一种基于信誉机制的科学文献影响力评价方法

冯 磊[1,2] 冀俊忠[1,2] 吴晨生[3]

(北京工业大学多媒体与智能软件技术北京市重点实验室 北京 100124)[1]

(北京工业大学信息学部 北京 100124)[2] (北京市科学技术情报研究所 北京 100048)[3]

摘 要 学术影响力评价一直是文献计量学领域的一个研究热点。已有的大多基于数据挖掘的学术影响力评价方法忽略了恶意活动产生的影响,导致评价结果欠佳。为解决这一问题,提出了一种名为 ReputeRank 的新方法,该方法采用信誉机制来评估引文网络中出版物的有效性。信誉机制包括 3 个阶段:种子集选择阶段、信誉传播阶段和集成计算阶段。首先,ReputeRank 利用 SCI 期刊分区信息选择引文网络中潜在的好种子和坏种子;然后,根据信誉传播思想,信誉度良好的种子指向的论文通常具有更高的可信度,而信誉度不好的种子指向的论文通常具有较低的可信度,该方法使用 TrustRank 和 Anti-TrustRank 评价公式在引文网络中迭代传播信任值和不信任值;最后,根据引文网络中的信任值和不信任值,利用综合集成公式对每篇论文计算评分,并根据评分结果对所有论文降序排列。在 KDD cup 2003 数据集上的实验结果表明,与 3 种影响力评价方法 PageRank,CountDegree 和 SPRank 相比,ReputeRank 能够获得更优的效果。

关键词 引文网络,学术影响力评价,信息传播,信誉度

中图法分类号 TP391 **文献标识码** A

New Method for Ranking Scientific Publications with Creditworthiness Mechanism

FENG Lei[1,2] JI Jun-zhong[1,2] WU Chen-sheng[3]

(Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology, Beijing University of Technology, Beijing 100124, China)[1]

(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)[2]

(Beijing Institute of Science and Technology Information, Beijing 100048, China)[3]

Abstract Evaluating the scientific value of publications has always been a research focus in the field of bibliometrics. However, some mainstream methods based on data mining overlook the influence of malicious activities and result in poor evaluation results. To solve this problem, this paper proposed a new method named ReputeRank, which employs a creditworthiness mechanism to evaluate the effectiveness of publications in the citation network. The creditworthiness mechanism consists of the seeds selection phase, the spread credit phase and the integrated computation phase. First, ReputeRank employs background information on the division of SCI Periodicals to select potential good seeds and bad seeds in the citation network. Then, in light of assumption that good credibility seeds pointing to papers which usually have a higher credible degree while bad credibility seeds pointing to papers which often have a lower credible degree, the method uses TrustRank and Anti-TrustRank evaluation formula to iteratively spread trust values and distrust values over the citation network. Finally, according to the trust and distrust values in the citation network, the method utilizes an integrated equation to comprehensively compute the score value of each paper and arranges all papers in the descending order of the score values. The experimental results on KDD cup 2003 datasets demonstrate that ReputeRank has good performance of effectiveness and robustness compared with PageRank, CountDegree and SPRank.

Keywords Citation network, Evaluation of academic influence, Information dissemination, Credibility

1 引言

科学文献影响力评价是文献计量学中一个重要的研究热点。随着科学文献出版物的不断发表,文献间的引用关系构成了一个大规模的复杂网络——引文网络。引文网络中,文献质量参差不齐,如何使读者识别高质量的论文成为一个挑战性的问题。文章质量也是基金申请和职位晋升的一个重要参考标准[1]。因此,科学文献影响力评价具有重要意义。

概括地说,文献影响力评价方法通常分为两类。第一类是基于数据统计的方法,第二类是基于网络数据挖掘的方法。一些基于统计的方法尝试利用直接引用次数,如论文的被引次数[2]、每篇论文的平均引用次数[3]、H 指数[5]等进行评价。这些方法认为文献的引用次数越多,文献相对来说影响力越大。然而,基于统计的方法的一个缺陷是把所有文献看作是影响力均等的,并不客观准确。由于引文网络中存在大量自引和错误引用[1],这种情形会造成基于引用次数的统计方法

本文受国家自然科学基金重点项目(613300194)资助。

冯 磊(1992—),男,硕士生,主要研究方向为复杂网络、数据挖掘;冀俊忠(1969—),男,博士,教授,CCF 会员,主要研究方向为机器学习、数据挖掘和群智能算法,E-mail:jjz01@bjut.edu.cn(通信作者);吴晨生(1967—),男,博士,研究员,主要研究方向为科技情报、科学普及。

