一种基于 k-核的社会网络影响最大化算法

曹玖新[1),2)] 董 丹[1),2)] 徐 顺[1),2)] 郑 啸[1),2),3)] 刘 波[1),2)] 罗军舟[1),2)]

[1)](计算机网络和信息集成教育部重点实验室(东南大学) 南京 211189)

[2)](东南大学计算机科学与工程学院 南京 211189)

[3)](安徽工业大学计算机学院 安徽 马鞍山 243002)

摘 要 社会网络中影响最大化问题是指在特定传播模型下,获取一个指定大小的节点集合,使得该集合在网络中的聚合影响力最大.针对贪心算法运用于大规模社会网络时存在效率低下且不可扩展的问题,文中提出基于核数层次特征和影响半径的启发式算法——核覆盖算法(Core Covering Algorithm,CCA).该算法首先引入 k-核概念,基于 k-核分解求出每个节点的核数,然后根据核数分布的层次性,引入节点的影响半径参数,最后综合核数和度数两个属性,找出影响力节点集合.文中在两个数据集和两种传播模型上进行了实验,结果表明:(1)在传播概率较大的独立级联模型(Independent Cascade Model,IC)下,CCA 能取得比现有启发式算法更优的影响效果;(2)在三价(TRIVALENCY Model,TR)模型下,CCA 的表现也同样优于其他启发式算法;(3)与其他启发式算法相比,CCA 的运行时间更少.

关键词 社交网络;影响最大化;独立级联模型;k-核;社会计算

中图法分类号 TP393 **DOI 号** 10.3724/SP.J.1016.2015.00238

A k-Core Based Algorithm for Influence Maximization in Social Networks

CAO Jiu-Xin[1),2)] DONG Dan[1),2)] XU Shun[1),2)] ZHENG Xiao[1),2),3)] LIU Bo[1),2)] LUO Jun-Zhou[1),2)]

[1)](*Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, Nanjing* 211189)

[2)](*School of Computer Science and Engineering, Southeast University, Nanjing* 211189)

[3)](*School of Computer Science, Anhui University of Technology, Maanshan, Anhui* 243002)

Abstract Influence maximization is the problem of obtaining a set of nodes with specified size in a social network to maximize their aggregate influence under certain influence diffusion model. Since greedy algorithms are inefficient and not-scalable, we propose a heuristic algorithm named Core Covering Algorithm (CCA) based on the coreness hierarchical characteristic and influence radius. Firstly, the algorithm introduces the concept of k-core and calculates the coreness of each node. Then, it introduces the influence radius parameter according to the hierarchy of coreness. Finally, it identifies influential nodes in accordance with the coreness and degree. Experiments are conducted on two datasets and two diffusion models. Experimental results show that (1) CCA performs better than other heuristic algorithms under Independent Cascade Model with a larger influence probability; (2) CCA also performs better than other heuristic algorithms under TRIVALENCY model; (3) Compared with other heuristic algorithms, CCA has lower running time.

收稿日期:2013-06-20;最终修改稿收到日期:2014-11-21.本课题得到国家自然科学基金(61272531,61202449,61272054,61370207,61370208,61300024,61320106007,61472081)、国家"八六三"高技术研究发展计划项目(2013AA013503)、国家"九七三"重点基础研究发展规划项目基金(2010CB328104)、江苏省科技计划项目(SBY2014020139-10)、高等学校博士点学科专项科研基金(2011009213002)、江苏省网络与信息安全重点实验室(BM2003201)资助.**曹玖新**,男,1967 年生,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为服务计算、网络安全、社会计算.E-mail:jx.cao@seu.edu.cn.**董 丹**,女,1989 年生,硕士研究生,主要研究方向为社会网络.**徐 顺**,男,1987 年生,硕士研究生,主要研究方向为社会网络.**郑 啸**,男,1975 年生,博士,教授,中国计算机学会(CCF)会员,主要研究领域为服务计算、无线局域网.**刘 波**,女,1975 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为普适计算、社会计算.**罗军舟**,男,1960 年生,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究领域为下一代网络体系结构、协议工程、网络安全、网络与云计算、无线局域网.

Keywords　social networks; influence maximization; independent cascade model; k-core; social computing

1　引　言

近年来，大规模在线社会网络（如 FaceBook、Twitter 和微博等）迅速发展. 作为当前重要的传播媒介，社会网络在社会个体间信息传播、相互影响和思想引导方面发挥着重大作用. 社会网络的流行，为大规模在线“病毒式营销”提供了机遇. 其中一个关键问题是如何在大规模社会网络中挖掘一个较小的节点集合，使得该集合在社会网络的影响最大化，即影响最大化问题. Domingos 和 Richardson 等人[1-2]首次把影响最大化问题定义为一个算法问题.

为解决影响最大化问题，Kempe、Kleinberg 和 Tardos[3]提出一种贪心爬山算法，该算法的影响范围能近似达到最优解的 63%. 然而，贪心算法时间复杂度比较高，不适用于大规模在线社会网络. 针对此问题，文献[4-5]对贪心爬山算法进行了优化，提出了 CELF、NewGreedy 以及 MixedGreedy 算法. 实验证明上述算法在时间效率上有了数百倍的提高，然而在中等规模的网络（如 15 000 个节点和 31 000 条边）中选择 50 个节点在一台现代服务器上运行需要数小时. 因此对于较大规模的网络（如 500 000 条边）来说，改进的贪心算法仍然不能解决时间复杂度高这一问题.

为了克服贪心算法的低效性，人们提出了多种启发式算法[5-11]，其运行时间比贪心算法降低了多个数量级，然而这些算法在不同的社会网络和传播模型下影响范围不稳定，或与贪心算法相比，传播影响范围相对较小，因此研究影响范围接近贪心算法的高效启发式算法是本文的主要内容.

针对上述问题，本文提出了一种基于核数的启发式算法 CCA. 实验结果表明 CCA 在传播概率较大的 IC 模型和 TR 模型下，影响范围明显优于其他启发式算法，且 CCA 的时间复杂度低于除 Degree 算法之外的其他启发式算法.

本文第 2 节介绍现有解决影响最大化问题的相关工作；第 3 节介绍两种最常用的传播模型；第 4 节提出核覆盖算法 CCA；第 5 节介绍本文的实验设计，包括实验数据集和对比算法；第 6 节验证本文提出的算法并对结果进行分析；第 7 节总结并探讨将来的研究工作.

