软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(Suppl.(2)):1−11

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

基于社交网络弱连接属性的影响力最大化算法*

易秀双, 胡金林, 王兴伟

(东北大学 信息科学与工程学院,辽宁 沈阳 110819)

通讯作者: 易秀双, E-mail: xsyi@mail.neu.edu.cn

摘 要: 首先研究了目前影响力最大化问题的解决方案,并总结了这些解决方案的优缺点.对社交网络中弱连接的研究之后发现,弱连接可以有效地打通社交网络中不同社团之间的信息壁垒,使得信息在不同社区间流通.利用弱连接的这一作用,同时基于贪心思想,提出 BWTG(base-on weak tie greedy)算法来解决影响力最大化问题,并根据解空间的不同,把 BWTG 算法分为 BCWTG(base-on complete weak tie greedy)和 BNCWTG(base-on not complete weak tie greedy)两种算法.影响力最大化问题的传统评价指标有两种:时间复杂度和最终激活节点数,但考虑到实际情况,定义了 ANNI(actived nodes/node influence)这一新的评价指标,用于衡量回报与付出之比.为了验证 BCWTG 和 BNCWTG 算法的性能,在不同类型、不同规模的真实数据集中对算法进行实验验证,在时间复杂度、最终激活节点数和 ANNI 这 3 个方面与经典的 Greedy 算法进行对比,实验结果表明,BCWTG 算法和 BNCWTG 算法在运算时间和 ANNI 方面有所提高,最终激活节点数方面却弱于 Greedy 算法,但当满足一定条件时,BCWTG 和 BNCWTG 算法在最终激活节点数方面也能接近 Greedy 算法.

关键词: 社交网络;弱连接;影响力最大化;节点影响力;关系强度

中文引用格式: 易秀双,胡金林,王兴伟.基于社交网络弱连接属性的影响力最大化算法.软件学报,2016,27(Suppl.(2)):1−11. http://www.jos.org.cn/1000-9825/16012.htm

英文引用格式: Yi XS, Hu JL, Wang XW. Influence maximization algorithm based on the weak tie in social network. Ruan Jian Xue Bao/Journal of Software, 2016,27(Suppl.(2)):1−11 (in Chinese). http://www.jos.org.cn/1000-9825/16012.htm

Influence Maximization Algorithm Based on the Weak Tie in Social Network

YI Xiu-Shuang, HU Jin-Lin, WANG Xing-Wei

(School of Information Science and Engineering, Northeast University, Shenyang 110819, China)

Abstract: This thesis introduced the solution of influence maximization and analyzed the advantages and disadvantages of those solutions. After studying the weak tie in social network, it is found that weak ties can effectively break the information barriers between different societies in social network and make information circulate in different societies. Making use of the weak tie's advantages, this thesis proposes a new solution, the BWTG algorithm, based on the greedy thought to resolve influence maximization problem. According to different solution spaces, the BWTG algorithm is divided into two different types: BCWTG and BNCWTG algorithm. There are two traditional evaluation indexes of influence maximization problem, namely, time complexity and the final activated nodes number. But considering the practical situation, a new evaluation index named ANNI is proposed to measure the ratio of profit and pay. Besides, in order to verify the performance of the proposed algorithm, different scales and types of data are used to carry out the experiment. The time complexity, the final activated nodes number and ANNI are compared with the classical Greedy algorithm. The experimental result finds

* 基金项目: 国家杰出青年科学基金(61225012); 国家自然科学基金(61070162, 71071028, 70931001); 高等学校博士学科点专项科研基金(20120042130003, 20100042110025, 20110042110024); 中央高校基本科研业务费专项资金(N110204003, N120104001)

Foundation item: National Science Fund for Distinguished Young Scholars of China (61225012); National Natural Science Foundation of China (61070162, 71071028, 70931001); Specialized Research Fund for the Doctoral Program of Higher Education of China (20120042130003, 20100042110025, 20110042110024); Fundamental Research Funds for the Central Universities (N110204003, N120104001)

收稿时间: 2015-05-31; 采用时间: 2016-01-05

that BCWTG and BNCWTG algorithm have lower time complexity and higher ANNI, but lower final activated nodes number than Greedy algorithm. But under some certain conditions, BCWTG and BNCWTG can be almost equal to Greedy in activated nodes number.

