基于 LT$^+$ 模型的社交网络影响力最大化研究

蔡国永 裴广战
(桂林电子科技大学广西可信软件重点实验室 桂林 541004)

摘 要 影响力最大化问题的目标是寻找社交网络中一组种子结点集合,在给定的传播模型下,使得这些结点最终传播的影响范围最大。Kempe 和 Kleinberg 提出的贪心算法可以获得很好的影响范围,但是因复杂度太高而并不适用于大型社交网络。Chen 和 Yuan 等人基于线性阈值(LT)模型提出了构造局部有向无环图的启发式算法,但是 LT 模型只考虑了邻居结点的直接影响力,忽略了结点之间存在的间接影响力。因此,在 LT 模型的基础上,结合网络中结点之间存在的间接影响力,提出了 LT$^+$ 影响力模型,并利用构造局部有向无环图的启发式算法求解 LT$^+$ 模型的影响力最大化,称为 LT$^+$ DAG 算法。真实数据集上的对比实验表明,LT$^+$ DAG 算法具有更好的影响范围以及较好的可扩展性。

关键词 社交网络,影响力最大化,贪心算法,传播模型
中图法分类号 TP311.13 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.9.018

Influence Maximization Based on LT$^+$ Model in Social Networks

CAI Guo-yong PEI Guang-zhan
(Guangxi Key Lab of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China)

Abstract Influence maximization is a problem of finding a small group of seed nodes in a social network, so that the influence scope of spread in the network is maximized. Kempe and Kleinberg proposed a greedy algorithm which has a wide influence, but its high complexity makes it unsuitable for large social network. Chen and Yuan proposed a heuristic algorithm called local directed acyclic graphs based on linear threshold (LT) model. But LT model only considers the direct influence of neighbors nodes, and ignores the indirect influence between the settled nodes. Therefore, combining with the indirect influence between nodes in the network, we proposed LT$^+$ influence model based on LT model. We also used the local directed acyclic graphs (DAGs) heuristic algorithm to solve the problem of influence maximization, known as LT$^+$ DAG algorithm. Extensive experiments were done on real-world dataset to compare the proposed method with other influence maximization algorithms. The result shows that the proposed method can achieve better influence scope and extensibility.

Keywords Social network, Influence maximization, Greedy algorithms, Propagation model

1 引言

影响力最大化问题是研究在给定的传播模型下如何选择社交网络中一组初始种子结点集,使得这些结点最终导致传播的影响范围最大。Kempe 和 Kleinberg 等人在独立级联模型(IC model)和线性阈值模型(LT model)下研究了影响力最大化问题,证明了社交网络影响力最大化问题是 NP-hard 的[1];他们提出了一个贪心算法,通过依次选择具有最大边际影响力(marginal influence)的结点,使得该贪心算法最终可以求得接近 1−1/e 的近似解。然而,该贪心算法的复杂度仍然太高,并不适用于大型的社交网络。

为了解决贪心算法效率不高的问题,研究者们提出了一系列改进的贪心算法和新的启发式算法[2-4]。Chen 和 Wang 等人研究了 LT 模型下的影响力传播,提出了构造局部有向无环图的启发式算法 LDAG[5],但是 LDAG 算法中的 LT 模型只考虑了邻居结点的直接影响力,而忽略了结点之间存在的间接影响力,因此 LDAG 算法得到的传播结果与实际情况偏离较大。针对这个问题,本文提出了考虑间接影响力的 LT$^+$ 传播模型,并在 LT$^+$ 模型下应用类似 LDAG 的思想求解,实验结果表明,在 LT$^+$ 模型下 LDAG 算法可以计算得到更精确的传播范围。

2 问题背景描述

给定社交网络图 $G=(V,E,p)$,V 表示网络中的结点集合,$E\subseteq V\times V$ 表示结点之间存在的边集合,p 表示边上的权重,令 $\sigma_m(S)$ 表示传播模型 m 下传播过程结束后激活结点的总数目,则社交网络影响最大化问题可以形式化为:对给定的社交网络图 G,传播模型 m 和 $k\leqslant|V|$,寻找 k 个结点的种子集合 S,使得 $\sigma_m(S)$ 最大。