2　相关工作

Domingos 和 Richardson 等人[1-2]首次把影响最大化作为一个算法问题引入到社会网络领域进行研究. Kempe、Kleinberg 和 Tardos[3]首次把影响最大化问题建模为在特定传播模型上寻找影响力最大的 k 个节点的离散优化问题，并证明在多种传播模型下，影响最大化问题是一个 NP-hard 问题. 在此基础上，提出了性能近似达到最优解 63% 的贪心爬山算法，该算法每轮选择边际收益最大的节点，然而计算给定种子集合的影响范围是 NP-hard 问题，因此贪心算法运用多次 Monte-Carlo 模拟[12]获得近似的影响范围. 然而 Monte-Carlo 模拟非常耗时，因此贪心算法不适用于大规模网络.

为了改善原始爬山算法的低效性，Leskovec 等人[4]提出了 CELF 算法，它利用影响传播函数的子模特性，延迟边际收益计算，与贪心爬山算法相比，其时间效率提高了数百倍. 另外，Chen 等人[5]提出了两种改进的贪心算法 NewGreedy 和 MixedGreedy[5]，进一步对传统贪心算法进行了优化. 在独立级联模型下，NewGreedy 以 $1-p$ 的概率去掉原图中的每条边，从而得到一个更小的子图，然后在子图中考虑影响最大化问题. MixedGreedy 第一轮采用 NewGreedy，其余轮采用 CELF，实验表明 MixedGreedy 的性能略优于 NewGreedy. 然而，上述改进的贪心算法的计算复杂度仍然非常高，不适用于大规模社会网络.

针对贪心算法的低效性，近年来涌现出了大量启发式算法. Chen 等人[5]在度的基础上提出了 DegreeDiscount 算法，该算法的思想是当一个节点 v 有邻居节点被选为种子节点时，在计算它的度时要打一定的折扣. 它优于简单的 Degree[13]算法. SCG 算法[6]为避免选取度数最大的节点而产生邻居重叠现象，尽可能把这些种子节点分散. PageRank[14]是 Google 用于用来标识网页等级和重要性的一种方法，也应用在社会网络中寻找影响力节点. 基于最短路径计算的 SPM/SP1M[15]时间复杂度同样非常高. PMIA[7]算法是一种较好的启发式算法，该算法有稳定的影响范围并且运行速度比贪心算法提高了 3 个数量级. PMIA 算法通过计算每个节点的本地树结构，加快影响传播的计算和更新. 然而，PMIA 算法需要耗费较大的内存. IRIE[11]算法被认为在运

行时间、耗费内存以及影响范围等三个方面综合排名第一的算法.该算法首先通过全局影响力排名算法(Influence Ranking)计算每个节点的全局影响力排名,根据排名次序选择影响力节点.为避免节点影响力重叠问题,在选择一个种子节点之后,利用影响力评估算法(Influence Estimation)估计网络中剩余节点的影响力变化,再根据节点变化后的影响力更新节点影响力排名.如此循环往复直至找到 k 个影响力节点.在 2010 年《Nature》物理版上,Kitsak 等人[16]通过实验表明在影响力传播方面,核数比度数和介数等节点属性具有更稳定的传播力,并且提出了基于覆盖的最大核算法和最大度算法.总之,启发式算法通常比贪心算法在时间性能方面能够提高数个数量级,却在不同的社会网络和传播模型下表现不稳定,或者与贪心算法相比,只能取得一个相对小的影响范围,这是因为启发式算法并没有考虑到传播模型的约束.贪心算法由传播模型出发,寻找当前能获得最大影响范围的节点.虽然贪心算法的传播效果非常好且稳定,但其时间复杂度非常高,并不适合大规模社会网络.

3 传播模型

对于特定的社会网络,在网络中寻找影响力节点集,需要借助于相应的传播模型.社会网络通常表示成由 n 个节点和 m 条边组成的有向图 $G(V,E)$,其中节点表示个体,有向边表示个体之间的社会关系.独立级联模型(Independent Cascade Model)[3,17]和线性阈值模型(Linear Threshold Model)[3,18]是两种最基本的传播模型.

在这两种模型中,节点有活跃和不活跃两种状态.节点可以从不活跃状态转变成活跃状态,反之则不可.随着节点的活跃邻居数越来越多,节点也越倾向于活跃.

3.1 独立级联模型

独立级联模型是一个概率模型.对于每条边 $\langle u,v\rangle\in E$,需指定一个影响概率 $p_{uv}\in[0,1]$,p_{uv} 表示节点 u 通过边 $\langle u,v\rangle$ 影响节点 v 的概率.给定初始活跃节点集合 A,传播过程以如下的方式进行:当不活跃节点 u 在时间步 t 变成活跃节点,那么 u 在时间步 t 有单次机会去激活当前每一个不活跃的邻居节点 v,其激活成功概率为 p_{uv}.不论 u 是否能成功激活 v,在以后的时间步中 u 不再激活 v.如果激活成功,则 v 在时间步 $t+1$ 变成活跃节点.如果在时间步 t,v 有多个父节点变为活跃状态,则这些活跃的父节点以任意顺序尝试激活 v,但所有的这些尝试都发生在时间步 t.当不存在激活的可能时,传播过程结束.在简单的独立级联传播模型中,p_{uv} 为一个系统常量.

TRIVALENCY 模型(简称 TR 模型)是 IC 模型的一个变型,图 $G(V,E)$ 中每条边 $\langle u,v\rangle$ 的传播概率从集合 $\{0.1,0.01,0.001\}$ 中随机选取,它们依次代表从高到低的传播影响力.

3.2 线性阈值模型

任一节点 v 都有一个阈值 $\theta_v\in[0,1]$,该阈值表示节点 v 受影响的难易程度,阈值越小,越容易受影响,即越容易被激活.在现实社会网络中,节点的阈值很难被度量,因此 θ_v 通常为一个随机值(一些情况下,指定为一个常量,如 1/2),服从[0,1]均匀分布.$\forall w\in N(v)$($N(v)$ 表示 v 的邻居节点集),$b_{v,w}$ 表示节点 w 对节点 v 的影响力,且满足邻居节点的影响力之和不超过 1,即 $\sum_{w\in N(v)} b_{v,w}\leq 1$.给定初始种子节点集 A,传播过程以如下的方式进行:在时间步 t,一个不活跃的节点 v 受每一个活跃的邻居节点 u 的影响.如果节点 v 的活动邻居节点集 $AN(v)$ 的影响力之和大于或等于 θ_v,即 $\sum_{u\in AN(v)} b_{u,v}\geq\theta$.则 v 在时间步 $t+1$ 变成活跃节点.当不存在激活的可能时,传播过程结束.线性阈值模型体现了影响力的累积效应.

