基于作者权威值的论文价值预测算法

刘大有 [1] 薛锐青 [1] 齐 红 [1]

摘 要 论文引用网络是一个动态变化的网络, 不断有新的论文加入引用网络中. 传统的论文评价标准如引用次数、PageRank 值等 "终身评价标准" 存在排挤新结点的问题, 如何在海量论文中寻找有价值、被持续关注的论文, 成为人们感兴趣的问题. Sayyadi 提出了 FutureRank 算法, 该算法通过预测论文未来 "一段时间" 的被引次数排名和 PageRank 值排名来达到这一目的. 但 FutureRank 算法需提前计算 PageRank 值, 要耗费大量运算时间. 据此, 我们尝试在不计算论文现有 PageRank 值的条件下, 从论文的撰写者以及引用者的权威值的角度来预测论文未来的被引次数排名和 PageRank 值排名. 实验结果表明, 我们的算法与 FutureRank 相比, 不但缩短了运算时间, 而且提高了预测准确率.

关键词 引用网络, 排名预测, FutureRank, PageRank

引用格式 刘大有, 薛锐青, 齐红. 基于作者权威值的论文价值预测算法. 自动化学报, 2012, **38**(10): 1654−1662

DOI 10.3724/SP.J.1004.2012.01654

The Paper Value Prediction Algorithm Based on the Author′s Authority Value

LIU Da-You[1] XUE Rui-Qing[1] QI Hong[1]

Abstract Citation network is a dynamic network and new papers are added to it every day. The traditional literature evaluation criteria like citations number and PageRank are unfair to the new node. How to retrieve the valuable papers of continuous concerns has become an interesting focus. To solve this problem, Hassan Sayyadi proposed the FutureRank algorithm, but it needs to calculate the PageRank value, which takes a lot of time. Accordingly, we proposed a paper value prediction algorithm without computing the PageRank value. We predict paper′s rank of citations number and PageRank value in the future by writers′ authority value and citer′s authority value. Experimental results show that as compared with FutureRank, our algorithm not only shortens the computing time but also improves the forecast accuracy.

Key words Citation network, ranking prediction, FutureRank, PageRank

Citation Liu Da-You, Xue Rui-Qing, Qi Hong. The paper value prediction algorithm based on the author′s authority value. *Acta Automatica Sinica*, 2012, **38**(10): 1654−1662

论文引用和互联网络一样, 都可以表示为结点间的链接网络. 如何在海量的论文中寻找有价值、被持续关注的论文, 成为人们感兴趣的问题. 论文的价值可以通过 "流行度" 和 "权威度" 来评价.

流行度评价标准: 1970 年, Garfield[1] 提出了这种评价方法, 用某杂志前两年发表的论文在今年的被引用次数, 除以该杂志前两年发表论文的数量, 得到该杂志在今年的影响因子. Sayyadi 等[2] 搜集了一系列影响因子的改进及变种方法.

权威度评价标准: 1998 年, Page 等[3] 在互联网领域革命性地提出了 PageRank 算法, 在算法中引入了随机浏览的概念. 一个网页的得分取决于其被多少网页链入, 以及这些被链入的网页的质量.

PageRank 算法自推出以来, 吸引了大批学者的关注, 一些学者将 PageRank 算法引入论文引用网络中, 对论文及期刊进行评分. 文献 [4] 比较了科学期刊的 ISI 影响因子 (Impact factor, IF) 和加权 PageRank 值, 并综合这两种评价标准提出了 Y-factor. 文献 [5] 用 PageRank 算法和引用次数分析了 *Physical Review* 期刊族在 1893 年 ∼ 2003 年间发表的所有论文. 类似地, 文献 [6] 提出了一种基于 PageRank 的评价论文重要程度的算法.

Eigenfactor[7] 是一个类似 PageRank 的评分算法. 与 PageRank 不同之处在于 Eigenfactor 处理的对象不是网页超链接而是期刊间的引用链接, 评分对象不是网站而是科学期刊. 汤姆森路透科技集团在 2009 发布的 Journal Citation Reports (JCR) (R) 中增加了该计量指标.

1999 年, Kleinberg[8] 提出了 HITS 算法, 与 PageRank 算法不同, HITS 算法将网页分为两类: 中心网页和权威网页, 同时每个网页拥有两种评分 — 中心值和权威值. 算法通过迭代获得网页的中心值和权威值.

收稿日期 2011-09-15 录用日期 2012-01-10
Manuscript received September 15, 2011; accepted January 10, 2012
国家自然科学基金 (61133011, 61170092, 60973088, 60873149), 中央高校基本科研业务费专项资金 (200903181, 200903192) 资助
Supported by National Natural Science Foundation of China (61133011, 61170092, 60973088, 60873149) and the Fundamental Research Funds for the Central Universities (200903181, 200903192)
本文责任编委 刘成林
Recommended by Associate Editor LIU Cheng-Lin
1. 吉林大学计算机科学与技术学院 长春 130012
1. College of Mathematics, Jilin University, Changchun 130012

1) 依赖时间的改进算法. 无论是 PageRank、HITS 还是引用计数都存在排挤新结点的问题, 这是因为新节点由于存在时间短, 没有足够的时间获得足够的链接以提升名次. 对于这个问题, 人们提出了大量改进方法.

文献 [9−11] 中, 基于链接存在的时间对链接加权, 较新的链接拥有较高权重. 与此类似, 本文算法也对论文的引用链接进行加权处理, 但不同的是, 本文算法对作者和论文间的撰写和引用链接进行加权, "撰写" 和 "引用" 事件发生时间越近, 权值越高.

针对一些论文排序算法对新论文评价不公平的问题, Walker 等[11] 引入了 CiteRank 算法, CiteRank 实际上是一种随机游走模型的变形, 在基本的随机游走模型中随机冲浪者以等概率在结点间跳转, 而在 CiteRank 中, 随机冲浪者优先访问存在时间短的结点.

