软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(2):326−340 [doi: 10.13328/j.cnki.jos.005117]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于重叠社区搜索的传播热点选择方法*

单 菁[1,2], 申德荣[1], 寇 月[1], 聂铁铮[1], 于 戈[1]

[1](东北大学 计算机科学与工程学院,辽宁 沈阳 110819)

[2](沈阳建筑大学 信息与控制工程学院,辽宁 沈阳 110168)

通信作者: 单菁, E-mail: mavis0129@126.com

摘 要: 随着社交网络的蓬勃发展,信息传播问题由于具有广泛的应用前景而受到广泛关注,影响力最大化问题是信息传播中的一个研究热点.它致力于在信息传播过程开始之前选取能够使预期影响力达到最大的节点作为信息传播的初始节点,并且大多采用基于概率的模型,如独立级联模型等.然而,现有的影响力最大化解决方案大多认为信息传播过程是自动的,忽略了社交网站平台在信息传播过程中可以起到的作用.此外,基于概率的模型存在一些问题,如无法保障信息的有效传播、无法适应动态变化的网络结构等.因此,提出了一种基于重叠社区搜索的传播热点选择方法.该方法通过迭代式推广模型根据用户行为反馈逐步选择影响力最大化节点,使社交网站平台在信息传播过程中充分发挥控制作用.提出了一种基于重叠社区结构的方法来衡量节点影响力,根据这种衡量方式来选择传播热点.提出了解决该问题的两种精确算法(包括一种基本方法和一种优化方法)以及该问题的近似算法.通过大量实验验证了精确及近似算法的效率、近似算法的准确率以及迭代式传播热点选择方法的有效性.

关键词: 影响力最大化;信息传播;重叠社区;社交网络

中图法分类号: TP311

中文引用格式: 单菁,申德荣,寇月,聂铁铮,于戈.基于重叠社区搜索的传播热点选择方法.软件学报,2017,28(2):326−340. http://www.jos.org.cn/1000-9825/5117.htm

英文引用格式: Shan J, Shen DR, Kou Y, Nie TZ, Yu G. Approach for hot spread node selection based on overlapping community search. Ruan Jian Xue Bao/Journal of Software, 2017,28(2):326−340 (in Chinese). http://www.jos.org.cn/1000-9825/5117.htm

Approach for Hot Spread Node Selection Based on Overlapping Community Search

SHAN Jing[1,2], SHEN De-Rong[1], KOU Yue[1], NIE Tie-Zheng[1], YU Ge[1]

[1](School of Computer Science and Engineering, Northeastern University, Shenyang 110819, China)

[2](Information and Control Engineering Faculty, Shenyang Jianzhu University, Shenyang 110168, China)

Abstract: With the development of social network, information diffusion problem has received a lot of attention because of its extensive application prospects, and influence maximization problem is a hot topic of information diffusion. Influence maximization aims at selecting nodes that maximize the expected influence as initial nodes of information diffusion, and most work on influence maximization adopts probabilistic models such as independent cascade model. However, most existing solutions of influence maximization view the information diffusion process as an automatic process, and ignore the role of social network websites during the process. Besides, the probabilistic models have some issues in that, for example, they cannot guarantee the information to be delivered effectively, and they cannot adapt the dynamic networks. To tackle the problem, this paper proposes an approach for hot spread node selection based on overlapping community search. This approach selects influence maximized nodes step by step through the iterative promotion model

* 基金项目: 国家重点基础研究发展计划(973)(2012CB316201); 国家自然科学基金(61472070)

Foundation item: National Program on Key Basic Research Project of China (973) (2012CB316201); National Natural Science Foundation of China (61472070)

收稿时间: 2016-01-21; 修改时间: 2016-04-25; 采用时间: 2016-06-14; jos 在线出版时间: 2016-10-11

CNKI 网络优先出版: 2016-10-12 16:26:49, http://www.cnki.net/kcms/detail/11.2560.TP.20161012.1626.014.html

according to users' behavior feedback, and makes the social network websites play the controller's role during information diffusion process. The paper also proposes a new method to measure the influence of nodes based on overlapping community structure, and utilizes this information measure method to select hot spread nodes. Two exact algorithms are proposed including a basic algorithm and an optimized algorithm, as well as an approximate algorithm are presented. Comprehensive experiments demonstrate the performance and accuracy of both exact and approximate algorithms, and the effectiveness of the iterative hot spread node selection method.

Key words: influence maximization; information diffusion; overlapping community; social network

近年来,随着互联网信息技术的蓬勃发展,社交网络平台因其能够灵活、有效地传播和交换各种信息而受到广泛的关注,社交网络用户呈爆炸式增长.由于在病毒式市场营销(viral marketing)上具有可观的应用前景,信息传播问题(information diffusion)是社交网络领域被广泛研究的一个问题.它旨在解决在社交网络上如何快速、有效地传播信息.

为了使信息在社交网络中进行广泛的传播以使其影响力达到最大,影响力最大化问题(influence maximization)成为信息传播问题中的一个研究热点,该问题由 Kempe 等人在 2003 年首次提出[1].影响力最大化问题被定义为一个离散的优化问题:给定一个社交网络图和一个整数 k,要找到一个含有 k 个点的集合 S,使得以该集合为初始化节点在某种传播模型下预期的影响力达到最大化.例如,为了推广新产品,公司会在社交网站平台上向用户免费发放试用品,使这些初始用户在试用产品后可以将产品宣传给他们的朋友,以此达到宣传目的.由于宣传存在花费代价,投放给初始用户的试用品和广告的数量有限,因此最关键的问题是如何选取初始用户,使得他们具有最大的影响力,以致最终会选择购买新产品的用户数量达到最大,这就是影响力最大化问题在病毒式市场营销背景下的一个典型应用.然而,影响力最大化问题致力于在信息传播过程开始之前选取关键的节点作为初始节点,认为信息在选定初始化节点后会自动传播下去,信息传播过程是不受控制的,社交网站平台在信息传播过程中可以起的作用被忽略.实际上,在信息传播的过程中,可以根据用户对传播信息的反馈对传播策略加以调整,因而可以让社交网站平台参与控制信息传播过程,以达到更好的传播效果.为此,本文提出一种迭代式推广模型:并非一次性选择出所有初始节点,而是分批次选择传播节点,根据本轮对传播节点进行信息推送后用户的行为反馈来选取下一轮传播节点,以此来控制信息的传播过程.

影响力最大化问题的两种经典模型分别为独立级联模型(independent cascade mode)和线性模型(linear model),皆由 Kempe 等人提出[1].其中,独立级联模型被后续研究者广泛采用并在此基础上衍生出各种变种模型.该模型设定:一旦某用户 u 接受了一条信息(如点击了一条广告),则该用户会进行一次独立尝试以试图影响其每个邻居 v.每次尝试影响成功具有一个概率 $p_{u,v}$,若某节点被影响成功,则该节点变为激活节点,且所有节点的状态仅可以从未激活变为激活,反之则不可以.然而,现有的基于概率的影响力最大化模型在实际应用中存在一些问题,会影响到信息传播的效果:首先,基于概率的传播模型无法保障信息从一个用户传播到另一个用户,传播本身是一个概率化的过程;其次,概率模型的概率赋值本身在实际应用中不便于衡量;第三,该模型的种子节点选择方法是一种静态方法,然而实际中社交网络的拓扑结构是不断动态变化的,因此该模型无法很好地适用于动态变化的网络结构.鉴于以上原因,本文并未采用基于概率的模型,而是提出了一种基于重叠社区搜索的方式来衡量节点的影响力.