虽然该问题是 NP 难问题,但 Kempe 和 Kleinberg 等人

到稿日期:2015-07-11 返修日期:2015-08-19 本文受广西研究生教育创新计划资助项目(YCSZ2015147)资助。
蔡国永(1971−),男,博士,教授,主要研究领域为社会媒体挖掘、可信网络计算;裴广战(1987−),男,硕士生,主要研究领域为社交网络信息传播、机器学习,E-mail:930275779@qq.com。

证明了影响力最大化在 IC 和 LT 传播模型下满足单调性和子模性，因此他们提出了一种贪心算法来近似求解影响力最大化问题，并可以获得很好的近似解，贪心算法如算法 1 所示。

算法 1 贪心算法

输入：G，k，$\sigma_m(S)$

输出：种子结点集合 S

1. S←∅
2. while|S|≤k do
3. $u \leftarrow \underset{w \in V-S_0}{\operatorname{argmax}}(\sigma_m(S+w)-\sigma_m(S))$
4. S←S+u //插入集合
5. output S

由算法 1 可以看出，贪心算法每一步都需要计算将所有未激活结点作为种子结点而带来的激活范围增量，即贪心算法需要选择的是当前具有最大边际影响力的结点作为种子节点，该算法可获得很好的近似结果，但很耗时，不适用于大型的社交网络。

在影响力最大化问题求解过程中要学习结点之间的传播影响力，也即传播模型中的一些参数。已有许多机器学习算法能从真实数据集中学习影响力传播的模型参数[6-9]，通过真实数据集学习出的模型参数，可以用于生成影响力最大化问题中的影响力传播图。

3 影响力模型及影响最大化

3.1 影响力模型 LT⁺

Kempe 和 Kleinberg 提出的 LT 模型是最基本的传播模型之一，但是 LT 模型只考虑邻居结点的直接影响，忽略了结点之间的间接影响。然而在社交网络中，如果存在结点 u 到 t 的一条边和 t 到 v 的一条边，则结点 u 可以通过 t 对 v 产生某些影响，即网络中存在间接影响的特点，因此结合网络中间接影响的特点，提出了 LT⁺ 影响力模型，定义如下。

LT⁺ 是一个带权的社交网络图 $G=(V,E,p)$，V 是社交网络中结点的集合，$E\subseteq V\times V$ 是边的集合，$p:V\times V\rightarrow[0,1]$ 表示边上的权重，且 $\sum_{u\in N^{in}(v)} p(u,v)\leqslant 1$，其中 $N^{in}(v)$ 表示结点 v 的入度结点。如果存在结点 u 到 v 的一条边，令 $F_{u,v}$ 表示结点 u 对 v 的影响力，则 $F_{u,v}$ 的计算公式可以表示为：

$$F_{u,v}=\sum_{w\in N^{in}(v)\cap u\in S} F_{u,w}\cdot p(w,v) \tag{1}$$

式(1)的递归出口条件为 1，即一个用户对自身的影响力为 1。令 $K_{S,v}$ 表示种子集合 S 传播到 v 的激活概率，当 $v\in S$ 时，即结点 v 是激活结点，此时集合 S 传播到结点 v 的激活概率为 1，则 $K_{S,v}$ 的计算公式如下：

$$K_{S,v}=\begin{cases}1, & \text{if } v\in S\\ \sum_{u\in S}F_{u,v}, & \text{其他}\end{cases} \tag{2}$$