但是, 这种理想化的随机游走模型只考虑了论文的发表时间. 实际上, 科技工作者在检索论文的时候, 会考虑多种因素, 如作者权威度、所发表杂志影响因子、论文被引次数, 论文发表时间只是其中之一. 所以 CiteRank 算法的预测值与未来真实被引次数的相关度相对较低, 只有 0.57[2], 并且 CiteRank 算法缺乏对论文权威度的预测能力.

2) 融合多种网络的改进算法. 这类方法综合考虑了多种结构的网络, 从多个角度出发对网络中的结点进行排名.

文献 [12] 依据施引论文所在杂志的影响因子以及引用时间, 为引用链接赋予不同的权值, 进而评价论文的权威值, 文献 [13−14] 在文献 [12] 的基础上进一步综合考虑了施引杂志和施引作者. 利用三种不同的网络结构: 作者 – 论文网络、论文 – 论文网络、论文 – 杂志网络对论文的权威值进行评分[15]. 利用 HITS 算法的思想, 通过三种网络: 作者 – 论文引用网络、论文 – 论文引用网络以及作者共著网络, 来评价论文的价值以及作者的价值.

2009 年, Sayyadi 等[2] 提出了 FutureRank 算法, FutureRank 算法考虑了论文的发表时间、作者权威度和论文已有的 PageRank 值, 来预测论文未来的被引数量排名和 PageRank 值排名, 从实验结果看, 相对 CiteRank 算法, FutureRank 算法大幅提高了引用数量的预测准确率 (与未来真实数据的相关度为 0.75) 并且增加了对论文权威值的预测. 但是 FutureRank 算法需要计算当前论文的 PageRank 值, 如果不考虑优化算法, 当结点数量很大时, 计算 PageRank 值是非常耗费时间的, 并且 FutureRank 算法预测值与未来 PageRank 值排名的相关度还比较低, 仅为 0.59.

针对以上研究的不足, 本文提出一种新的论文权威度和流行度预测方法. 在不计算当前论文 Pagerank 值的前提下, 预测论文的价值. 本算法主要考虑了以下几个因素: 1) 论文被引时间; 2) 撰写者权威度; 3) 引用者权威度; 4) 论文发表时间.

1 论文被关注度

PageRank 值和 Cited counts 可以看作是对论文的终身评价, 体现了科研工作者群体对一篇论文的认可程度, 一篇论文发表时间越长, 对其评价越准确. 如果将范围缩小, 考察 "一个时间段内" 论文获得的引用链接 (例如论文在 2001 年 ∼2003 年间获得的引用链接), 就能计算出论文在 "该段时间内"(由这些新形成的引用链接) 所形成的 PageRank 值和 Cited counts 值. 它们体现了科研工作者群体在 "这一时间段内" 对这篇论文的认可程度.

反过来, 如果要预测论文在未来 "一段时间内" 的 PageRank 值排名和 Cited counts 值排名, 可以将问题转化为预测论文在 "这段时间内" 被科研工作者群体关注的程度.

我们给出 4 点假设, 用来判断什么样的论文可以在未来获得更多的关注:

1) 与老论文相比, 新发表的论文更可能吸引科研工作者的关注;

2) 与一般作者相比, 领域权威作者所撰写的论文更可能吸引同行的关注;

3) 无论是新论文还是老论文, 如果它在近段时间内被大量作者引用, 那么该论文在未来一段时间内将会被持续引用;

4) 有一些作者对其所在的研究领域非常熟悉, 他们能够鉴别出该领域的优秀作者和优秀论文, 被这些作者引用的论文, 优秀的可能性很高, 这些论文也更可能吸引同行的关注.

需要说明的是, 假设 4) 中的作者, 是指那些引用过大量优秀论文的作者, 即如果一个作者历史上引用过大量优秀论文, 那么他再次引用优秀论文的可能性就很大.

2 论文被关注度预测算法

2.1 引用网络

本文中的引用链接网络中含有两类结点 (论文和作者) 以及两类边: 作者 – 论文边用来链接论文和论文的作者, 代表论文的撰写关系; 论文 – 论文边用来链接论文和论文的引文, 代表论文间的引用关系.

图 1 是一个作者、作者论文间的撰写和引用网络. 网络中有 5 位作者、8 篇论文. 图 1 包含两种边 — 无向边以及有向边, 其中从作者 A_i 到论文 P_j 间的无向边表示作者 A_i 撰写了论文 P_j; 从论文 P_i 到论文 P_j 的有向边表示论文 P_i 引用了论文 P_j. 在

图 1 的基础上, 我们定义作者 – 论文引用链接、作者 – 作者引用链接.

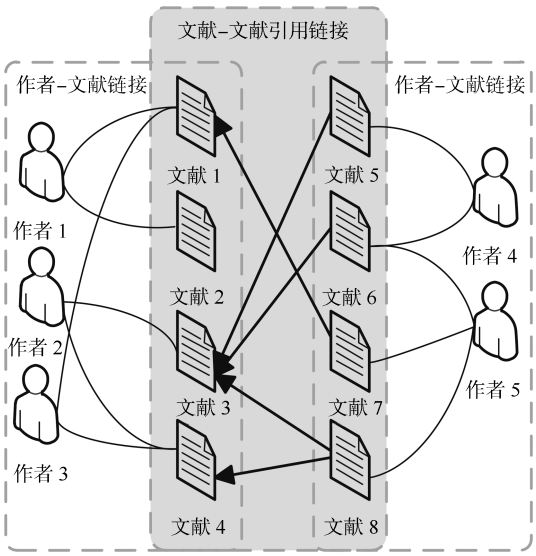


图 1 作者 – 论文的撰写关系

Fig. 1 Writing relationships between authors and papers

图 2 是一个作者论文间的引用网络. 虚箭头从作者 A_i 指向论文 P_j 表示该作者引用过该篇论文.