4 基于核数的影响力最大化算法

4.1 k-核(k-core)

k-核概念由Seidman[19]于1983年在论文"Network Structure and Minimum Degree"中提出,它可用来描述度分布所不能描述的网络特征,揭示源于系统特殊结构的结构性质和层次性质.

给定网络 $G(V,E)$,其中 V 为节点集合,E 为边集合.相关定义如下.

定义 1. k-核(k-core).集合 $C\subseteq V$ 的任一节点 v 的度数不少于 k,由它所推导出的最大子图 $G_C(C,E|C)$ 称为 k-核,即递归移去图中度数小于 k 的节点及与其连接的边后所得到的子图为图的 k-核.

定义 2. 核数(coreness).若节点 v 属于 k-核,而不属于 $(k+1)$-核,则节点 v 的核数为 k.

k-核的一个重要特征是它的连通性[20],若图的 k-核为 k-连通,那么 k-核中的任意两个节点之间存在 k 条不相交的路径,这意味着核数越大的节点,其连通性就越好.

4.2 节点影响力重叠分析

已有的影响最大化算法(如 Degree、PageRank)

都没有考虑节点影响区域的重叠问题. 给定一组相互连接的节点，它们都有很高的度数或者 PageRank 值，如果它们彼此的影响区域存在大部分重叠，那么一个节点所带来的影响范围和多个节点的影响范围相近，使得影响力难以得到有效的扩散，即后续部分种子节点选择所带来的边际效益非常少. 如图 1 所示，节点 u 和 v 有着较大的重叠区域. 贪心算法提供了一个较优的解决方案，但它选择一个种子节点需要进行大量计算，不适用于大规模社会网络. 一个更好的解决方案是避免识别出的影响力节点在彼此的影响区域内.

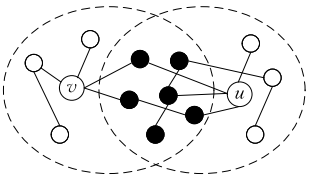


图 1　节点影响区域重叠

节点核数越大，节点间聚集程度就越高，因此简单地选择核数最大的 k 个节点并不能使传播信息得到有效的扩散，反而使得它们的影响范围产生大幅度重叠. 结合节点核数分布的层次特征，本文引入覆盖距离参数 d 使得种子节点间保持一定距离，影响力能够在网络上得到扩散，能有效地克服影响重叠问题. 核数的层次分布指的是从最高核开始，节点由高核到低核，由内层至外层分布，核数越高的节点，其网络分布越集中，也就越能受到彼此影响.

4.3　CCA 算法

核覆盖算法(Core Covering Algorithm，CCA)的基本思想：若节点 u 被选为种子节点，则与 u 距离小于等于 d(d 为自定义参数)的所有节点标识为覆盖状态，被标记为覆盖状态的节点不能被选为种子节点，每轮选择核数最大或在核数相等的情况下选择度数最大且未被覆盖的节点作为种子节点. 算法描述如算法 1 所示. 在核数相同的情况下，本文兼顾了节点的度数，这使得节点有更多的机会去影响邻居节点.

算法 1.　核覆盖算法.

输入：社会网络 $G(V,E)$，种子节点个数 k，覆盖距离 d

输出：选取的种子节点集合 S

算法描述：

1. initialize $S=\varnothing$;
2. *ComputeCores*(G);
3. FOREACH vertex $v\in V$ DO
4. CO_v=false;
5. END FOR
6. FOR i=1 to k DO
7. $w=\underset{v}{\operatorname{argmax}}\{C_v \mid v\in V\backslash S, CO_v=\text{false}\}$;
8. $u=\underset{v}{\operatorname{argmax}}\{d_v \mid v\in V\backslash S, C_v=C_w, CO_v=\text{false}\}$;
9. $S=S\cup\{u\}$;
10. FOREACH vertex v in $\{v \mid d_{u,v}\le d, v\in V\}$ DO
11. CO_v=true;
12. END FOR
13. END FOR
14. RETURN S;

其中 C_v 为节点 v 的核数，CO_v 表示节点覆盖属性，若节点未被覆盖则为 false，否则为 true；$d_{u,v}$ 表示节点 u 和节点 v 的距离.

算法第 2 行：根据网络图的核分解算法[21]计算每个节点的核数；第 3～5 行：设置每个节点的初始覆盖状态为未覆盖状态；第 7～8 行：所选节点为当前环境中核数为第一关键字，度数为第二关键字且未被覆盖的节点. 在核数相等的情况下，度数最大的节点有更好的连通性和传播能力；第 10～12 行：根据覆盖距离 d，以节点 u 为中心，标记与它距离小于等于 d 的所有节点为覆盖状态，这些所标记的节点可以理解为节点 u 的影响范围.

假设社会网络 $G(V,E)$ 有 n 个节点，m 条边，CCA 算法时间复杂度分析如下：第 2 行计算网络中节点核数的时间复杂度为 $O(m)$，第 3～5 行的时间复杂度为 $O(n)$，第 7～8 行选择目标节点的时间复杂度为 $O(n)$，第 10～12 行标记节点的 d 跳邻域的时间复杂度为 $O(m)$，则第 6 行到第 13 行的时间复杂度为 $O(km)$. 因此算法的总时间复杂度为 $O(km)$.

4.4　CCA 算法实例分析

图 2 是一个简单有向图，以出度作为节点的度. 通过 k-核解析，易得到每个节点的核数，图中节点的核数从外至里依次递增，最外层和最里层节点的核数分别为 0 和 3. 从图中可以清晰地看出节点的核数分布呈现层次结构. 假定 $k=3$，$d=1$，在初始阶

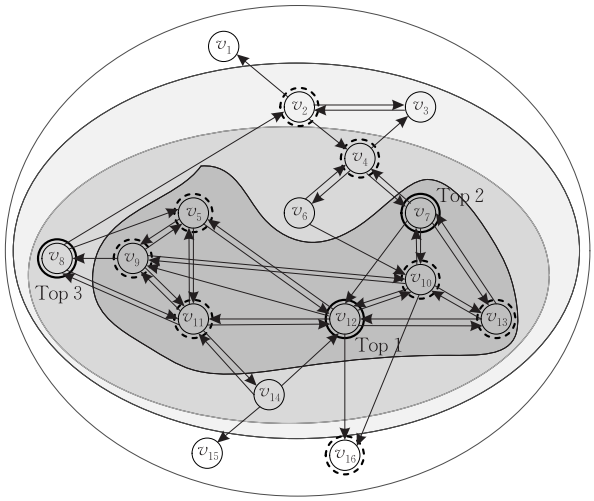


图 2　CCA 算法节点选择模拟图

段，每个节点都处于未被覆盖状态. 第 1 轮：选取核数为 3，出度为 6 的节点 v_{12}，并标记它及它的所有出边邻居集合 $\{v_5, v_9, v_{10}, v_{11}, v_{13}, v_{16}\}$ 的节点为覆盖状态；第 2 轮，选择当前核数最大并且未被覆盖的节点 v_7，标记 v_7 和 v_4 为覆盖状态；第 3 轮，选择当前核数最大并未被覆盖的节点 v_8，标记 v_2 和 v_8. 当选择过程结束后，所选择的种子集 $S=\{v_{12}, v_7, v_8\}$.

