文章编号:1009-2552(2015)03-0117-04 DOI:10.13274/j.cnki.hdzj.2015.03.030

一种改进的社交网络影响力最大化算法

周莲英，朱 锋

（江苏大学计算机科学与通信工程学院，江苏 镇江 212013）

摘 要：社交网络影响力最大化即是在社交网络中如何挑选包含 K 个节点的种子节点集，去激活整个网络，使网络中最终被激活的节点数最大化。基于 IC 模型研究了社交网络影响力最大化问题。Diffusion degree 算法提出了节点潜在影响力的概念，即一个节点的邻居节点的影响力也可以作为当前节点的影响力的一部分。基于 Diffusion degree 算法做出了改进，在考虑潜在影响力的时候进一步考虑了节点潜在影响力的有效性，更加准确地判断节点的影响力，再综合了算法 SingleDiscount 中的核心思想，从而选出更加优质的种子节点。仿真结果表明，该算法在影响范围上接近 KK 贪婪算法的影响范围，同时在时效性上优于 Diffusion degree 算法，较适合大型社交网络。

关键词：影响力最大化；IDD；社交网络；启发式

中图分类号：TP393 **文献标识码**：A

An improved algorithm for influence maximization in social networks

ZHOU Lian-ying, ZHU Feng

(**School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang** 212013, **Jiangsu Province, China**)

Abstract: Influence maximization is the problem of selecting a seed set of top k seed nodes in a social network to maximize their influence spread under certain influence propagation models. This paper focuses on the problem of influence maximization in social network based on independent cascade model. The diffusion degree algorithm proposes that the influence of the neighbors of nodel can be considered as part of influence of the nodel. It makes improvements to the diffusion degree algorithm, considering which neighbors should be used as the potential influence and combining the core idea of the algorithm SingleDiscount to select better quality seed node. The experimental results show that the heuristics match the influence spread with the KK greedy algorithm and have a better performance than diffusion degree in efficiency.

Key words: influence maximization; IDD; social network; heuristics

0 引言

近年来，很多大型在线社交网络（SNS）不断流行起来，比如国外的 Facebook，Twitter 和 Google + 等，国内的人人，新浪微博等等。因此，社交网络影响力最大化问题也逐渐成为了业界研究热点，越来越多的学者开始这方面的研究。社交网络是由海量的用户及用户间复杂的关系（包括亲属关系，朋友关系，同学关系，工作关系等）组成，不同于传统网络，社交网络中的信息传播与扩散依赖于用户间的关系。假设有一个小公司开发了一个新的应用，但是由于资金缺乏，无法进行大量的广告宣传，这时公司在社交网络中选择一些影响力大的用户，并给他们提供应用的免费试用，希望这些试用用户在肯定

收稿日期：2014-03-12

作者简介：周莲英（1964-），女，教授，研究方向为网络安全，网络性能，无线（移动）网络，电子商务和物联网等。

该应用的同时，会向自己的好友或粉丝推荐该应用，然后他的朋友或粉丝中肯定该应用的人又向他们的朋友推荐，如此不断迭代下去，最后这个应用会被更多的人接受，直到没人再肯定它。这时如何选择最初的试用者能使这个应用最终能在网络中被更多的用户肯定，得到更大的传播扩散就成了研究的热点，即社交网络影响力最大化问题。

Richardson 和 Domingos 将这个问题归纳成一个算法问题，即在社交网络中选择 K 个最具影响力的初始节点，使得影响力在整个网络中的传播最大化。Kemple 和 Kleinberg 证明了社交网络影响力最大化问题是一个 NP - hard 问题，并提出了一种贪婪爬山算法。这种算法虽然保证了 $1-1/e$ 范围内接近最优解，但是其时间复杂度太高，并不适合大型的在线社交网络。后续的研究重点主要可以分为两个部分：一是对于贪婪算法的时间复杂度的优化，在保持其良好的影响范围的基础上提高算法的时效性；二是对于启发式算法的影响范围的优化，在保持其良好的时效性的基础上提高其算法的影响范围。对于这两方面的优化，一些学者（如 Wei Chen）认为后者更具研究价值，因为后者能更好的适应大型社交网络。

Diffusion degree 算法提出了节点潜在影响力的概念，即一个节点的邻居节点的影响力也可以作为当前节点的影响力的一部分。本文基于 Diffusion degree 算法做出了改进，首先过滤了网络中对种子节点选择无用的边，更加准确地判断节点的影响力，从而选出更加优质的种子节点。本文的主要贡献有：①基于 Diffusion degree 算法做出了改进；②改进后的算法在影响范围上得到了改进；③在数据集上进行了实验仿真，分析并验证了算法的有效性。

