上能影响的节点集合$Cover_{S'}$，$Cover_v$(2～8)，然后求这两个

集合的并集就可得到将 v 加入 S' 后的覆盖范围(13).这样只需对每个候选节点及 S' 的影响范围搜索一次,可大量减少算法的运行时间,且候选节点数量有限,无需大量的存储空间.

算法 3. 全静态算法(AS).

输入:有向图 $G=(V,E,P)$,含有 k 个种子节点的集合 S,未激活种子节点数目 t,含有 $k-t$ 个能激活的种子节点集合 S_1,候选节点集合 C,快照图数目 r

输出:含有 k 个节点的新种子节点集合 S'

```
Initialize S'=S1
FOR j=1 to r do
    对G中的每一条边(u,v)以 p_{u,v} 的概率保留生成 G'_j
    Cover_{(G'_j,S')}=BFS(G'_j,S')
      //Cover_{(G'_j,S')} 为在 G'_j 中从 S' 出发能到达的节点集合
      //BFS 表示宽度优先搜索
    FOR all v∈C\(S∪S')
      Cover_{(G'_j,v)}=BFS(G'_j,v)
    END FOR
END FOR
FOR i=1 to t
    Set Add_v=0 for all v∈C\(S∪S')
      //Add_v 是在 r 个快照图上{S'∪v}的影响力总和
    FOR all v∈C\(S∪S')
      FOR j=1 to r
        Add_v=Add_v+|Cover_{(G'_j,S')}∪Cover_{(G'_j,v)}|
      END FOR
    END FOR
    Inf_Max_v= argmax_{v∈C\(S∪S')} Add_v
    S'=S'∪{Inf_Max_v}
    FOR j=1 to r
      Cover_{(G'_j,S')}=Cover_{(G'_j,S')}∪Cover_{(G'_j,Inf_Max_v)}
    END FOR
END FOR
RETURN S'
```

4.2.4 算法复杂性分析

该节主要对算法 1 和算法 3 进行复杂性分析,算法 2 与算法 1 复杂度在同一数量级上.算法 1 包含两部分:生成 r 个快照图和在快照图上寻找替补节点.生成 r 个快照图的时间复杂度为 $O(rm)$,m 为 G 中边的数量.在替补节点选择部分,算法 1 要在 r 个子图上对 $c+k-t$ 个节点搜索 t 次,所以算法 1 的时间复杂度为 $O(rm)+O((c+k-t)rtm')$,m' 为快照图 G' 边数的平均值,c 为候选节点的个数.算法 1 每次只需读取一个快照图进行搜索,空间复杂度为为 $O(n+m)$.算法 3 也需生成 r 个快照图,但只需在 r 个快照图上对 $c+k-t$ 个节点搜索一次并存储搜索结果.因此算法 3 的时间复杂度为 $O(rm)+O((c+k-t)rm')$,空间复杂度为 $O(n+m)+O((c+k-t)l)$,l 为候选节点或种子节点在一个快照图上能到达的边数的平均值.

5 实 验

5.1 实验设置

本文在两个真实数据集上进行了实验.数据集 1 为腾讯微博数据,从 KDDCUP2012 数据集①中抽取获得,其中节点为微博中的用户,有向边代表用户之间的关注关系,该数据集包含 6322 个节点和 148044 条边,还包括一些其他信息,如行为信息、用户属性等.数据集 2 为 Epinions 信任网络②,节点代表网站会员,u 到 v 的有向边代表 u 信任 v,该数据集包含 49287 个节点和 487182 条边.

