基于社区划分和改进PageRank的影响力最大化算法

吴海林

（广州杰赛科技股份有限公司，广东 广州 510310）

【摘　要】　为了解决传统贪心算法不能有效解决大规模社会网络影响力最大化的效率问题，采用模块度将大规模的通信网络划分成较小的社区模块，并通过改进PageRank排名算法来评价有向复杂网络节点的传播能力，然后再利用KK算法挑选当前带来最大影响范围的剩余种子节点，提出基于社区划分和改进PageRank的影响力最大化算法。实验证明，该方法具有一定的扩展性和有效性。

【关键词】　社区划分　改进PageRank　KK算法　影响力最大化　传播能力

doi:10.3969/j.issn.1006-1010.2017.10.017　中图分类号：TP391.4　文献标志码：A　文章编号：1006-1010(2017)10-0083-04

引用格式：吴海林. 基于社区划分和改进PageRank的影响力最大化算法[J]. 移动通信, 2017,41(10): 83-86.

An Influence Maximization Algorithm Based on Community Partition and Improved PageRank

WU Hailin

(GCI Science & Technology Co., Ltd., Guangzhou 510310, China)

[Abstract]　As the traditional greedy algorithm can not effectively deal with the influence maximization of large-scale social network, the large-scale communication network is divided into smaller community modules and the improved PageRank ranking algorithm is used to evaluate the propagation ability of directional complex network nodes. Then, KK algorithm is used to select the remaining seed nodes with the largest influence. Finally, an algorithm based on community partition and improved PageRank is proposed in this paper. Experimental results show that the proposed method is effective and scalable.

[Key words]　community partition　improved PageRank　KK algorithm　influence maximization　propagation ability

1　引言

随着各类移动社交服务如Facebook、Twitter、微博、微信、QQ等对人类生活、社交的渗透，社交网络在信息沟通、信息共享、信息传播扩散等方面起到了不可忽视的作用。伴随移动社交网络节点的增长，核心节点作为信源，其传播影响作用不容小觑，其中影响力最大化研究能够对社会安防、市场营销等领域起到重要作用，因此得到了不少学者的追捧。比如：Luo[1]等人基于幂定律法则的影响力分布下，提出了PageRank的启发式算法来寻找种子节点；Chen[2]等人考虑节点的拓扑结构，提出了基于节点、最邻近节点以及次近节点的多级邻居指标的节点重要性排序；Wang[3]等人提出了基于社区发现求解影响力的CGA算

收稿日期：2017-05-11

责任编辑：袁婷　yuanting@mbcom.cn

法；郭进时[4]等人选取影响力传播范围和影响力传播时延这两个指标衡量节点的影响力，提出了社区结构的影响力最大化算法。在上述研究的基础上，本文提出了基于社区划分和改进PageRank的影响力最大化算法。该算法首先通过模块度进行社区划分，然后通过改进PageRank算法选取移动通信有向加权复杂网络的种子节点，再利用KK算法局部最优的特性选取当前带来最大影响范围的剩余种子节点，以期尽可能地扩展算法的传播影响范围。

2 网络影响力最大化研究

网络影响力最大化这个问题要追溯到2002年，Richardson和Domings如何寻找社会网络中最具影响力的成员，并发放免费样品，通过这些最具影响力的成员口口相传，使这些样品的相关信息在社会网络中顺利传播，以达到波及范围广和成本低廉的营销目的。

影响力最大化初始定义是在所有网络节点中选取最具有影响力的节点作为初始活跃节点，使其经过影响传播，网络最终被影响的节点数最多。但在实际应用过程中，处理网络影响力的关键包括如下：

（1）在有限的资源下种子节点的选取问题；

（2）如何在节点的影响力信息的级联下，通过贪心算法求出k个节点的影响力最大范围。

传统的种子节点选取使用静态的启发式节点选择策略，一般采用度中心性、介数、紧密度等指标进行衡量，通过这些指标或者组合指标来确定节点的影响力，然后采用影响力排序的方法选取种子节点，以便实现网络传播最大化的目的；另外一种是动态的影响力最大化算法，首先构建网络传播模型，然后通过贪心算法求出最大传播范围的k个种子节点。