对给定的种子结点 $S\subseteq V$，LT⁺ 模型的传播机制如下：1）每一个结点 $v\in V$ 随机地选择一个阈值 $\theta_v\in[0,1]$；2）对 $t\geqslant 1$ 的每个时刻，令 $S_t=S_{t-1}$，对任意的非激活结点 $v\in V\backslash S_{t-1}$，如果结点 v 的激活邻居结点传播到 v 的影响力之和大于或等于 θ_v，即 $K_{S,v}\geqslant\theta_v$，则结点 v 被激活，把结点 v 加入集合 S_t 中。图 1 给出了 LT⁺ 模型传播过程的一个例子，在图 1 中，结点括号中的值表示该结点的阈值，边上的值表示权重，灰色的结点表示初始的种子结点，黑色的结点表示在相应时刻被激活的结点。

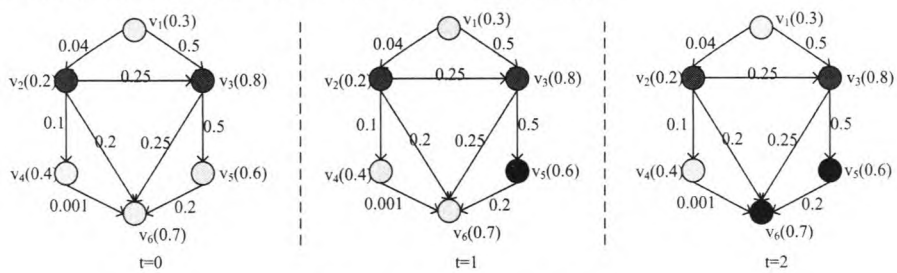

图 1 模型传播过程的例子

从图 1 中可以看出，在 $t=0$ 时刻，结点 v_2 和 v_3 为初始的种子结点，在 $t=1$ 时刻，由式(2)可以计算出结点 v_2 和 v_3 传播到 v_5 的激活概率：

$$\begin{aligned}K_{S,v_5}&=\sum_{u\in S}F_{u,v_5}\\&=F_{v_2,v_5}+F_{v_3,v_5}\\&=F_{v_2,v_3}\times p(v_3,v_5)+F_{v_3,v_3}\times p(v_3,v_5)\\&=0.25\times 0.5+1\times 0.5=0.625>0.6\end{aligned}$$

因此结点 v_5 被激活，在图 1 中用黑色表示；而结点 v_2 和 v_3 传播到 v_6 的激活概率为：

$$K_{S,v_6}=0.5125<0.7$$

所以结点 v_6 未被激活；在 $t=2$ 时刻，因为结点 v_5 已经处于激活状态，所以 v_2，v_3 和 v_5 传播到 v_6 的激活概率为：$K_{S,v_6}=0.8375>0.7$，则结点 v_6 被激活。

3.2 LT⁺ 模型的影响力最大化

$K_{S,v}$ 表示的是集合 S 传播到结点 v 的激活概率，令 $\sigma_{LT^+}(S)$ 表示影响力在 LT⁺ 模型下的传播范围，则 LT⁺ 模型下影响力传播可以表示为：

$$\sigma_{LT^+}(S)=\sum_{v\in V}K_{S,v} \tag{3}$$

LT⁺ 模型的影响力最大化可以表示为：对给定网络图 $G=(V,E,p)$，LT⁺ 模型和 $k\leqslant|V|$，寻找集合 $|S|=k$，使得 $\sigma_{LT^+}(S)$ 最大。可以证明 $\sigma_{LT^+}(S)$ 满足单调性和子模性，Wei Chen 等人证明了在有向无环图下计算影响力的传播可以在线性时间内得到求解。因此，我们也用构造局部有向无环图(local directed acyclic graph)的方法来近似 LT⁺ 模型的影响力最大化。

3.3 局部有向无环图的构造

在 LT⁺ 模型中，给定一个社交网络图 $G=(V,E,p)$，结点 $v\in V$ 和一个阈值 $\delta\in[0,1]$ 作为输入，需要计算出一个局部的有向无环图 $G'=(V',E',p)$，使得：1）G' 是图 G 的一个子图；2）$v\in V'$；3）对所有的 $u\in V'$，满足 $p(u,v)\geqslant\delta$。令 $LT^+DAG(v,\delta)$ 表示构造的结点 v 的局部有向无环图，如算法 2 所示。