1 相关工作

一般将社交网络抽象成一张有向（或无向）图 $G=(V,E)$，其中 V 是网络中节点集合，E 是节点间关系集合。V 中每一个节点有两种状态：激活与未激活状态。处于未激活状态的节点被激活后成为激活状态，但是激活状态的节点无法变回未激活状态，即激活过程不可逆。对于每一条边 $(u,v)\in E(u,v\in V)$，$p(u,v)$ 代表节点 u 激活节点 v 的概率（若 $(u,v)\in E$，则 $p(u,v)=0$），其中 u 是已经被激活的节点。

两种经典的影响力传播模型是独立级联（IC）模型和线性阈值（LT）模型，本文的研究是基于独立级联模型，对于种子节点集 $S\subseteq V$，影响力在 IC 模型中的传播过程如下：$A_t\subseteq V$ 表示在步骤 $t\geqslant 0$ 时网络中被激活的节点集，其中 $A_0=S$。在步骤 $t+1$，每一个节点 $u\in A_t$ 会去激活它的邻居节点 $v\in V$，激活概率为 $p(u,v)$。当 $A_t=\phi$ 时，传播过程结束。需要注意的是，每个处于激活状态的节点有且仅有一次机会去激活它的邻居节点，并且保持自身处于激活状态。$\sigma_I(S)$ 表示种子节点集 S 的影响力，即整个传播过程结束时，网络中被 S 激活的节点个数总和。社交网络影响力最大化问题就是在网络中寻找一个子集 $S^*\subseteq V$ 且 $|S^*|=k$，使得 $\sigma_I(S^*)=\max\{\sigma_I(S)\mid S\subseteq V,|S|=k\}$。

由于经典的贪婪算法每一步都要计算所有未激活节点的边际影响，所以时间复杂度过高，不适合大型在线社交网络，因此一些学者建议将研究重点放在如何提高启发式算法的影响范围。Diffusion degree 算法提出了节点潜在影响力的概念，即一个节点的邻居节点的影响力也可以作为当前节点的影响力的一部分，这一定程度上使节点影响力得到了更准确的判断，但是原有算法还可以有进一步提升的空间。

2 改进的启发式算法 IDD

传统 IC 模型中对于节点影响力的判断一般基于节点出度和节点对邻居节点的激活概率这两个属性，Diffusion degree 算法提出的潜在影响力重新定义了节点影响力，在计算节点影响力的时候将该节点的邻居节点的影响力也进行加权计算。这样解决了这样一种情况，即有些节点虽然自身影响力（即节点出度）不足以作为种子节点，但是该节点的邻居节点都拥有相当的影响力；而另外一些节点自身拥有相当的影响力，但是该节点的邻居节点的影响力都不高，从而加权计算以后，前者的综合影响力反而超过了后者。由此根据节点潜在影响力来判断节点是否被选作种子节点会更合理。

度中心算法这样定义节点影响力：

$$C_D(v)=\sum_{i=1}^{n}\sigma(u_i,v) \tag{1}$$

当 $(u_i,v)\in E$，则 $\sigma(u_i,v)=1$；当 $(u_i,v)\in E$，则 $\sigma(u_i,v)=0$。$p(u_i,v)$ 是节点 u_i 被节点 v 激活的概率。在影响力传播过程中，节点 v 的影响力为：

$$C'_{DD}(v)=p(u,v)*C_D(v) \tag{2}$$

在 Diffusion degree 算法中，节点影响力要被重新定义，需要综合考虑节点潜在影响力（即当前节点邻居节点的影响力）。邻居节点的影响力为：

$$C^*_{DD}(v)=\sum_{i\in neighbors(v)}C'_{DD}(i) \tag{3}$$

Diffusion degree 算法中的节点综合影响力这样定义：

$$C_{DD}(v) = C'_{DD}(v) + C^{*}_{DD}(v) = p(u,v) \times C_D(v) + \sum_{i \in neighbors(v)} C'_{DD}(i) = p(u,v) \times C_D(v) + \sum_{i \in neighbors(v)} p(u,v) \times C_D(v) \quad (4)$$