在社交网络中,研究发现重叠社区(overlapping community)结构普遍存在[2].社区结构由一些关联紧密的用户节点组成,同一社区内节点间的相互连接要比它们与社区外的节点连接紧密得多.因此,在社交网络中,一个社区可以代表一个有共同点(如兴趣、背景等)的用户群体,由于社区内的用户联系更紧密,直观上可以看出,同一社区内的用户之间的影响力要更大,且这些影响不仅存在于线上网络世界,也存在于线下实际世界.比如,用户 u 看了某部电影并对其评价非常好,则用户 u 可能会在电影网站上给出评分,u 的好友会在网站上看到 u 的评分,同时好友们也会在生活中接受到 u 对该电影的多次强烈推荐,这种实际的影响力可能要远大于网络上的影响力.此外,社区之间是存在重叠的,即一个用户可以隶属于多个与其兴趣、社交活动、朋友等相关的社区,并且隶属于多个社区的用户与多个社区内的用户关系都十分紧密,因此,这种用户的推荐范围更广,推荐能力更强.此外,由于社交网络中的信息量巨大,用户每天都会面对海量信息,因此,用户的关注度是有限的,大量的广告推广

信息会给用户造成不友好的用户体验,从而降低用户对社交网站的好感,基于重叠社区搜索的影响力衡量方法可以有效利用用户间的潜在影响力,如用户间线下多次的推荐行为,这种方式更具说服力,且不会破坏用户对网站的体验感.鉴于以上原因,本文利用社交网络中的重叠社区的结构来衡量用户的影响力,提出了一种基于重叠社区搜索的传播热点选择(hot spread node selection,简称 HSNS)方法.其主要贡献点如下.

1) 提出一种迭代式推广模型.该模型基于用户行为反馈的实时更新,迭代地选择最有影响力的目标用户进行信息推荐,这与现有研究的一次性影响力最大化策略存在巨大差异.

2) 提出一种基于重叠社区结构的方法来衡量节点的影响力,并据此提出传播热点选择问题,给出该问题的精确和近似解决方法.

3) 通过丰富的实验验证了精确和近似算法的效率和近似算法的准确率以及迭代式传播热点选择方法的有效性.

本文第 1 节介绍与本文相关的研究工作.第 2 节提出迭代式推广模型,并对传播热点选择问题给出形式化定义.第 3 节提出传播热点选择问题的精确方法,其中包括基本方法和优化方法.第 4 节提出基于打分机制的传播热点选择问题的近似方法.第 5 节通过大量实验来验证精确和近似算法的有效性和准确率.第 6 节对文章进行总结,并展望下一步的研究工作.

1 相关工作

本文主要研究了一种基于重叠社区搜索的传播热点选择方法.该方法是一种基于重叠社区来衡量影响力的影响力最大化问题解决方法.本节主要介绍与本文研究最相关的 3 个研究方向:影响力最大化问题、社交化广告以及重叠社区发现问题.

社交网络中的影响力最大化问题近年来引起了广泛的关注.Kempe 等人[1]将此问题定义为:选择一组种子节点来进行影响力传播,使这些种子节点所能影响到的预期节点数量达到最大化.他们证明了该问题是 NP 难的,且提出一种贪心算法来近似地解决.该方法重复地选择可以引起最大边际效应增量的节点.计算给定种子节点的影响力传播问题是难解的[3],Kempe 等人[1]提出了影响力传播仿真过程,用仿真结果的均值来估计影响力传播,然而该方法仿真过程计算量巨大.因此,Leskovec 等人[4]提出了一种称为 CELF 的机制来减少计算次数.Chen 等人提出了两种称为 DegreeDiscount[3]和 PMIA[5]的快速启发式算法.Jung 等人[6]用一种迭代方法来估计影响力传播.文献[7,8]提出基于影响力传播路径的方法来估计受影响的节点.曹玖新等人提出了一种基于 k-核的核覆盖率启发式算法[9].影响力最大化模型大多基于概率的级联模型,一些研究[1,10−12]给出多种方法来模型化一个用户受其朋友影响而购买商品的概率.影响力最大化问题的传统解决方法忽略了对信息传播过程的控制,并且是一个静态过程,无法很好地适应动态变化的网络结构.Lin 等人[13]提出一种推送驱动级联模型(push-driven cascade),可以强化社交网站对信息传播过程的控制.Tong 等人[14]提出一种动态独立级联模型(dynamic independent cascade)以自适应地动态选择种子节点.然而这两种方法仍以独立级联模型为基础,并没有充分利用社交网络拓扑结构所蕴含的用户之间潜在的深层社会关系.

社交化广告通过对社交网络中潜在信息的挖掘,对广告进行个性化定制以及定向目标投放.现有的一些社交化广告研究通过观察用户的朋友对一则广告所采取的行为来衡量该用户对同一则广告采取行为的可能性,以此验证社交化广告的意义和可行性.例如,Bakshy 等人[15]在 Facebook 数据上运行的实验验证了朋友对用户对待广告行为的影响,实验发现社交网络中朋友对广告留下的痕迹增加了用户点击广告的概率,最近有一些对社交化广告策略的最新研究.Aslay 等人[16]站在社交网络平台的角度,提出了一种基于预算花费的社交化广告推广策略.该策略权衡了广告客户预算花费和广告推广平台的利益,试图从二者之间找到最佳的平衡.Abbassi 等人[17]提出了一种社交化广告推广策略,它根据社交网络中用户的印记来制定目标用户以及广告推广次序,使广告的预期点击量最大化.然而,上述社交广告推广策略均采用基于概率的传播模型来衡量用户之间的相互影响,对社交网络拓扑结构所蕴含的潜在信息并未充分加以利用,并且现有的社交化广告策略大多是一次性的推广策略,并没有考虑用户的动态行为对广告推广策略产生的影响.

重叠社区结构被发现广泛存在于社交网络中,Palla 等人[2]在 2005 年首先提出了重叠社区问题,提出一种团渗透方法(clique percolation method)来发现重叠社区结构,并提出一种称为 CFinder[18]的工具来探测生物网络中的重叠社区.除了团渗透方法以外,其他一些方法也被提出来,用以发现重叠社区,如基于链接划分的方法[19−21]、基于标签传播的方法[22,23]等.胡云等人[24]提出了一种专门针对微博社交网络的,以用户-话题关系为主要划分原则的多模网络重叠社区表达模型以及相应的重叠社区结构发现算法.此外,探测重叠社区还有一类基于统计模型的方法[25−28].然而上述重叠社区探测问题以整个网络作为输入,一次性地探测出网络上所有重叠社区,这种解决方法具有计算量巨大、无法适应动态图、灵活性差等局限性,因此 Cui 等人[29]提出了重叠社区搜索问题.该问题以单个查询点作为输入,仅在网络中查找包含该点的所有社区,是一种局部搜索方法,重叠社区搜索方法是轻量级的、灵活的,能够很好地适应动态网络.Shan 等人[30]在此基础上提出了以多个查询点作为输入的重叠社区搜索方法.该方法也是本文所采用的方法.本文利用重叠社区结构来选择传播热点.

2 迭代式推广模型及传播热点选择问题定义

2.1 迭代式推广模型

在介绍模型之前,首先考虑一件产品的普通传播模式:当一件新产品 A 上市时,它会被少数一群最初始的用户购买,随着初始用户对产品的使用,体验感很好的用户会将 A 推荐给周围的朋友,且用户在推荐时已经有预筛选,会把 A 推荐给对它有需求的朋友,这些被推荐者对朋友的推荐信息会比普通广告更优先考虑,因此他们购买 A 的几率更高,这些经过初始用户推荐而购买 A 的人自动加入了用户群,并把产品 A 推荐给他们的朋友,如此层层迭代,最终产品 A 会被广泛地传播出去,并且找到最合适的受众.然而,传统影响力最大化问题的解决方案普遍采用独立级联模型,在传播过程开始之前预先选定预计可产生最大影响力的种子节点,这种推广模型是一次性的静态选择策略,没有考虑信息传播过程中用户的行为反馈对整个传播效果所产生的影响,这与现实生活中的传播过程存在差异,因此,本文提出一种根据用户行为反馈进行动态调节的迭代式推广模型,如图 1 所示.