与实际论文影响力存在偏差。第二类方法起源于网络数据挖掘方法,该方法的核心思想是使用数据挖掘方法分析引文网络拓扑结构,提取网络结构相关的信息和知识。最典型的案例是将搜索引擎中万维网网页排序的 PageRank 算法应用于引文网络[3],将文献看作网页,文献间的引用关系看作网页间的链接。基于网络挖掘的方法利用参考文献间的有向链接构造相关矩阵,然后模拟马尔科夫随机游走过程,迭代计算节点评分直至算法收敛。很多科学家受 PageRank 启发,发表了大量基于网络科学视角的学术影响力评价方法,如 CiteRank[6],FutureRank[7],PrestigeRank[8] 等。以上方法都有一定的创新性,但它们忽视了网络中存在的恶意活动(Malicious Activities)[9],如不合理的引用等[17]。为了增强计算的合理性,Yao 将 PageRank 算法中的线性公式改为非线性公式,从不同节点聚合资源来提升重要论文的作用[9]。Zhou 提出一种改进 PageRank 算法的优先机制来增强相似节点的有效性[15]。以上方法都将引文网络看作同质性网络,没有充分利用引文网络中的特征信息。网络结构中,节点之间的边通常具有正负特性[11],节点通过调用其他节点来完成一种有目的的指向和表达。传统方法只考虑节点入度的有效性,却忽视了节点出度的有效性。

本文提出一种基于信誉机制的科学文献影响力评价方法——ReputeRank 方法。该方法包括 3 个阶段:种子集选择阶段、信誉传播阶段和集成计算阶段。在种子集选取阶段,利用 PageRank 和 Inverse PageRank 算法分别初始化节点的入度评分和出度评分,然后结合文献期刊背景信息人工标记部分种子集为好种子集、一般种子集和坏种子集,然后利用 ASE(Auto Seed Expansion)策略[11]扩展引文网络中潜在的好种子集和坏种子集。在信誉传播阶段,基于好种子集指向的节点有着更高的可信度和坏种子集指向的节点有着较低的可信度的思想,使用 TrustRank 和 Anti-TrustRank 算法迭代计算信任值和不信任值。在集成计算阶段,利用网络拓扑结构扩散信任值和不信任值,综合计算论文的信誉评分并按照分数降序排列。

2　相关工作

2.1　引文网络

引文网络是一种特殊的社会网络,是典型的复杂网络。引文网络模型的构建,以论文为网络节点,文献间的引用关系为网络中的有向边。引文网络表示为 $G(V,E)$,有 N 个节点(论文),V 代表节点集$\{v_1,v_2,v_3,\cdots,v_n\}$,$E$ 代表边集(引用关系),$e_{i,j}=\{v_i,v_j\}$。如果顶点 v_i 指向 v_j 的有向边存在,则边元素 $e_{i,j}$ 为 1;若顶点 v_i 指向 v_j 的有向边不存在,则边元素 $e_{i,j}$ 为 0。提取出引证数据的局部特征,N_i^{out} 表示顶点 v_i 的出度数量,N_i^{in} 表示顶点 v_i 的入度数量;V_i^{out} 表示顶点 v_i 出度指向的顶点集合(即论文 v_i 引用过的文献集合),V_i^{in} 表示顶点 v_i 入度指向的顶点集合(即论文 v_i 被引用过的文献集合)。

2.2　PageRank

PageRank[12]是一种基于网页链接来计算网页重要程度的经典算法。PageRank 分数集合定义为 $P=\{P_1,P_2,P_3,\cdots,P_n\}$,公式如下:

$$P_j(t)=c\cdot\sum_{i=1}^{N}\left[\frac{e_{ij}}{N_i^{\text{out}}}\cdot(1-\delta_{N_i^{\text{out}},0})\right]P_i(t-1)+\frac{1-c}{N}$$
$$i,j\in\{1,2,\cdots,N\}\tag{1}$$

其中,t 表示迭代循环次数;$P_j(t)$ 表示节点 v_j 的第 t 次 PageRank 值,$P_j\in[0,1]$;c 为阻尼系数,代表指标中所占的比重,$c\in[0,1]$。

$$\delta_{N_i^{\text{out}},0}=\begin{cases}1, & N_i^{\text{out}}=0\\ 0, & N_i^{\text{out}}\neq 0\end{cases}$$

2.3　Inverse PageRank

Inverse PageRank[13]需要迭代计算,使网络中每个节点的 Inverse PageRank 值趋于稳定,即 $IP=\{IP_1,IP_2,IP_3,\cdots,IP_n\}$,公式如下:

$$IP_j(t)=c\cdot\sum_{i=1}^{N}\left[\frac{e_{ji}}{N_i^{\text{in}}}\cdot(1-\delta_{N_i^{\text{in}},0})\right]IP_j(t-1)+\frac{1-c}{N}\tag{2}$$