5 实验设计

5.1 数据集

本文从社会网络中选择两个真实的作者合作网络数据集，网络中节点表示作者，边表示两个作者之间存在合作关系. NetPHY①[5] 是预印本 e-print arXiv② 物理领域的论文全文列表，包括 37 154 个节点，231 584 条边；DBLP③ 合作网络包括 317 080 个节点，1 049 866 条边. 节点度的幂律分布是社会网络的一个重要规律，节点的核数同样也满足幂律分布，其中在 NetPHY 网络中核数分布的幂指数为 1.5093；DBLP 网络中核数分布的幂指数为 2.3756.

k-核解析与节点的度值有关，但节点的核数与度值之间并无明显的关系，度大的节点核数不一定高，核数高的节点度也不一定大.

5.2 对比算法和参数设置

为验证本文所提算法的性能，将它们与当前最有代表性的算法进行比较. 实验对比的算法及参数设置如表 1 所示. 贪心爬山算法的时间复杂度较高，虽然后续算法对其进行了优化，使得时间效率提高了数百倍，但面对十万个节点的社会网络，仍需要数天的时间才能得到结果，因此所比较的算法中不包含贪心算法. 相关研究表明，节点度数比介数和最短路径等节点属性或者网络属性更能表征节点的影响力，因此本文所对比的算法中包括 Degree 算法，而没有包含介数中心算法和距离中心算法. Chen 等人提出的 DegreeDiscount 算法，较好地衡量了一个节点在当前邻域的影响力，并在多数实验中有可观的表现. Kitsak 等人提出的基于覆盖的最大核算法（本文称之为 MaxCoreCover）和最大度算法（本文称之为 DegreeCover）具有较好的传播范围. 目前很多重要的链接分析算法和社会网络中主题排序算法都是从 PageRank 算法基础中衍生出来的，如 Topic-Sensitive PageRank 算法[22] 和 TwitterRank 算法[23]. PMIA 算法是在独立级联模型中表现较好的启发式算法，它具备传播范围的稳定性和运行时间的快速性等特征. IRIE 算法是迄今为止被认为在运行时间、消耗内存、传播范围等方面综合排名第一的启发式算法.

表 1 实验中对比算法列表

算法	算法描述
Degree	选取 k 个度最大的节点的启发式算法
DegreeDiscount	对种子节点的邻居节点进行度折扣的启发式算法，简称为 DD
MaxCoreCover	选取 k 个核数最大的节点，且每选择一个种子节点后，将该种子节点的邻居标记为覆盖状态，简称为 MCC
DegreeCover	选取 k 个度数最大的节点，且每选择一个种子节点后，将该种子节点的邻居标记为覆盖状态，简称为 DC
PageRank	Google 用于用来标识网页等级/重要性的算法，阻尼因子设为 0.15，简称为 PR
PMIA	基于本地树结构的启发式算法，其中 theta 设置为 1/320
IRIE	影响力排名和影响力评估相结合的启发式算法
CCA(d)	基于核数层次特征和影响半径的启发式算法，d 设为 1 和 2

种子节点数 k 的取值范围从 1 到 50，为便于阅读，算法是根据种子数为 50 时各算法的影响范围大小而排列的.

影响最大化算法 A 和 B 在种子数为 i 时的差异定义为

$$Diff(A,B,i)=\frac{\sigma(A,i)-\sigma(B,i)}{\sigma(B,i)} \quad (1)$$

其中，$\sigma(A,i)$ 为算法 A 在种子数为 i 的影响传播范围. 则影响最大化算法 A 和 B 的平均差异定义为

$$Diffavg(A,B)=\frac{1}{k}\sum_{i=0}^{k}Diff(A,B,i) \quad (2)$$

在我们的实验分析中，所有影响范围的百分比差异指的是种子数从 1 到 50 时，两算法之间影响范围大小的平均差异，平均差异更能体现出算法之间的差异性. 其中差异性的比较以 CCA(1) 为基准，即在式(2)中 A 为 CCA(1). 本文通过 Monte-Carlo 模拟 10 000次传播过程来获得一个较为精确的影响范围.

为验证各算法的影响传播效果，本文在两个真实的社会网络数据集和两种传播模型上进行了实验. (1) IC 模型. 假定在 IC 模型下社会网络中的传播概率相同，在不同的传播概率（$PP\in\{0.01, 0.02, \cdots, 0.05, 0.06\}$）下分析对比各算法的性能；(2) TR 模型. 网络中每条边 $\langle u, v\rangle$ 的传播概率从集合 $\{0.1, 0.01, 0.001\}$ 随机选取，它们依次代表着从高到低的传播影响力.

① http://research.microsoft.com/en-us/people/weic/graph-data.zip
② http://www.arXiv.org
③ http://snap.stanford.edu/data/com-DBLP.html

所有的实验在装有 Microsoft Windows 2008 系统的 IBM System X3755 M3 上运行，其硬件配置为 2.00 GHz Quad-Core 处理器、16 GB 内存.

