第50卷第8期 中国科学技术大学学报 Vol. 50, No. 8
2020年8月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Aug. 2020
文章编号:0253-2778(2020)08-1058-06

基于全局的引文网络影响力最大化算法

张文静[1,2],班志杰[1]
(1.内蒙古大学计算机学院,内蒙古自治区社会计算与数据处理重点实验室,内蒙古呼和浩特 010000;
2.呼和浩特市规划展览馆,内蒙古呼和浩特 010000)

摘要:从大量的期刊论文中搜寻出最具有影响力的若干篇论文对于学术研究具有重要意义,但现有影响力最大化算法需要结合贪心算法,时间复杂度较高.依据论文引用网络中引用关系的时间单向性和无环特征,提出一种基于节点全局影响力的影响力最大化算法.该算法主要包括:①计算所有节点的全局影响力.结合引文网络的发表时间特性,构造上三角稀疏影响方阵.在线性阈值传播模型的基础上,利用节点间的直接、间接路径影响以及累积计算规则模拟影响力在网络上的传播过程.方阵每进行一次运算,会将全部节点的影响向下传播一跳,得到下一个路径的影响,并统计全部影响,最终得到表示所有节点全局影响力的方阵;②将全部节点按全局影响力排序.选择前 n 个节点作为候选节点来选取 k 个种子节点,在选取的过程中避免影响力较大节点的聚集情况.以真实的学术引文网络数据集为实验数据,将提出的算法与两种基准算法从激活范围和运行时间两个方面进行对比.实验结果表明,该算法大大降低了时间复杂度,且激活范围接近于贪心算法.

关键词:引文网络;社交网络;影响力最大化;传播模型
中图分类号:TP391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2020.08.003

引用格式:张文静,班志杰.基于全局的引文网络影响力最大化算法[J].中国科学技术大学学报,2020,50(8):1058-1063.
ZHANG Wenjing, BAN Zhijie Citation network's influence maximization algorithm based on global influence[J]. Journal of University of Science and Technology of China, 2020,50(8):1058-1063.

Citation network's influence maximization algorithm based on global influence

ZHANG Wenjing[1,2], BAN Zhijie[1]
(1. *Inner Mongolia A.R. Key Laboratory of Data Mining and Knowledge Engineering, College of Computer, Inner Mongolia University, Hohhot* 010000, *China*;
2. *Hohhot Historical and Cultural City and Intangible Cultural Heritage Protection Center, Hohhot* 010000, *China*)

Abstract: It is of great significance for academic researches to search out the most influential papers from a huge number of Journal papers. However, the existing algorithms for maximizing influence need to be combined with greedy algorithm, which increases the time complexity. According to the time unidirectional and acyclic features of the citation relationship in the citation network, an algorithm is proposed to maximize the influence based on the global influence of nodes. The algorithm mainly includes: ①Calculating the global influence of all nodes. Combined with the publication time characteristics of the citation network, the upper triangular sparse influence matrix is constructed. On the basis of the linear threshold propagation model, the direct and indirect path effects between nodes and the cumulative calculation rule are used to simulate the propagation process of influence on the network. Every time the square matrix is calculated, the influence of all nodes will be propagated down one hop to get the influence of the next path, and all the influences will be counted to finally get the square matrix representing the global influence of all nodes; ②All nodes will be ranked according to the global influence, and the first n nodes will be selected as candidate nodes to select k seed nodes. By the cumulative calculation rule, the proposed algorithm avoids the overlapping of influence among nodes during the process of selecting seed nodes. The real academic citation network data set is taken as the experimental sample, and our algorithm is compared with the two benchmark algorithms in terms of activation range and running time. Experimental results show that the proposed algorithm greatly reduces the time complexity, and that the activation range is close to the greedy algorithm.