其中,i,j,t,c,δ 的含义同上。$IP_j(t)$ 表示顶点 v_j 的第 t 次 Inverse PageRank 值,$IP_j\in[0,1]$。

2.4　TrustRank

TrustRank[13]是雅虎公司为搜索引擎反作弊而提出的网页排序方法,用于计算每个站点的信任评分。首先,初始化每篇论文的信任评分,令各节点的初始信任值等于前面求出的 PageRank 值 $Tr=P$。然后,初始化信任好种子集向量 GS:

$$GS_i=\begin{cases}\dfrac{1}{N_{GS}}, & V_i\in\text{好种子集}\\ 0, & V_i\notin\text{好种子集}\end{cases}\tag{3}$$

其中,N_{GS} 表示好种子集数量。最后,将 GS 和常量参数代入 TrustRank 算法公式,迭代计算 TrustRank 值。

$$Tr_j(t)=\omega\cdot\sum_{i=1}^{N}\left[\frac{e_{ij}}{N_i^{\text{out}}}\cdot(1-\delta_{N_i^{\text{out}},0})\right]Tr_i(t-1)+$$
$$(1-\omega)\cdot GS_j\tag{4}$$

其中,N_i^{out},t 和 $\delta_{N_i^{\text{out}},0}$ 参数的含义同上。

2.5　Anti-TrustRank

Anti-TrustRank[13]根据 TrustRank 算法逆向思维,从坏种子集中逆向传递不信任值,计算网页的不信任值。坏种子集向量 BS 定义为:

$$BS_i=\begin{cases}\dfrac{1}{N_{BS}}, & V_i\in\text{坏种子集}\\ 0, & V_i\notin\text{坏种子集}\end{cases}\tag{5}$$

不信任指数 Dr 的计算公式为:

$$Dr_j(t)=u\cdot\sum_{i=1}^{N}\left[\frac{e_{ji}}{N_i^{\text{in}}}\cdot(1-\delta_{N_i^{\text{in}},0})\right]Dr_i(t-1)+(1-u)\cdot$$
$$BS_j\tag{6}$$

其中,N_i^{in},t 和 $\delta_{N_i^{\text{in}},0}$ 参数的含义同上。

3　基于信誉机制的文献影响力评价方法

引文网络是一种特殊的社会网络,包含丰富的特征信息,因此,我们认为结合论文的背景信息来衡量一篇论文引用与被引用的作用是有必要的。论文一旦被发表,历史特征信息就被确定,如作者、期刊、研究领域和参考文献等。唯一的变化是论文会随着时间的发展而被不断地引用。一般地,权威性的作者更倾向于向某一研究领域具有权威性的期刊投稿。因此,我们尝试利用利用文章的背景信息来推测其潜在的价值,并结合网络数据挖掘方法对论文进行评分排序。由斯坦福大学提出的 TrustRank 和 Anti-TrustRank 算法被应用于解决网页作弊问题,一些网页通过发出大量链接来帮助一个原本质量很差的目标网页提升搜索引擎排名[12-13]。这个问题与引文网络中存在大量自引和错误引用的情形类似。受以

上两种方法的启发，本文提出一种基于信誉机制的文献影响力评价方法，用于衡量各节点入链和出链的信誉度。

基于信誉机制的文献影响力评价方法的实现流程如图 1 所示。首先，输入引文网络结构和引文数据信息，对数据进行预处理和相关变量初始化操作；然后进行基于信誉机制的影响力计算阶段；最后根据评分降序输出结果。

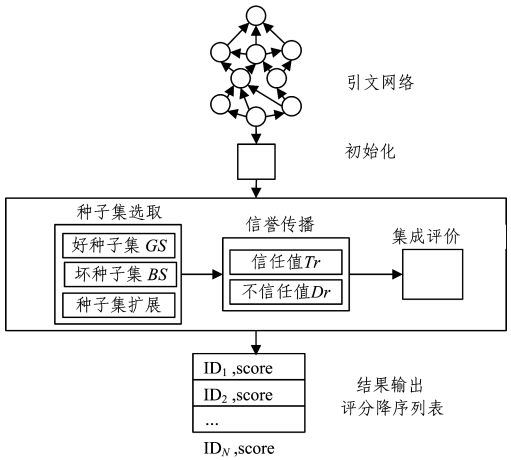

图 1 基于信誉机制的文献影响力评价方法的流程图

基于信誉机制的科学文献影响力评价方法包括 3 个阶段：种子集选择阶段、信誉传播阶段和集成计算阶段。

3.1 种子集选取阶段

网络中的论文包含很多属性，如作者、期刊、合作关系和参考文献等，它们在研究领域的发展和知识的传播中扮演着重要的角色，我们可以利用数据集中的特征属性来识别潜在的高质量的论文和低质量的论文。