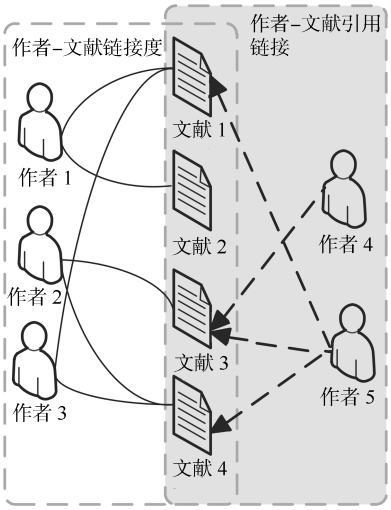


图 2 作者 – 论文间引用链接

Fig. 2 Citing links between authors and papers

图 3 是作者间的一个引用网络. 从作者 A_i 到作者 A_i 的虚箭头表示该作者 A_i 引用过作者 A_j 撰写的论文.

网络通常是由邻接矩阵表示的, 用以下四个矩阵表示上述网络结构 (其中 A 表示作者, P 表示论文).

首先定义论文 – 论文间的引用矩阵 M^{pp}:

$$M^{pp}_{i,j}=\begin{cases}1, & \text{如果论文 } P_i \text{ 引用了论文 } P_j\\ 0, & \text{否则}\end{cases}$$

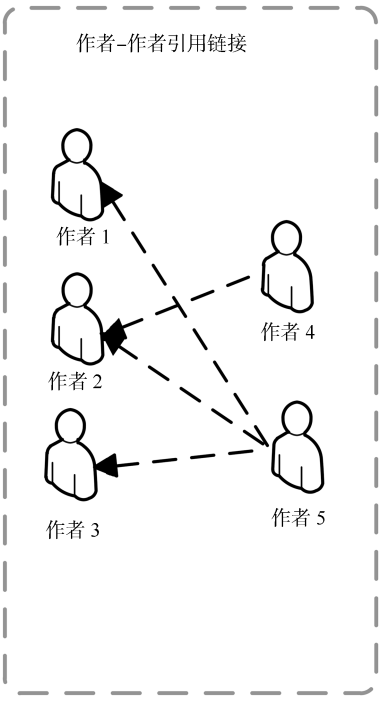


图 3 作者 – 作者间引用链接

Fig. 3 Citing links between authors and authors

FutureRank 算法为所有初度为 0 的结点添加指向所有其他结点的链接. 与 FutureRank 算法不同, 由于不需要计算论文现有的 PageRank 值, 所以对于悬挂结点, 也就是没有引文的论文, 我们不做任何特殊处理. 然后, 定义另外 3 个矩阵: M^{ac}, M^{aa} 和 M^{ap}:

$$M^{ac}_{i,j}=\begin{cases}1, & \text{如果作者 } A_i \text{ 引用了论文 } P_j\\ 0, & \text{否则}\end{cases}$$

$$M^{aa}_{i,j}=\begin{cases}1, & \text{如果作者 } A_i \text{ 引用了作者 } A_j \text{ 的论文}\\ 0, & \text{否则}\end{cases}$$

$$M^{ap}_{i,j}=\begin{cases}1, & \text{如果作者 } A_i \text{ 是论文 } P_j \text{ 的作者}\\ 0, & \text{否则}\end{cases}$$

2.2 预测算法

定义一篇论文的两类相关者: 撰写者 (撰写了该论文) 和引用者 (引用过该论文), 可通过这两类相关者来衡量一篇论文的质量.

优秀的撰写者指其撰写过大量、高质量论文, 体现了该作者在某领域的科研水平; 优秀的引用者指其引用过大量、高质量论文, 体现了发现、鉴别高质量论文的能力. 优秀的撰写者和优秀的引用者并不

互斥, 一个人既可以是优秀的撰写者, 又可以是优秀的引用者.

高质量的论文通常指那些同时拥有优秀的撰写者和大量优秀的引用者的论文. 一篇高质量的论文往往具有高被引次数、高 PageRank 值, 被某领域的科研工作者高度评价, 体现了该论文对相关领域发展做出的贡献.

但一篇高质量的论文, 不能保证其持续获得高被引频率. 原因有很多, 例如某些曾经流行的理论方法, 可能被在它基础上发展起来的新方法所代替, 或在科研工作者探索学科前沿的动机下, 减少了对发表时间久远的论文的关注程度等. 所以, 还需要设计一个评价论文被关注潜力的方法. 为此, 我们通过论文的发表时间和被引用时间来定义一篇具有被关注潜力的论文. 此类论文系指其撰写者优秀且在 "近期内" 被大量优秀引用者引用的论文. 可以认为这种具有被关注潜力的论文会引来大量科研工作者的关注, 在未来一段时间会被大量引用.

2.2.1　作者的价值

每位作者拥有两类权威值: 撰写权威值以及引用权威值. 可以通过如下迭代过程计算每位作者的撰写权威值以及引用权威值.

$$\boldsymbol{H} = M^{aa} \times \boldsymbol{A} \tag{1}$$

$$\boldsymbol{A} = (M^{aa})^{\mathrm{T}} \times \boldsymbol{H} \tag{2}$$

其中, $\boldsymbol{H}$ 是作者的引用权威值向量, 记录了作者的引用权威值, $\boldsymbol{A}$ 是作者的撰写权威值向量, 记录了作者的撰写权威值.

向量 $\boldsymbol{H}$ 和向量 $\boldsymbol{A}$ 初始时都是长度为 n 的全 1 向量, 其中 n 是网络中的作者结点的数量. 即初始时, 所有作者的撰写权威值和引用权威值是相等的.

类似 HITS 算法, 作者的撰写权威值和引用权威值是一对彼此相互定义的评分标准, 即一位撰写者的撰写权威值高低取决于其撰写的文章的引用者数量, 以及这些引用者的引用权威值高低情况. 引用者越多、引用者的引用权威值越高, 撰写者的撰写权威值就越高. 反之, 一位引用者的引用权威值高低取决于其引用的文章撰写者的撰写权威值高低情况. 撰写者的撰写权威值越高, 引用者的引用权威值就越高.