Key words: social network; weak ties; influence spread maximization; node influence; tie strength

1 相关背景

随着互联网技术的发展,在线社交网站迅速崛起并吸引着数以亿计的用户,如成立于 2004 年的社交网站 Facebook,截止到 2014 年 1 月,Facebook 的月活跃用户已达到 12 亿,日活跃用户达 7.5 亿.其他著名社交网站还有国外的 Twitter.国外大型在线网站有 YouTube,雅虎等,国内的如新浪微博,QQ 等.同时还有像 Netflix,豆瓣这样的网站,可以对电影进行评价,以供他人选择影片.这些在线社交网站不但丰富了人们的生活,拓宽了人们的交友方式,更为商业市场营销以及学术研究提供了有利条件.

欧莱礼媒体公司总裁兼 CEO 提姆•奥莱理提出了病毒营销(viral marketing)的营销方式,这种方式是在新产品销售之前,选定几个初始体验用户,通过人与人之间的“口碑效应(word-of-mouth effect)”进行产品的宣传,这种传播方式主要利用用户之间直接的、可信任的特点进行信息扩散.社交网络中的“六度分割理论”认为一个人和任何一个陌生人之间所间隔的距离不会超过 5 个,也就是说,最多通过 5 个中间人你就能够认识任何一个陌生人.这样就大大增加了病毒营销的传播范围.社交网站出现后使得病毒营销发挥了巨大的优势.

Richardson 和 Domingos 第一次提出了把“病毒营销”应用在社交网络中后,影响力最大化得到了很多学者的关注.他们首先提出了一个这样的算法问题:如果我们选出一些初始体验者来体验一个新的产品或新概念,目标是能触发一个大的信息级联来推广新产品或概念,那么我们应该怎样选取这些初始体验者集合?

问题提出后,Kempe[1]等人首次提出了运用贪婪思想解决这一问题,Kempe 等人的主要贡献有:(1) 首次正式地定义了影响力最大化问题,并详细描述了两个经典的信息扩散模型,LT 模型和 IC 模型;(2) 首次证明了影响力最大化问题是 NP 难问题;(3) 运用子模块特性,文献[1]的贪婪求解策略在多个经典模型上验证都至少到达最佳解的 63%.

但是,该 Kempe 贪婪算法存在的主要问题是算法时间复杂度高,不能适应大型社交网络.针对 Kempe 贪婪算法的不足之处,一些研究者对其进行了改进.Leskovec 等人[2]证明了很多问题,如影响力最大化问题,其最大化目标具有“子模性”,“子模”特性是当向初始集合 S 添加一个节点 v 后,激活节点个数会比加入 v 之前集合 S 所获得的激活节点多.他们利用“子模”特性提出 CELF(cost-effective lazy forward)算法,经实验观察发现,在迭代计算时,大多数情况下,节点 v 的最大收益,即 $\sigma(v|S)=\sigma\left(S\bigcup\{v\}\right)-\sigma(s)$ 在连续多轮都没有明显的改变,所以 CELF 算法使用惰性评估的方法而不是多次迭代来计算候选激活节点的 $\sigma(v|S)$ 平均值.通过在真实数据集上的实验,该算法获得了近似最优的结果,更重要的是,CELF 算法比 Kempe 贪婪算法在时间上快近 700 倍,大大提高了算法时间复杂度.

这之后,Chen 等人[3]提出了对 Kempe 贪婪算法进行改进的算法:NewGreedy 算法和 MixGreedy 算法,并提出基于度数的启发式算法 DegreeDiscount.其中,MixGreedy 算法混合了 CELF 算法和 DegreeDiscount 算法.改进前,Kempe 贪婪算法时间复杂度是 $O(knRm)$,其中,k 是要选择的初始节点数,n 为网络中节点个数,m 为边的个数,R 为仿真次数.改进后,NewGreedyIC(基于 IC 信息扩散模型)算法的时间复杂度是 $O(kRm)$,NewGreedyWC(基于 WC 信息扩散模型)算法的时间复杂度是 $O(kRTm)$,其中,T 是迭代次数,而 DegreeDiscountIC 算法时间复杂度到达了 $O(k\log n+m)$.说明改进后的算法在时间复杂度方面确实得到降低.通过在真实社交网络数据集的实验显示:MixGreedy 算法相对于 CELF 算法有 15%~34%的优化.最近,Chen 等人[4]又提出了 PMIA 算法,该算法使用于 IC 模型.在实验中,该算法比 DegreeDiscount 算法获得了多于 3.9%~6.6%的影响力扩散,但时间复杂度却高于 DegreeDiscount 算法.LDAG 算法[5]与 PMIA 算法相似,但只使用于 LT 模型.除了利用贪婪思想,还有研究者提出了基于社区的算法[6,7].

在之前的算法中,都没有考虑到以下两点:

(1) 弱连接在信息级联中的作用,弱连接(weak tie)这一词是由美国社会学家 Granovetter[8,9]于 1973 年首次提出来的.