种子集选取阶段包括两个过程：种子集预处理和种子集扩展。种子集分为好种子集和坏种子集。

1)种子集预处理

$G(V,E)$表示引文网络图，V 代表论文集合，E 代表论文间的引用关系。首先，论文的初始评分通过式(1)、式(2)计算，分别得到论文集的初始优质评分和劣质评分。然后，根据评分结果对论文集合降序排序。最后，结合论文期刊背景信息，人工标记选取 top-H 和 top-L 个优秀种子集。

2)种子集扩展

种子集扩展过程主要利用以下公式：

$$Reputation(v_i)=Reputation(v_i)+Neigh(v_i).scalar \quad (7)$$

其中，$Reputation$ 表示节点 v_i 的信誉度，$Neigh.scalar$ 表示领域节点的标量值，种子集扩展过程的思想是好种子集指向的节点通常具有较高的声誉，具有很高的可信度，而指向坏种子集的节点通常具有较高的不可信度。算法的实现过程如下。

算法 1 选取种子集的算法

输入：引文网络 G(V,E)，好种子阈值 g 以及坏种子阈值 b

输出：好种子集合 GS 和坏种子集合 BS

Begin

1. 根据式(1)、式(2)迭代计算节点 v_i 的 PageRank 评分和 Inverse PageRank 评分
2. 根据步骤 1 得到的评分 P_i 和 IP_i 分别降序排列顶点集 v_i，结合期刊背景信息选取 top-H 个好种子集和 top-L 个坏种子集
3. 结合 ASE 策略的种子集扩展过程
   3.1 好种子集扩展过程
      3.1.1 Let i←0；
      3.1.2 While i<H
         1)统计 GS_i 指向 BS 的链接数，标记为弱参考数目 weakoutlinkNum
         2)If weakoutlinkNum>坏种子阈值 b
            惩罚 GS_i 每个出度的标量，outlinkscalar 从 1 降为 1/2 ；
         3)Else
            Let j←0
            While j<GS_i 的出度数
               提取 GS_i 指向的第 j 个节点，标记为 V_q
               If the node of V_{ij} is not in GS and BS
               Reputation(V_q) ←Reputation(V_q)+outlinkscalar;
               If Reputation(V_q)>好种子阈值 g
                  Add V_q to GS，H←H+1；
            j←j+1；
            End while
         i←i+1；
         End while
   3.2 坏种子集扩展过程
      3.2.1 Let i←0；
      3.2.2 While i<L
         1)统计 BS_i 被 GS 指向的链接数，标记为强引用数目 stronginlinkNum
         2)If stronginlinkNum>好种子阈值 g
         3)惩罚 BS_i 的每个入度标量，inlinkscalar 从 1 降为 1/2 ；
         Else
            Let j←0
            While j<BS_i 的入度数
               提取 BS_i 被指向的第 j 个节点，标记为 V_p
               If the node of V_{ij} is not in GS and BS
               Reputation(V_p)←Reputation(V_p)+ inlinkscalar;
                  If Reputation(Vp)>坏种子阈值 b
                     Add Vp to BS，L←L+1；
            j ←j+1；
            End while
         i←i+1；
         End while
Return GS，BS；
End

3.2 信誉传播阶段

信誉传播阶段主要分为两个过程：好种子集传播信任过程和坏种子集传播不信任过程。假设好种子集发出的链接网页信任值更高，坏种子集发出的链接不信任值较高，利用上一步中的好种子集和坏种子集模拟引文网络随机游走过程，根据网络中传播的影响力计算信任评分和不信任评分。首先，初始化每篇论文的信任评分，令各节点的初始信任值等于前面求出的 GS_i，不信任值等于 BS_i。然后，从好种子集按照有向链接在网络中传播信任值，从坏种子集逆向传递不信任值。最后，循环迭代计算直到收敛，强引用的论文会得到更高的信任值，而弱引用的论文将会得到更高的不信任值。

算法 2 信誉传播算法

输入：引文网络 G(V,E)，好种子集 GS 和坏种子集 BS，节点数 N，最大迭代次数 iterMax

输出：信任值集 Tr 和不信任值集 Dr

Begin

1. 初始化参数信任值集 Tr 和不信任值集 Dr；所有节点置未收敛状态。
   If G ≠Ø
      Foreach i∈V
      根据式(3)、式(5)标准化节点 V_i 的静态种子集得分分布

```
    Foreach i∈V
      Let Tr_i←GS_i；Dr_i←BS_i
  End If
2. 根据网络结构传播信誉度
  Let i←0
    While i<iterMax
      Let j←0
      While j<N
      对 V 中的每个节点 V_j 作如下计算：
      {
      根据式(4)、式(6)迭代计算节点 V_j 的 TrustRank 评分和 Anti-TrustRank 评分；
      判断节点 V_j 是否收敛，若弱收敛则改变收敛状态；
      若未收敛，继续迭代；
    }
    j++
    End While
  i++
  End While
3. Return 信任值集 Tr 和不信任值集 Dr
End
```