式 (1) 和式 (2) 说明, 一位作者的引用权威值是其引用过的文章的作者的撰写权威值之和; 反之, 一位作者的撰写权威值是引用过其撰写的文章的作者的引用权威值之和.

2.2.2　论文的价值

我们从作者的撰写权威值和引用权威值的角度评价一篇论文的好坏.

1) 作者的撰写权威值高, 论文优秀的可能性越大.

2) 引用该论文的作者的引用权威值越高, 该论文优秀的可能性越大.

$$\boldsymbol{R} = C \times (M^{ap})^{\mathrm{T}} \times \boldsymbol{A} + (M^{ac})^{\mathrm{T}} \times \boldsymbol{H} \tag{3}$$

其中, C 是常数, $\boldsymbol{H}$ 是作者的引用权威值向量, $\boldsymbol{A}$ 是作者的撰写权威值向量.

2.2.3　论文未来的价值

一般来说, 论文的质量越高, 其被引用的可能性越大, 但一篇高质量的论文的被引频率并不是一成不变的. 论文的理论方法会不断地被后继论文修改、提高, 曾经经典的论文现在可能并不被广泛引用, 同时, 科研工作者需要跟踪其所在领域的前沿论文, 这就导致近两年的论文被大量引用.

我们统计了在数据集 arXiv (hep-th) 中论文平均的施引频率. 结果表明, 论文趋向于引用新发表的论文, 1 年内发表的论文最多, 1 年～2 年内发表的论文次之; 总体呈现递减趋势. 统计结果如图 4 所示.

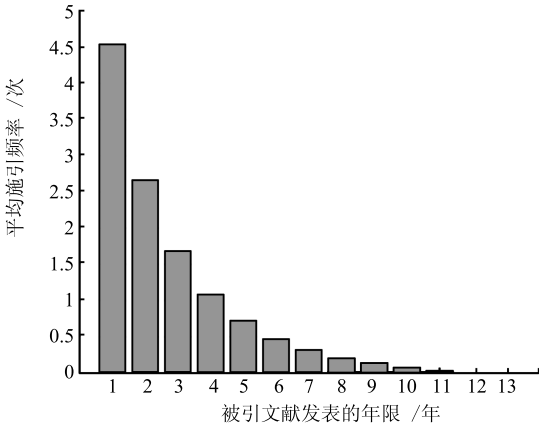


图 4　论文平均施引频率

Fig. 4　The average frequency of papers′ citing

基于此, FutureRank 在排名中引入论文的发表时间, 发表时间短的论文具有较高权重. 但实际上有一些经典论文会长时间保持高被引频率 (尽管论文的发表时间已相当久), FutureRank 未考虑这样一些论文的特殊性. 与 FutureRank 算法不同, 我们认为一篇论文的发表时间很重要, 但这篇论文在近期内是否拥有高的被引频率也很重要, 若有则其在未来的一段时间内还会被频繁引用, 而不管其发表的时间长短. 这一假定既符合科研工作者跟踪领域研究热点的实际情况, 又解决了 FutureRank 对经典论文评分不够恰当的问题.

首先, 类似 FutureRank 处理论文发表时间的方

法, 定义代表论文的发表时间的时间衰减量 R_i^{Time}:

$$R_i^{\mathrm{Time}} = \mathrm{e}^{-\rho\times(T_c-T_i)^2} \tag{4}$$

其中, T_c 是当前时间, T_i 是第 i 篇论文的发表时间, $R_i^{\mathrm{Time}}\in(0,1]$, ρ 为常数. 发表时间越短的论文 R_i^{Time} 值越高, 它的作用是给网络中的边加权, 存在时间较短的边有较高的权值.

计算论文 P 价值的预测分数 R:

$$\begin{aligned} R_i =& C\times\sum(A_m\times R_i^{\mathrm{Time}})+\\ &\sum(H_n\times\sum R_j^{\mathrm{Time}}) \end{aligned} \tag{5}$$

其中, R_i 是对第 i 篇论文 P_i 的价值预测分数. A_m 是第 i 篇论文 P_i 的作者, 一篇论文可能有多位撰写者; H_n 是引用过第 i 篇论文 P_i 的作者, 一篇论文可能会有多位引用者; 假设作者 H_n 在其撰写的论文 P_j 中引用了论文 P_i, 则 R_j^{Time} 就是论文 P_j 的 R_i^{Time} 值.

论文评分 R_i 综合考虑了作者的撰写权威值、引用权威值以及时间因素的影响 (与第 1 节的假设相对应):

1) 作者的撰写权威值越高论文评分越高;

2) 作者的引用权威值越高论文的评分越高;

3) 相对于老论文, 新发表的论文评分较高 (因为新论文具有较高的 R^{Time} 值);

4) 论文的被引数量越多, 评分越高. 但是引用链接的权值是不同的, 施引论文发表时间越近, 权值越高, 对被引论文评分的贡献也越大. 即近期被大量引用的论文, 会得到较高评分.

2.3 预测值计算

我们需要上文定义的两个矩阵 M^{pp} 和 M^{ap}:

$$M_{i,j}^{pp} = \begin{cases} 1, & \text{如果论文 } P_i \text{ 引用了论文 } P_j \\ 0, & \text{否则} \end{cases}$$

$$M_{i,j}^{ap} = \begin{cases} 1, & \text{如果作者 } A_i \text{ 是论文 } P_j \text{ 的作者} \\ 0, & \text{否则} \end{cases}$$

在它们的基础上定义矩阵 MT^{pp}, MT^{ap} 以及 MT^{ac}.

$$MT_{i,j}^{pp} = \begin{cases} 1\times R_i^{\mathrm{Time}}, & \text{如果论文 } P_i \text{ 引用了 } P_j \\ 0, & \text{否则} \end{cases}$$

$MT_{i,j}^{pp}$ 是加权后的矩阵 M^{pp}, 其中 $R_i^{\mathrm{Time}}\in(0,1]$, 论文$p_i$发表时间越近, R_i^{Time}越大, $MT_{i,j}^{pp}$越大.