(2) 在选择初始节点集时,我们只考虑到选择的节点要达到最终激活节点数最大化,但没有考虑到激活这些初始节点要付出的代价.

本文将基于以上两点提出基于弱连接的影响力最大化算法 BWTG(base-on weak tie greedy).其创新点总结如下:

(1) 考虑到实际应用情况,提出了 ANNI(actived nodes/node influence)评价指标,该指标用于评价影响力最大化算法中最终收益与付出的成本比值.可表示净收益的概念;

(2) 分析弱连接在信息级联中的作用,并基于弱连接在信息扩散中的作用,提出新的影响力最大化算法,并通过真实社交网络数据集实验验证了算法性能.

2 弱连接及节点影响力

我们知道“病毒营销”这一营销模式,是首先在众多用户中选择 k 个初始用户对新产品进行试用,然后通过“口碑效应”使新产品在用户中自行扩散(推广).其中存在两个问题:一是如何选择这 k 个初始用户,也就是我们要解决的问题,即影响力最大化问题;二是信息(新产品)是如何在用户间进行扩散(推广)的,什么情况下信息级联过程终止.

我们先假设信息传播规则:假设用户是否使用某种新产品的关键在于其朋友中使用这种产品所占的比例 p_i,当 p_i 大于或等于某个阈值 θ 时,该用户被激活,即尝试使用这个新产品,故激活规则公式如下:

$$p_i \geqslant \theta \tag{1}$$

其中,$\theta \in [0,1]$,p_i 可按如下公式计算:

$$p_i = \frac{\text{激活邻居节点数}}{\text{所有邻居节点数}} \tag{2}$$

由于社交网络不同于随机网络,它有一些重要的性质,如社团性.

我们可以看出,如图 1 所示的社交网络可以被分成 3 个社区:社区 C_1,社区 C_2 和社区 C_3.社区 C_1={1,2,3},社区 C_2={4,5,6,7,8,9,10},社区C_3={11,12,13,14,15,16,17}.信息扩散结果图中的结果显示信息只在社区 C_2 中得到了扩散,其他两个社区 C_1 和 C_3 都没有节点被激活.同质性是解释这一现象的原因,同质性想要表达在社交网络中个体及其朋友之间往往具有相似的特点这一现象.在交友过程中我们的朋友并不是随机得到的,而是有一定的挑选条件,比如:性格、年龄、种族、文化背景等,当我们处于同一年龄段时,可能有着相似的经济收入、兴趣爱好等.当有这些相似的文化背景时,造成误解的可能性就会减小,容易相处融洽.这些相似之处往往可能使我们彼此相互理解、彼此感兴趣,最后会成为朋友.

如图 1 所示,由于同质性现象,每个社区中的个体都表现出了相似性,也就是说,具有相同特征的个体会自动结合成一个社区,具有不同特征的个体组成了不同的社区,因而当网络中某个社区中的个体体验了某种新产品,个体的同质性使得新产品很容易在本社区内推广.同理,个体的异质性使得信息很难在不同社区内进行传递,因为相同社区内的个体交流见面的机会远远高于不同社区之间成员交流见面的机会.这一结果验证了信息容易在同一个社区中进行扩散,但很难在不同社区之间进行扩散这一理论.

综上所述,我们知道同质性和社交网络中的社团结构是信息级联扩散过程终止的主要因素,因为人们倾向于与相识的人互动交流,而信息新鲜事物往往来自不同社区,因而相同的社区易于扩散信息,而不同社区间不易于信息的交流.在了解了信息扩散停止的原因后,考虑如何使得信息能跨越不同的社区就成为我们的研究重点.

社交网络中每个节点在信息扩散的过程中起到的作用是不同的,可以通过节点影响力这一概念说明节点在一个社区中的作用.社交网络中节点影响力的计算早已成为一个研究热点.

网络中节点的度数是节点影响力的一个衡量标准.在无向网络中,节点的入度和出度相同,不加以区分.在有向网络中,节点的出度和入度意义不同,例如在微博网络中,节点的入度表示哪些节点“关注”了该节点,节点的

出度表示该节点被哪些节点关注,即该节点的粉丝数,在这样的网络中一般选用节点的出度,也就是说用节点的粉丝数来衡量节点的影响力,其计算公式如下:

$$IF(i)=d_i \tag{3}$$

其中,$IF(i)$表示节点 i 的影响力值,当网络为有向图时,d_i 为节点出度值;当网络为无向图时,d_i 为节点度数值.