独立级联模型中的参数 $p_{u,v}$ 要提前给出,这个概率过大会使影响力传播对算法不敏感,过小则导致传播效果不明显,因此这个概率一般不超过 0.1[5],常见做法是直接设为 0.01.在微博数据集中,在设置激活概率时我们将边的权重 $w_{u,v}$ 考虑了进去,即 $p_{u,v}=w_{u,v}\times\lambda$,$\lambda$ 为全局传播概率.边的权重表示节点之间关系的强度,当节点 v 与节点 u 之间的关系更紧密时,v 更易受到 u 的影响,所以这样设置可以提高模型的合理性.边的权重包括了三部分内容:两个节点之间的内容相似性(共同关键词的比例)、行为亲密度(u 对 v 评论、转发所占比例)、共同粉丝数量.我们将所有边的权重 $w_{u,v}$ 规约到 0～1 之间.为使激活概率在一个合理的范围内,在微博数据集中我们将 λ 设置为 0.1,因此整个网络的传播概率 $p_{u,v}$ 在 0～0.1 之间.在 Epinions 数据集中,激活概率均设为 $p_{u,v}=\lambda=0.01$.

在实验中我们将求节点影响力值的模拟次数设为 20000 次,算法中快照图个数 $r=100$.S 是由文献[6]中提出的 StaticGreedyCELF 算法获得,且在 GIA 中也用该算法预选取替补节点.不能激活的种子节点集合 S_2 则是从 S 中随机选出的.

本文比较了 5 种替补节点的选择策略,由于 GSG 算法运行时间过长,在实验中我们对 GSG 进行了 CELF 优化,称优化算法为 GSG-CELF.除了带 CELF 优化的全局静态贪心算法(GSG-CELF)、

① http://www.kddcup2012.org/c/kddcup2012-track1
② http://snap.stanford.edu/data/soc-Epinions1.html

预选式贪心算法(GIA)和全静态算法(AS)外，还有Degree、Random 算法参与了比较. Degree 是一种基于节点度的启发式算法，只考虑候选节点的邻居节点数量，即按照每个候选节点在一步之内能到达的节点个数排序，选择前 t 个节点作为替补节点. Random 方法则是从候选节点中随机选取 t 个节点作为替补节点.

5.2　实验结果

5.2.1　保留候选节点度的取值 d 不同对各算法挖掘出的替补节点的影响

为了验证两个数据集的无尺度特性，我们对两个数据集的度分布进行了分析. 在图 2、图 3 中，横坐标代表节点的度，纵坐标代表该度值所对应的节点个数与总节点个数的比值，以 log 的形式表示. 从图 2、图 3 可以看出，两个数据集的节点的度均服从幂律分布，即少量节点的度很大，大部分节点的度都很小. 据此我们可以根据节点的度对节点进行过滤，只保留度较大的节点作为候选节点.