$$MT_{i,j}^{ap} = \begin{cases} 1\times R_i^{\mathrm{Time}}, & \text{如果作者 } A_i \text{ 是论文 } P_j \text{ 的作者} \\ 0, & \text{否则} \end{cases}$$

类似地, MT^{ap} 是加权后的矩阵 M^{ap}, 其中 $R_i^{\mathrm{Time}}\in(0,1]$, 论文 P_j 发表时间越近, R_j^{Time} 越大, $MT_{i,j}^{ap}$ 越大.

$$MT_{i,j}^{ac} = \begin{cases} 1\times\sum R_m^{\mathrm{Time}}, & \text{如果作者 } A_i \text{ 在论文 } P_m \text{ 中引用了 } P_j \\ 0, & \text{否则} \end{cases}$$

MT^{ac} 是加权后的矩阵 M^{ac}, 作者 A_i 可能在多篇论文中引用了论文 P_j, 将这些论文的 R^{Time} 相加, 最后得到矩阵 MT^{ac}.

下面计算这三个矩阵:

首先定义过渡矩阵 MT^{Time}:

$$MT_{\mathrm{Time}} = I\times R^{\mathrm{Time}} \tag{6}$$

其中, I 是单位矩阵, R^{Time} 是论文发表的时间量. MT^{Time} 矩阵实际上是将 R^{Time} 上的值拷贝到单位矩阵的主对角线上.

1) 生成矩阵 MT^{pp}, 根据定义可知:

$$MT^{pp} = MT^{\mathrm{Time}}\times M^{pp} \tag{7}$$

2) 生成矩阵 MT^{ap}:

$$MT^{ap} = MT^{\mathrm{Time}}\times((M^{ap})^{\mathrm{T}})^{\mathrm{T}} \tag{8}$$

3) 生成矩阵 MT^{ac}:

$$MT^{ac} = M^{ap}\times MT^{pp} \tag{9}$$

通过矩阵 MT^{ap}、MT^{ac} 以及向量 $\boldsymbol{A}$, $\boldsymbol{H}$, 得到式 (5) 的矩阵计算公式:

$$\boldsymbol{R} = C\times(MT^{ap})^{\mathrm{T}}\times\boldsymbol{A}+(MT^{ac})^{\mathrm{T}}\times\boldsymbol{H} \tag{10}$$

3 实验结果

首先比较一系列算法的执行效率, 包括 PageRank 算法、需要预先计算 PageRank 值的 FutureRank 算法, 以及同样采用 HITS 思想进行论文和作者排序的 Co-Ranking 算法 [15]. 之后对比本文算法和 FutureRank 算法在排名预测上的准确率.

3.1 数据集

实验数据集取自 arXiv 中的 hep-th 部分 — 保存高能物理学论文的引用数据以及作者信息上的真实数据. 这个数据集包含了 1992 年 ~ 2003 年发表的所有高能物理论文, 共有 29 555 篇论文、352 807

个引用链接、超过 15 000 位作者. 如果两位作者姓名相同, 则认为他们是同一位作者.

将数据集分为两部分: 1999 年～2000 年和 2001 年～2003 年, 第一部分为查询数据, 第二部分为验证数据. 查询数据中, 包含 22 071 篇论文, 217 849 个引用链接, 12 429 个作者.

我们的目标是在 2001 年 1 月 1 日预测 1992 年～2000 年发表的所有论文在今后 3 年 (2001 年～2003 年) 被关注的程度, 进而预测它们在 2001 年～2003 年的 PageRank 值排名和 Cited counts 排名.

3.2　算法收敛速度与比较

实验平台为 Core2 四核, 2.66 GHz, 4 GB 内存, Windows 7, 64 位系统.

我们将从算法运行效率的角度讨论我们的算法、FutureRank 算法以及 Co-Ranking 算法[15]. FutureRank 算法需要计算论文现有的 PageRank 值. 所以, 作为对比, 需要考察计算论文现有 PageRank 值的效率.

选用 Numerical Computing with Matlab 工具箱计算论文 PageRank 值, 该工具箱充分运用稀疏矩阵的特点, 实现了一种不进行矩阵乘法的幂法迭代求解 PageRank 值计算过程. 其运算效率相当高, 对于 2.2 万个节点的图, 该方法能够在 15 秒内完成 84 次迭代, 最终求得 PageRank 值.

本文算法用作者的权威值替代了 FutureRank 和 PageRank 中的论文权威值. 这种替换带来的一个好处是使转移矩阵变得更加稀疏 (因为作者的数量远远小于论文的数量), 而更加稀疏的矩阵会在迭代过程中带来更小的运算量, 这可以解释为什么我们的算法会大大提高运算效率. 实际上, 在实验中本文算法用 1.2 秒完成中间矩阵的生成后, 仅花费 0.1 秒时间进行 15 次迭代就可以使算法收敛. 相对于需要预先计算 PageRank 值的 FutureRank 算法来说, 这是一个巨大的性能提升. 在图 5 中, 分别比较了本文算法和 PageRank 算法的计算时间和迭代收敛次数. 图 6 是本文算法与 Co-Ranking 算法在运行效率上的对比.

Co-Ranking 算法利用 HITS 算法的思想, 通过三种网络: 作者 – 论文引用网络、论文 – 论文引用网络网络以及作者共著网络, 评价论文的价值以及作者的价值. 一位作者的权威值来源于其共著作者的权威值以及其所著论文的权威值; 一篇论文的权威值来源于该篇文章作者的权威值以及引用过它的论文的权威值. 算法的一次迭代涉及三部分随机游走过程: 共著网络中的随机游走过程, 论文引用网络中的随机游走过程, 以及作者 – 论文引用网络中的随机游走过程. 可以看出, 复杂的迭代过程造成其算法的运行效率低下, 在实验中, 对于相同的数据, 我们花费 113.7108 秒完成 58 次迭代后算法收敛. 与 Co-Ranking 算法不同, 本文算法的迭代过程只涉及作者 – 作者引用网络, 迭代过程大大简化, 速度大为提升. 实验数据对比如图 6 所示.