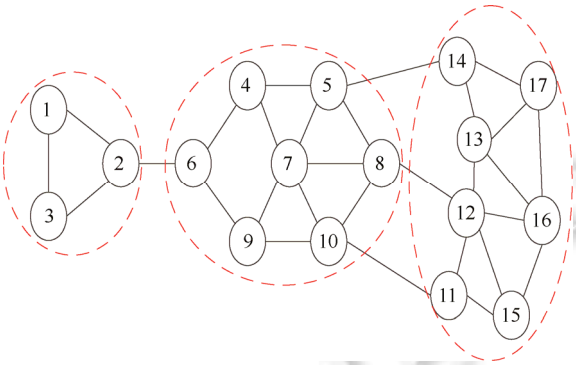

图 1 社区结构

节点影响力的计算方法有很多,本文将在信息扩散模型的研究中使用 PageRank 算法[10]计算节点影响力的值.任一页面 p_j 的 PageRank 值可以概括为

$$PR(p_j)=d\times\sum_{p_j\in M(p_j)}\frac{PR(p_j)}{L(p_j)}+\frac{1-d}{N} \tag{4}$$

其中,网页集 $P=\{p_1,p_2,\ldots,p_N\}$,N 为所有网页数,$M(p_j)$表示网页 p_j 中所包含的链接集合,$L(p_j)$表示网页 p_j 所包含的链接数.d 为阻尼系数.

我们完全可以把 PageRank 算法应用在社交网络中计算节点的影响力上,在微博网络中,微博用户 v_j 可以代替公式(4)中的网页 p_j,N 可以表示网络中的节点数,集合$M(p_j)$被替换为节点 v_j 的关注关系集合 $F(v_j)$.故公式(4)可替换成如下公式:

$$IF(v_j)=d\times\sum_{v_i\in F(v_j)}\frac{IF(v_i)}{L(v_i)}+\frac{1-d}{N} \tag{5}$$

分析节点的影响力有很多应用价值,可以基于节点影响力对在线社交网络作进一步的分析,还可以促进对用户属性的研究.

3 BWTG 影响力最大化算法

3.1 算法基本思想

总的来说,阻止信息级联发生的因素有两点.

(1) 社交网络中具有社区结构.

(2) 社交网络中的个体表现出同质性.

综合以上两种原因可以看出,社交网络中同一社区内的个体表现出同质性,同一个社区内的成员一般在某个属性方面有着相似性,如兴趣爱好、工作性质等,一个社区内的成员交流的信息往往是成员们感兴趣的内容,因而信息容易在同一个社区内进行扩散.而不同社区之间表现出异质性,信息不容易在不同社区中进行扩散,通常整个社交网络是由不同的社区构成,由于社区间的异质性差异使得信息不易在不同社区内传播,从而信息级联不会扩散至整个社交网络.

因而提出的 BWTG 影响力最大化的基本思想是:打通各个社区之间的弱连接,从而使得信息可以在各社区乃至整个网络中得到最大化的扩散.基于此思想提出基于完全弱连接的贪婪(base-on complete weak tie greedy,

简称 BCWTG)算法.该算法首先将这个网络划分成若干个相互独立的社区并找到弱连接集合;然后计算整个社区中节点的影响力,影响力的计算参照公式(3),并找到整个社区中影响力最大的节点;之后利用贪心思想在各社区的影响力最大的节点中找到能产生最大信息级联的节点,加入到初始节点集中,这个节点所在的社区被称为激活社区;最后找到激活社区中的弱连接集合,在该集合中利用贪心思想,选择能产生最大信息级联的节点放入到初始集合中,直到初始集合中元素个数为 k,算法停止.算法中涉及到的符号表示见表 1.

表 1 符号描述表

	符号解释
n	网络中节点的个数
m	网络中边的个数 C
C	整个网络所划分的社区集合, $C=\{C_1,C_2,\ldots,C_{n_c}\}$
C'	激活社区集, $C_i \in C'$
n_c	社区个数
R	节点重复模拟扩散的次数
W	弱连接集合, $w_{uv} \in W$
I	每个社区中影响力最大节点集合

但有时仅仅利用弱连接点无法使信息扩散得到最大化,弱连接可以理解为一个社区网络中的边缘节点,仅仅激活边缘节点,在社区直径较大的情况下很可能得不到理想的信息扩散结果,在一个社区中意见领袖往往在信息扩散过程中起着重要的作用.因此提出另一种基于弱连接的算法,基于不完全弱连接的贪婪(base-on not complet weak tie greedy,简称 BNCWG)算法.该算法利用同时社区间的弱连接节点和社区中的最具影响力的节点共同作用来达到影响力最大化的目的.

3.2 算法描述

下面给出 BCWTG 算法和 BNCWTG 算法的伪代码描述.