从上述对于 Diffusion degree 算法的描述中不难发现一个问题，即在考虑节点潜在影响力的时候，该算法加权计算了该节点的所有邻居节点。但这明显存在一定问题，即当节点 v 的一个邻居节点 u 具有相当的影响力，但是 $p(u,v)$ 却非常小，这时由于节点 u 被节点 v 激活概率过小，节点 u 的影响力被加权计算进节点 v 的综合影响力并不合理，由此得知不是每一个邻居节点的影响力都应该被计算进该节点的综合影响力，所以为了更准确计算节点的综合影响力，应该对那些被加权进去的邻居节点进行过滤，本文定义了一个阈值 θ，来对 $p(u,v)$ 进行进一步的过滤，只有 $p(u,v) > \theta$ 时，才将 u 的影响力计算进 v 的综合影响力。

$$C_{DD}(v) = p(u,v) * C_D(v) + \sum_{i \in neighbors(v)} p(u,v) * C_D(v), p(u,v) > \theta \quad (5)$$

同时，本文综合算法 SingleDiscount 中的思想，即每当一个节点被选作种子节点，就更新该节点的邻居节点的出度，这样就更准确地计算了节点综合影响力。

$$t_v = \sum_{ies}^{n} \sigma(u_i, v) \quad (6)$$

$$C'_D(v) = C_D(v) - t_v \quad (7)$$

最后，节点综合影响力为：

$$C_{DD}(v) = p(u,v) \times C_D(v) + \sum_{i \in neighbors(v)} p(u,v) \times C'_D(v), p(u,v) > \theta \quad (8)$$

Improved diffusion degree 算法伪代码如下：

Algorithm Improved diffusion degree: $IDD(G = (V,E), k, \theta)$

1: initialize$S = \phi$

2: foreach v do

3: $C_D(v) = \sum_{i=1}^{n} \sigma(u_i, v)$

4: $t_v = 0$

5: end for

6: for $i = 1$ to k do

7: select $u = \text{argmax}_v \{C_{DD}(v) | v \in H \backslash S\}$

8: $S = S \cup \{u\}$

9: for each neighbor v of u and v and V \ S do

10: $t_v = \sum_{i \in s}^{n} \sigma(u_i, v)$

11: $C'_D(v) = \sum_{i=1}^{n} \sigma(u_i, v) - \sum_{i \in s}^{n} \sigma(u_i, v)$

12: $C_{DD}(v) = p(u,v) * C'_D(v) + \sum_{i \in neighbors(v)} p(u,v) * C'_D(v), p(u,v) > \theta$

13: end for

14: end for

15: output S

3 实验和评估

本文在两个真实网络的数据集上对 IDD 算法进行了仿真实验，并且对实验结果进行分析对比，比较对象是贪婪算法的影响范围及其他启发式算法的时间效率。实验目标是，在保持良好的时效性的基础上，尽可能的提升其影响范围。

3.1 实验场景

本文针对的是大型社交网络，因此选取了 2 个大型真实网络的数据集（Epinions 和 DBLP），这两个数据集的规模如表 1 所示。

表 1 数据集 Epinions 和 DBLP 的数据统计

数据集	Epinions	DBLP
节点数量	75k	655k
边的数量	655k	2.0M
度平均值	13.4	6.1
度最大值	3079	588

在这 2 个数据集的基础上，本文分别对 IDD 算法，贪婪算法（Greedy）和另外 2 种启发式算法 PMIA 和 Diffusion degree 算法进行了评估，其中 PMIA 是目前启发式算法中表现最好的算法之一。以贪婪算法（Greedy）的影响范围作为影响范围的基准，以另 2 种启发式算法的时效性作为时效性的基准。本文中 θ 采用的是网络中节点间激活概率的均值。

3.2 仿真结果分析

本文主要是在影响范围和时效性两方面对 4 种算法进行了对比，即种子节点集的质量和运行时间。贪婪算法的耗时太多，需要三天的时间，无法在两个数据集上实时地完成计算，所以只将其影响范围作为评估基准。

3.2.1 影响范围

影响范围即种子节点集的质量的评估主要是基于最终网络中被激活的节点数。本文首先完成网络 $G = (V,E)$ 的初始化。然后在 $G = (V,E)$ 上分别采用 4 种算法来仿真计算出种子节点集 S。如图 1 所示，在数据集 Epinions 上，IDD 算法得到的种子节点集的质量要优于 Diffusion degree 算法，其影响范围与 PMIA 算法基本相当，也接近贪婪算法的影响范围。如图 2 所示，在数据集 DBLP 也得到了一致的结果。（图中 X 坐标为种子节点集 S 的大小，Y 坐标为网络中被激活节点数）。