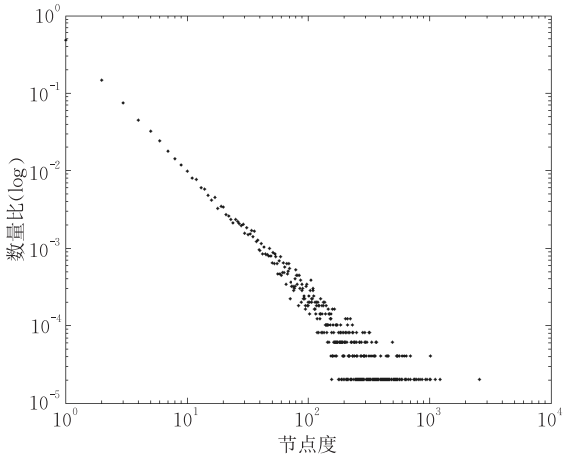

图 2　Epinions 数据集节点的度分布

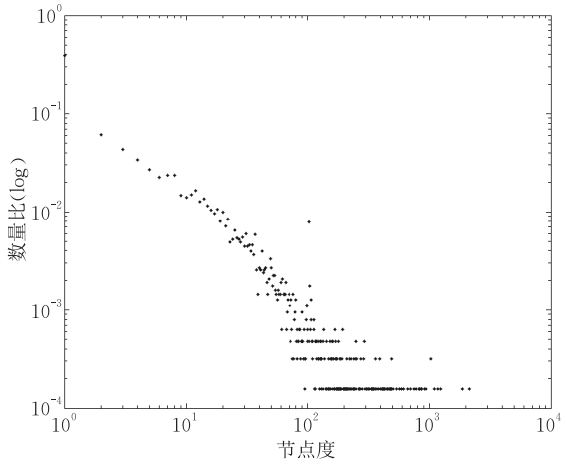

图 3　微博数据集节点的度分布

为了评估保留度值不同的候选节点集对各个算法选出的替补节点的影响力及运行时间的影响，我们按节点的度过滤保留了不同规模的候选节点集，再从候选节点集中运行不同算法选取替补节点，与能激活种子节点组成新的种子节点集 S'. 在图 4～图 7 中，横坐标 d 代表节点度的取值，即保留度值大于等于 d 的节点作为候选节点. 随着保留候选节点度的增大，保留的候选节点个数会减少. 图 4 中横坐标对应的候选节点个数分别为 3701，2012，1332，950，702，541，435，363，图 5 中横坐标对应的候选节点个数为 1089，827，666，555，469，424，378，343. 图 4、图 5 的纵坐标表示不同算法获得的新种子节点集合 S' 的影响力值，图 6、图 7 的纵坐标表示算法的运行时间. 在图 4、图 5 中，$R(S)$ 为原来种子节点集 S 的影响力值. 由于 S' 中部分节点为 S 中的可

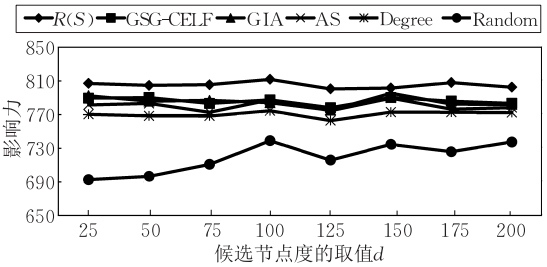

图 4　Epinions 数据集中不同算法获得的 S' 的影响力随候选节点度的取值 d 变化的情况($k=20$，$t=10$，$\lambda=0.01$)

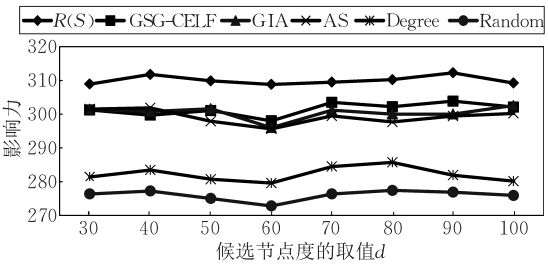

图 5　微博数据集中不同算法获得的 S' 的影响力随候选节点度的取值 d 变化的情况($k=20$，$t=10$，$\lambda=0.1$)

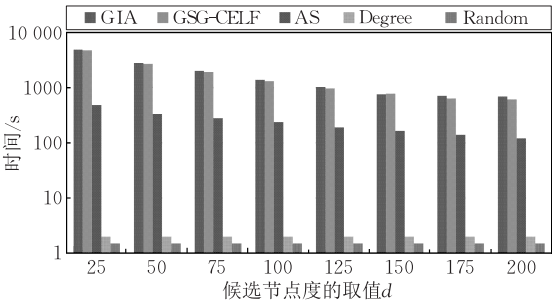

图 6　Epinions 数据集中不同算法获得 S' 的运行时间随候选节点度的取值 d 变化的情况($k=20$，$t=10$，$\lambda=0.01$)

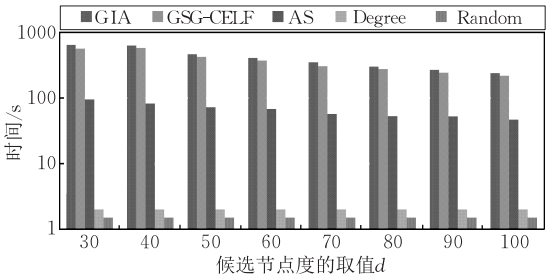

图 7 微博数据集中不同算法获得 S'的运行时间随候选节点度的取值 d 变化的情况($k=20, t=10, \lambda=0.1$)

激活种子节点,为了尽量排除这部分种子节点的影响能力不同对 S' 的结果造成的影响,我们保留了 S 中影响力排前 10 的节点作为可激活节点,各算法获得 S'的影响力值越接近 $R(S)$越好.