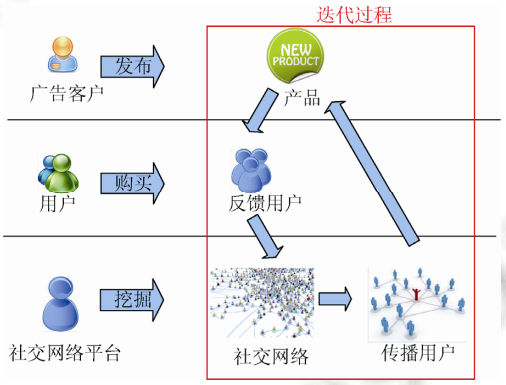


Fig.1 Iterative advertising model

图 1 迭代式推广模型

首先,本文中设定社交网络是一个无向图,其中节点代表用户,边代表用户间的关系.在迭代式推广模型中,第 i 次传播热点的选择取决于第 i−1 次传播推广后的反馈用户(我们将所有购买产品的用户称为反馈用户),第 1 次传播热点的选择取决于产品上市后未经推广时自发选购商品的原始用户.假设广告客户提供的推广总预算为 B,且每人次广告推送的花费为 1.由此,我们给出传播模式和传播策略的概念.

定义 1. 传播模式 $P=(k_1,\ldots,k_n)$是一系列非负整数,代表在第 i 次迭代选择 k_i 个传播热点.其中,$k_1,\ldots,k_n$ 满足 $\sum k_i=B$.

定义 2. 给定一个社交网络 G,一个传播模式 $P=(k_1,\ldots,k_n)$,与之对应的一个传播策略 $S_P^G=(s_1,\ldots,s_n)$,由一系列

点的集合组成,其中$|s_i|=k_i$, s_i 是第 i 次迭代选择出的传播热点集合.

基于传播模式和传播策略的定义,整个迭代式推广模型的执行框架如算法 1 所示.

算法 1. 迭代式推广模型框架.

输入:图 $G(V,E)$;推广总预算 B;反馈用户集合 $F=(b_0,\ldots,b_{n-1})$.

输出:传播策略 $S_P^G=(s_1,\ldots,s_n)$.

1) $S_P^G \leftarrow \varnothing$;
2) $k_1=Compute(b_0)$;
3) $sum \leftarrow 0$;
4) $i \leftarrow 1$;
5) while $sum+k_i \leqslant B$ do
6) $sum \leftarrow sum+k_i$;
7) $s_i \leftarrow SelectNodes(b_{i-1},k_i)$;
8) $S_P^G \leftarrow S_P^G \cup \{s_i\}$;
9) $Spread(s_i)$;
10) $k_{i+1}=Compute(b_i)$;
11) $i \leftarrow i+1$;
12) if $sum \neq B$
13) $s_i \leftarrow SelectNodes(b_{i-1},B\text{-}sum)$;
14) $S_P^G \leftarrow S_P^G \cup \{s_i\}$;
15) $Spread(s_i)$;
16) return S_P^G;

从算法 1 中可以看出,迭代式推广模型的输入有推广总预算 B,以及反馈用户集合 $F=(b_0,\ldots,b_{n-1})$,其中,b_0 是传播开始之前的原始用户,$b_i(1\leqslant i\leqslant n-1)$是经过第 i 次传播热点选择并进行信息传播之后对推广产品进行购买的反馈用户,虽然它是在传播过程中产生的,但它是已知的,因此我们将它看成是整个推广模型框架的一个输入.传播模式 $P=(k_1,\ldots,k_n)$中的元素 k_i 与反馈用户 b_i 相关,第 2.2 节将给出其具体定义方法.首先,根据原始用户 b_0 计算出第 1 次迭代要选择的传播热点数量 k_1(第 2 行),对于进行的第 i 次迭代,当前 i 次的传播代价之和 sum 不大于推广总预算 B 时,根据当前的反馈用户 b_{i-1} 选择 k_i 个传播热点作为传播策略 s_i,并将其加入传播策略集合(第 5 行~第 8 行),然后进行传播推广(第 9 行),传播推广包括社交网站对传播热点用户的推广以及用户之间的相互推广,待本轮传播结束后,根据新的反馈用户 b_i 计算下一轮迭代的传播热点数量 k_{i+1}(第 10 行),当包含最后一轮的传播热点数量之和超过推广总预算时,选择满足预算所剩数量的传播热点,将所得传播策略加入传播策略集合并进行传播推广,最终返回传播策略集合(第 12 行~第 16 行).

2.2 传播热点选择问题定义

在迭代式推广模型中,最重要的环节是根据反馈用户,从社交网络中挖掘出最具推广价值的用户,因此,需要定义何种用户可称为“具有推广价值”.直觉上来说,具有推广价值应该具备两个基本条件:(1) 用户对产品感兴趣,有很大的可能购买;(2) 用户本身具有很强的影响能力,可以把产品广泛地推广出去,并使信息接收者容易被说服.研究发现,社区结构可以很好地反映用户的共性,即具有共同兴趣的用户会处于同一社区中,利用这一特点,根据反馈用户可以挖掘出对产品具有同样兴趣的用户.此外,同一社区内的用户彼此之间联系紧密,因此,传播的广告更具有效性,容易被接受.另外,社区之间存在重叠,即一个用户可以隶属于多个社区,所在的社区数量越多,用户的传播能力越强.鉴于以上特点,我们利用重叠社区结构来定义具有推广价值的用户,其中,重叠社区我们采用 Palla 等人[2]提出的基于 k 团(k-clique)的定义方法.

定义 3. 一个社区,准确地说,一个 k 团社区,是由一组互相可达的 k 团连接而成.其中,若两个 k 团共享 k–1

个节点,则称这两个 k 团相邻接.因此,一个 k 团社区是一个 k 团连通分支.

定义 4. 给定图 G,反馈用户集合 b 和整数 k,已知反馈用户集合 b 中的节点分布于社区 $C_1,\ldots,C_m$ 中,传播热点选择问题是要从社区 $C_1,\ldots,C_m$ 中找出 k 个节点,使得 $\sum_{i=1}^{k}|n_i|$ 最大,其中,$|n_i|$是节点 i 所隶属的社区数量,这些节点称为传播热点.

以图 2 为例,假设反馈用户节点为点 b,可得到其所隶属的两个 4 团社区$\{a,h,c,b,i\}$,$\{b,c,l,k,j\}$,要从这两个社区的所有节点中选出两个传播热点(k=2),则选出的节点是$\{a,c\}$,因为它们都隶属于两个社区(其中,a 隶属于$\{a,d,f,g\}$和$\{a,h,c,b,i\}$,b 隶属于$\{a,h,c,b,i\}$和$\{b,c,l,k,j\}$),而其他节点 h,i,j,l,k 只隶属于 1 个社区.

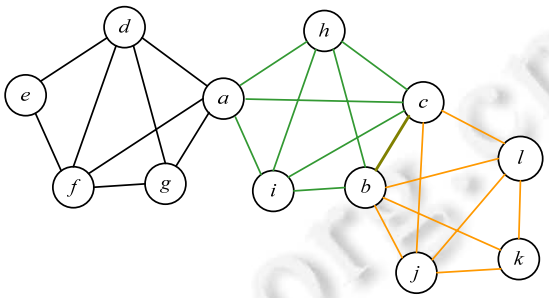


Fig.2 An example of hot spread node selection

图 2 传播热点选择问题举例

对于传播热点选择问题中要选取的节点个数 k,它与反馈节点所分布的社区 $C_1,\ldots,C_m$ 有关,k 的计算方法如公式(1)所示.

$$k=\left\lceil \alpha\left|\bigcup_{i=1}^{m}C_i\right|\right\rceil \tag{1}$$

其中,$\left|\bigcup_{i=1}^{m}C_i\right|$ 表示反馈节点所分布的所有社区 $C_1,\ldots,C_m$ 中的节点个数,α 为比例参数.该公式表示从反馈用户所分布的所有社区节点中选择一定比例数量的节点作为传播热点.