Key words: citation network; social network; influence maximization; propagation model

收稿日期:2020-06-05;**修回日期**:2020-07-28
基金项目:国家自然科学基金(61662053)资助.
作者简介:张文静,女,硕士,研究方向:数据挖掘. E-mail:2501648350@qq.com
通讯作者:班志杰,博士/副教授. E-mail:banzhijie@imu.edu.cn

0　引言

随着人类知识总量的不断增长，学术引文网络的规模日益扩大. 在超大规模的学术引文网络中快速准确地寻找最具价值、有影响力的若干论文，选取最具权威性和参考价值的文献，有助于不同学术领域学者快速了解专业领域内容，为今后的学术研究提供有益帮助. 目前，社会网络影响最大化算法采用线性阈值传播模型时[1]，大部分均为启发式算法，但启发式算法需要结合时间复杂度高的贪心算法[2]. 本文研究基于线性阈值模型的学术引文网络的影响力最大化问题，结合引文网络节点发表时间特性，提出一种基于全局的引文网络影响力最大化算法. 该算法不使用贪心算法，而是利用网络本身的无环及后发表论文只能引用先发表论文的特征，通过影响稀疏方阵来快速计算引文网络节点的全局影响力，从而降低时间复杂度. 实验结果表明，本文算法的影响范围优于现有表现较好的启发式算法，且与贪心算法接近，同时时间复杂度低于启发式算法.

1　相关工作

Richardson 和 Domingos 等[3-4]将影响力最大化问题定义为一个算法的问题，并应用于社会网络. 随后，Kempe 等[2]提出将影响力最大化问题看作网络节点的影响力在规定传播模型上传播，求取影响力最大的 K 个节点的离散最优问题，并证明该问题是一个 NP-hard 问题. 同时，他们还提出了多次使用 Monte-Carlo 模拟达到最优解 63% 的爬山贪心算法. 爬山贪心算法虽然效果较优，但是它在选取种子节点的每一轮都对所有节点计算影响力增量，这使贪心算法的时间复杂度较高. 为了改进爬山贪心算法时间复杂度过高的问题，研究者基于独立级联和线性阈值两种传播模型[1,5]以及社会网络的特点开展了很多探索. 基于独立级联模型的改进算法有 CELF[6]、CELF++[7]、MixGreedy[8]等. 基于线性阈值模型的改进算法有初始节点选取策略[9]、基于阈值的社会网络影响最大化算法[10]等，改进办法均是启发式算法与贪心算法分阶段使用. 这些改进算法均降低了最大化算法的时间复杂度. 此外，还有一些算法也进行了一定程度的改进[11-16]，但是在影响范围上与贪心算法仍有差距.

本文提出基于全局的学术引文网络影响力最大化算法，结合引文网络的发表时间特性，不仅提高了时间效率，也得到了很好的结果.

2　理论基础

2.1　引文网络

引文网络是由论文和论文间的引用关系构成的网络[17-19]，引文网络有很多区别于其他网络的特点[20-21]. 首先，引文网络是静态的，文章一旦发表，它的引用文献关系就确定不变，即论文间的引用关系保持固定状态. 其次，引文网络中节点间的相互引用具有时间单向性，只存在后发表的论文引用先发表论文的情况. 最后，引文网络中的节点无自引，论文节点不能引用自身，且引文网络中不存在循环引用关系.

2.2　影响力最大化问题及传播模型

影响力最大化问题是在特定传播模型下的网络中选取目标数量的种子节点，使种子节点能激活最多数量的节点. 基本的传播模型有线性阈值模型和独立级联模型. 本文算法将线性阈值模型作为传播模型.

在线性阈值模型下，对于一个有向网络 $G(V, E)$，$\mathrm{IN}(v)$是节点 v 的入边节点集合，影响权重代表了节点之间的影响大小，$A(v)$为节点 v 已经激活的入边节点集合. 线性阈值模型下网络中节点的影响力传播过程如下：

假设 v 为未被激活的一个节点，则在 T 时刻，v 的被激活的节点入边集 $A(v)$对 v 的影响之和大于等于它的激活阈值，则在 $T+1$ 时刻，节点 v 变为激活状态. v 节点可以继续影响它的出边邻居节点. 若未被激活则影响累积，不断重复上述过程，直到网络 G 不再有节点变为激活状态.

3　基于全局的引文网络影响力最大化算法

3.1　节点间的多路径影响

经过探究影响力在网络上的传播过程，我们发现在线性阈值模型下一个节点对另一个节点存在影响时，大多数情况下是由直接路径的影响和间接路径的影响组成. 这就是节点间的多路径影响. 下面给出多路径影响的定义.

定义 3.1　直接路径的影响. 指激活节点对其直接出边邻居节点的影响. 图 1 中，节点 v_1、v_2 和 v_3 是节点 u 的直接出边邻居节点，节点 u 被激活后对这 3 个节点的影响为直接路径的影响.