3.3　集成计算阶段

集成 TrustRank 和 Anti-TrustRank 算法的评分结果，利用网络结构传播信誉的影响力，信誉聚集越高的点信誉评分越高，同时结合不信任值对于最终的影响力进行综合计算，公式如下：

$$ReputeRank_i=\alpha\cdot Tr_i+\beta\cdot Dr_i+\gamma\cdot\frac{1}{N} \tag{8}$$

其中，$ReputeRank_i$ 表示节点 v_i 的最终信誉评分，Tr_i 表示节点 v_i 的信任值，Dr_i 表示节点 v_i 的不信任值。实现过程为：首先，初始化相关参数；然后，提取节点 v_i 的信任值和不信任值；最后利用信誉评分公式综合计算论文的影响力。

算法 3　综合计算信誉评分算法

```
输入：节点集 V，信任评分 Tr，不信任评分 Dr，阻尼系数 α，β 和 γ
输出：信誉评分集 ReputeRank
Begin
1. 初始化最大迭代次数 iterMax、信誉评分集 ReputeRank
  iterMax ← V.length
2. If V≠Ø
    Let i←0
    While i<iterMax
    提取 V_i 的信任评分∈Tr
    提取 V_i 的不信任信任评分∈Dr
    根据集成公式计算 V_i 的综合评分：
        ReputeRank_i=α·Tr_i+β·Dr_i+γ·1/N
    i←i+1
    End while
  End If
3. 返回信誉评分集 ReputeRank
4. 根据 ReputeRank 评分结果，按照 QuickSort 算法对节点集 V 降序排列
End
```

3.4　时间复杂度分析

基于算法 3 个阶段的描述对算法进行时间复杂度分析。首先，在种子集选取阶段，迭代次数为 k，初始化信任评分的时间复杂度为 $O(k*n^2)$，种子集扩散过程取决于初始种子集规模和节点局部范围，平均计算时间复杂度为 $O(n*\log n)$。然后，在信誉传播阶段，初始化参数操作的时间复杂度为 $O(n)$，信誉扩散过程为 $O(n^2)$。最后，集成计算的时间复杂度为 $O(n)$。因此，总的时间复杂度为 $O(k*n^2+n*\log n+n^2+n)$。

4　实验结果分析

4.1　数据集

实验选取 KDD cup 2003 数据集[18]，该数据集收集了 1992－2003 年的论文，包括 27773 篇论文和 352807 条引用关系，其中包含了 *Physical Review*，*Physical Letters B*，*Physical Report*，*Physical Review D* 等 SCI 期刊以及非 SCI 期刊。

4.2　参数的设置和选取

本文方法涉及几个重要的参数，如好种子集数目 H 和坏种子集数目 L，以及弱参考数目 *weakoutlinkNum* 和强引用数目 *stronginlinkNum*。弱参考数目 *weakoutlinkNum* 被用于好种子集扩展过程中，作为判断一个节点是否成为好种子的阈值；强引用数目 *stronginlinkNum* 被用于坏种子集扩展过程中，作为判断一个节点是否成为坏种子的阈值。参数的选取要根据实验来确定。实验首先统计了数据集的度分布，如图 2、图 3 所示。

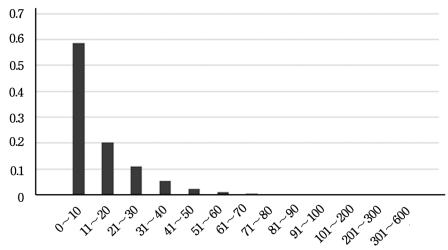

图 2　数据集出度区间分布

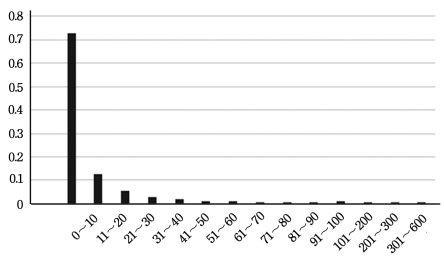

图 3　数据集入度区间分布

观察图 2、图 3 可知，58.7％的论文引用文献数目小于 10，72.6％的论文被引用小于 10 次，因此大多数论文的引用次数和被引用次数小于 10，大多数论文的影响力覆盖 10 篇左右。