6　结果分析

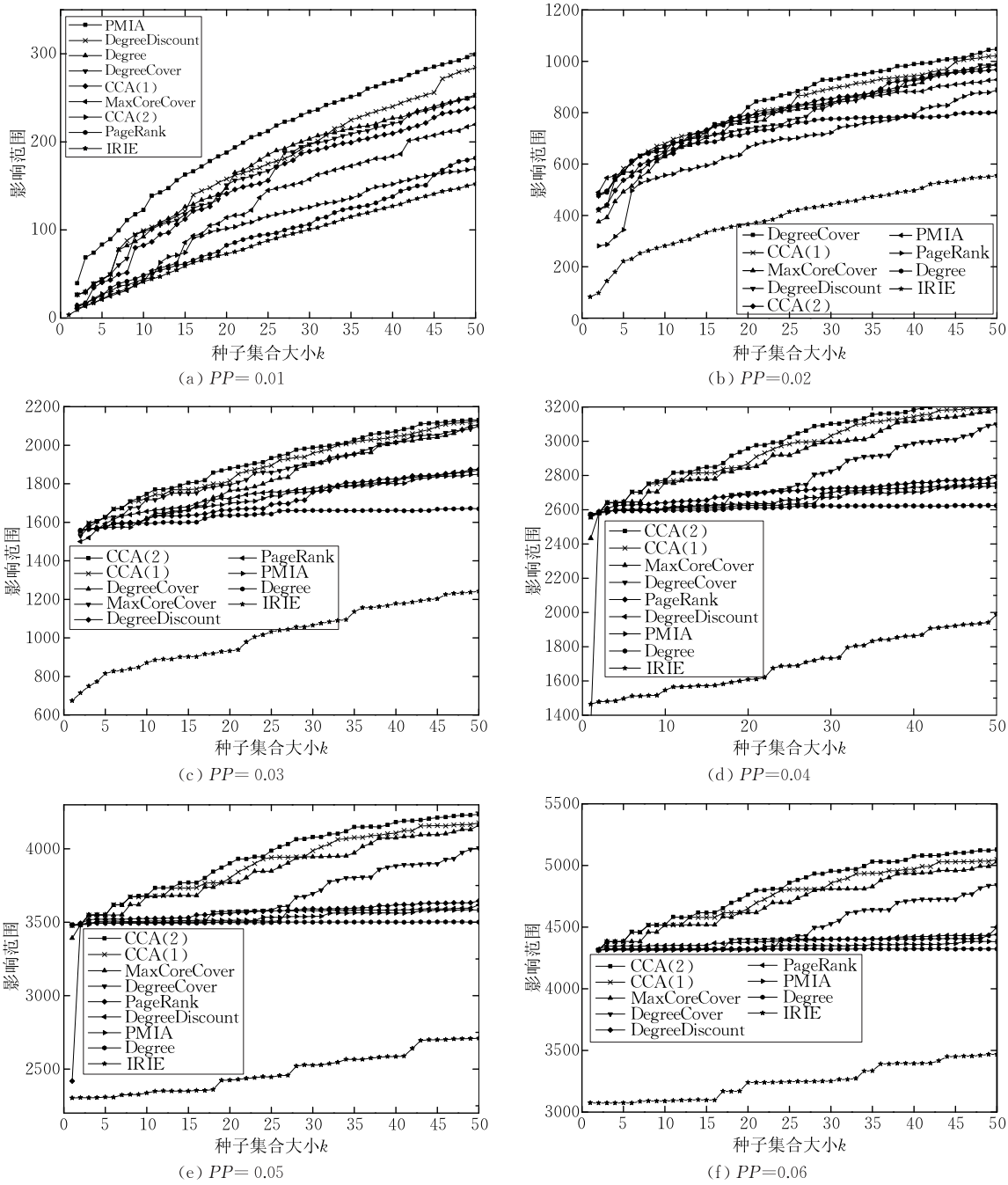
社会网络的影响最大化算法的评价指标可概括为如下两点：(1) 时间效率. 如何用较短的时间找出初始种子节点集合；(2) 影响效果. 在此初始集合的影响下，使得社会网络中最终受影响的节点数最多.

在传播影响图中，X 轴表示种子节点数，Y 轴表示在给定种子数的影响范围.

6.1　IC 模型上的数据结果与分析

6.1.1　NetPHY 数据集上算法性能分析

NetPHY 数据集是一个中等规模大小的网络，图 3 显示了在 IC 模型和 NetPHY 数据集上影响力算法的传播结果，由此可以清晰地看出 CCA 算法在传播概率大于 0.01 时优于其他启发式算法；并随着种子数 k 和传播概率的增大，算法之间效果的差异性越来越大，体现了 CCA 算法的有效性.



(a) PP= 0.01　(b) PP=0.02　(c) PP= 0.03　(d) PP=0.04　(e) PP= 0.05　(f) PP=0.06

图 3　不同算法在 NetPHY 数据集和不同传播概率上的影响效果

如表 2 所示，在传播概率为 0.01 时，PMIA 的影响范围显著优于其他算法. 当传播概率大于 0.01 时，在算法的传播范围方面，CCA(1)算法平均优于 IRIE、PageRank、Degree、PMIA、DegreeDiscount、DegreeCover和MaxCoreCover算法 77.26%、14.72%、12.54%、9.00%、8.84%、3.9% 和 2.36%. 且在传播概率大于 0.02 时，CCA(1)的影响范围平均比 CCA(2)低 1.19%.

表 2 在 NetPHY 数据集上 CCA(1)与其他算法影响范围的百分比差异

传播概率	CCA(2)/%	PMIA/%	IRIE/%	DD/%	DC/%	MCC/%	Degree/%	PR/%
0.01	42.53	−27.95	79.56	−13.01	−7.77	31.73	−10.51	59.64
0.02	3.15	5.01	121.04	4.78	−1.60	6.38	13.03	31.52
0.03	−1.15	9.50	88.05	9.85	3.64	1.91	15.33	12.43
0.04	−1.21	10.89	73.82	10.82	5.93	1.28	13.04	11.04
0.05	−1.23	10.15	57.15	9.99	6.00	1.18	11.36	9.86
0.06	−1.17	9.45	46.27	8.74	5.49	1.06	9.94	8.73

如图 3 所示，CCA(1)算法的性能始终优于 MaxCoreCover，因此可以得出这样的结论：以核数作为第一关键字，度数作为第二关键字选择影响力节点的算法肯定优于只考虑核数这一节点属性的选择算法.

当传播概率大于等于 0.05 时，其他启发式算法的影响范围(除 CCA 外)都基本上停止增长，即它们的曲线与 X 轴趋于平行，这表明传统的启发式算法都存在种子节点聚集现象，或者种子节点处于一个强连通图中. 因此考虑选择的节点间的距离是必要的. 而且从图 3 可以看出随着影响概率 p 的增大，CCA(2)的影响范围的增长速度快于 CCA(1)，由此可得出定性结论：当影响概率 p 较大时，选择的节点间距离 d 也应该较大. 因此在特定的传播概率下如何通过定量关系合理选择节点间距离是一个值得探讨的问题.