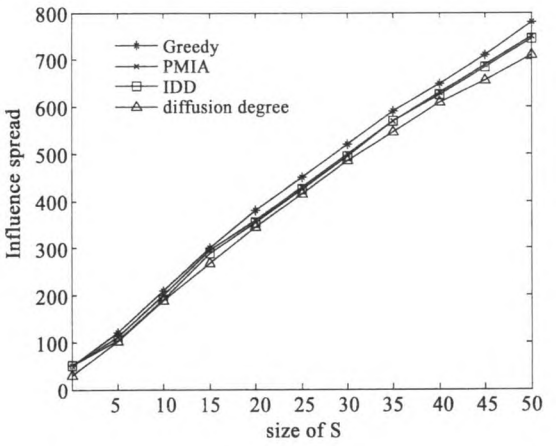

图1 Epinions 上的种子节点集质量对比

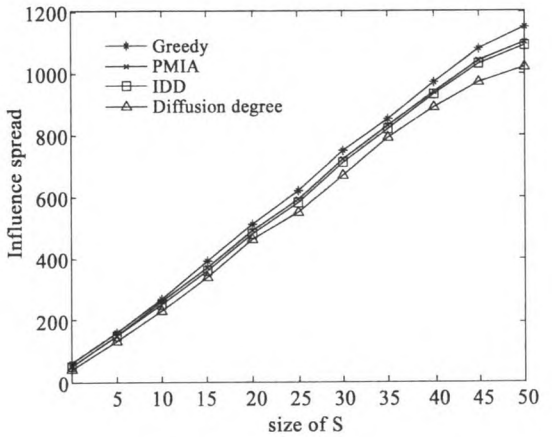

图2 DBLP 上的种子节点集质量对比

3.2.2 时效性

时效性即找出种子节点集所需运行时间。各算法的运行时间如表2所示,贪婪算法的运行时间太久,因此不在比较范围内。由于在对节点影响力计算时,进行了进一步的过滤,还综合了 SingleDiscount 算法的更新节点影响力的操作,IDD 算法的运行时间相较于 Diffusion degree 算法有一点提高,但是相较于 PMIA 算法具有相当的优势,拥有更好的时效性。

表2 运行时间

算法	Epinions	DBLP
Greedy	—	—
PMIA	15s	3m
diffusion degree	5s	15s
IDD	6s	20s

4 结束语

本文基于 diffusion degree 算法做出了改进,在考虑潜在影响力的时候进一步考虑了节点潜在影响力的有效性,更加准确地判断节点的影响力,再综合了算法 SingleDiscount 中的核心思想,提出了 IDD 算法,该算法在影响范围上更接近 KK 贪婪算法的影响范围,基本与 PMIA 算法相当,超越了 Diffusion degree,并且在时效性上有较好的表现。相较于贪婪算法,启发式算法更适合于大型的在线社交网络。对于 IDD 算法还有许多值得改进与进一步研究的地方,例如如何更加动态的选择种子节点,能否高效地模拟传播过程。同时,本文的研究只针对 IC 模型,可以将 IDD 算法应用于线性阈值模型进行进一步的研究。

参 考 文 献:

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

[2] Chen W, Wang Y, Yang S. Efficient influence maximization in social networks[C]//Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2009.

[3] Chen W, Wang C, Wang Y. Scalable influence maximization for prevalent viral marketing in large-scale socialnetworks[C]//KDD, 2010:1029 - 1038.

[4] Chen W, Lu W, Zhang N. Time-critical influence maximization in social networks with time-delayed diffusion process (full technical report)[R]. CoRR abs/1204.3074. 2012.

[5] Chen W, Collins A, Cummings R, et al. Influence maximization in social networks when negative opinions may emerge and propagate [C]//SDM, 2011.

[6] Chen W, Yuan Y, Zhang L. Scalable influence maximization in social networks under the linear threshold model[C]//ICDM, 2010.

[7] Goyal A, Lu W, Lakshmanan L V S. Celf++: optimizing the greedy algorithm for influence maximization in social networks[C]//WWW (Companion Volume), 2011.

[8] Goyal A, Bonchi F, Lakshmanan L V S. A data-based approach to social influence maximization[C]//PVLDB, 2011.

[9] Goyal A, Bonchi F, Lakshmanan, L V S. Learning Influence Probabilities in Social Networks[C]//Proceedings of the Third International Conference on Web Search and Web DataMining, New York, USA,2010:241 - 250.

[10] Goyal A, Lu W, Lakshmanan L V S. SIMPATH: An Efficient Algorithm for Influence Maximization under the Linear Threshold Model[C]//Proceedings of the 11th IEEE International Conference on Data Mining, Vancouver, Canada, 2011:211 - 220.

责任编辑:薛慧心