需要提到的是,本文提出的传播热点选择问题目标是找到满足定义 4 的约束条件中的 k 个传播热点,在多种节点组合都满足约束条件的情况下,只需找到其中的 1 种即可.因此,针对多种满足条件的节点组合,本文的算法随机选择其中一种作为传播热点.此外,本文提出的基于重叠社区搜索的模型来自于实际应用,例如,用户听到 1 个熟人推荐某产品和听到多个熟人推荐该产品所受到的影响力强度是不一样的.因此,即使下一轮传播热点和上一轮传播热点影响到的节点相同,下一轮传播热点对所影响节点进行的传播行为也会加深上一轮传播热点的影响效果.所以,即便传播热点所影响的节点有所重叠,也并不算对总的推广预算 B 造成浪费.

3 传播热点选择精确方法

本节介绍传播热点选择的两种精确算法,一种是简单的基本算法,另一种是经过优化的算法.

3.1 基本方法

从图 2 的例子中可以看出,点 a 所隶属的$\{a,h,c,b,i\}$是已知社区,而它所隶属的$\{a,d,f,g\}$是未被发现的社区,因为该社区没有包含反馈节点 b.因此,要想统计出已知社区中每个节点所隶属的社区数量,最直接也是最准确的办法是对每个节点都进行重叠社区搜索(overlapping community search).重叠社区搜索是根据给定的查询点,局部地发现查询点所隶属的社区.该方法的优点是灵活、轻量,本文利用 Shan 等人提出的支持多查询点的重叠社区搜索方法[30]来得到相关用户的重叠社区结构.算法 2 给出了传播热点选择的基本方法的执行过程.

算法 2. 传播热点选择基本算法.

输入:图 $G(V,E)$;社区 $C_1,\ldots,C_m$;参数 k_i,k_c.

输出:k_i 个传播热点 $n_1,\ldots,n_{ki}$.

1) N=min*heap*(k_i); //初始化大小为 k_i 的小顶堆
2) for C_i in $C_1,\ldots,C_m$ do
3) for n_i in C_i do
4) $|n_i|\leftarrow$OCS(G,n_i,k_c); //重叠社区搜索
5) N.*update*($n_i,|n_i|$); //更新小顶堆
6) return N;

算法 2 解决的问题是从已知社区 $C_1,\ldots,C_m$ 中选择出 k_i 个隶属社区数量最多的节点,其中 k_i 为迭代推广模型第 i 次迭代要选择的节点数量.具体做法是,在已知社区 $C_1,\ldots,C_m$ 中,我们对每个点 n_i 进行重叠社区搜索(k_c 是 k 团的参数),得到其所隶属的社区数量$|n_i|$,并用小顶堆 N 来维护 k_i 个社区数量最多的节点.然而,由于重叠社区搜索是一个 NP 难问题[30],因此对每个节点进行重叠社区搜索的时间代价太高,在输入的已知社区包含节点数量过多的情况下是不适用的,为此,我们提出一种优化的精确方法.

3.2 优化的精确方法

由于基本方法中对每个点进行重叠社区搜索的方法会出现大量不必要的冗余计算,Shan 等人[30]提出了一种方法,根据节点的类别进行单点重叠社区搜索优化.在该方法中,将节点针对某一个社区分为内部节点、边界节点和外部节点.受到该方法的启发,本文对节点分类的定义加以改进,将节点针对某一组已知社区进行分类,使其适应传播热点选择问题这一应用背景.给定已知社区 $C_1,\ldots,C_m$,节点根据其邻点和已知社区的关系分为内部点、边界点和外部点.这 3 种节点定义如下.

定义 5. 节点 n_i 属于已知社区 $C_1,\ldots,C_m$,若其所有邻点也都属于已知社区 $C_1,\ldots,C_m$,则节点 n_i 是内部点.

定义 6. 节点 n_i 属于已知社区 $C_1,\ldots,C_m$,若它有 1 个或多个邻点不属于已知社区 $C_1,\ldots,C_m$,则节点 n_i 是边界点.

定义 7. 不属于已知社区 $C_1,\ldots,C_m$,的节点称为外部点.

以图 2 为例,已知反馈节点 b 和两个 4 团社区{a,h,c,b,i},{b,c,l,k,j},则已知社区中的点 c,h,i,j,k,l 是内部点,只有节点 a 是边界点.显而易见,内部点不会存在于新的社区中,因此,我们只需对已知社区中的边界点进行重叠社区搜索,计算其隶属社区数量;对于已知社区中的内部节点,只需统计其隶属社区数,然后选出 k_i 个隶属社区最多的节点.具体优化的精确算法如算法 3 所示.

算法 3. 传播热点选择优化算法.

输入:图 $G(V,E)$;社区 $C_1,\ldots,C_m$;参数 k_i,k_c.

输出:k_i 个传播热点 $n_1,\ldots,n_{ki}$.

1) N=min*heap*(k_i); //初始化大小为 k_i 的小顶堆
2) for C_i in $C_1,\ldots,C_m$ do
3) for n_i in C_i do
4) if n_i is interior node then //对于内部点
5) $|n_i|\leftarrow$Count($C_1,\ldots,C_m,n_i$); //统计其社区数
6) N.*update*($n_i,|n_i|$);
7) else //对于边界点
8) add n_i into B; //将边界点存起
9) for n_i in B do
10) $|n_i|\leftarrow$OCS(G,n_i,k_c); //对边界点进行重叠社区搜索,计算社区数
11) N.*update*($n_i,|n_i|$);
12) return N;

如算法 3 所示,仍然采用小顶堆来维护 k_i 个社区数量最多的节点.对于已知社区中的每个节点 n_i,判断是内

部点还是边界点,若是内部点,则直接统计它的社区数然后更新排序;若是边界点,则对其进行重叠社区搜索,计算出社区数目,然后更新排序,最终将 k_i 个隶属社区数最多的节点返回.针对图 2 中的例子,我们只需对点 a 进行重叠社区搜索,这极大地降低了计算代价.具体优化效果将在实验中详细体现.

4 传播热点选择近似方法

由于两种精确方法中采用的对节点进行重叠社区搜索的方法是 NP 难问题,因此,尽管经过优化,其时间复杂度仍然较高,因此,我们提出一种有效降低时间复杂度的近似方法来进行传播热点选择.

在本节中,利用给已知社区中每个节点打分的方式来衡量节点所属的社区数量,以此方法来替代对每个节点进行时间复杂度较高的重叠社区搜索,从而可以快速选出传播热点,因此,一种可以较好地反映隶属社区数目的近似打分方法是该近似方法的核心.近似方法的执行过程与算法 2 基本相同,只需将对每个节点进行重叠社区搜索(第 4 行)替换成对每个节点计算其得分即可.

由社区的定义可知,社区是由 k 团连接而成的,因此,隶属于多个社区的点肯定存在于多个 k 团中,所以一个点最多可以隶属的 k 团数可以用来衡量它所隶属的社区数.然而,即使一个点隶属于多个 k 团中,若这些 k 团彼此相邻接,则该点仍然属于同一个社区中.因此,我们引入局部集聚系数(local clustering coefficient)[31]来解决这个问题.该系数量化衡量一个节点和邻点之间可以构成一个团的连接紧密程度,当节点的所有邻点之间互相连接时,局部集聚系数为 1,节点的所有邻点之间没有边相连,局部集聚系数为 0.因此,若一个节点隶属于多个 k 团,且该节点的局部集聚系数较低,则认为它所隶属的这些 k 团分布较为分散,因此,这些 k 团可能属于不同的社区.由此我们引入社区数目估计函数作为节点的打分函数.该函数表示如公式(2)所示.

$$Score_{CN}(n_i)=\frac{k\text{-}clique_{\max}(n_i)}{lcc(n_i)} \tag{2}$$

其中,$k\text{-}clique_{\max}(n_i)$表示点 n_i 最多可以隶属的 k 团数,$lcc(n_i)$是点 n_i 的局部集聚系数.一个点的局部集聚系数定义如公式(3)所示.

$$lcc(n_i)=\frac{2|e_i|}{d_i(d_i-1)} \tag{3}$$

其中,$|e_i|$是点 n_i 的邻居节点之间存在的边的数量,d_i 是点 n_i 的度.