无论采取上述哪种算法，都有本身的缺陷。静态启发式的算法虽然处理速度较高、计算复杂度低，但是不能保证种子节点的传播影响力最大化；贪心算法由于计算复杂度太高，所以并不擅长处理大型的复杂网络。

因此，本文在总结上述两种算法的优点的基础上，结合社区划分的方法、改进PageRank以及KK局部最优算法来解决有向加权复杂网络影响力最大化的问题。

3 基于社区划分和改进PageRank的影响力最大化算法

3.1 有向加权复杂网络模型

本文研究的数据是移动通信网络的用户业务数据，因此在介绍算法之前，必须先介绍移动通信网络模型。

移动通信网络模型可以认为是有向加权网络，其模型可用$\mathbf{G}$表示，$\mathbf{G}=(\mathbf{V}, \mathbf{E})$。其中，$\mathbf{V}$表示网络节点集合，可表示为$\mathbf{V}=\{v_1, v_2, \cdots, v_n\}$；$\mathbf{E}$表示网络有向边集合，可表示为$\mathbf{E}=\{e_1, e_2, \cdots, e_n\} \subseteq \mathbf{V} \times \mathbf{V}$。$w(v_i, v_j)$表示有向边$(v_i, v_j)$的权值（或称连接强度）。本文基于文献[5]定义，$S_{\text{in}}(v_i)$表示节点$v_i$的入强度，入强度通常说明节点的指向或者传播的强度，公式如下：

$$S_{\text{in}}(v_i)=\sum_{v_j \in \mathbf{V}_{\text{in}}(v_i)} w(v_j, v_i) \quad (1)$$

同理，$S_{\text{out}}(v_i)$表示节点v_i的出强度为：

$$S_{\text{out}}(v_i)=\sum_{v_j \in \mathbf{V}_{\text{out}}(v_i)} w(v_i, v_j) \quad (2)$$

其中，$\mathbf{V}_{\text{in}}(v_i)$表示指向节点$v_i$的所有节点集合；$\mathbf{V}_{\text{out}}(v_i)$表示节点$v_i$所指向的所有节点集合。

把有向加权网络模型运用于移动通信网络，$\mathbf{V}$表示移动通信网络的移动用户集合；$\mathbf{E}$表示移动通信网络用户之间相互联系的集合。每个用户的传播影响力强度用$S_{\text{out}}(v_i)$表示，则用户v_i的信息传播影响力强度为：

$$S_{\text{out}}(v_i)=\sum_{v_j \in \mathbf{V}_{\text{out}}(v_i)} w(v_i, v_j) \quad (3)$$

其中，$\mathbf{V}_{\text{out}}(v_i)$表示用户$v_i$拨打给其他所有用户的集合。

3.2 基于模块度的社区划分方法

众所周知，复杂网络的节点之间的信息得以传播是因为节点在某种程度上具有相似性。基于模块度划分社区的思想是社区内部的节点相似度较高，而在社区外部的节点相似度较低。因此，本文在社区划分的基础上，将社区的逐渐扩散等效于信息传播的过程，借助于连接紧密的社区内部节点具有相似性这一思想，信息传播在连接紧密的相似节点进行传播。