4.2.1　种子集初始化参数

首先执行式(1)、式(2)中的 PageRank 算法和 Inverse PageRank 算法，限于篇幅，先列出排名前 10 的实验结果。

表 1　PageRank 算法排名前 10 的文章

No.	ID	出度	入度	*PR* 值	期刊
1	9407087	9	1299	0.003010	*Nucl. Phys. B*
2	9503124	10	1114	0.002210	*Nucl. Phys. B*
3	9510017	10	1155	0.002080	*Phys. Rev. Lett*
4	9402044	0	257	0.001890	*Phys. Rev. D*
5	9711200	54	2414	0.001665	*IInt. J. Theor. Phy*
6	9410167	25	748	0.001627	*Physics*
7	9408099	7	1006	0.001545	*Nucl. Phys. B*
8	9402002	14	282	0.001430	*Int. J. Mod. PhyA*
9	9610043	19	1199	0.001318	*Phys. Rev. D*
10	9205027	0	191	0.001242	*Phys. Rev. D*

表 2 Inverse PageRank 算法排名前 10 的文章

No.	ID	出度	入度	*IPR* 值	期刊
1	0303256	136	1	0.000770	*Physics*
2	0304263	51	0	0.000709	*J. Math. Phys*
3	0304271	58	0	0.000589	*Adv. Math. Phys*
4	0303207	42	5	0.000588	*JHEP*
5	0304187	38	0	0.000527	*Phys. Lett. B*
6	0302075	68	0	0.000498	*Phys. Rev. D*
7	0303144	67	1	0.000433	*Phys. Lett. B*
8	0304131	20	0	0.000423	*Phys. Lett. B*
9	0111258	29	1	0.000416	*Physics*
10	0303256	136	1	0.000770	*Physics*

实验中阻尼系数 c 选择通用常数 0.85，迭代执行 30 次。对比表 1 和表 2 可以发现，对于排名前 10 的文章，被引用次数越多且参考文献越少的文章 PageRank 评分越高，引用次数越少且参考文献越多的文章 Inverse PageRank 评分越低。我们假设评分阈值为 0.0001，大于阈值内的节点且结合期刊的影响因子和期刊分区信息，影响因子高的论文权威性相对较高，影响因子较低且是非核心刊物的一些孤立节点的相对影响力较低。实验在 KDD cup 2003 数据集的基础上，按照种子集预处理阶段，根据评分结合论文期刊背景信息，人工选取好种子数目 H 为 564，坏种子数目 L 为 296。

4.2.2 种子集扩展参数的选取

结合 ASE 策略的种子集扩展过程，在不同弱参考数目 *weakoutlinkNum* 和强引用数目 *stronginlinkNum* 参数变量下的扩展结果如表 3、表 4 所列。

表 3 不同参数下的好种子集扩展结果

weakoutlinkNum	*stronginlinkNum*	扩展后的数目	*Tr*>*Dr* 的数目	准确率/%
5	5	145	136	93.79
5	6	76	71	93.40
5	7	17	17	100
5	8	6	6	100
5	9	3	3	100

由于 50%以上的节点出入度分布在 10 以内，因此采取折初衷方法，扩展好种子集时弱参考数目 *weakoutlinkNum* 定量为 5，观察在不同强引用数目 *stronginlinkNum* 取值下，好种子集的扩展结果。当强引用数目取值为 5 时，扩展了 145 个种子，但是信任值评分 *Tr* 大于不信任值评分的数目却只有 136 个，准确率为 93.79%。当逐步提高强引用数目的取值时，种子集的扩展数目相对减少，但是准确率却相对较高。

表 4 不同参数下的坏种子集扩展结果

weakoutlinkNum	*stronginlinkNum*	扩展后的数目	*Tr*>*Dr* 的数目	准确率/%
0	5	704	624	88.63
1	5	703	624	88.76
2	5	237	209	88.19
3	5	1	1	100
4	5	0	0	—

坏种子集的扩展相对要求严格，强引用数目 *stronginlinkNum* 的取值少于 5，且弱参考数目 *weakoutlinkNum* 也在相对较小的范围内才有可能被添加入坏种子集合中，此策略是按照影响力一般的论文相对孤立且传播范围有限的思想设置的。扩展坏种子集时，强引用数目 *stronginlinkNum* 定量为 5，观察在不同弱参考数目 *weakoutlinkNum* 取值下，坏种子集的扩展结果。当弱参考数目 *weakoutlinkNum* 为 1 时，扩展了 704 个种子，但是信任值评分 *Tr* 小于不信任值评分的数目有 624 个，准确率为 88.63%。当逐步提高弱参考数目 *weakoutlinkNum* 的取值时，种子集的扩展数目相对减少，但是准确率却相对较高，但当弱参考数目 *weakoutlinkNum* 提高到 5 时，扩展后的数目为 0，这是因为数据集中大部分还是较为优秀的期刊，指向影响力一般的文献相对较少。