候选节点度的取值 d 越大,候选节点集包含越少的节点,理论上选出的种子节点效果也应越差,即 $R(S)$越小.但从图 4、图 5 中可以看出,$R(S)$并没有随候选节点的减少而呈下降趋势,只是出现了轻微的波动.这是由于绝大部分种子节点出自度较高的节点集合,去掉度较小的节点不会对种子节点的选择造成影响,所以没有出现下降.另外,由于无法精确计算各节点的影响力值,只能通过大量的蒙特卡洛模拟来估算,而模拟存在一定的不确定性,所以,同一节点在不同次的模拟中可能出现不同的影响力值.这种不确定性不仅会影响节点影响力的模拟结果,还会影响种子节点的选择.例如一个节点在从候选节点集 $C1$ 选择种子节点的过程中模拟出的效果很好而被选为种子节点,而在从另一个候选节点集 $C2$ 的选择中模拟出的效果不好而没有被选为种子节点,被另一个节点代替.尽管增加模拟的次数可大大降低这种不确定性,但仍不能完全消除,所以会出现波动.在图 4、图 5 中,各个算法挖掘出的含有替补节点的 S'的影响力也没有随候选节点个数的减少而发生下降.各算法挖掘出的 S'也存在波动现象,一是由于 S'中包含 S 中的部分节点,受 S 波动的影响 S'也会出现波动,从图中可以看出各算法结果的波动情况与 $R(S)$的波动情况有一定的相似性.二是替补节点选择的过程也受到节点影响力模拟的不确定性影响,但各算法出现的波动都比较轻微.图 4 中候选节点度的取值由 25 增加为 200,保留的候选节点个数由 3701 降低到 363,各算法(Random 除外)从这两个候选集中选取的替补节点的效果差值基本都在 10 之内.这其中 GSG-CELF、GIA 和 SC 的结果重叠在一起,都比较接近原始种子节点集的结果.作为对照的 Degree 和 Random 效果要差于本文提出的 3 种方法,Degree 要好于 Random. Degree 方法选择网络中除 S 中的节点外度最大的 t 个节点,因此它不受候选节点度值变化的影响,它的影响力变化是由保留的可激活种子节点不同和模拟结果不同造成的,影响力变化的幅度也是最小的,不超过 5 个节点. Random 算法波动幅度最大,因为随着保留的候选节点的度变大,Random 方法可能会从候选节点集中选出度更大的替补节点,效果会更好.

在运行时间上,从图 6、图 7 可以看出,GIA、AS、GSG-CELF 算法的运行时间都会随着候选节点数目的减少而减少.其中 GIA、GSG-CELF 在一个数量级上,都需要较长的运行时间,且运行时间下降得十分明显;相比之下 AS 算法的运行时间就要少的多.实验说明根据节点的度对节点进行过滤的方法能有效地降低运算时间,尤其是对 GSG-CELF、GIA 这类较耗时的贪心算法来说.

根据节点的度对节点过滤的方法有很强的应用价值,由于网络用户数量巨大,企业在进行产品推广时可以直接从有一定知名度的网络用户中选择种子节点或者替补节点,而不必从全部网络用户中寻找,可以省去大量的计算时间,且不会影响产品的传播效果.