由公式(2)可知,如何计算给定节点最多可隶属的 k 团数是打分函数的关键问题.该问题可以形式化定义为:给定节点 n_i,已知其度为 d_i,邻点之间的边数为$|e_i|$,求 n_i 最多可以存在于多少个 k 团中.该问题可以转化为:给定 d_i+1 个节点,$|e_i|+d_i$ 条边,在保证这些节点相连的情况下,最多可以构成多少个 k 团,为了表述方便,令 $d=d_i+1$,$|e|=|e_i|+d_i$,即给定 d 个点,$|e|$条边,最多能构建多少个 k 团.

对于给定已知点数和边数求最多可以构建的 k 团数问题,可以采用贪心策略来得到该问题的最优解:首先,从给定的点中取出 k 个点和 $k(k-1)/2$ 条边构建出第 1 个 k 团,接下来每次从剩下的点中引入 1 个点和 x 条边,要保证新加入的这 1 个点和 x 条边可以引入最多的 k 团个数,且剩下的点和边数可以满足点数大于等于边数,以保证整个图的连通性.概括来说,贪心策略的宗旨是在保证图的连通性的情况下,每引入一个点要引入最多的 k 团.下面我们通过图 3 举例说明贪心策略的执行过程.

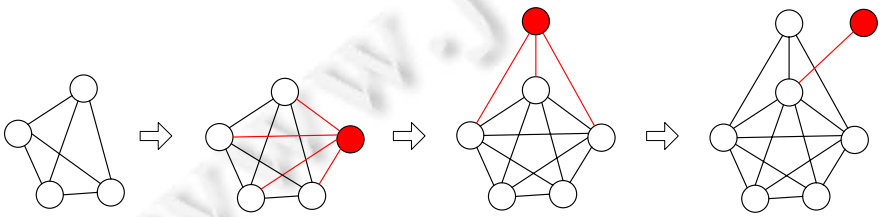


Fig.3 An example of greedy strategy for d=7, $|e|$=14

图 3 贪心策略举例,d=7,$|e|$=14

图 3 的例子中给定 7 个点,14 条边,求最多可以构建多少个 4 团.首先,选取 4 个点和 6 条边构建出第 1 个 4 团,接下来引入一个新点和 4 条边,则新构建出 4 个 4 团,然后引入一个新点,此时若引入全部余下的 4 条边,则还剩 1 个点和 0 条边,形成孤立点,因此本次最多可以引入 3 条边,构建出 1 个 4 团,最后引入 1 个新点和 1 条边,执行过程结束,因此,最多可以构建出 1+4+1 共 6 个 4 团.贪心策略求解最多 k 团构建问题的具体执行过程如算法 4 所示.

算法 4. 贪心策略最多 k 团构建求解算法.

输入:点数 d,边数 e,参数 k.

输出:最多 k 团数 c.

```
1)  c=0;
2)  e=e–k(k–1)/2;          //构建第 1 个 k 团
3)  d=d–k;
4)  if e<d then            //如果剩余的点和边无法保证图的连通性
5)     c=0;                //则构成的 k 团数为 0
6)  else
7)     c=1;
8)     x=k;                //x 为当前最多可以添加的边数,初始化是 k 条边
9)     while e≥d+k–2       //当剩余的点和边数不满足终止条件时
10)       e=e–x;           //引入新的 1 个节点和 x 条边
11)       d=d–1;
12)       if e<d then      //如果剩余的点和边无法保证图的连通性
13)          e=e+x;        //则还原剩余点和边数,并将最多可加边数减 1
14)          d=d+1;
15)          x=x–1;
16)       else             //否则,计算新引入 k 团数,并将最多可加边数增 1
17)          c=c+ C_x^{k-1} ;
18)          x=x+1;
19) return c;
```

如算法 4 所示,给定点数 d 和边数 e,求最多可以构建多少个 k 团.首先取 k 个点和 $k(k–1)/2$ 条边构建首个 k 团(第 2 行~第 3 行),然后检查剩余的点和边数是否满足图的连通性,若不满足,则直接返回 k 团数为 0(第 4 行~第 5 行).接下来,当剩余的点和边数不满足终止条件 $e≥d+k–2$ 时(第 9 行),从剩余点中引入一个新点和 x 条边,保证新引入的点和边可以构成最多的 k 团(第 10 行~第 11 行),若此时剩余点和边数不满足图的连通性,则回溯并将最多可添加的边数 x 减 1(第 12 行~第 15 行);若剩余点和边数满足图的连通性,则本次引入的点和边数方案成立,计算新引入的 k 团数 C_x^{k-1}, 并将最多可增加边数 x 增 1(第 16 行~第 18 行).

下面分析贪心过程的终止条件.假设过程进行到添加最后一个可以引入 k 团的节点和边数,此时剩余点数 d 中包括即将引入新 k 团的 1 个节点,以及其余无法引入新 k 团的 $d–1$ 个节点,而要想引入新 k 团,则至少要引入 $k–1$ 条边,要使剩余 $d–1$ 个节点保证和查询点相连,则至少需要 $d–1$ 条边,因此,剩余边数 e 要满足:$e≥(d–1)+(k–1)=d+k–2$,终止条件得证.

抽象地看待整个贪心算法过程,它主要分两个步骤.

Step 1. 构建首个 k 团.

Step 2. 每次引入 1 个新点和 x 条边以引入最多的 k 团,直到剩余点和边数违背终止条件.

下面我们证明本文提出的贪心策略所得的解即最优解.要证明这个问题,我们需要证明:(1) 有一个最优解以贪心选择开始,即最优解包含初始选择 Step 1;(2) 最优子结构性质,即一个问题的最优解包含其子问题的最优解.

定理 1. 给定 d 个节点和 e 条边,在保证图连通性的情况下,通过每引入一个点要引入最多个 k 团的方式构建图可以得到最多数量的 k 团.

证明:首先证明贪心选择性,假设最多可以构造出 c 个 k 团,若 $c>0$,则第 1 步可以得到 1 个 k 团,必然包含在 c 中;若 $c=0$,则第 1 步无法构造出 k 团,因而得到 0 个 k 团,也包含在 c 中.接下来证明最优子结构性质.假设 c 是所对应的最多 k 团数,c'是 d 个点 e 条边的上一步 d'个点 e'条边所能构成的 k 团数,因此 $c=c'+C_x^{k-1}$,若 c'不是最多所能构成的 k 团数,即存在 $c''>c'$,则 $c''+C_x^{k-1}>c$,即 c 不是最优解,与原假设矛盾,因此,c'是 d'个点和 e'条边的最优解,最优子结构得证.因此,本贪心策略可以得到最优解. □

贪心策略最多 k 团构建求解算法的时间复杂度为 $O(d)$,其中 d 为每个点的度,因此,近似算法中对每个点进行打分这一执行过程的时间复杂度是线性的,而精确算法中对每个点进行重叠社区搜索的时间复杂度是 $O(C_N^k)$,其中 N 是被搜索社区中所有点的数量.由此可以看出,近似算法的时间代价被极大地降低.此外,由于贪心策略的输入是每个点的邻点个数和邻点之间的边数,因此可以对图进行预处理,将这些信息预先存储在每个节点中,以降低实时计算的时间代价.我们将通过实验来衡量近似算法的效率和准确度.

5 实验结果与分析

5.1 实验设置与数据集

实验环境采用 Intel Core2 2.67GHz 处理器,4G 内存,32 位 Windows 7 操作系统,所有算法均采用 C++语言实现.