影响最大化问题可以归结为一个概率最大覆盖问题. 我们希望选择那些影响力节点可以触发激活网络中的很大一部分节点而又尽可能的避免重叠激活. CCA 启发式算法利用核数高的影响力节点又使它们保持一定的距离，从而减少重叠激活的可能性.

算法在 NetPHY 数据集上的运行时间如表 3 所示，从表中可以看出 CCA、Degree、DegreeDiscount、DegreeCover、MaxCoreCover 算法运行时间都在 1 s 以下，IRIE 算法的运行时间在 2.3 s 左右，PMIA 算法在传播概率为 0.06 时上升为 24 s，并随着概率的增大，所耗内存也越大.

表 3 不同算法的运行时间 (单位：s)

传播概率	Degree	DD	DC	MCC	PR	PMIA	IRIE	CCA(1)	CCA(2)
0.01	0.006	0.0392	0.007	0.007	1.618	0.936	2.365	0.008	0.014
0.02	0.006	0.0390	0.007	0.007	1.645	1.013	2.347	0.008	0.014
0.03	0.006	0.0392	0.007	0.007	1.620	1.096	2.344	0.008	0.014
0.04	0.006	0.0393	0.007	0.007	1.602	0.909	2.319	0.008	0.014
0.05	0.006	0.0392	0.007	0.007	1.762	0.915	2.309	0.008	0.014
0.06	0.006	0.0595	0.007	0.007	1.633	24.042	2.199	0.008	0.014

6.1.2 DBLP 数据集上算法性能分析

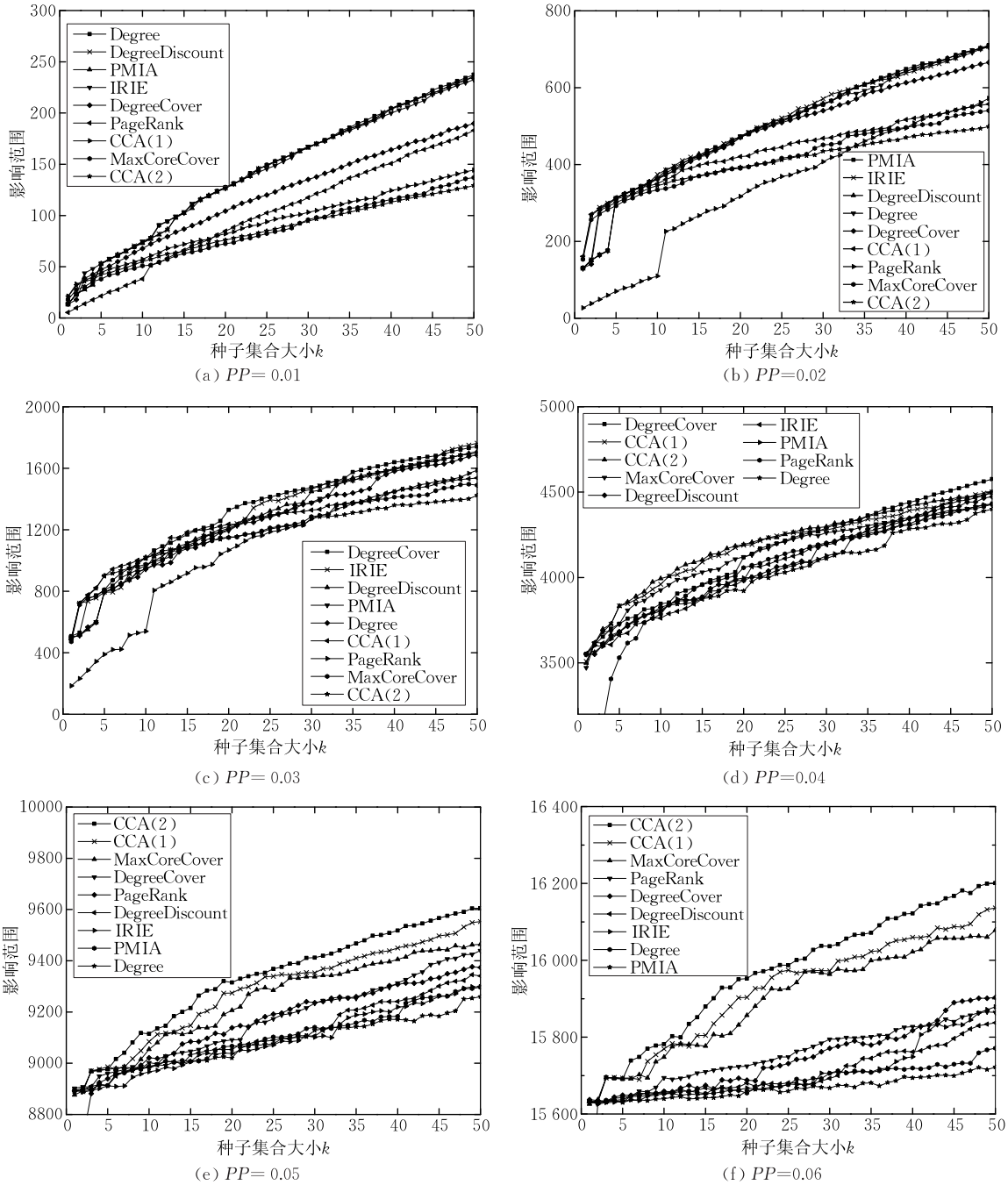
DBLP 数据集是有着百万边数的一个较大规模的网络，从图 4 的实验结果显示：当传播概率小于等于 0.03 时，PMIA、DegreeDiscount、DegreeCover、IRIE 等算法影响范围要大于 CCA 算法；当传播概率属于[0.04,0.06]时，与 PMIA、IRIE、DegreeCover、DegreeDiscount、PageRank 和 Degree 等启发式算法相比，CCA 算法的表现效果更优. 然而，我们可以发现：无论传播概率多大，CCA(1)算法的影响范围始终大于 MaxCoreCover 算法的影响范围.

如表 4 所示，传播概率小于 0.03 时，PMIA、IRIE、DegreeDiscount、DegreeCover、Degree 等算法的传播效果更好，随着传播概率的增大，CCA 算法有更好的表现.