算法 2 计算 $LT^+DAG(v,\delta)$

输入：G=(V,E,p)

输出：G′=(V′,E′,p)

1. V′←∅;E′←∅;
2. ∀u∈V,$F_{u,u}$=1
3. while $\max_{u\in V\setminus V'} p_{u,v}\geqslant\delta$ do
4. t=$\operatorname{argmax}_{u\in V\setminus V'} p_{u,v}$
5. E′=E′∪{(t,v)} //增加边
6. V′=V′∪{t} //增加结点
7. for x∈N^{in}(t) do
8. E′=E′∪{(x,t)}
9. V′=V′∪{x}
10. end for
11. end while

图 2 给出了对结点 v_6 构造局部有向无环图的一个例子，在图 2 中，选取了 δ=1/320，与 LDAG 算法中选取的 δ 相同，Wei Chen 等人在实验过程中发现了 δ 在 1/320 和 1/640 之间所有的传播过程可以获得相似的结果，当选取 δ=1/320 时可以获得更好的传播结果，因此在所有的测试中同样选取 δ=1/320[5]。图 2(a)表示原图 G，图 2(b)表示对结点 v_6 构造的一个局部有向无环图 G'，其中结点内的数字表示结点在局部有向无环图中加入的顺序，边上的数字表示边被加入无环图中的顺序。

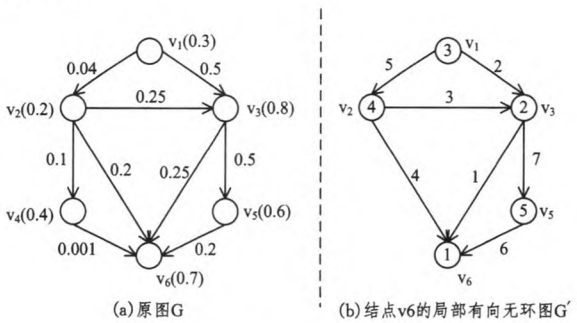

图 2 构造局部有向无环图的一个例子

算法 2 给出了构造 LT⁺DAG 的过程，由于网络中一个图可以包含多个有向无环图，算法的第 3 行表示选择图 G 中最大的有向无环图 G'。在构造好 LT⁺DAG 之后，可以用算法 3 计算影响力的传播并选择 k 个种子结点。

算法 3 LT⁺DAG 算法选择种子节点

输入：G,k,σ_{LT^+}(S)

输出：种子节点 S

1. S←∅
2. for v∈V do
3. 用算法 2 计算 LT⁺DAG(v,δ)
4. 选择有向无环图中结点数目和边最多的一个，用 D 表示
5. end for
6. ∀u 在 D 中，用算法 4 计算 $F'_{u,v}$
7. for u∈D do
8. incInf(u) +=$F'_{u,v}$
9. end for
10. while |S|≤k do
11. s=$\operatorname{argmax}_{v\in V\setminus S}$ {incInf(v)}
12. S=S∪{s}
13. end while

算法 3 的 1—8 行为准备阶段，首先对每个结点 v 构造局部有向无环图，如果结点 u 被选为种子结点，需要更新结点 u 增加的影响力 $incInf(u)$，$F'_{u,v}$ 的计算方法由算法 4 给出。算法的 9—13 行为主循环，进行种子结点的选取，结点 s 是每次选取的当前具有最大影响力增量的节点，即当一个结点被选为种子结点后，该结点到邻居结点增加的影响力最大。

算法 4 对所有的 $u\in LT^+DAG$，计算 $F'_{u,v}$

输入：LT⁺DAG(v,δ)