在对BCWTG算法进行详细的描述之前,先介绍算法的输入:社交网络图 $G(V,E)$,要选择的初始节点数为 k($k \geqslant 1$),初始节点集合为 A,初始时 A 为空集.

算法 3.1. BCWTG 算法.

```
1.  getUserInfluece(G); //计算每个节点的影响力
2.  C=getGNCommunity(G); //把整个网络划分成n_c个社区
3.  FOR EACH  c∈C  do
4.  InfluenceSet.add(getMaxInfluence();
5.  WeakTieNodesSet.add(getWeakTieNodes);
6.  END FOR
7.  max=0;
8.  FOR EACH u:InfluenceSet do
9.  FOR 1 TO R
10. sum=simmulate(A U u);
11. END FOR
12.  u_0 ← max(sum/R);
13. END FOR
14. A=A U u_0; //向初始节点集中添加节点
15. Update(max,ActivedCommunity,k); //更新激活节点数和激活社区
16. WHILE k !=0 DO //利用贪婪思想,在弱连接中选取初始激活节点
17. FOR EACH u∈weakTieSet:ActivedCommunitySet DO
18. FOR 1 TO R DO
19. sum=simmulate(A U u);
```

20. END FOR
21. END FOR
22. $u_i \leftarrow \max(sum/R)$;
23. $A=A \cup u_i$;
24. *Update*(max,*ActivedCommunity*());
25. END WHILE
26. RETURN max;

算法 3.1 第 1 步是计算网络 $G(V,E)$中节点的影响力,在本算法中将利用公式(3)进行计算,同时会把计算结果保存在文件中以备以后重复计算.

第 2 步,利用 GN 快速社区发现算法[11−13]把整个网络划分成 n_c 个社区,GN 社区发现算法是社区发现算法中的经典算法,该算法利用“边介数(edge betweenness)”进行社区发现,算法过程如下:

(1) 计算网络中所有边的介数;

(2) 移除具有最高介数的边;

(3) 重新计算所有受影响的边的介数;

(4) 重复第 2 步,直到没有边可移除.

得到网络中的社区之后同样要保存在文件中,以防影响整个算法的时间复杂度.

注意算法第 1、2 步属于算法的前期准备工作,只需完成一次,把得到的结果保存在文件中后,就无需重复计算.

第 3~6 步,用于发现社区中影响力最大节点和每个社区中的弱连接节点,分别保存在 *InfluenceSet* 和 *WeakTieNodesSet* 中.

第 8~13 步,是在所有影响力最大节点中找到激活节点数最多的节点,每个节点都模拟 R 次,并取其平均值进行比较.

第 14~15 步,把第 8~13 步中选中的节点加入到初始节点集 A 中,并更新最终影响节点个数 max 值、激活社区集和 k 值.

第 16~25 步,利用贪心思想,在激活社区集中含有的弱连接中找到激活节点数最多的节点,加入到集合 A 中.

最后返回激活节点数 max 的值.

BNCWTG 与 BCWTG 算法大体相似,不同之处在于,BNCWTG 算法在选择初始节点集时不仅仅利用弱连接节点,同时利用每个社区中影响力最大的节点加以辅助,以希望最终得到的激活节点数达到最大化. BNCWTG 算法描述如下.

算法 3.2. BNCWTG 算法.

1. *getUserInfluece*(*G*); //计算每个节点的影响力
2. *C* = *getGNCommunity*(*G*); //把整个网络划分成n_c个社区
3. FOR EACH $c \in C$ do
4. *InfluenceSet.add*(*get*Max*Influence*();
5. *WeakTieNodesSet.add*(*getWeakTieNodes*);
6. END FOR
7. max=0;
8. FOR EACH *u*:*InfluenceSet* do
9. FOR 1 TO R
10. *sum*=*simmulate*($A \cup u$);
11. END FOR
12. $u_0 \leftarrow \max(sum/R)$;
13. END FOR
14. $A=A \cup u_0$; //向初始节点集中添加节点

15. *Update*(max,*ActivedCommunity*,*k*); //更新激活节点数和激活社区
16. WHILE *k* !=0 DO //在弱连接和影响力最大节点集中选取初始激活节点
17. FOR EACH $u \in weakTieSet \cup InfluenceSet$:*ActivedCommunitySet* do
18. FOR 1 TO RDO
19. *sum*=*simmulate*(*A* U *u*);
20. END FOR
21. END FOR
22. $u_i \leftarrow \max(sum/R)$;
23. $A=A \cup u_i$;
24. *Update*(max,*ActivedCommunity*());
25. END WHILE
26. RETURN max;

3.3 ANNI评价指标

就算法本身而言,时间复杂度和最终激活节点数足以评价影响力最大化算法,但考虑其应用,如病毒营销问题,时间复杂度和最终激活节点数不足以完全评价影响力最大化算法,因为考虑到实际网络,如果我们在微博中进行产品的病毒营销,要在网络中选择 k 个初始节点进行扩散,但微博中影响力大小分化及其严重,仅从粉丝数情况而言,每个人的粉丝数差别极大.