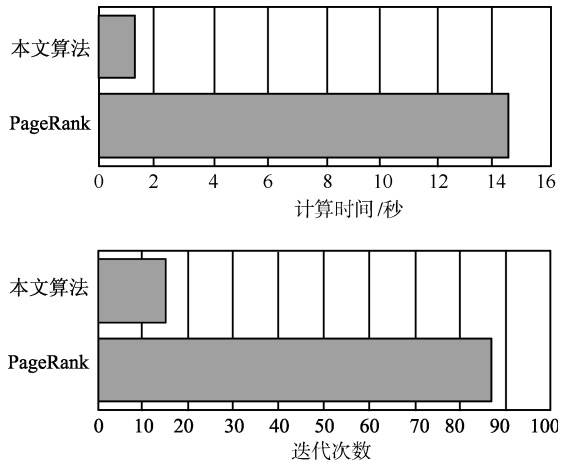


图 5　与 PageRank 算法执行效率的对比

Fig. 5　The efficiency of our algorithm and PageRank

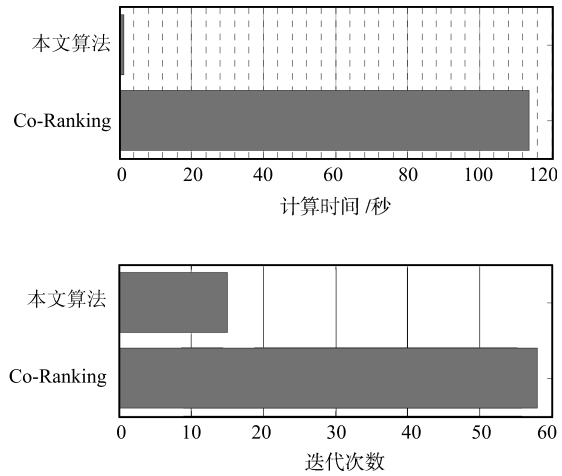


图 6　与 Co-Ranking 算法执行效率的对比

Fig. 6　The efficiency of our algorithm and Co-Ranking

3.3　实验结果与比较

首先, 我们定义评价标准 Top100:

$$\text{Top100} = \left|\text{预测排名}_{\text{Top100}} \bigcap \text{真实排名}_{\text{Top100}}\right| \tag{11}$$

Top100 是预测排名与真实排名在前 100 位交集的数量, 这个评价标准用来检验预测算法对排名靠前的论文的预测准确度.

在比较过程中, 我们还用到了 Spearman′s rank, 它是一种评价等级序列相关性的标准.

下面利用 Spearman′s rank 和 Top100 来评价

预测排名与未来真实 PageRank 排名、被引次数排名之间的相关程度.

图 7 为预测结果随参数变化的情况, 其中, 图 7 (a) 是预测排名与未来真实 PageRank 排名的相关度 $SPPR$, 随参数 ρ 和 C 的变化情况; 图 7 (b) 是预测排名与未来真实 Cited counts 排名的相关度 $SPCC$, 随参数 ρ 和 C 的变化情况; 图 7 (c) 是预测排名与未来真实 PageRank 排名前 100 交集的数量 $TOPPR$, 随参数 ρ 和 C 的变化情况; 图 7 (d) 是预测排名与未来真实 Cited counts 排名前 100 交集的数量 $TOPCC$, 随参数 ρ 和 C 的变化情况.

从图 7 中可以看出, 预测结果与 ρ 紧密相连. 对于 $SPPR$ 和 $SPCC$, 当 ρ >1.0 时, 基本保持稳定.

在 ρ 确定的情况下, 预测准确率主要受参数 C 的影响, C 是度量 $\sum(A_m \times T_i^{\mathrm{Time}})$ 以及 $\sum(H_n \times \sum R_i^{\mathrm{Time}})$ 权重的参数.

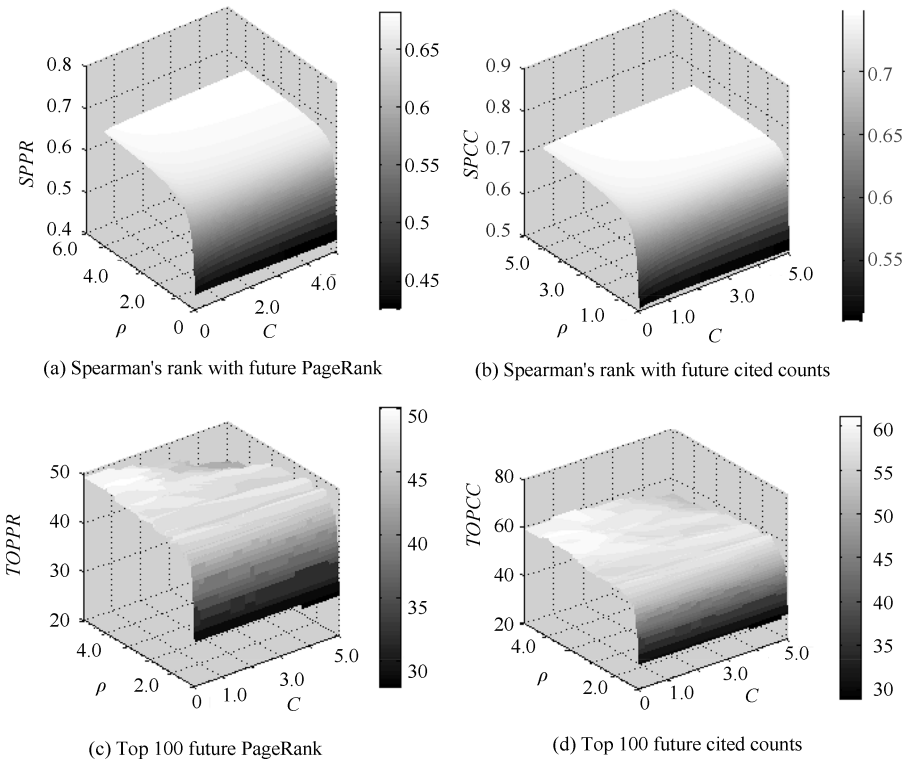
由于参数 C 不是线性反映上述两项的权重变化情况的, 所以在图 8 中我们采用对数坐标, 当 $C \in (0,1]$ 时, $\sum(A_m \times T_i^{\mathrm{Time}})$ 项起主要作用, 当 C >1 时, $\sum(H_n \times \sum R_i^{\mathrm{Time}})$ 项起主要作用. 从图中可以看出, 当 $C = 10$ 左右, $SPPR$ 和 $SPCC$ 达到峰值.