本文采用 4 个真实的社交网络作为实验数据集,这些网络的数据规模统计见表 1.豆瓣网是国内的一个社交网站,该网站提供图书、影视、音乐等作品信息,用户可以对这些作品进行收藏、评论、打分等操作,豆瓣数据集中包括用户的关系网络,以及 1 周之内用户对电影的评价(包括时间戳),该数据集由爬虫程序抓取.DBLP 是一个学术领域的合作者网络,该数据来源于近期公布的 DBLP 数据库快照,DBLP 网络中的节点代表文章的作者,边代表作者间的合作关系.LiveJournal 提供了 LiveJournal 网站中用户间的朋友关系.该网站是一个在线博客社区,用户在网站上会建立朋友关系.Friendster 是一个在线游戏网站.该网站曾经是一个社交网站,用户之间也会建立朋友关系.我们从斯坦福大型网络数据集集合中获取了 LiveJournal 和 Friendster 的数据集.我们用豆瓣网数据集评价本文所提出方法的有效性,用 DBLP,LiveJournal 和 Friendster 数据集来评价算法的性能.

Table 1 Social networks for experiments

表 1 实验所用社交网络

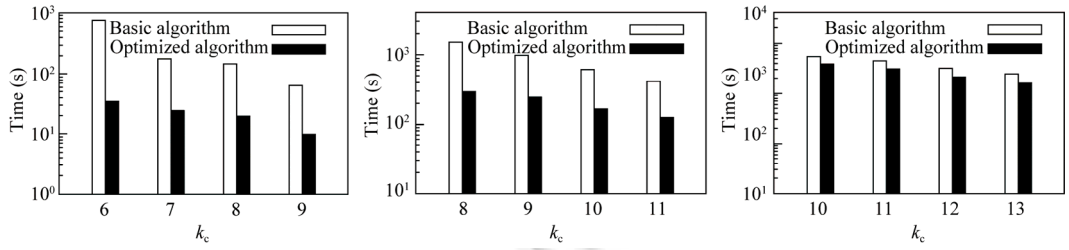
数据集	点数	边数	平均度
豆瓣网	35 216	425 328	12.1
DBLP	968 956	4 826 365	9.96
LiveJournal	3 997 962	34 681 189	17.4
Friendster	4 856 981	78 051 642	32.1

5.2 实验结果与分析

本文通过在 DBLP,LiveJournal 和 Friendster 这 3 个数据集上进行大量的实验来评估传播热点选择精确算法和近似算法的执行效率以及近似算法的准确率.从算法 2、算法 3 中可以看出,算法的性能与两个输入相关:一是已知社区的规模,二是 k 团的参数 k_c.因此,我们分别通过调节这两个输入来测试算法的性能.已知社区的规模用已知社区所包含的所有节点来定义,其用公式可以表示为 $N=\sum_{i=1}^{m}|C_i|$. 由于传播节点数量参数 k 对算法性能影响基本可以忽略不计,因此下面每组实验我们固定选取 5 个传播热点.

首先调节参数 k_c 来比较基本方法和优化方法这两种精确方法的执行效率.实验设置方法是:对每个 k_c,随机选取 10 个节点,将这 10 个节点所隶属的社区作为已知社区,对于 3 个数据源,分别选取社区规模 N 为 50 个,70 个,90 个节点(±5 个节点)的已知社区作为实验的已知社区输入,每个数据源选取 5 组已知社区数据进行重复实

验.对于基本算法和优化算法中涉及的对每个节点进行的重叠社区搜索操作,我们限定每个节点的搜索执行时间为 60s,若超过 60s 搜索仍未结束则强行终止.实验结果如图 4 所示.优化方法的效率要优于基本方法:在 DBLP 数据源中,优化方法的效率比基本方法提高了 85%~95%;在 LiveJournal 数据源中,优化方法的效率提高了 70%~81%;在 Friendster 中,优化方法的效率提高了 28%~34%.由于每个点的最大执行时间被设置为 60s,而 Friendster 数据集点和边数分别是百万级和千万级,且每个点的平均度高达 32.1,因此在该数据集中,执行过程有被终止的情况,所以实际上的优势可能要高于这个值.然而可以注意到,从图 4(a)~图 4(c),优化算法的效率提高呈衰减趋势,这是由于随着数据集单点平均度的增加,大部分节点可能都同时隶属于多个社区,因此社区的内部节点相对减少,所以优化算法的优化机制受限.

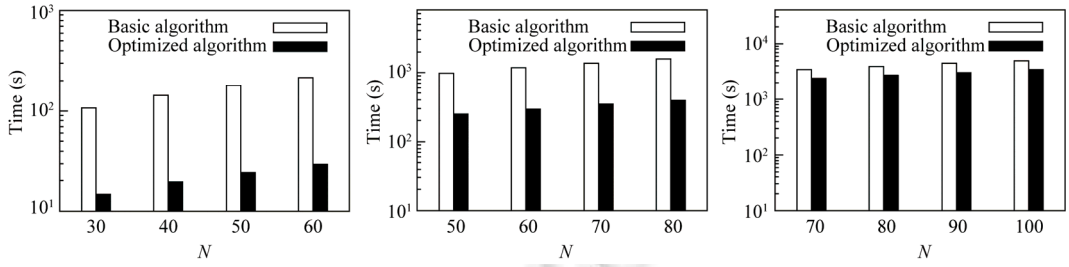


(a) DBLP (N=50) (b) LiveJournal (N=70) (c) Friendster (N=90)

Fig.4 Efficiency of basic algorithm and optimized algorithm with varying k_c

图 4 不同 k_c 下基本方法和优化方法的效率

接下来调节已知社区规模 N 来比较两种精确方法.实验设置的方法是:对于 3 个数据源,分别固定其参数 k_c 为 7,9,11,然后将每个数据源的已知社区规模 N 划分为 4 个区间,每个区间允许相差±5 个节点,且每个区间选取 5 组已知社区,这些已知社区的获取方法与图 4 相同,同样限定每个节点的搜索执行时间为 60s.实验结果如图 5 所示.优化方法的效率仍优于基本方法:在 DBLP 数据源中,优化方法的效率比基本方法提高了 86%左右;在 LiveJournal 数据源中,优化方法的效率提高了 75%左右;在 Friendster 中,优化方法的效率提高了 31%左右.另外,随着已知社区规模 N 的增大,优化方法的耗时增长要比基本方法的耗时增长缓慢,这说明了我们省略内部节点的优化策略的有效性.



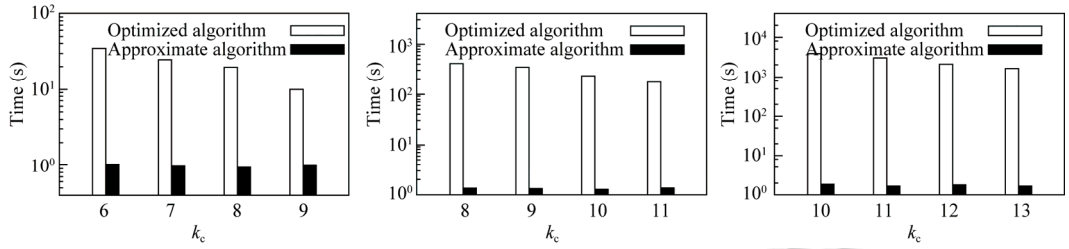
(a) DBLP (k_c=7) (b) LiveJournal (k_c=9) (c) Friendster (k_c=11)

Fig.5 Efficiency of basic algorithm and optimized algorithm with varying N

图 5 不同 N 下基本方法和优化方法的效率

下面比较优化的精确算法和近似算法的效率.首先调节参数 k_c,已知社区规模的设置方法与图 4 相同,对 3 个数据源仍选取 50,70,90 大小的社区作为已知社区输入,实验结果如图 6 所示.近似方法的执行效率要远远优于优化算法:在 DBLP 数据源中,近似方法的效率提高了 90%~97%;在 LiveJournal 数据源中,近似方法的效率提高了 99.2%~99.6%;在 Friendster 数据源中,近似方法的效率提高了近 100%.与优化方法不同,近似方法的效率并不受 k_c 影响,并没有像优化算法一样随着的 k_c 增大而减小;并且近似方法也不受数据源规模的影响,对于平均度

亦不敏感,无论是平均度为 10 的 DBLP 数据源还是平均度为 32 的 Friendster 数据源,近似方法的执行时间都基本相等.由此可以看出,近似算法极大地提高了执行效率.