模块度的基本思想是：首先认为每个节点是独立节点，然后将某个节点并入到其中一个社区后，计算并入后形成的社区内部边数与并入前内部边数的差值

占整个网络边数的比值[6]。公式如下：

$$Q=\sum_i[e_{ii}-a_i^2] \quad (4)$$

其中，e_{ii}为每个社区i内部节点间的连边数量；a_i为另一端与社区i中节点相连的连边数量。

3.3 基于PageRank算法的节点重要性评估算法

PageRank算法[7]是用于搜索引擎中网页排名的经典算法，该算法的主要思想是从优质网页链接而来的网页必定还是优质网页。如果网页A是优质网页，A指向B，那么B的重要性程度取决于A的重要性以及网页A指向的网页总数。由于网络中网页链接的相互指向，该分值的计算为一个迭代过程，公式如下：

$$PR(B)=\frac{(1-\sigma)}{n}+\sigma\sum_{i=1}^{n}\frac{PR(Y_i)}{C_{\text{out}}(Y_i)} \quad (5)$$

其中，$PR(B)$为网页B的PageRank值；$PR(Y_i)$为链接到网页B（可理解为指向网页）的网页Y_i的值；$C_{\text{out}}(Y_i)$为网页Y_i的出度数量；σ为阻尼系数；n为网页总数。根据公式（5）可知，指向网页B的链接数目越大，则B越重要。

借鉴上述表达，本文PageRank算法将网页排名思想引入到通信有向加权复杂网络的用户传播能力评估来寻找种子节点，公式如下：

$$PR(v)=\frac{(1-\sigma)}{n}+\sigma\sum_{i=1}^{n}\frac{w(v_i,v)\times PR(v_i)}{S_{\text{out}}(v_i)}=\frac{(1-\sigma)}{n}+\sigma\sum_{i=1}^{n}\frac{w(v_i,v)\times PR(v_i)}{\sum_{j=1}^{m_i}w(v_i,z_j)} \quad (6)$$

其中，$\sum_{j=1}^{m_i}w(v_i,z_j)$为用户源$v_i$的出强度$S_{\text{out}}$；$w(v_i, v)$为用户$v$获得用户源$v_i$的边权重，公式如下：

$$w(v_i, v)=T(v_i, v)\times L(v_i, v) \quad (7)$$

其中，$T(v_i, v)$为用户v获得用户源v_i的通话次数占比；$L(v_i, v)$为用户v获得用户源v_i的通话时长占比。

3.4 KK算法

Kempe和Kleinberg首次采用一种贪心算法来解决社会网络影响力最大化的问题，通过KK算法[8]来寻找种子节点组合。KK算法是一种局部最优的算法，通过遍历每个节点的影响范围，然后采用影响力降序的排名算法来挑选当前影响力最大范围的节点作为种子节点，但是这种局部改进的算法并不能保证最终得到的种子节点的影响范围最优。因此，本文采用改进PageRank算法来选择k-[ck]个种子节点，然后通过节点来激活由模块度划分的重要社区，在激活阶段利用贪心算法来选取剩余[ck]个种子节点。

$$\inf(u)=\sum_{v\in\text{out}(u),\text{active}(v)=0}w(u,v) \quad (8)$$

其中，$w(u,v)$表示用户v获得用户源u的边权重。如果$w(u,v)$大于所设定的阈值，那么说明用户v被激活。

根据上述步骤对划分的重要社区进行激活节点的影响范围计算，再对各个种子节点的影响范围进行排序，进而确定最终的k个种子节点。

4 实验分析

本文采用某市移动运营商的用户通话数据，验证基于社区划分和改进PageRank的影响力最大化算法的有效性。该复杂网络共由5 000个节点组成，有90 663条有向边，有向边表示用户之间满足特定的通话关系和通话时长的条件（一个月内大于4次通话，或每次平均通话时长约为30 s）。以用户之间通话次数的比例乘以通话时长的比例作为权值。

首先对该数据进行社区划分，得到8个社区；然后采用PageRank算法对每个社区进行用户传播能力排名，再用KK算法选择剩余种子节点来激活的社区，最终算出每个社区的k（$k\leqslant10$）个种子的影响范围，得到每个社区选取的种子数量如表1所示。