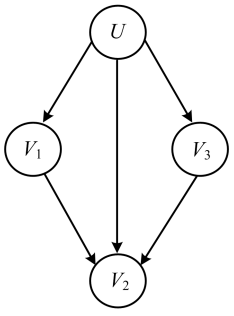

图 1　引文网络示意图

Fig. 1　Citation network diagram

定义 3.2　间接路径的影响. 指激活节点通过激活中间节点对其他节点产生的影响. 图 1 中，节点 u 通过激活节点 v_1 和 v_3 间接影响节点 v_2，节点 v_1 和 v_3 为中间节点.

定义 3.3　节点的多路径影响. 指影响力在网络上传播的过程中，一个节点对另一个节点的影响由直接路径的影响和间接路径的影响组成. 图 1 中，

节点 u 对节点 v_2 的影响路径有 3 条，节点 u 对节点 v_2 的一条直接影响路径，即 $u->v_2$. 节点 u 通过节点 v_1 和节点 v_3 对节点 v_2 的两条间接影响路径，即 $u->v_1->v_2$ 和 $u->v_3->v_2$.

在线性阈值模型下，以图 1 为例给出节点 u 对节点 v_2 的多路径影响过程，具体描述为：

(Ⅰ)假设节点 v_1，v_2 和 v_3 的激活阈值 $\theta_{v1}=\theta_{v2}=\theta_{v3}=0.5$，则节点 u 被激活后，它首先以 $b_{uv1}=0.6$，$b_{uv2}=0.3$ 和 $b_{uv3}=0.5$ 影响它的直接相邻节点 v_1，v_2 和 v_3，此时 $b_{uv1}>\theta_{v1}$、$b_{uv3}>=\theta_{v3}$，那么节点 v_1 和 v_3 被节点 u 激活. 由于 $b_{uv2}<\theta_{v2}$，节点 u 未激活节点 v_2，且线性阈值模型是积累模型，所以节点 u 对节点 v_2 的影响积累下来，此时节点 v_2 的剩余激活阈值为 $\theta_{v2}-b_{uv2}=0.2$.

(Ⅱ)在(Ⅰ)中，节点 u 未通过直接影响路径激活节点 v_2，但是直接激活了节点 v_1 和 v_3. 这一轮中，被 u 直接激活的节点 v_1 和 v_3 继续影响未激活的节点 v_2. 假设节点 v_1 和 v_3 对节点 v_2 的影响分别为 $b_{v1v2}=0.3$，$b_{v3v2}=0.4$，由(Ⅰ)可知节点的剩余阈值 $\theta_{v2}=0.2$，而 $b_{v1v2}+b_{v3v2}>\theta_{v2}$，则节点 u 通过节点 v_1 和 v_3 的间接路径影响激活节点 v_2.

节点 u 被激活后通过直接路径激活了节点 v_1、节点 v_3，未能直接激活节点 v_2，随后节点 u 通过它直接激活的节点 v_1 和 v_3 间接激活了节点 v_2. 任意两节点之间可能存在直接路径的影响和间接路径的影响，这就是节点间的多路径影响. 一个节点被激活后可能直接激活若干节点，也可能间接激活若干节点.

3.2 直接路径影响和间接路径影响的计算

直接路径的影响只需计算节点对直接出边邻居节点的影响. 间接路径影响的计算较复杂，若能快速准确地计算全部节点之间间接路径的影响，那么就可以得到全部节点单独激活时的全局影响力. 计算节点的全局影响力，实际上是计算单个节点被激活后至影响完毕时能直接与间接激活节点的总数目.

(Ⅰ)直接路径影响的计算

我们用方阵表示节点之间的直接影响路径. 假设一个网络中有 n 个节点，构建 $n\times n$ 方阵，n 行和 n 列分别代表 n 个节点，节点的激活阈值为(θ_1，θ_2，θ_3，…，θ_n)，网络中的任意一个节点 u 对任意一个节点 v 的影响权重为 b_{uv}，方阵中每一项的值是行节点对列节点的影响权重与列节点激活阈值的比值. 当比值大于 1，即影响权重大于节点的激活阈值，则表示激活成功. 统计每一行大于等于 1 的项就可以得到行节点的激活节点数. 若比值小于 1，代表列节点受到的累计影响.