4.2.3 不同阻尼系数 α，β 和 γ 下的评分结果

针对 3 个阻尼系数进行调参，3 个系数的绝对值之和为 1；α 是信任值 *Tr* 的系数，从 0.1 个区间开始变动；β 是不信任值 *Dr* 的系数；γ 是公式平滑系数，所以按照 PageRank 公式平滑系数的赋值，一般赋值为 0.05。根据不同参数下 ReputeRank 方法的运行结果，选取排名前 100 的论文进行评价，结合论文期刊背景，根据 top-100 的 SCI 论文覆盖率进行观察，发现当 α 为 0.5、β 为 −0.45、γ 为 0.05 时实验效果最好。当 α 为 0、β 为 −1、γ 为 0 时，方法退化为 Anti-TrustRank 方法，而当 α 为 1、β 为 0、γ 为 0 时，方法变为 TrustRank 方法。

表 5 ReputeRank 方法在不同参数下的实验结果

α	β	γ	SCI 覆盖率
0	−1	0	0.485
0.1	−0.85	0.05	0.864
0.2	−0.75	0.05	0.835
0.3	−0.65	0.05	0.930
0.4	−0.55	0.05	0.773
0.5	−0.45	0.05	0.930
0.6	−0.35	0.05	0.765
0.7	−0.25	0.05	0.812
0.8	−0.15	0.05	0.883
0.9	−0.05	0.05	0.870
1	0	0	0.890

4.3 与其他相关方法的对比分析

在迭代运行 TrustRank 和 Anti-TrustRank 算法 30 次后，取 α 为 0.5、β 为 −0.45、γ 为 0.05，运行集成评价公式 ReputeRank，并将该算法得到的 Top-n 论文的 SCI 覆盖率与 PageRank，CountDegree 和 SPRank 算法中评价方法得到的 Top-n 论文的 SCI 覆盖率进行对比，结果如图 4 所示。

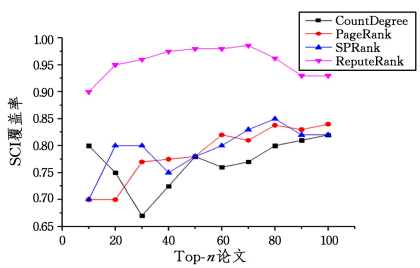

图 4 4 种评价方法的 Top-n 的 SCI 论文覆盖率

观察图 4 发现，采用 ReputeRank 算法得到的 Top-n 论文的 SCI 覆盖率明显高于其他算法，说明 ReputeRank 算法能够反映权威性论文的传播能力，信誉值越高的论文互相引用从而使排名更高。SPRank 算法有着相对较好的覆盖率，CountDegree 得到的 Top-n 论文的 SCI 覆盖率最低，说明基于统计频次的方法不能完全反映真实论文的影响力。综上所述，ReputeRank 算法能够有效排除孤立的论文节点对评分结果的干扰，排序结果更加合理。

根据文献[15]的阐述，引文网络中存在大量自引。因此本文随机选择了两个入度为 0 的孤立节点，把它们看作试图提升评分排名的目标论文。人工添加 n 个虚拟节点到引文网络，每个节点带有 m 个出链，$m\in[10,20]$。让新添加的节点引用以上两个目标节点 0205176 和 9912286，新添加的 m 条链接随机指向网络中的其他节点。由于 CountDegree 是基于

被引用频次的排序方法，不具有鲁棒性，对比意义不大，因此本文比较 3 种评价方法的排序变化，实验结果如图 5、图 6 所示。

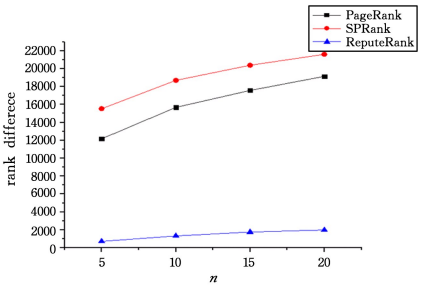

图 5　3 种评价方法在论文 0205176 排序变化上的对比

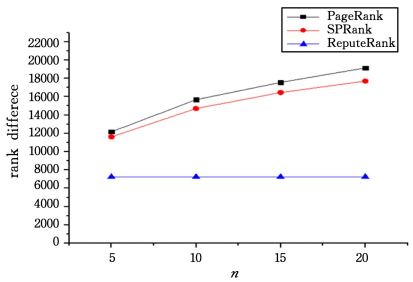

图 6　3 种评价方法在论文 9912286 排序变化上的对比

如图 5、图 6 所示，为两篇未被引用过的孤立论文节点分别添加 5，10，15 和 20 个虚拟节点链接，使用 PageRank 和 SPRank 评价方法得到的论文排名与添加虚拟节点前的排名变化逐步变大，而使用 ReputeRank 方法得到的排名更趋于平稳。