如果给出微博网络的网络结构图,影响力最大化算法应该能选中类似这样的用户,如果选择这种影响力最大的用户来激活,其带来的最终激活用户也一定会增加得非常明显,但问题是这样的用户是否能成功被激活(乐于尝试一种新产品)?即使被激活,那么付出的成本(例如,金钱)应该会很高.但是考虑到普通用户,如果选中普通用户去尝试新产品,则所付出的代价会小得多,而且很可能乐于尝试该产品.本文提出如下假设:节点影响力越小,激活该节点所付出的代价越小;影响力越大,激活该节点所付出的代价越大.因而我们提出了 ANNI(active num/average influence)的概念.

ANNI 可以定义为最终激活用户数与选择的 k 个节点的平均影响力之比,其定义公式如下:

$$ANNI = \frac{\max(\sigma(A))}{\overline{IF(v_{k_i})}} \tag{6}$$

其中,$\overline{IF(v_{k_i})}$ 表示经过影响力最大化算法所选出的 k 个初始用户的平均影响力,$\max(\sigma(A))$ 表示最终激活节点数.当 $\overline{IF(v_{k_i})}$ 越小时,$\max(\sigma(A))$ 越大,ANNI 值越大,说明以最小的代价得到最大的影响力,从而使最终收益最大化;反之,若 ANNI 值越小,则最终收益会受到影响.ANNI 可以理解为最终收获的激活节点数与付出的激活成本之比,即“净收入”,这里的激活成本可由节点的影响力决定.

综上所述,我们可以在影响力最大化算法中加入 ANNI 这一评价标准,使得影响力最大化算法能够保证最终激活节点数最多、算法时间复杂度低的同时,也能应对 ANNI 所有要求.

4 实 验

4.1 数据集

为了在不同的社交网络中验证 BNCWTG 与 BCWTG 算法,实验选取了在线社交网络 Facebook 网络数据集[14]和美国电网网络数据集[15]作为实验对象.Facebook 数据集构成的社交网络是由 Facebook 注册用户和好友关系构成,每一行表示一对朋友关系,即表示社交网络图的一条边.

表 2 为 Facebook 社交网络数据集相关属性表,包括网络所包含的节点数、边数、平均度数、最大度数和网络聚类系数.

表 2 数据集相关参数

	Facebook
图形类型	无向无权图
节点数	63 731
边数	817 090
平均度数	25.642
最大度数	1 098
聚类系数	$1.477\ 208\times10^{-1}$

4.2 结果分析

本次实验所采用的硬件环境为:台式电脑,八核 3.4GHz 的 Intel Core i7-2600,4GB 内存.软件平台为:操作系统为 64 位 Win7,编程语言采用 Java.下面分别介绍美国电网数据集和 Facebook 数据集的实验结果.

(1) 美国电网数据集

BWTG 算法的两种算法首先都需要利用 GN 社区发现算法进行社区发现.GN 社区发现算法是社区发现算法中的经典算法.表 3 为由美国电网数据集构成的社交网络社区发现情况.

一般网络模块度在 0.3~0.7 之间.从表 3 可以看出,美国电网社交网络的模块度为 0.93,说明该网络具有强社区结构.该网络的平均社团规模为 120.51,说明美国电网网络社区分布均匀.

如图 2 所示为不同激活节点数 k 与最终激活节点数关系图,激活概率 IC 取值 0.6.图中横坐标为选择的激活节点数,取值范围为 5~50,纵坐标为最终激活节点数.从图中可以看出,3 种算法的最终激活节点数都随着 k 值的增加而增加,但 Greedy 算法的激活节点数要大于其他两种算法,这是因为 Greedy 算法基于贪婪思想,每次都能选取激活节点数最多的节点,而 BCWTG 算法仅利用弱连接节点在网络的社区间扩散节点,相对于 Greedy 算法,其扩散能力较弱,BNCWTG 算法同时在弱连接和影响力强的节点中利用贪心思想选择节点,因而会比 BCWTG 算法得到更多的激活节点数.