由于本文研究的是学术引文网络的影响力，论文节点不能引用自己，且发表时间早的节点只能影响发表时间晚于它的节点，并假设 n 篇论文节点发表年份均不同(真实引文网络中发表年份相同的论文节点非常多，本例为了更好地介绍多路径影响的计算，设定 n 篇论文节点的发表年份均不相同). 按发表年份先后排序，影响方阵 A 就变为一个上三角方阵. 若网络中共 4 个节点，其影响方阵为

$$A=\begin{bmatrix}0 & \frac{b_{12}}{\theta_2} & \frac{b_{13}}{\theta_3} & \frac{b_{14}}{\theta_4}\\ 0 & 0 & \frac{b_{23}}{\theta_3} & \frac{b_{24}}{\theta_4}\\ 0 & 0 & 0 & \frac{b_{34}}{\theta_4}\\ 0 & 0 & 0 & 0\end{bmatrix}.$$

(Ⅱ)间接路径影响的计算

方阵 A 表示各节点的直接路径影响. 各节点通过直接路径试图激活其相邻节点，被直接激活的节点可以继续向下传播影响，即为间接路径的影响，下一时刻的间接影响路径可以表示为方阵 A 乘方阵 A，即

$$A^2=\begin{bmatrix}0 & 0 & \frac{b_{12}}{\theta_2}\cdot\frac{b_{23}}{\theta_3} & \frac{b_{12}}{\theta_2}\cdot\frac{b_{24}}{\theta_4}+\frac{b_{13}}{\theta_3}\cdot\frac{b_{34}}{\theta_4}\\ 0 & 0 & 0 & \frac{b_{23}}{\theta_3}\cdot\frac{b_{34}}{\theta_4}\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}.$$

A^2 代表路径中被激活的节点继续向下影响，可以用 A^3 表示，A^3 为

$$A^3=\begin{bmatrix}0 & 0 & 0 & \frac{b_{12}}{\theta_2}\cdot\frac{b_{23}}{\theta_3}\cdot\frac{b_{34}}{\theta_4}\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{bmatrix}.$$

对比 A、A^2 与 A^3，3 个方阵的相同位置都表示了同一个节点对另一个节点的影响，其中 A 表示直接路径的影响，A^2 与 A^3 是表示间接路径的影响. 如 $A[0][3]$、$A^2[0][3]$和 $A^3[0][3]$ 是单独激活节点 1 后，节点 1 对节点 4 的直接影响和间接影响.

3.3 全局影响力的计算

全局影响力的计算是求各个路径的影响之和. 例如，在 3.2 节的 A，A^2 和 A^3 中，若和大于等于 1，则表示节点单独激活后的影响传播结束后成功激活节点；若和不大于 1 则表示未激活成功，需要将影响积累. 由于线性阈值模型是积累模型，所以计算全局影响力时需注意如下 3 个问题：

(Ⅰ)间接影响路径可以继续向下传播影响力的前提是中间影响力传播节点被激活. 判断每一时刻的路径影响是否有新的节点被激活，被激活的节点才可以继续向下传播影响力.

(Ⅱ)间接影响路径的计算需考虑节点阈值的变化并及时判断节点的状态. 路径影响的发生使节点的阈值可能变化，且可能被激活. 若经过某时刻一路径影响后某节点被激活，由于节点只能从未激活状态变为激活状态，反之则不可以，所以下一时刻该节点所受到的其他路径的影响已经失去意义，需将这些路径的影响去除掉.

(Ⅲ)间接影响路径的计算边界问题. 为了将节点的全部影响路径统计出来，需要得到影响路径的计算边界，即全部节点影响力传播结束的计算上界.

为了解决上述 3 个问题，我们将引文网络数据中的全部 n 个节点按发表年份先后排序，将全部节点之间的影响权重与被影响节点阈值的比值写成一个 $n\times n$ 的方阵 A，得到的 A 方阵为

$$A=\begin{bmatrix} 0 & \frac{b_{12}}{\theta_1} & \frac{b_{13}}{\theta_1} & \frac{b_{14}}{\theta_1} & \frac{b_{15}}{\theta_1} & \cdots & \frac{b_{1n}}{\theta_1} \\ 0 & 0 & \frac{b_{23}}{\theta_2} & \frac{b_{24}}{\theta_2} & \frac{b_{25}}{\theta_2} & \cdots & \frac{b_{2n}}{\theta_2} \\ 0 & 0 & 0 & \frac{b_{34}}{\theta_3} & \frac{b_{35}}{\theta_3} & \cdots & \frac{b_{3n}}{\theta_3} \\ 0 & 0 & 0 & 0 & \frac{b_{45}}{\theta_4} & \cdots & \frac{b_{4n}}{\theta_4} \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & \frac{b(n-1)n}{\theta(n-1)} \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