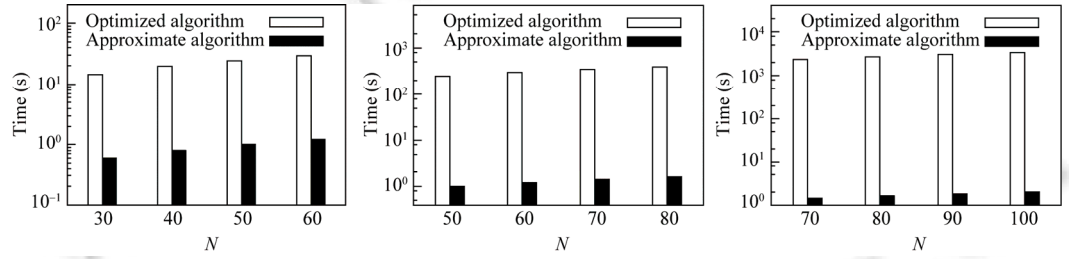


(a) DBLP (N=50) (b) LiveJournal (N=70) (c) Friendster (N=90)

Fig.6 Efficiency of optimized algorithm and approximate algorithm with varying k_c

图 6 不同 k_c 下优化方法和近似方法的效率

接下来调节已知社区规模 N 比较优化算法和近似算法的效率.对于 3 个数据源,分别固定其 k_c 为 7,9,11,已知社区规模的调节方法与图 5 相同,实验结果如图 7 所示.近似方法的执行效率仍远高于优化算法:在 DBLP 数据源中,近似方法比优化方法提高了 4 个数量级;在 LiveJournal 数据源中,近似方法比优化方法提高了 5 个数量级;在 Friendster 数据源中,近似方法比优化方法提高了 6 个数量级.由于优化方法受节点平均度的影响较大,因此近似方法提高的效率在 3 个数据源上依次递增.从图中可以看出,近似算法随已知社区规模 N 呈线性增长.这组实验进一步验证了近似方法的效率得到了显著的提高.



(a) DBLP (k_c=7) (b) LiveJournal (k_c=9) (c) Friendster (k_c=11)

Fig.7 Efficiency of optimized algorithm and approximate algorithm with varying N

图 7 不同 N 下优化方法和近似方法的效率

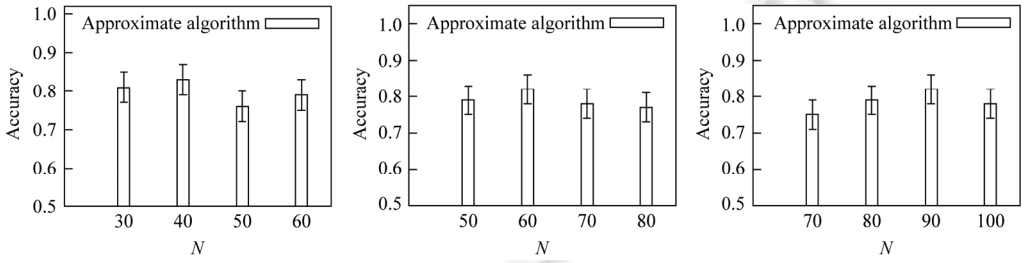
然后验证近似算法的准确率.对于 3 个数据源,固定其 k_c 为 7,9,11,已知社区规模的调节方法与图 5 相同,近似算法平均和偏差准确度如图 8 所示.在 DBLP 数据源中,近似方法的准确度在 76%~83%之间;在 LiveJournal 数据源中,近似方法的准确度在 77%~82%之间;在 Friendster 数据源中,近似方法的准确度在 75%~82%之间.实验数据显示,近似算法的准确度至少在 70%以上,且波动较小,基本稳定,与数据集规模和已知社区规模基本不相关,因此健壮性较好.与近似算法效率提高的幅度相比,其准确度是较为理想的,可以被接受.图 7 和图 8 的实验验证了近似方法打分策略的合理性和可行性.

本文通过豆瓣数据集来验证迭代式推广模型下基于重叠社区搜索的传播热点选择策略的有效性.对比的两种方法分别是:Kempe 等人提出的传统的解决影响力最大化问题的贪心方法(Greedy)[1],该方法基于独立级联模型,采用爬山算法在传播过程开始前一次性选取传播热点,本实验中每次选取节点进行 10 000 次仿真;迭代式推广模型下基于重叠社区搜索的传播热点选择方法,该方法以天为循环周期,对每天的数据进行一次迭代计算选取传播热点.

由于本文提出的影响力最大化模型没有采用基于概率的模型,因此无法用激活点期望值来衡量影响力最

大化模型的有效性.因此,对于上述两种方法,采用如下方法来评价算法选取的传播节点的影响力.在此简单认为,若一个用户的邻居在该用户之后观看了相同的电影,则是受到了该用户的影响,此方法虽不够严谨,但用来作为标准统一评价两种算法,亦不失公平性.对于用户网络中的每个节点,对不同的电影进行分别统计,统计的是一个用户 u 的所有在该用户对某部电影 m 进行评价之后对同一部电影进行评价的邻居节点数量,记为 I_u^m, 对于给定预算值 B,将 B 个 I_u^m 值最大的节点之和作为理论最佳影响力值,将算法实际选出的 B 个节点的 I_u^m 之和与之相比,该比值作为影响力衡量比(IMR),如公式(4)所示.

$$IMR_m=\frac{\sum_{u\in R}I_u^m}{\sum_{u\in top-B}I_u^m} \tag{4}$$



(a) DBLP (k_c=7) (b) LiveJournal (k_c=9) (c) Friendster (k_c=11)

Fig.8 Accuracy of approximate algorithm with varying N

图 8 不同 N 下近似方法的准确率

两种算法的有效性对比实验结果如图 9 所示.可以看出,基于重叠社区搜索的传播热点选择方法的影响力衡量比要高于基于独立级联模型的贪心方法,并且随着预算值 B 的增大,传播热点选择方法的影响力衡量比逐渐增大,最大可达到 65%左右,而贪心方法的影响力衡量比并没有随着预算值而变化,基本维持在 40%左右.该结果验证了本文提出的基于重叠社区搜索的迭代式传播热点选择方法的有效性要优于传统的影响力最大化方法.

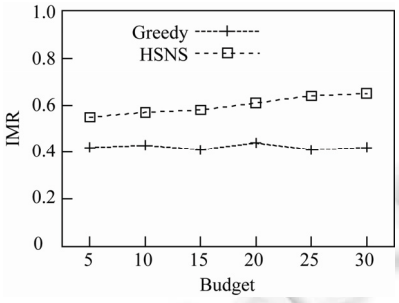


Fig.9 Comparing the effectiveness of Greedy and HSNS methods under different B

图 9 比较不同 B 下贪心算法和传播热点选择方法的有效性

6 结 论

本文研究了社交网络背景下信息传播问题中的影响力最大化问题.与现有的大多数基于概率的独立级联模型一次性选出所有目标推广节点不同,本文提出了一种基于重叠社区结构的方法来衡量节点的影响力,基于这种影响力衡量方法的迭代式推广模型根据用户行为反馈逐步选择影响力最大化节点,使社交网站平台在信息传播过程中充分发挥控制作用,为广告客户从用户中选取对其最有价值的推广目标.为此,我们提出了迭代式推广模型以及基于重叠社区结构的传播热点选择方法.该方法的精确方法包括基本方法以及优化方法.由于精

确方法是 NP 难问题,其时间复杂度较高,应用性较差,因此,我们提出了一种基于打分机制的近似方法,该方法能够快速、有效地选取传播热点.同时,通过大量实验验证了精确和近似算法的效率和准确度以及迭代式传播热点选择方法的有效性.