5.2.2 未激活种子节点占种子节点总数的比值 t/k 不同对各算法挖掘出的替补节点的影响

为了评估不同比例的未激活种子节点对不同算法的效果及运行时间的影响,我们在含有 20 和 50 个种子节点的集合中分别随机选择不同比例的未激活种子节点,然后利用各个算法寻找对应数目的替补节点.在图 8～图 11 中,图中横坐标为未激活种子节点占种子节点总数的比值,即 t/k,图 8、图 9 的纵坐标为算法选出的含替补节点的新种子节点集 S'的影响力值,$R(S)$为原始种子节点集 S 的影响力值,各算法选出的 S'的影响力越接近 $R(S)$,说明选择的替补节点效果越好.在图 8 和图 9 中,我们还分别使用各自的种子节点集多次求 $R(S)$,用来验证影响力模拟是否会对各算法结果造成严重干扰.图 10、图 11 的纵坐标代表各算法选择替补节点的运行时间.

在图 8、图 9 中,因为分别对各自的同一组种子节点 S 模拟了多次,所以 $R(S)$出现波动完全是由影响力模拟造成的,从图中可以看出波动非常轻微,说明充分大的模拟次数(20000 次)可以使模拟结果保持相对稳定,不会对各算法的结果造成太大的干

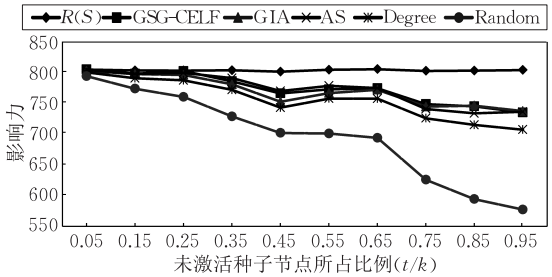

图 8　Epinions 数据集中不同算法获得的 S' 的影响力随未激活种子节点所占比例 t/k 变化的情况($k=20$, $\lambda=0.01$, $|c|=950$)

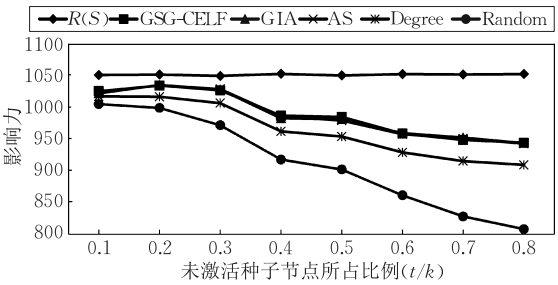

图 9　Epinions 数据集中不同算法获得的 S' 的影响力随未激活种子节点所占比例 t/k 变化的情况($k=50$, $\lambda=0.01$, $|c|=950$)

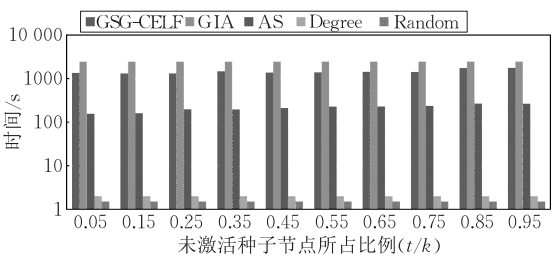

图 10　Epinions 数据集中不同算法的运行时间随未激活种子节点所占比例 t/k 变化的情况($k=20$, $\lambda=0.01$, $|c|=950$)

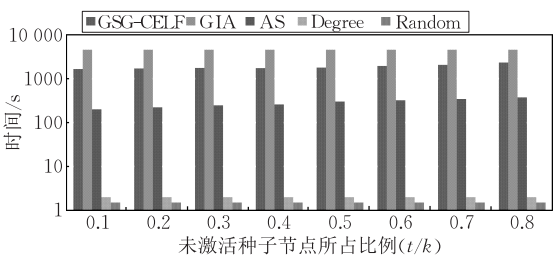

图 11　Epinions 数据集中不同算法的运行时间随未激活种子节点所占比例 t/k 变化的情况($k=50$, $\lambda=0.01$, $|c|=950$)