A 代表直接路径的影响，间接路径的影响可以用 A 的 n 次方来计算. 为了解决计算全部路径影响之和需要注意的 3 个问题，首先，本文定义了运算规则 $\sharp$，用以解决问题(Ⅰ)和(Ⅱ)；其次，得到路径影响的计算边界，用以解决问题(Ⅲ)；最终得到单个节点的全局影响力方阵.

定义 3.4　$\sharp$ 运算规则. 某一路经的影响方阵为 R，全局影响方阵 E 表示全部路径的影响，$\sharp$ 规则是对 R 和 E 相同位置的值进行运算. 设 E 方阵中某项的值为 pre，R 方阵中相同位置的值为 post. E 和 R 方阵值的变化分以下 3 种情况：

(Ⅰ)当 pre≥1 时，表示行节点的路径累积影响已将列节点激活，此时令 pre=1 表示该列节点已被激活，由于累积影响已经将列节点激活，R 中对该列节点的路径的影响 post 没有机会激活该列节点，不能继续向下传播影响，所以令 post=0.

(Ⅱ)当 pre<1，并且 pre+post≥1，E 中行节点的路径累积影响未将列节点激活，但是加入某一路经的影响 post 后，列节点被激活，则 pre=1 表示列节点已激活，post=1，使该路径继续向下传播影响.

(Ⅲ)当 pre<1，并且 pre+post<1，即 E 中行节点的路径累积影响未将列节点激活，且加入某一路经的影响 post 后，列节点仍未被激活，所以使 pre=pre+post，更新累积影响. 由于 R 中的路径影响未将列节点激活，不能继续传播影响力，使 post=0.

将 $\sharp$ 运算规则应用于计算全局影响力中，得到计算单个节点全局影响力的方法. E 为所要求得的单独激活各节点时全部节点受到的所有路径的累积影响，R 为某一个路径的影响. 初始状态时 E 为一个 $n\times n$ 的全 0 方阵，某一路径的影响为 $n\times n$ 的方阵 R，直接影响矩阵为 $n\times n$ 的方阵 A. 首先令 $R=A$，接着进行 t 轮 E 和 R 的 $\sharp$ 运算. 每进行一次，下一时刻新的路径影响为 $\sharp$ 运算后得到的 R 与直接影响方阵 A 相乘得到的方阵. t 轮计算后得到的 E 为全部行节点单独激活时通过各路径对列节点的总影响力，即单个节点的全局影响力. 将累积影响 E 与某时刻路径的影响 R 进行 $\sharp$ 运算，目的是将某时刻的路径影响 R 加入累积影响 E 中，得到新的 E，并及时判断是否有节点在加入这一路径的影响后被激活，使在这一路径被激活的节点可以继续向下传播影响力；若未被激活则影响不可以继续向下传播. 某路径激活节点继续向下传播的下一时刻的路径影响为 $\sharp$ 运算后的 R 乘以 A，继续将新的路径的影响 R 与新的 E 进行 $\sharp$ 运算.

t 为计算轮数，决定路径影响的计算什么时候完成，用以解决问题(Ⅲ). 每循环一次全部节点向下传播一跳. 由于本文认为同一年发表的论文节点不存在引用关系，只存在先发表的论文影响后发表的论文的情况，因此设引文网络中节点所有不同发表年份的总数为 m，则循环次数 t 应为 $m-1$. 因为节点的影响路径最长就是从最早发表的节点通过所有中间年份发表的节点影响到最晚发表的节点，所以循环 $m-1$ 次就可以覆盖到所有的影响路径. 循环计算 t 次后，E 方阵的每一行是行节点的全局影响力，即单独激活行节点后全部列节点所受到的直接路径和所有间接路径的累积影响，为 1 的项表示激活，小于 1 的项是未被激活后的累计影响.