References:

[1] Kempe D, Kleinberg J, Tardos E. Maximizing the spread of influence through a social network. In: Proc. of the 9th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2003. 137−146. [doi: 10.1145/956750.956769]

[2] Palla G, Derényi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 2005,435(7043):814−818. [doi: 10.1038/nature03607]

[3] Chen W, Wang Y, Yang S. Efficient influence maximization in social networks. In: Proc. of the 15th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2009. 199−208. [doi:10.1145/1557019.1557047]

[4] Leskovec J, Krause A, Guestrin C, Faloutsos C, VanBriesen J, Glance N. Cost-Effective outbreak detection in networks. In: Proc. of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2007. 420−429. [doi: 10.1145/1281192.1281239]

[5] Chen W, Wang C, Wang Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: Proc. of the 16th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2010. 1029−1038. [doi: 10.1145/1835804.1835934]

[6] Jung K, Heo W, Chen W. Irie: Scalable and robust influence maximization in social networks. In: Proc. of the 12th Int'l Conf. on Data Mining. 2012. 918−923. [doi: 10.1109/ICDM.2012.79]

[7] Kim J, Kim SK, Yu H. Scalable and parallelizable processing of influence maximization for large-scale social networks? In: Proc. of the 29th Int'l Conf. on Data Engineering. 2013. 266−277. [doi: 10.1109/ICDE.2013.6544831]

[8] Liu B, Cong G, Xu D, Zeng Y. Time constrained influence maximization in social networks. In: Proc. of the 12th Int'l Conf. on Data Mining. 2012. 439−448. [doi: 10.1109/ICDM.2012.158]

[9] Cao JX, Dong D, Xu S, Zheng X, Liu B, Luo JZ. A *k*-core based algorithm for influence maximization in social networks. Ji Suan Ji Xue Bao/Journal of Computers, 2015,38(2):238−248 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2015.00238]

[10] Domingos P, Richardson M. Mining the network value of customers. In: Proc. of the 7th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2001. 57−66. [doi: 10.1145/502512.502525]

[11] Hartline J, Mirrokni V, Sundararajan M. Optimal marketing strategies over social networks. In: Proc. of the 17th Int'l Conf. on World Wide Web. 2008. 189−198. [doi: 10.1145/1367497.1367524]

[12] Mossel E, Roch S. Submodularity of influence in social networks: From local to global. Society for Industrial and Applied Mathematics Journal on Computing, 2010,39(6):2176−2188. [doi: 10.1137/080714452]

[13] Lin SY, Hu QB, Wang FJ, Yu PS. Steering information diffusion dynamically against user attention limitation. In: Proc. of the 2014 IEEE Int'l Conf. on Data Mining. 2014. 330−339. [doi: 10.1109/ICDM.2014.131]

[14] Tong GM, Wu WL, Tang SJ, Du DZ. Adaptive influence maximization in dynamic social networks. IEEE/ACM Trans. on Networking, 2016,PP(99):1−14. [doi: 10.1109/TNET.2016.2563397]

[15] Bakshy E, Eckles D, Yan R, Rosenn I. Social influence in social advertising: Evidence from field experiments. In: Proc. of the 13th ACM Conf. on Electronic Commerce. 2012. 146−161. [doi: 10.1145/2229012.2229027]

[16] Aslay C, Lu W, Bonchi F, Goyal A, Lakshmanan LVS. Viral marketing meets social advertising: Ad allocation with minimum regret. Proc. of the VLDB Endowment, 2015,8(7):814−825. [doi: 10.14778/2752939.2752950]

[17] Abbassi Z, Bhaskara A, Misra V. Optimizing display advertising in online social networks. In: Proc. of the 24th Int'l Conf. on World Wide Web. 2015. 1−11. [doi: 10.1145/2736277.2741648]

[18] Adamcsek B, Palla G, Farkas I J, Derényi I, Vicsek T. Cfinder: Locating cliques and overlapping modules in biological networks. Bioinformatics, 2006,22(8):1021−1023. [doi: 10.1093/bioinformatics/btl039]

[19] Ahn YY, Bagrow JP, Lehmann S. Link communities reveal multiscale complexity in networks. Nature, 2010,466(7307):761−764. [doi: 10.1038/nature09182]

[20] Evans T, Lambiotte R. Line graphs, link partitions, and overlapping communities. Physical Review E, 2009,80(1):No.016105. [doi: 10.1103/PhysRevE.80.016105]

[21] Lim S, Ryu S, Kwon S, Jung K, Lee JG. Linkscan*: Overlapping community detection using the link-space transformation. In: Proc. of the 30th Int'l Conf. on Data Engineering. 2014. 292−303. [doi: 10.1109/ICDE.2014.6816659]

[22] Gregory S. Finding overlapping communities in networks by label propagation. New Journal of Physics, 2010,12(10):No.103018. [doi: 10.1088/1367-2630/12/10/103018]

[23] Šubelj L, Bajec M. Unfolding communities in large complex networks: Combining defensive and offensive label propagation for core extraction. Physical Review E, 2011,83(3):No.036103. [doi: 10.1103/PhysRevE.83.036103]

[24] Hu Y, Wang CJ, Wu J, Xie JY, Li H. Overlapping community discovery and global representation on microblog network. Ruan Jian Xue Bao/Journal of Software, 2014,25(12):2824−2836 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4721.htm [doi: 10.13328/j.cnki.jos.004721]

[25] Ball B, Karrer B, Newman MEJ. An efficient and principled method for detecting communities in networks. Physical Review E, 2011,84(3):No.036103. [doi: 10.1103/PhysRevE.84.036103]

[26] Shen HW, Cheng XQ, Guo JF. Exploring the structural regularities in networks. Physical Review E, 2011,84(5):No.056111. [doi: 10.1103/PhysRevE.84.056111]

[27] Gopalan PK, Blei DM. Efficient discovery of overlapping communities in massive networks. Proc. of the National Academy of Sciences, 2013,110(36):14534−14539. [doi: 10.1073/pnas.1221839110]

[28] Sun BJ, Shen HW, Cheng XQ. Detecting overlapping communities in massive networks. Europhysics Letters, 2014,108(6):No. 68001. [doi: 10.1209/0295-5075/108/68001]

[29] Cui W, Xiao Y, Wang H, Lu Y, Wang W. Online search of overlapping communities. In: Proc. of the 2013 ACM SIGMOD Int'l Conf. on Management of Data. 2013. 277−288. [doi: 10.1145/2463676.2463722]

[30] Shan J, Shen DR, Nie TZ, Kou Y, Yu G. An efficient approach of overlapping communities search. In: Proc. of the 20th Int'l Conf. on Database Systems for Advanced Applications. 2015. 374−388. [doi: 10.1007/978-3-319-18120-2_22]

[31] Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature, 1998,393(6684):440−442. [doi: 10.1038/30918]

附中文参考文献:

[9] 曹玖新,董丹,徐顺,郑啸,刘波,罗军舟.一种基于 *k*-核的社会网络影响最大化算法.计算机学报,2015,38(2):238−248. [doi: 10.3724/SP.J.1016.2015.00238]

[24] 胡云,王崇骏,吴骏,谢俊元,李慧.微博网络上的重叠社群发现与全局表示.软件学报,2014,25(12):2824−2836. http://www.jos.org.cn/1000-9825/4721.htm [doi: 10.13328/j.cnki.jos.004721]

单菁(1986−),女,辽宁沈阳人,博士,讲师,CCF 学生会员,主要研究领域为社交网络,个性化 推荐.

聂铁铮(1980−),男,博士,副教授,CCF 专业会员,主要研究领域为数据质量,数据集成.

申德荣(1964−),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为分布式数据管理,数据集成.

于戈(1962−),男,博士,教授,博士生导师,CCF 会士,主要研究领域为数据管理理论与技术,分布与并行系统.

寇月(1980−),女,博士,副教授,CCF 专业会员,主要研究领域为实体搜索,数据挖掘.