通过上述结果可知，如果需要对该网络选取10个种子，那么根据表1选取的种子用户为15、101、19、132、187、55、199、54、34、487。

本文采用度中心性算法进行种子节点的选取，得到的影响节点数量与本文设计的算法进行对比，具体结果如图1所示。

由图1可知，采用基于社区划分和改进PageRank的影响力最大化算法具有较大的传播范围。并且本文提出的算法与度中心性算法在不同规模的种子节点上的扩散速率是相吻合的，从而验证了该算法的有效性。

表1　本文采用算法的种子节点以及影响范围

社区名称	种子序号	影响范围	种子序号	影响范围	种子序号	影响范围	种子序号	影响范围	种子序号	影响范围	种子序号	影响范围
社区1	15	82	19	57	34	33	39	26	17	18	25	8
社区2	54	37	76	32	79	25	81	17	97	6		
社区3	101	66	132	54	187	49	199	39	27	9	22	4
社区4	55	48	33	15	281	3						
社区5	487	32	576	26	388	18	421	5				
社区6	683	18	771	8	891	5						
社区7	1 029	15	1 784	7	1 954	3						
社区8	3 218	26	3 781	5								

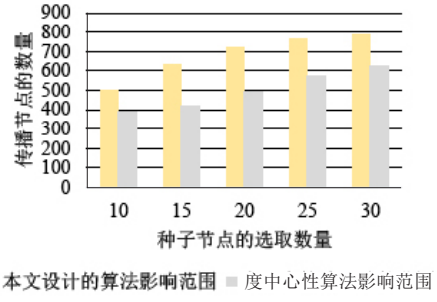

图1　不同算法的传播范围对比

5　结束语

本文提出了一种基于社区划分的移动通信有向加权复杂网络影响力最大化算法——基于社区划分和改进PageRank的影响力最大化算法，与之前关于影响力最大化研究不同的是首先基于模块度将移动通信网络划分成社区，实现了化整为零的效果，然后采用改进PageRank算法评估用户传播能力，并以此来选择k-[ck]个种子节点，再利用KK算法挑选当前带来最大影响范围增量的剩余[ck]个种子节点，以达到影响力最大化的目标。经过实验验证了该算法的有效性和高效性，通过社区划分能够在一定程度上提高算法的效率，可以很好地适应大规模的社会网络环境。

参考文献：

[1] Luo Z L, Cai W D, Li Y J, et al. A PageRank-Based Heuristic Algorithm for Influence Maximization in the Social Network[C]//Recent Progress in Data Engineering and Internet Technology. Berlin Heidelberg: Springer, 2012: 485-490.

[2] Chen D, Lv L, Shang M, et al. Identifying influential nodes in complex networks[J]. Physica A Statistical Mechanics & Its Applications, 2012,391(4): 1777-1787.

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

[4] 郭进时,汤红波,吴凯,等. 基于社区结构的影响力最大化算法[J]. 计算机应用, 2013,33(9): 2436-2439.

[5] 张益. 一种定量评估复杂网络节点重要度的算法[J]. 计算机工程, 2011,37(20): 87-88.

[6] 李建华,汪晓锋,吴鹏. 基于局部优化的社区发现方法研究现状[J]. 中国科学院院刊, 2015,30(2): 238-247.

[7] 张琨,李配配,朱保平,等. 基于PageRank的有向加权复杂网络节点重要性评估方法[J]. 南京航空航天大学学报, 2013,45(3): 429-434.

[8] Kempe D, Kleinberg J, Tardos E. Maximizing the spread of influence in a social network[C]//Proceeding of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM New York, 2003: 137-146.

[9] 王双,李斌,刘学军,等. 基于社区划分的影响力最大化算法[J]. 计算机工程与应用, 2016,52(19): 42-47.

[10] 吴萍. 移动社交网络中的信息传播最大化问题研究[D]. 济南: 济南大学, 2015.★

作者简介

吴海林：学士毕业于安徽省合肥学院，现任职于广州杰赛科技股份有限公司，目前主要从事客户推广的工作，擅长于移动通信用户行为分析、移动通信网络的用户影响力等方面的研究。