3.4　种子节点的选取

通过 3.3 节的计算得到了单个节点的全局影响力方阵 E，方阵 E 的每行为 1 的项表示列节点被激活，小于 1 的项是列节点未被激活时受到的累积影响. 选取种子节点时对方阵 E 的每行进行计算，在已选种子节点已有影响的基础上，每次选择能激活最多节点的节点加入种子节点中，计算影响力增量. 在每一轮选取种子节点时，不必计算全部节点的影响力增量. 由于学术引文网络的时间单向性，在引文网络中发表时间越晚的节点所能影响节点的范围越小，且已选取的种子节点越多被激活的节点越多，剩余的未激活的节点就越少，所以只需要将单个节点的影响力排序，每次在选择激活节点数较多的节点中选择. 每选取一个种子节点后需去掉候选节点中种子节点激活的节点. 种子节点的选取步骤如下：

(Ⅰ)当选择 k 个种子节点时，统计方阵 E 每行中 1 的总数，并按数量排序，1 的个数即行节点激活的列节点的数目. 选取激活节点数最多的前 p 个节点作为候选节点. 在 p 个节点中，选取一个激活最多数目的节点为种子节点，将该行数据加入一个 1 行 n 列的全零方阵 S 中，方阵 S 表示种子节点的累积影响.

(Ⅱ)去掉候选节点中被种子节点激活的节点，然后将 S 中不为 1 的项分别与其他候选节点所在行的对应位置项相加. 若有两项相加大于等于 1，则表示被激活，统计和大于等于 1 的个数，选取候选节点中能激活最多未激活节点的节点作为种子节点，并将新选取的种子节点的影响加入种子节点的累计影响中. 在非种子节点及其激活节点中，重新获取 p 个候选节点.

(Ⅲ)重复进行(Ⅱ)，直到选取出目标数量的种子节点.

为了便于计算和表示，在选取种子节点阶段引入 3.3 节定义的 # 运算规则. 若要选择 k 个种子节点，令 S 为表示种子节点累积影响的 1 行 n 列的零矩阵，将方阵 E 的 p 个候选行设为 B_1 到 B_p. 当选择一个种子节点时，循环对 S 和 B_i 进行 # 运算，# 运算后的 B_1 到 B_p 矩阵中 1 数目最多的被选为种子节点；此时使种子节点 B_i 加到 S 中，并将候选节点中的种子节点的激活节点去除，按全局影响力排序补齐候选节点后继续选择下一个种子节点，直到选出全部种子节点，基于全局的影响力最大化算法见算法 3.1.

算法 3.1　基于全局的影响力最大化算法

输入：初始网络中直接路径影响方阵 A，全局影响方阵 E 为 0，初始路径方阵 R 为 0，种子节点选取数 k，种子节点集 V 为空，t 为计算的轮数，其值为引文网络中不同年份总数减 1.

输出：种子节点集 V.

```
1 R=A
2 for j=1 to t /* 计算单个节点的全局影响力 */
3 E#R
4 R=R*A
5 end for
6 S 为 1 行 n 列的非零矩阵 /* 种子节点的累积影响 */
7 根据 E，选取激活节点数最多的前 p 个节点作为候选节点，记为 B1 到 Bp；影响集 Q 为空
8 for  j=1 to k /* 选取 k 个种子节点 */
9 i=1;
10 while  i<=p  do
11     temp_S[i]=S;temp_B[i]=Bi
12     temp_S[i] #temp_B[i]
13 i=i+1
14 end while
/* 选取 # 运算后 1 总数最多的 Bi 对应的节点 vi 加入种子节点 V 中，影响加入到 S */
15 V=V+vi;S#Bi；影响加入 Q 中
16 根据 E 和 Q，选取激活节点数最多的前 p 个节点作为候选节点 B1 到 Bp
17 end for
```

4　实验与结果

本节对本文提出的算法以及基线算法在真实的引文网络数据集上进行大量的实验，从影响范围和运行时间两方面衡量算法.

4.1　实验数据及预处理

实验数据来源于学术社会网络分析与挖掘系统 Aminer[22]. 引文数据均从 DBLP、ACM、MAG (microsoft academic graph) 和其他来源中提取，每篇论文都包含了摘要、作者、发表年份、出版地、标题和引文等特征. 在预处理时选取了 2006 到 2011 年期间发表的论文数据. 我们使用 Pycharm 工具对数据集进行处理，按时间特征对论文及其引用文献进行排列，并且论文节点中只保留了发表年份和引文这两个特征. 实验采用的数据集中的节点总数为 47764，边总数为 69999，孤立节点数为 10334，其中孤立节点在引文网络中既不会受到影响也不能影响其他节点，并不能对影响力的传播做出贡献，因此需要去除网络中的孤立节点.