表 4 在 DBLP 数据集上 CCA(1)与其他算法与影响范围的百分比差异

传播概率	CCA(2)/%	PMIA/%	IRIE/%	DD/%	DC/%	MCC/%	Degree/%	PR/%
0.01	8.71	−27.96	−29.82	−28.11	−16.47	11.31	−28.53	18.24
0.02	7.10	−7.34	−10.63	−7.46	−7.49	5.27	−6.82	88.13
0.03	5.73	1.11	−2.83	0.29	−4.96	4.72	3.91	33.81
0.04	−0.36	3.11	2.68	2.65	0.64	1.07	3.91	4.69
0.05	−0.53	2.01	2.12	1.85	1.22	0.40	2.24	1.50
0.06	−0.30	1.63	1.37	1.37	1.13	0.17	1.50	1.10

如图 4(a)所示，Degree、DegreeDiscount、IRIE 和 PMIA 的曲线之所以拟合在一起，是因为它们有着 90%以上的公共种子节点. PageRank 在种子数小于 10 时，种子节点集合的影响范围非常有限，导致其平均差异低于 CCA 算法. 如图 4(b)所示，Degree、DegreeDiscount、IRIE 和 PMIA 的影响范围仍处于领先. 如图 4(c)所示，CCA(1)在种子数小于 20 时，比其他算法有较小优势. 如图 4(d)～4(f)所示传播概率在[0.04，0.06]区间时，CCA 具有比其他启发式算法更大的传播影响力.



(a) PP= 0.01　(b) PP=0.02　(c) PP= 0.03　(d) PP=0.04　(e) PP= 0.05　(f) PP=0.06

图 4　不同算法在 DBLP 数据集和不同传播概率上的影响效果

度数是节点的一个关键属性，当传播概率较小时，节点所能影响到的邻居深度有限，这时度越大，能够激活的节点也就越多；当传播概率较大时，由于节点影响所产生的重叠效果，使得节点集合的影响力无法得到有效地扩散，并随着种子节点数的增大，影响边际效益却降低. CCA 算法较优的影响效果是由社会网络中节点核数分布的层次性和种子节点间有效的距离所致的.

算法在 DBLP 数据集上的运行时间如表 5 所示，从表中可以看出 CCA、Degree、DegreeCover、MaxCoreCover 和 DegreeDiscount 算法运行时间仍都在 1s 以下，PageRank 运行时间在 12s 左右，IRIE 算法的运行时间在 15 s 左右，PMIA 算法在传播概率为 0.06 时上升为 197 s.

表 5 在 DBLP 数据集上不同算法的运行时间 (单位：s)

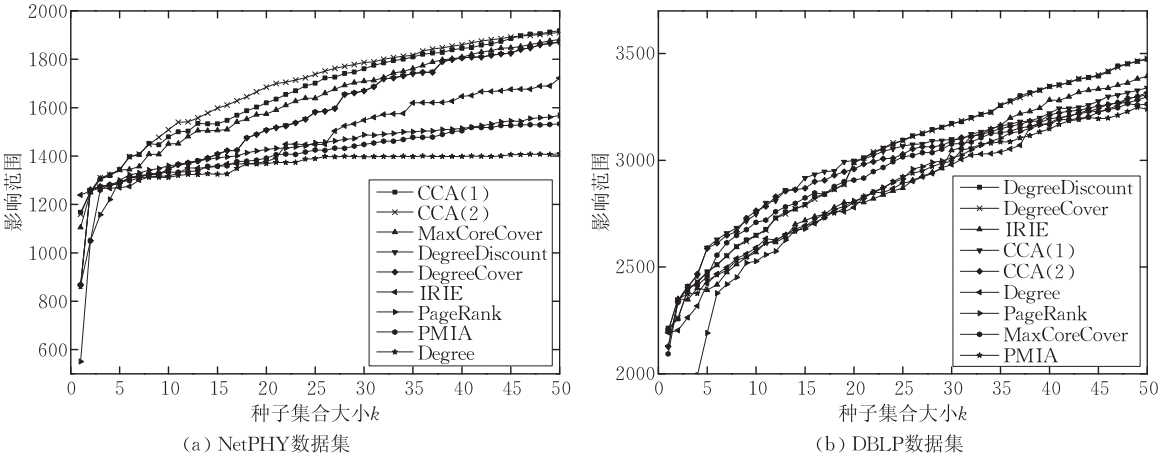
传播概率	Degree	DD	DC	MCC	PR	PMIA	IRIE	CCA(1)	CCA(2)
0.01	0.034	0.270	0.055	0.067	12.272	4.666	15.2510	0.095	0.102
0.02	0.034	0.306	0.055	0.067	11.980	5.658	14.5600	0.095	0.102
0.03	0.034	0.287	0.055	0.067	12.224	4.658	15.4523	0.095	0.102
0.04	0.034	0.289	0.055	0.067	11.103	5.157	14.9696	0.095	0.102
0.05	0.034	0.289	0.055	0.067	12.192	4.466	15.1513	0.095	0.102
0.06	0.034	0.240	0.055	0.067	11.161	196.685	14.9629	0.095	0.102

本文在这两个数据集上同样进行了传播概率在[0.07,0.10]区间时的实验，各算法的传播效果与传播概率为 0.06 时相比，影响传播效果相似.

随着传播概率的增大，覆盖距离越大其影响力传播的越远，当覆盖距离 d 为 3 时，CCA(3)的传播效果略优于 CCA(2). 值得注意的是，距离并非越大越好，当覆盖距离 d 为 4 时，CCA(4)的表现并不理想，这是因为网络的直径有限，覆盖距离的增大使得后续所选择的节点可能处于核心区域之外，从而边际效益不高.

6.2 TR 模型上数据结果与分析

如图 5(a)所示，CCA 在 NetPHY 数据集和 TR 模型上表现很好. 与其他启发式算法相比，CCA 算法有显著的优势，Degree 几乎没有增长，PageRank、IRIE 和 PMIA 增长缓慢. CCA(1)的影响范围大幅度优于 Degree、PageRank、PMIA 和 IRIE(22.37%、18.08%、18.17%和 12.21%).



(a) NetPHY数据集 (b) DBLP数据集

图 5 TRIVALENCY 模型

CCA、Degree、DegreeCover 和 MaxCoreCover 算法在 TR 模型下的运行时间和在 IC 模型下的运行时间相同，DegreeDiscount、PageRank、IRIE 和 PMIA 的运行时间分别为 0.82 s、1.33 s、2.26 s 和 1.556 s. 除了简单的 Degree 算法之外，CCA 的运行时间都低于其他算法，却取得较好的影响范围.

如图 5(b)所示，在种子数 k 小于 20 时，CCA(1)和 CCA(2)比其他启发式算法有着更好的传播效果. 而在种子数 k 大于 20 时，DegreeDiscount 和 DegreeCover 算法的效果比 CCA 好. CCA(2)接近 CCA(1)的影响范围，两者之间只有 0.59%的差异.