输出：$F'_{u,v}$

1. ∀u∈LT⁺DAG(v,δ)，$F'_{u,v}$=0
2. 对所有可以到达 v 的结点进行拓扑排序并存入序列 ρ 中，结点 v 在序列最前。
3. for each node u∈ρ\S do
4. $F'_{u,v}=\sum_{w\in N^{in}(u)\cap\rho} F_{w,u}\cdot p(u,v)$
5. end for

算法 4 给出了计算 $F'_{u,v}$ 的方法，$F'_{u,v}$ 表示的是当结点 u 被选为种子结点后结点 u 对 v 的影响力增加了多少，首先使用拓扑排序的方法将 LT⁺DAG 中可以到达 v 的结点保存至序列 ρ 中，然后对每个可以到达 v 的结点更新 u 对 v 增加的影响力，其中 $N^{in}(u)$ 表示结点 u 的入度结点，$F_{w,u}$ 由式(1)计算得出。

4 实验与分析

4.1 实验数据集

实验用到的数据集的统计信息如表 1 所列。

表 1 数据集的统计信息

序号	数据集	结点数	边	平均度
1	Epinions	75879	508837	6.7
2	Flickr	15229	68979	6.1

第 1 个数据集是电子商务网站 Epinions 的数据集，在 Epinions 数据集中，结点 v 到 u 的有向边表示 v 信任 u（因此 u 可以影响 v），该数据集来自 Stanford 大学的大型社交网络数据收集网站[1)]。

第 2 个数据集是 Flickr[2)] 网站的数据，Flickr 是一个著名的照片分享平台，用户可以在 Flickr 中向朋友分享自己的照片，Flickr 数据集使用的是文献[8]中的数据。

4.2 实验结果分析

为了验证 LT⁺模型的有效性，对比了不同算法影响力最大化，对比的算法如下。

(1)$LT^+DAG(\delta)$：算法 2 中的参数 δ 用于控制构造有向无环图的大小，在实验中选取 δ=1/320，与 LDAG 算法选取参数相同。

(2)Greedy 贪婪算法：对每个种子结点的选择，运行了 20000 次的蒙特卡洛模拟以获得影响力传播的准确估计。

(3)PageRank 算法：选取了 k 个具有最高 PageRanks 值的结点作为种子结点。

(4)$LDAG(\delta)$算法：LT 模型构造局部有向无环图，参数 δ=1/320。

1) http://snap.stanford.edu/data/index.html

2) https://www.flickr.com/groups

在所有的实验中，选取了 50 个结点作为种子结点，图 3 和图 4 分别表示 Flickr 和 Epinions 数据集的影响力转播。

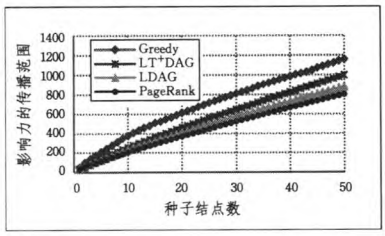

图 3　Flickr 数据的影响力传播

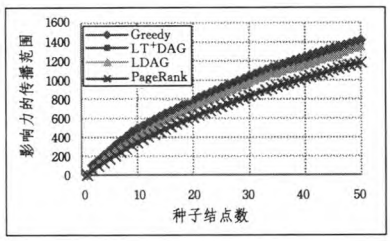

图 4　Epinions 数据的影响力传播

图中的横轴表示种子结点的个数，纵轴表示在相应种子个数下各个算法计算得出的影响传播范围。为了验证 LT^+ DAG 算法的有效性，以经典的贪婪算法为基准，由 CELF 优化算法实现，CELF 优化算法可以获得 $1-1/e$ 的精确解[10]。从图 1 和图 2 中的实验结果可知，相比 LDAG 和 PageRank 算法，LT^+ DAG 算法在相同种子结点个数下均具有更准确的影响范围估算，即本文的方法更靠近贪心算法。