扰. 随着未激活种子节点所占比例的上升，各算法与原始种子节点集传播效果之间的差距都会增大. 原始种子节点集是网络中可以产生最优传播效果的一个节点组合，替补节点可以弥补因部分种子节点不能激活导致的传播效果方面的损失，但一般情况下不能与原始种子节点相媲美，所以未激活种子节点越多，新种子节点集 S' 中含有的替补节点越多，与 $R(S)$ 差距越大. 在各算法中，Random 效果最差，Degree 次之，Random 和 Degree 与 GSG-CELF 等算法的差距十分明显. GSG-CELF、GIA 和 AS 三者重叠在一起. 理论上 GIA 效果应比 GSG-CELF、AS 差，但实验结果显示并非如此. 经过分析我们认为造成这种现象的原因是：网络规模大，网络的传播概率很小，各个节点的影响力范围有限，节点之间影响力重叠情况不严重；有着相似影响力的节点较多，节点之间的可替代性强. 在图 8 中，当 $t/k=0.95$ 时，20 个种子节点中有 19 个种子节点被替换，贪心算法 GSG-CELF 选出的节点集的影响力由原始种子节点集的 804.8 降至 735.6，每替换一个节点只带来平均 3.6 个节点的影响力损失，可说明有影响力节点之间的可替代性强. Degree 算法完全没有考虑节点之间影响力的重叠，当 $t/k=0.95$ 时，Degree 比 GSG-CELF、AS 等算法效果下降了 5%左右，Degree 算法的表现也比理论上要好，说明节点间影响力的重叠较少. 由于只保留了度较大的节点作为候选节点，Random 方法也有很好的表现，但效果仍比 GSG、GIA 等算法差了 10%～20%.

图 10、图 11 展示了不同算法的运行时间. 由于 GIA 是在选择种子节点的同时选择替补节点，所以默认它的运行时间不变. GSG-CELF 和 GIA 的运行时间在同一数量级上. GSG-CELF 的运行时间随 t/k 的上升而上升，由于进行了 CELF 优化，上升幅度不大. AS 运行时间只需 GSG-CELF、GIA 的 10%～50%，在近 50 000 节点规模的网络中寻找 40 个替补节点只需 6 min，而 GSG-CELF 算法需要 39 min. Random 和 Degree 的运行时间很少，但是效果比 GSG-CELF、AS 等下降了 5%～20%不等. 对于营销问题来说，提高 5%～20%的影响力就能增加 5%～20%的客户，有很大的社会和经济效益. 且在实际传播过程中，网络规模更大，传播概率可能要远大于实验设置值，所以尽管 GSG-CELF、GIA、AS 算法耗时长一些，但更有应用价值. 这其中 AS 算法实用性最强，既可取得不逊于 GSG-CELF、GIA 的效果，运行时间却少得多.

5.2.3　全局传播概率 λ 不同对各算法挖掘出的替补节点的影响

为了评估不同的全局传播概率对不同算法的运行结果及运行时间的影响，我们在不同的 λ 取值下寻找替补节点. 图 12、图 13 的横坐标均为微博数据集中全局传播概率 λ 的取值，图 12 的纵坐标为算法选出的含替补节点的新种子节点集的影响力，图 13

的纵坐标为算法的运行时间.从图 12 可以看出,GSG-CELF、GIA、AS 的运行结果依旧重叠在一起,均随 λ 的增大而增大,且都非常接近原始种子节点集的影响力 $R(S)$.这说明随着激活概率的增加,文中提出的方法仍可以有效地弥补因部分种子节点无法激活带来的损失.图 13 展示了各算法的运行时间随 λ 变化的情况.随着 λ 的增加,GSG-CELF、GIA 所需运行时间均大幅度增加,其中 GIA 变化幅度最大,当 $\lambda=0.1$ 时,用 GIA 寻找 10 个替补节点只需运行几分钟,当 $\lambda=0.55$ 时,GIA 需运行 1.2 h.由于对 GSG 进行了 CELF 优化,GSG 变化幅度较小,但传播概率较大时仍需要很长的运行时间,与 GIA 在同一数量级,而 AS 的运行时间只需 GSG-CELF 的一半甚至更低.Degree、Random 拥有良好的时间性能,不受 λ 变化的影响,但 Degree 和 Random 选出的替补节点的影响力稍差.