4.2　基线算法及参数设置

本文选择了两个对比算法：贪心算法和 TBH[10] 算法. 其中，贪心算法的优点是影响范围大，可达到最优解的 63%，且效果稳定，缺点是时间复杂度高. TBH 算法是一种启发式算法，选取一部分最具潜在影响力的节点，使用贪心算法选取剩余部分的种子节点，其优点为时间复杂度低，缺点是影响范围不及贪心算法. 这两个算法中的参数 $c=0.2$.

实验发现，同年发表论文之间的引用很少，所以我们忽略同年发表论文之间的引用关系，因此对于本文提出的算法，t 的取值为最新论文的出版年份 2011 减去最早论文的出版年份 2006，再减 1，即 t 的取值为 5. p 为候选集中候选种子节点的个数，实验发现，当 p 取值为 $k/4$ 时，具有最好的效果.

4.3　实验和结果

我们比较了 3 种算法在种子节点选取以 50 为间隔从 50 到 500 的数量时的影响范围. 图 2 给出了影响范围的实验结果，由图可以看出本文所提出的 GAIM 算法所选取种子节点的影响范围与贪心算法差距很小，且一直优于启发式算法 TBH.

我们还比较了 3 种算法在选取相同数量的种子节点时所花费的时间，如图 3 所示. 从图 3 可以看出，本文提出的 GAIM 算法所花费的时间远少于贪心算法，所花费的时间仅为 TBH 算法的 20% 到 40%.

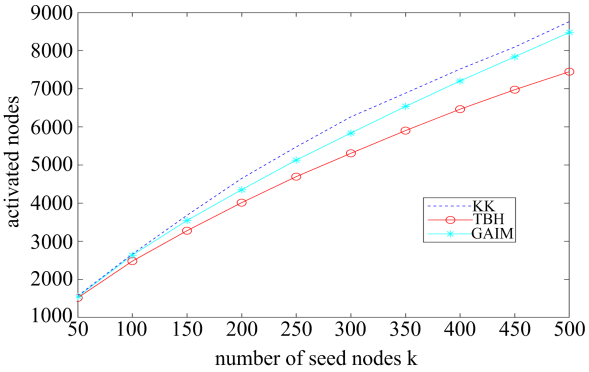

图 2　影响范围的对比

Fig. 2　Comparison of number of influence node

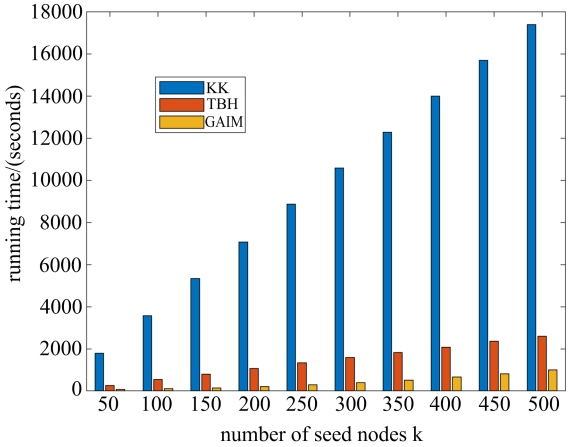

图 3　运行时间对比

Fig. 3　Running time comparison

图 2 和图 3 说明本文提出的算法在影响范围和花费时间上具有明显的优势，影响节点数接近贪心算法，且远高于 TBH 算法所影响的节点数，而所花费的时间远远小于贪心算法，同时远低于 TBH 算法. 本文提出的算法能够取得很好的性能，主要是因为可以得到全部节点单独激活时的全局影响力及累积影响，并将全局影响力的大小排序并选取种子节点，所以所选取种子节点在影响范围上有较好的表现. 同时，在计算单个节点的全局影响力时结合论文节点的引用单向性构建稀疏矩阵模拟影响力的传播，每次只在一部分全局影响力最大的节点中选择种子节点，所以时间效率较高.