表 2 给出了不同算法运行时间的对比，从表 2 中可以看出，贪婪算法运行时间最长，算法的可扩展性最差；PageRank 算法选择最高 PageRank 值的结点，不需要进行迭代，所以其运行时间最短。然而 PageRank 不能获得更好的影响范围，LT^+ DAG 算法的运行时间和 LDAG 的时间相近，LT^+ DAG 算法能获得较好的可扩展性。

表 2　算法的运行时间对比

数据集	运行时间(min)			
	Greedy	LT^+ DAG	LDAG	PageRank
Epinions	157.87	19.27	18.25	9.65
Flickr	102.3	10.8	7.6	7

4.3　算法复杂度分析

令 m 和 n 分别表示图 G 中结点的个数和边数，$\overline{t_1}$ 表示贪心算法每个种子结点的平均激活范围，m_θ 和 n_θ 分别表示局部有向无环图 G' 中结点的个数和边数，$\overline{t_2}$ 表示构造一个局部有向无环图的平均时间，则算法 2 中构造局部有向无环图时 LT^+ DAG 花费的时间为 $O(n\overline{t_\theta})$，在构造局部有向无环图后，用算法 3 选择种子结点，算法 3 中更新 $incInf(u)$ 的时间为 $O(\lg n)$，则算法的总时间复杂度为 $O(n\overline{t_\theta}+km_\theta n_\theta(m_\theta+\lg n))$。

LDAG 算法的复杂度为：$O(n\overline{t_\theta}+kn_\theta m_\theta \lg n)$。贪心算法的复杂度为：$O(n\overline{t_1}km)$。Pagerank 算法的复杂度为：$O((n+1)m)$。

$\overline{t_\theta}$ 和 $\overline{t_1}$ 是比较接近的值，而 n_θ 和 m_θ 值远小于 m，因此可以得出 LT^+ DAG 算法与 LDAG 算法的复杂度接近，而比贪心算法的复杂度低很多。

结束语　本文在 LT 模型的基础上，结合社交网络中存在间接影响的特点，提出了 LT^+ 的影响力模型，并通过构造有向无环图的方法近似 LT^+ 模型的影响力最大化，实验表明，提出的方法在有较好的可扩展性的同时能获得较好的激活范围。然而提出的算法依然有需要改进的地方，文中只考虑了静态的网络结构，未来工作是将文中的模型扩展到动态社交网络的影响力最大化中。

参 考 文 献

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

[2] Kimura M, Saito K. Tractable models for information diffusion in social networks[M]// Knowledge Discovery in Databases: PKDD 2006. Springer Berlin Heidelberg, 2006: 259-271

[3] Leskovec J, Krause A, Guestrin C, et al. Cost-effective outbreak detection in networks[C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2007: 420-429

[4] Goyal A, Lu W, Lakshmanan L V S. Celf++: optimizing the greedy algorithm for influence maximization in social networks [C]// Proceedings of the 20th International Conference Companion on World Wide Web. ACM, 2011: 47-48

[5] Chen W, Yuan Y, Zhang L. Scalable influence maximization in social networks under the linear threshold model[C]// 2010 IEEE 10th International Conference on Data Mining (ICDM). IEEE, 2010: 88-97

[6] Anagnostopoulos A, Kumar R, Mahdian M. Influence and correlation in social networks[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2008: 7-15

[7] Tang J, Sun J, Wang C, et al. Social influence analysis in large-scale networks[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2009: 807-816

[8] Goyal A, Bonchi F, Lakshmanan L V S. Learning influence probabilities in social networks[C]// Proceedings of the Third ACM International Conference on Web Search and Data Mining. ACM, 2010: 241-250

[9] Xiang B, Liu Q, Chen E, et al. Pagerank with priors: An influence propagation perspective[C]// Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. AAAI Press, 2013: 2740-2746

[10] Goyal A, Lu W, Lakshmanan L V S. Celf++: optimizing the greedy algorithm for influence maximization in social networks [C]// Proceedings of the 20th International Conference Companion on World Wide Web. ACM, 2011: 47-48