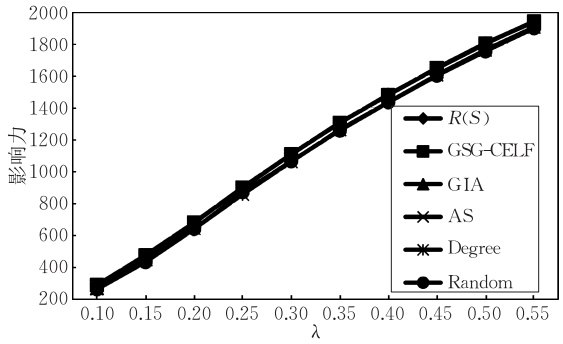

图 12 微博数据集中不同算法获得的 S' 的影响力随全局激活概率 λ 变化的情况($k=20$,$t=10$,$|c|=666$)

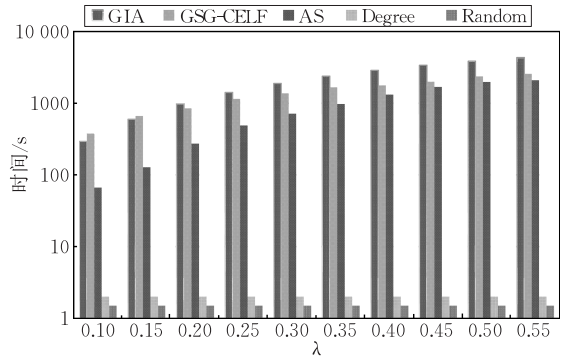

图 13 微博数据集中不同算法的运行时间随全局激活概率 λ 变化的情况($k=20$,$t=10$,$|c|=666$)

6 总 结

本文介绍并定义了在影响力最大化问题中,当部分种子节点无法激活时,如何从全局社会网络中寻找替补节点,减少因种子节点无法激活带来损失的问题.通过对问题进行分析,我们提出了解决方法:首先对网络中的节点进行过滤,保留度较大的节点作为候选节点;然后提出了 GSG、GIA 和 AS 这 3 种算法分别从候选节点中选择替补节点.在两个真实数据集上的实验表明:(1)根据节点的度对节点过滤的方法不会影响算法的效果且可以有效地减少运行时间;(2)提出的 3 种算法选出的替补节点均可以很好地代替未激活的种子节点.这其中,GSG 算法有良好的精度保证;GIA 算法有很好的鲁棒性;AS 是对 GSG 的改进,运行时间只需 GSG 的 $1/r$ 而不影响所选替补节点的效果,具有很强的实用性.

在未来的工作中,我们将从问题实用性的角度出发,为每个节点赋予一个激活费用值.当有种子节点不能激活时,考虑如何在预算固定的情况下选择更合适数目的替补节点使产品传播地更广泛.