同理，CCA、Degree、DegreeCover 和 MaxCoreCover 算法在 DBLP 数据集上的运行时间参照 IC 模型下的时间. DegreeDiscount、PageRank、IRIE 和 PMIA 的运行时间分别为 13.32 s、16.2 s 和 8.22 s.

7 总 结

本文提出了一种基于核数的启发式算法来解决影响最大化问题. 在 TR 模型和 IC 模型的不同概率下，与目前具有代表性的启发式算法进行了比较，实验结果表明：在较大传播概率下 CCA 算法

的传播范围优于其他启发式算法，且时间复杂度非常低，并随着种子节点数的增加，影响范围保持增长势头.

将来的工作将在以下几个方面进行. 首先，我们将考虑节点间影响概率的差异性，通过数据挖掘和机器学习等方法，推算个体之间准确的影响概率. 其次，社会网络中存在社区结构，我们将研究社区结构这一特性对影响最大化问题的影响.

致　谢　审稿专家和编辑对本文提出了宝贵的意见和建议，在此表示感谢！

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

[3] 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 DC, USA, 2003: 137-146

[4] 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

[5] Chen Wei, Wang Ya-Jun, Yang Si-Yu. Efficient influence maximization in social networks//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 199-207

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

[7] Chen Wei, Wang Chi, Wang Ya-Jun. 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 DC, USA, 2010: 1029-1038

[8] Chen Wei, Yuan Yi-Fei, 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

[9] Chen Wei, Collins A, Cummings R, et al. Influence maximization in social networks when negative opinions may emerge and propagate//Proceedings of the 11th SIAM International Conference on Data Mining. Mesa, USA, 2011: 379-390

[10] Narayanam R, Narahari Y. A shapley value-based approach to discover influential nodes in social networks. IEEE Transactions on Automation Science and Engineering, 2010, 8(1): 130-147

[11] Jung Kyomin, Heo Wooram, Chen Wei. IRIE: Scalable and robust influence maximization in social networks//Proceedings of the 12th IEEE International Conference on Data Mining. Brussels, Belgium, 2012: 918-923

[12] MacKay D J C. Introduction to monte carlo methods//Proceedings of the NATO Advanced Study Institute on Learning in Graphical Models. Erice, Italy, 1996: 175-204

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

[14] Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Computer Networks and Isdn Systems, 1998, 30(1): 107-117

[15] 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

[16] Kitsak M, Gallos Lazaros K, Havlin S, et al. Identification of influential spreaders in complex networks. Nature Physics, 2010, 6(11): 888-893

[17] Watts Duncan J. A simple model of global cascades in random networks. Proceedings of the National American of Sciences, 2002, 99(9): 5766-5771

[18] Granovetter M. Threshold models of collective behavior. American Journal of Sociology, 1978, 83(6): 1420-1443

[19] Seidman S B. Network structure and minimum degree. Social Networks, 1983, 5(3): 269-287

[20] Carmi S, Havlin S, Kirkpatrick S, et al. A model of internet topology using k-shell decomposition. Proceedings of the National Academy of Sciences of the United States of America, 2007, 104(27): 11150-11154

[21] Batagelj V, Zaversnik M. An $O(m)$ algorithm for cores decomposition of networks. Advances in Data Analysis and Classification, 2011, 5(2): 129-145

[22] Haveliwala Taher H. Topic-sensitive PageRank//Proceedings of the 11th International Conference on World Wide Web. Honolulu, USA, 2002: 517-526

[23] Weng Jian-Shu, Lim Ee-Peng, Jiang Jing, He Qi. TwitterRank: Finding topic-sensitive influential twitterers//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York City, USA, 2010: 261-270

CAO Jiu-Xin, born in 1967, Ph. D., professor, Ph.D. supervisor. His research interests include service computing, network security and social computing.

DONG Dan, born in 1989, M. S. Her research interest is social networks.

XU Shun, born in 1987, M. S. His research interest is social networks.

ZHENG Xiao, born in 1975, Ph. D., professor. His research interests include service computing and wireless local area network.

LIU Bo, born in 1975, Ph. D., associate professor. Her research interests include pervasive computing and social computing.

LUO Jun-Zhou, born in 1960, Ph. D., professor. His research interests include next-generation network architecture, protocol engineering, network security, grid and cloud computing, and wireless local area network.

Background

In recent years, social network analysis has obtained a great attention with the popularization of the social network sites such as Sina Microblog, Facebook and Twitter. In the field of social network research, there are following research directions: theoretical analysis of social network structure, information propagation in social networks, community structure analysis and so on. The problem we study in this paper is influence maximization problem which is about information propagation. It is the problem of obtaining a set of nodes with specified size in a social network to maximize their aggregate influence under certain influence diffusion model. The problem is very meaningful and it has many practical applications such as marketing by identifying influential users, monitoring public opinion etc. In social networks, we can promote some products by making use of "word-of-mouth" effect which is a new effective marketing strategy.

The problem was proposed as a discrete optimization problem in 2003. The optimization problem is proved to be NP-hard and the balance of the running time and spread range is the main bottleneck for large social networks. So far, there are two kinds of algorithms to solve the problem. One is the greedy algorithms which have larger spread range and more running time. Another is the heuristic algorithms which run faster but have lower spread range. So in this paper, we mainly attempt to tackle the problem of influence maximization in terms of both efficiency and effectiveness. We propose a heuristic algorithm based on the coreness of nodes instead of degree, betweenness because the coreness has hierarchical and aggregate characteristics. We conduct experiments on two real datasets and two diffusion models. Experimental results show that (1) the algorithm we propose performs better than other heuristic algorithms under Independent Cascade Model with a larger influence probability, (2) it also performs better than other heuristic algorithms under TRIVALENCY model, (3) Compared with other heuristic algorithms, it has lower running time. We are working on social network for some time and our work is mainly focused on: community partition, influential nodes identification, Sina Microblog Information Diffusion Analysis, microblog expert identification etc.

This work is supported by National Natural Science Foundation of China under Grant Nos. 61272531, 61202449, 61272054, 61370207, 61370208, 61300024, 61320106007, 61472081, National Key Basic Research Program (973 Program) of China under Grant No. 2010CB328104, National High Technology Research and Development Program (863 Program) of China under Grant No. 2013AA01303, China Specialized Research Fund for the Doctoral Program of Higher Education under Grant No. 2011009213002, Jiangsu Provincial Key Laboratory of Network and Information Security under Grant No. BM2003201 and Key Laboratory of Computer Network and Information Integration of Ministry of Education of China under Grant No. 93K-9.