5　结论

本文提出了一种基于全局的学术引文网络影响力最大化算法. 通过间接路径和直接路径及所定义的计算规则来计算全局影响力，并依据每个节点的全局影响力的大小来选择种子节点，同时根据论文引用之间的单向性来减少计算次数，从而降低了时间复杂度. 实验结果表明，与已有算法对比，本文提出的算法在时间花费和影响范围上都具有较好的性能，适合于大规模学术引文网络.

本文提出的算法也适合对有向无环图网络(DAG)进行影响力最大化分析. 分析过程中，仅需在构建影响方阵时，对各节点进行拓扑排序，这样就可将本文算法应用到有向无环图中.

参考文献(References)

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

[2] KEMPE D, KLEINBERG J, TARDOS E. Maximizing the spread of influence through a social network [C]// Ninth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Washington DC: ACM, 2003:137-146.

[3] RICHARDSON M. Mining the network value of customers [C]// Seventh ACM SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco, CA: ACM, 2001: 57-66.

[4] RICHARDSON M, DOMINGOS P, GLANCE N. Knowledge-sharing sites for viral marketing [C]// Eighth ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Edmonton, AB: ACM, 2002: 61-70.

[5] WATTS D J. A simple model of global cascades on random networks [J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(9):5766-5771.

[6] LESKOVEC J, KRAUSE A, GUESTRIN C, et al. Cost-effective outbreak detection in network[C]// 13th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, San Jose California: ACM, 2007: 420-429.

[7] GOYAL A, LU W, LAKSHMANAN L V. Celf++: optimizing the greedy algorithm for influence maximization in social networks [C]//20th International Conference on World Wide Web, New York: Association for Computing Machinery, 2011: 47-48.

[8] CHEN W, WANG Y, YANG S. Efficient influence maximization in social networks [C]// 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris: ACM, 2009:199-208.

[9] 田家堂，王轶彤，冯小军. 一种新型的社会网络影响最大化算法[J]. 计算机学报，2011，34(10):1956-1965.

[10] 陈浩，王轶彤. 基于阈值的社交网络影响力最大化算法[J]. 计算机研究与发展，2012，49(10):2181-2188.

[11] AGARWAL S, MEHTA S. Social influence maximization using genetic algorithm with dynamic probabilities[C]// Seventh International Conference on Contemporary Computing, India: IEEE, 2018:1-6.

[12] WENG X, LIU Z, LI Z. An efficient influence maximization algorithm considering both positive and negative relationships[C]// 2016 IEEE TRUSTCOM/BIGDATASE/ISP, Tian Jin: IEEE, 2016:1931-1936.

[13] LIX, CHENG X, SU S, et al. Community-based seeds selection algorithm for location aware influence maximization [J]. NeuroComputing, 2018, 275:1601-1613.

[14] CUI L, HU H, SHUI Y, et al. DDSE: A novel evolutionary algorithm based on degree-descending search strategy for influence maximization in social networks [J]. Journal of Network & Computer Applications, 2018, 103:119-130.

[15] CHEN W, WAND C, WANG Y. Scalable influence maximization for prevalent viral marketing in large-scale social networks [C]// 16th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Washington DC: ACM, 2010: 1029-1038.

[16] JUNG K, HEO W, CHEN W. IRIE: A scalable influence maximization algorithm for independent cascade model and its extensions [J]. Rev Crim, 2011, 56(10):1451-455.

[17] RADICCHI F, FORTUNATO S, VESPIGNANI A. Citation Networks [J]. UnderstandingComplex Systems, 2012:233-257.

[18] DING Y, YAN E, FRAZHO A, et al. Pagerank for ranking authors in co-citationnetworks [J]. Journal of the American Society for Information Science & Technology, 2014, 60(11):2229-2243.

[19] DING Y. Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks [J]. Journal of Informetrics, 2011, 5(1):187-203.

[20] GUAN J, YAN Y, ZHANG J J. The impact of collaboration and knowledge networks on citations [J]. Journal of Informetrics, 2017, 11(2):407-422.

[21] GOLOSOVSKY M, SOLOMON S. Growing complex network of citations of scientific papers: Modeling and measurements [J]. Physical Review E, 2017, 95(1): 012324.

[22] 学术社会网络分析与挖掘系统[EB/OL]. [2018-03]. https://www.aminer.cn.