参 考 文 献

[1] Domingos P, Richardson M. Mining the network value of customers//Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2001: 57-66

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

[3] Leskovec J, Krause A, Guestrin C, et al. Cost-effective outbreak detection in networks//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, USA, 2007: 420-429

[4] Goyal A, Lu W, 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

[5] Chen W, Wang Y, Yang S. 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

[6] Cheng S, Shen H, Huang J, et al. Static greedy: Solving the apparent scalability-accuracy dilemma in influence maximization //Proceedings of the 22nd ACM International Conference on Conference on Information & Knowledge Management. San Francisco, USA, 2013: 509-518

[7] Luo Z L, Cai W D, Li Y J, et al. A pagerank-based heuristic algorithm for influence maximization in the social network// Recent Progress in Data Engineering and Internet Technology. Berlin Heidelberg, Germany, 2012: 485-490

[8] Chen Y, Chang S, Chou C, et al. Exploring community structures for influence maximization in social networks//Proceedings of the 6th SNA-KDD Workshop on Social Network Mining and Analysis Held in Conjunction with KDD. Beijing, China, 2012, 12: 1-6

[9] Wang Y, Cong G, Song G, et al. Community-based greedy algorithm for mining top-*k* influential nodes in mobile social networks//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 1039-1048

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

[11] Chen W, Wang C, Wang Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 1029-1038

[12] Jung K, Heo W, Chen W. IRIE: Scalable and robust influence maximization in social networks//Proceedings of the 12th International Conference on IEEE Data Mining (ICDM). Brussels, Belgium, 2012: 918-923

[13] Cheng S, Shen H, Huang J, et al. IMRank: Influence maximization via finding self-consistent ranking//Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. Gold Coast, Australia, 2014: 475-484

[14] 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

[15] Carnes T, Nagarajan C, Wild S M, et al. Maximizing influence in a competitive social network: A follower's perspective//Proceedings of the 9th International Conference on Electronic Commerce. Mineapolis, USA, 2007: 351-360

[16] Shirazipourazad S, Bogard B, Vachhani H, et al. Influence propagation in adversarial setting: How to defeat competition with least amount of investment//Proceedings of the 21st ACM International Conference on Information and Knowledge Management. Maui, USA, 2012: 585-594

[17] Bhagat S, Goyal A, Lakshmanan L V S. Maximizing product adoption in social networks//Proceedings of the 5th ACM International Conference on Web Search and Data Mining. Seattle, USA, 2012: 603-612

[18] 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

[19] Lappas T, Terzi E, Gunopulos D, et al. Finding effectors in social networks//Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2010: 1059-1068

[20] Li C T, Hsieh H P, Lin S D, et al. Finding influential seed successors in social networks//Proceedings of the 21st International Conference Companion on World Wide Web. Lyon, France, 2012: 557-558

[21] Tang L, Liu H. Community Detection and Mining in Social Media. California, USA: Morgan & Claypool Publishers, 2010

[22] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab, 1998, 9: 1-14

[23] Kleinberg J M. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999, 46(5): 604-632

[24] Wasserman S, Faust K. Social Network Analysis: Methods and Applications. Cambridge, UK: Cambridge University Press, 1994

MA Qian, born in 1989, Ph. D. candidate. Her research interests include social network analysis, influence analysis and information diffusion.

MA Jun, born in 1956, Ph. D., professor, Ph. D. supervisor. His research interests include information retrieval, data mining, parallel computing and natural language processing.

Background

The research problem in this paper belongs to influence analysis in social networks. The rapid development of online social networks provide vast realistic data for researchers to study the evaluation, spreading and modeling of the social

influence. Among them influence maximization problem has been studied extensively in recent years. Advertisers convince some influential people, called seeds in this problem, and create word-of-mouth advertising for the product. As the problem is NP-hard, most works are dedicated to proposing efficient algorithms to find the seeds which can trigger a large cascade of adoption. However, most of the works ignore the fact that not all the seeds can be convinced. In this paper, we define the substitutes discovery problem in influence maximization and propose three algorithms to discover the substitutes for the inactive seeds. The substitutes discovery problem has great applicable value.

This work is supported by the National Natural Science Foundation of China (61272240, 61103151). These projects aim to study the social media information processing techniques under the application backgrounds of multi-document summarization and social network analysis. The group is dedicated to doing research on new theories, algorithms and systems for information retrieval, data mining and social network. Related papers have been published in reputable domestic and international journals and conferences.