尽管当 C 在 10 左右取得 $SPCC$ 和 $SPPR$ 的峰值, 但在确定参数 ρ 和 C 时, 我们更倾向于保证前 100 篇的论文的预测准确率, 而不是全部论文排名. 也就是说, 在保证能较准确地预测前 100 位的论文的前提下, 再考虑全部论文的 Spearman′s rank 值. 这是因为, 本文的目的是为科研工作者的文献检索提供辅助, 所以排名靠前论文的预测准确率显然更加重要.

综合上述几点, 我们最终确定参数 $\rho = 4.45$ 和 $C = 1.6$, 此时 $SPCC = 0.746$, $SPPR = 0.678$, $TOPPR = 49$, $TOPCC = 60$.

图 9 比较了本文算法、CiteRank 以及 FutureRank 算法的预测准确率. 可以看出, FutureRank 算法得到的预测排名与未来 PageRank 值排名的相关度 $SPPR$ 为 0.59. 而本文算法得到的预测排名与未来 PageRank 值排名的相关度为 0.68, 较 FutureRank 有较大幅度提高. FutureRank 算法得到的预测排名与未来 Cited counts 排名的相关度为 0.75, 这与我们的算法持平. CiteRank 算法计算的排名与未来真实 Cited counts 排名的相关度为 0.57[2].

可以看出, 本文算法在保持对论文 “流行度” (Cited counts) 预测准确率的同时, 大幅提高了论文 “权威度” (PageRank) 的预测准确率.



(a) Spearman's rank with future PageRank (b) Spearman's rank with future cited counts

(c) Top 100 future PageRank (d) Top 100 future cited counts

图 7 预测结果随参数变化

Fig. 7 Prediction varies with the parameters

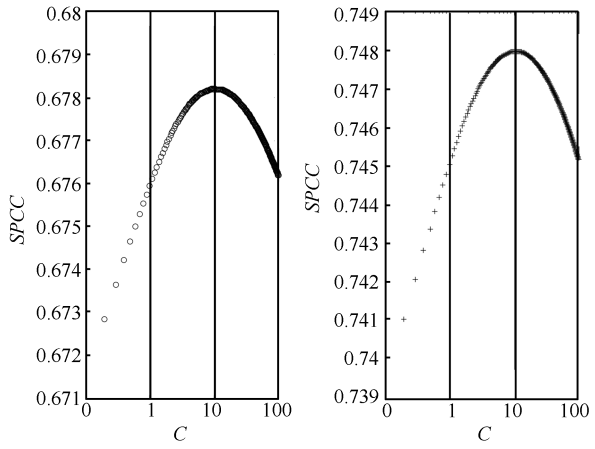


图 8　$\rho = 4.0$ 时相关度随 C 变化曲线

Fig. 8　Correlation coefficient varies with the parameters C ($\rho = 4.0$)

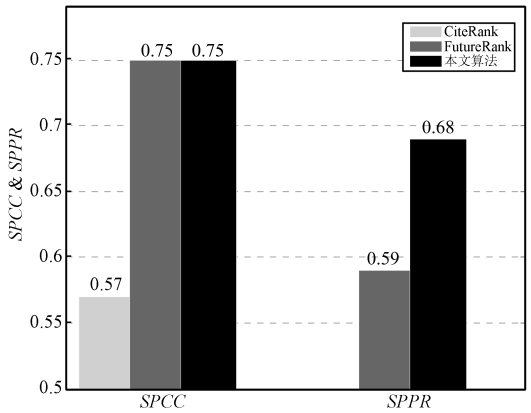


图 9　实验结果对比

Fig. 9　The experimental results contrast

表 1 直观地比较了预测排名和真实排名, 可以看出本文算法能够较准确的预测论文 (尤其是高排位论文) 的名次, 表 2 和表 3 分别列出了撰写权威值和引用权威值排名前 10 位的作者, 虽然作者的撰写权威值 (引用权威值) 与其撰写论文数 (引用论文数) 密切相关, 但二者不是完全重合的, 例如撰写权威值排名第 2 位的 Juan M. Maldacena 只撰写了 20 篇论文, 数量远低于排在其后的作者.

4　结语

本文基于作者–论文间的引用链接, 预测了论文未来被关注的程度. 考虑的因素包括撰写者权威值、引用者权威值、论文发表时间以及论文被引用时间. 实验结果表明, 本文的算法可以在不计算论文当前 PageRank 值的条件下, 更准确地预测论文未来被引数量排名和 PageRank 值.

我们的算法通过一篇论文过去一段时间的被引用情况来判断其是否具有继续被关注的潜力, 所以论文必须有被引用的记录, 才能加以判断. 对一些刚刚发表的论文, 由于没有引用链接, 对其的预测完全依赖于撰写者的权威值.

为了解决这个问题, 在未来的工作中可以考虑将论文所在期刊权威值加入到算法中, 具有高被引频率的期刊中发表的论文, 往往质量较高, 会得到科研工作者群体较高的评价. 预计在算法中加入期刊权重后, 对上述没有引用的论文的预测会更加准确.