本文同时对以上 4 种方法的运行效率进行了实验，结果如表 6 所列。相对于其他算法，ReputeRank 算法的执行时间相对较长，因为步骤相对较多；CountDegree 算法的执行时间最短，因为时间复杂度为 $O(n\log n)$；PageRank 和 SPRank 算法的执行时间居中。

表 6　4 种方法的运行时间

（单位：s）

算法	时间
ReputeRank	102.4
PageRank	92.9
SPRank	98.5
CountDegree	60.1

结束语　引文网络是由文献间引用关系构成的一种大规模的复杂网络，对于知识的传播和信息的扩散具有重要的研究价值。本文提出了一种基于信誉机制的科学文献影响力排序算法 ReputeRank，该方法包括 3 个阶段：种子集选择阶段、信誉传播阶段和集成计算阶段。该方法提出了一个信誉传播思想，利用引文网络拓扑结构扩散信任评分和不信任评分。在 KDD cup 2003 数据集上的实验结果证明，ReputeRank 算法相比 SPRank，PageRanK 和 CountDegree 算法能够取得更优的结果，ReputeRank 算法对个性化推荐、搜索引擎网页排序和社交网络舆论扩散与检测都有一定借鉴意义。

参 考 文 献

[1] ZHI L，QIN K P，CHE L. Two citation-based indicators to measure latent referential value of papers [J]. Scientometrics，2016，108(3)：1299-1313.

[2] BOYACK K W，BORNER K. Indicator-assisted evaluation and funding of re-search：Visualizing the influence of grants on the number and citation counts of research papers [J]. Journal of the Association for Information Science and Technology，2003，54(5)：447-461.

[3] BOLLEN J，RODRIQUEZ M A，VAN D S. Journal status [J]. Scientometrics，2016，69(3)：669-687.

[4] MAZLOUMIIAN A，EOM Y H，HELBING D，et al. How citation boosts promote scientific paradigm shifts and Nobel prizes [J]. PLoS ONE，2011，6(5)：e18975.

[5] HIRSCH J E. An index to quantify an individual's scientific research output[J]. Proceedings of the National Academy of Sciences of the United States of America，2005，102(46)：16569-16572.

[6] WALKER D，XIE H，YAN K，et al. Ranking scientific publications using a model of network traffic [J]. Journal of Statistical Mechanics：Theory and Experiment，2006，6(6)：P06010.

[7] SAYYADI H，GETOOR L. FutureRank：Ranking Scientific Articles by Predicting their Future PageRank[C]// Siam International Conference on Data Mining. USA，2009：533-544.

[8] SU C，PAN Y，ZHEN Y，et al. PrestigeRank：A new evaluation method for papers and journals [J]. Journal of Informetrics，2011，5(1)：1-13.

[9] YAO L，WEI T，ZENG A，et al. Ranking scientific publications：the effect of nonlinearity [J]. Scientific Reports，2014，4：6663.

[10] JOHN P I. A generalized view of self-citation：Direct，co-author，collaborative，and Coercive induced self-citation [J]. Journal of Psychosomatic Research，2015，78(1)：7-11.

[11] 蓝梦微，李翠平，王绍卿，等. 符号社会网络中正负关系预测算法研究综述[J]. 计算机研究与发展，2015，52(2)：410-422.

[12] KRISHNAN V，RAJ R. Web Spam Detection with Anti-TrustRank [C]// Proceedings of the 2nd International Workshop on Adversarial Information Retrieval on the Web. New York：ACM Press，2006：37-40.

[13] ZOLTAN G，HECTOR G M，JAN P. Combating Web Spam with TrustRank[C]// Proceeding of the 30th VLDB conference. Toronto，Canada，2004：576-587.

[14] ZHANG X C，LIANG W X，ZHU S P，et al. Automatic seed set expansion for trust propagation based anti-spam algorithms [J]. Information Sciences，2013，232(5)：167-187.

[15] ZHOU J L，ZENG A，FAN Y，et al. Ranking scientific publications with similarity-preferential mechanism[J]. Scientometrics，2016，106(2)：805-816.

[16] ELENI F，GEORGIOS E. Review of the indirect citations paradigm：theory and practice of the assessment of papers，authors and journals [J]. Scientometrics，2014，99(2)：261-288.

[17] JOHAN B，HERBERT V D S，ARIC H，et al. A Principal Component Analysis of 39 Scientific Impact Measures [J]. PLoS ONE，2009，4(6)：e6022.

[18] KDD Cup2003 datasets (Version2003) [DB/OL]. http：//www.cs. cornell. edu/projects/kddcup/datasets. html.

[19] 王振飞，朱静阳，郑志蕴，等. 基于 R-C 模型的微博社区用户影响力分析[J]. 计算机科学，2017，44(3)：254-258.