表 3 美国电网数据集社区发现结果

统计项	统计值
模块度	0.93
耗时/秒	1
社团数	41
最小社团规模	18
最大社团规模	251
平均社团规模	120.51

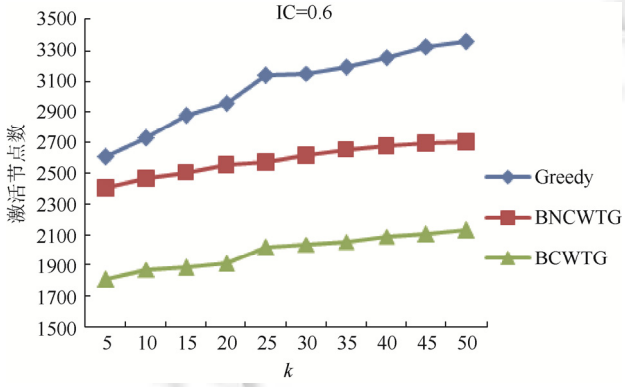

图 2 不同 k 值的最终激活节点数

ANNI 用于衡量最终激活节点数与激活 k 个节点所付出的代价的比值,其值越大,表示最终净收益越高.图 3 所示的 3 种算法在取不同 k 值时,ANNI 值趋势变化,从图中可以看出 BNCWTG 算法在 ANNI 上要高于其他两种算法,而 Greedy 算法 ANNI 值最小,这是因为 ANNI 计算公式中激活初始节点代价值由节点影响力决

定,Greedy 算法虽然得到的最终激活节点数多,但付出的激活代价更大,最终收益较小,而 BNCWTG 算法平衡最终激活节点数与激活代价,所以能获得更大的收益.

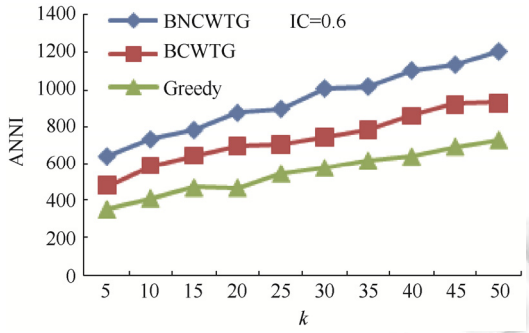

图 3 不同 k 值的 ANNI

(2) Facebook 数据集

从表 4 可以看出,Facebook 数据集提供的 Facebook 社交网络模块度为 0.51,在 0.3~0.7 之间,属于一般社团结构.另外,该网络中社区分布不均匀,最小社团规模为 2,而最大社团规模为 22 150.美国电网社交网络进行社区发现时只用时 1s,而 Facebook 社交网络用时 953s,Facebook 数据集提供的社交网络要比美国电网的网络大得多.因而耗时是美国电网数据集的百倍,这也说明 GN 社区发现算法不适合大型社交网络.

表 4 Facebook 数据集社区发现结果

统计项	统计值
社团数	824
模块度	0.51
耗时/秒	953
最小社团规模	2
最大社团规模	22 150
平均社团规模	73.87

图 4 显示出不同 k 值的最终激活节点数.算法中使用 IC 扩散模型,激活概率 IC 取值 0.6.图中横坐标为待选择的初始节点数 k,取值在 5~50 之间.从图中可以看出,随着 k 值的增加,最终激活节点数在不断增加,BNCWTG 算法和 BCWTG 算法在激活节点数方面要远大于随机算法 Random,且在相同 k 值条件下,BNCWTG 算法能激活最多的节点.

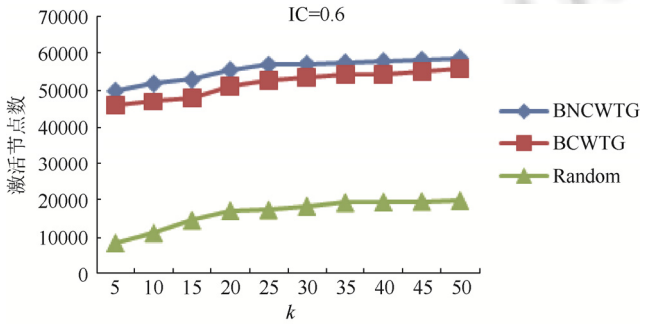

图 4 不同 k 值的激活节点数

图 5 为 Facebook 数据集在不同 k 值情况下 3 种算法的 ANNI 性能比较.从图中可以看出,随着 k 值的不断增加,3 种算法的 ANNI 值都在增加,总体上,BNCWTG 算法和 BCWTG 算法的 ANNI 值高于 Random 算法,说明

BNCWTG 算法和 BCWTG 算法在 ANNI 评价指标上要优于 Random 算法.