出于与 FutureRank 比较的目的, 本方法使用了与 FutureRank 相同的数据集 arXiv (hep-th). 我们考虑未来在更大、更多的数据集中验证本方法, 考察系数 C 和 ρ 的变化情况等.

表 1　预测排名与真实排名前 10 位的比较 ($\rho = 4.45$, $C = 0.01$)

Table 1　Comparison of the predict rank and the truly rank of the Top10 ($\rho = 4.45$, $C = 0.01$)

arXiv ID	日期	1993～2000 被引	2001～2003 预测	2001～2003 PageRank 排名	2001～2003 被引
9711200	1997-11-28	1	1	1	1
9908142	1999-08-23	14	2	3	2
9802150	1998-02-23	2	3	4	3
9802109	1998-02-17	4	4	5	5
9906064	1999-06-09	17	5	2	4
9711162	1997-11-21	16	6	7	7
9905111	1999-05-17	22	7	6	6
9711165	1997-11-24	37	8	34	20
10005031	2000-05-04	309	9	95	66
10003160	2000-03-20	299	10	19	21

表 2　撰写权威值排在前 10 位的作者及其撰写论文数 (截止到 2001 年)

Table 2　The Top10 writers and the number of papers that they writed (as of 2001)

排名	作者 ID	姓名	撰写论文数
1	1693	Edward Witten	83
2	2450	Juan M. Maldacena	20
3	1043	Joseph Polchinski	37
4	451	Michael R. Douglas	44
5	640	Andrew Strominger	52
6	704	Nathan Seiberg	34
7	1601	N. Seiberg	23
8	80	A. A. Tseytlin	92
9	749	Ashoke Sen	72
10	55	C. Vafa	31

表 3 引用权威值排在前 10 位的作者及其引用论文数 (截止到 2001 年)

Table 3 The Top10 citers and the number of papers that they cited (as of 2001)

排名	作者 ID	姓名	引用论文数
1	80	A.A. Tseytlin	922
2	8776	S.S. Gubser	662
3	3923	O. Aharony	660
4	1693	Edward Witten	639
5	5678	H. Ooguri	578
6	2235	Soo-Jong Rey	541
7	3303	J. Maldacena	570
8	6083	Y. Oz	569
9	1317	Yaron Oz	569
10	749	Ashoke Sen	669

References

1 Garfield E. Citation analysis as a tool in journal evaluation. *Science*, 1972, **178**(4060): 471−479

2 Sayyadi H, Getoor L. Future rank: ranking scientific articles by predicting their future PageRank. In: Proceedings of the 9th SIAM International Conference on Data Mining. Sparks, NV: SIAM, 2009

3 Page L, Brin S, Motwani R, Winograd T. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab, USA, 1998

4 Bollen J, Rodriguez M A, van de Sompel H. Journal status. *Scientometrics*, 2006, **69**(3): 669−687

5 Chen P, Xie H, Maslov S, Redner S. Finding scientific gems with Google′s PageRank algorithm. *Journal of Informetrics*, 2007, **1**(1): 8−15

6 Ma N, Guan J C, Zhao Y. Bringing PageRank to the citation analysis. *Information Processing & Management*, 2008, **44**(2): 800−810

7 Bergstrom C T, West J D, Wiseman M A. The Eigenfactor™ Metrics. *Journal of Neuroscience*, 2008, **28**(45): 11433−11434

8 Kleinberg J M. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 1999, **46**(5): 604−632

9 Berberich K, Vazirgiannis M, Weikum G. Time-aware authority ranking. *Internet Mathematics*, 2005, **2**(3): 301−332

10 Dong A L, Chang Y, Zheng Z H, Mishne G, Bai J, Zhang R Q, Buchner K, Liao C Y, Diaz F. Towards recency ranking in web search. In: Proceedings of the 3rd ACM International Conference on Web search and data mining. New York, USA: ACM, 2010. 11−20

11 Walker D, Xie H F, Yan K K, Maslov S. Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, 2007, **2007**(6): 06010

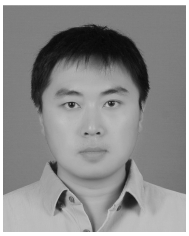
12 Yan E J, Ding Y. Weighted citation: an indicator of an article′s prestige. *Journal of the American Society for Information Science and Technology*, 2010, **61**(8): 1635−1643

13 Yan E J, Ding Y, Sugimoto C R. P-Rank: an indicator measuring prestige in heterogeneous scholarly networks. *Journal of the American Society for Information Science and Technology*, 2011, **62**(3): 467−477

14 Yan E, Ding Y. Measuring scholarly impact in heterogeneous networks. *Proceedings of the American Society for Information Science and Technology*, 2010, **47**(1): 1−7

15 Zhou D, Orshanskiy S A, Zha H Y, Yan K K. Coranking authors and documents in a heterogeneous network. In: Proceedings of the 7th IEEE International Conference on Data Mining. Omaha NE, USA: IEEE, 2007. 739−744

刘大有 吉林大学计算机科学与技术学院教授. 主要研究方向为知识工程, 专家系统与不确定性推理, 分布式人工智能, 多 Agent 系统.
E-mail: dyliu@jlu.edu.cn
(**LIU Da-You** Professor at the College of Computer Science and Technology, Jilin University. His research interest covers knowledge engineering, expert system and uncertainty reasoning, distributed artificial intelligence, and multi-agent systems.)



薛锐青 吉林大学科学与技术学院硕士研究生. 主要研究方向为数据挖掘.
E-mail: alvininchina@126.com
(**XUE Rui-Qing** Master student at the College of Computer Science and Technology, Jilin University. His main research interest is data mining.)

齐 红 博士, 吉林大学计算机科学与技术学院副教授. 主要研究方向为数据挖掘, 统计关系学习, 语义检索. 本文通信作者. E-mail: qihong@jlu.edu.cn
(**QI Hong** Ph. D., associate professor at the College of Computer Science and Technology, Jilin University. Her research interest covers data mining, statistical relational learning, and semantic retrieval. Corresponding author of this paper.)