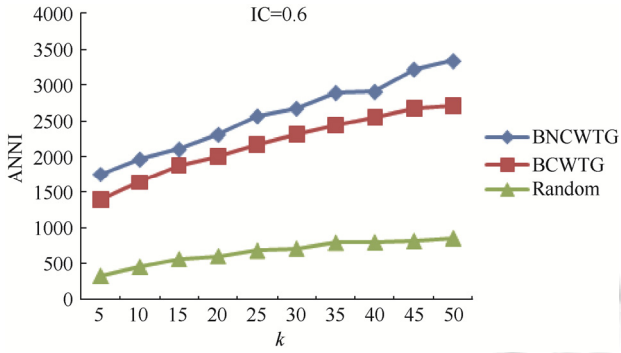

图 5 Facebook 数据集不同 k 值的 ANNI 值

通过在美国电网网络和 Facebook 网络的实验可以发现,提出的 BNCWTG 算法和 BCWTG 算法在计算时间方面要明显优于 Greedy 算法,在 ANNI 指标方面也优于 Greedy 算法和 Random 算法,但是,在最终激活节点数方面,BNCWTG 算法和 BCWTG 算法的性能不如 Greedy 算法.值得注意的是,当激活概率 IC 足够大或激活阈值 LT 足够小时,能取得和 Greedy 算法非常接近的激活节点数.这说明,BNCWTG 算法和 BCWTG 算法更适用于产品本身易于接受的环境中.

5 结 语

本文分析了弱连接在信息级联中的作用,介绍了节点影响力相关内容,并利用弱连接在信息级联中的作用提出了基于贪心思想的 BWTG 算法,根据解空间的不同把 BWTG 算法分成 BNCWTG 算法和 BCWTG 算法,前者的解空间同时包含影响力强的节点和弱连接节点,后者只在弱连接节点中搜索节点.同时提出了一个新的影响力最大化问题评价指标——ANNI 指标.最后,在不同规模不同类型的社交网络中分别实现了 BNCWTG 算法和 BCWTG 算法,并与 Greedy 算法和 Random 算法进行比较,发现 BNCWTG 算法和 BCWTG 算法虽然在激活节点数方面弱于 Greedy 算法,但在计算时间和 ANNI 方面要优于 Greedy 算法.另外,BNCWTG 算法和 BCWTG 算法在激活概率较大或激活阈值较小的环境中,在激活节点数方面能取得和 Greedy 算法相近的性能.

References:

[1] Kempe D, Kleinberg J, Tardos É. 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. ACM, 2003. 137–146.

[2] Leskovec J, Krause A, Guestrin C, *et al*. Cost-Effective outbreak detection in networks. In: Proc. of the 13th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2007. 420–429.

[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. ACM, 2009. 199–208.

[4] 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. ACM, 2010. 1029–1038.

[5] Chen W, Yuan Y, Zhang L. Scalable influence maximization in social networks under the linear threshold model. In: Proc. of the IEEE 10th Int'l Conf. on Data Mining (ICDM). IEEE, 2010. 88–97.

[6] Wang Y, Cong G, Song G, *et al*. Community-Based greedy algorithm for mining top-k influential nodes in mobile social networks. In: Proc. of the 16th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2010. 1039–1048.

[7] Cao T, Wu X, Wang S, *et al*. OASNET: An optimal allocation approach to influence maximization in modular social networks. In: Proc. of the 2010 ACM Symp. on Applied Computing. ACM, 2010. 1088–1094.

[8] Granovetter MS. The strength of weak ties. American Journal of Sociology, 1973, 1360–1380.

[9] Granovetter M. Getting a job: A study of contacts and careers [MS. Thesis]. University of Chicago, 1995.

[10] Page L, Brin S, Motwani R, *et al*. The PageRank citation ranking: Bringing order to the Web. 1999. http://www-db.stanford.edu/~backrub/pageranksub.ps

[11] Girvan M, Newman MEJ. Community structure in social and biological networks. Proc. of the National Academy of Sciences, 2002, 99(12):7821–7826.

[12] Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Physical Review E, 2004,70(6):066111.

[13] Newman MEJ. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. Physical Review E, 2001, 64(1):016132.

[14] The max Planck institute. WOSN2009 data set. 2009. http://socialnetworks.mpi-sws.org/data-wosn2009.html

[15] KONECT. US power grid social network data set. 1998. http://konect.uni-koblenz.de/networks/opsahl-powergrid

易秀双(1969－),男,内蒙古赤峰人,博士,教授,CCF 高级会员,主要研究领域为计算机网络,互联网及其应用技术,网络与信息安全,数据处理.

胡金林(1990－),男,硕士生,主要研究领域为大数据社交网络,云计算,机器学习,计算机网络.

王兴伟(1968－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为下一代互联网,自组织网络,IP/DWDM 光 Internet,移动无线 Internet,网络信息安全,网格计算.