首都师范大学学报(自然科学版)
第 41 卷 第 5 期 Journal of Capital Normal University No. 5
2020 年 10 月 (Natural Science Edition) Oct. ,2020

基于 PageRank 改进的文献价值排序算法*

孙泽锋 周 洁** 李忠义

(首都师范大学数学科学学院,北京 100048)

摘要:已有学者通过单一指标或者综合因素做了大量文献价值排序工作,但是新旧论文的价值评估不公平现象依旧存在. 为了解决这一问题,本文提出一种改进的文献价值排序算法,该算法结合了经典 PageRank 思想以及常用的被引量与下载量,同时将时间因子引入阻尼因子 (d),使得新旧文献在价值评估时有不同的侧重,提高了价值评估公平性. 实证分析结果表明,与传统的排序结果相比,新算法的排序更加符合搜寻者的需求.

关键词:文献价值;排序;PageRank;阻尼因子

中图分类号:TP391 **DOI**:10. 19789/j. 1004-9398. 2020. 05. 001

0 引 言

在信息时代,数据量的急剧增加使丰富的资源去创造价值成为可能,在繁杂的文献数据库中找到最适合的参考资料是关键的一步. 大多数科研工作者在撰写论文的时候,会查询大量文献资料。因此,如何快速而准确地找到最有价值又符合期望的文献是一个值得研究的课题.

PageRank 算法自推出以来就吸引了大批学者的关注[1],在对算法本身进行改进的同时,部分学者还将算法与文献价值评估相结合. 目前,对于如何改进文献排序,研究者提出了不同的解决方法. 刘大有等[2]对文献作者的权威性做了量化,基于作者与引用者二者的权威值对文献的影响力进行评估;Chen 等[3]用 PageRank 算法和引用次数分析了 *Physical Review* 期刊族在 1893—2003 年发表的所有论文;刘欣[4]提出了一种综合考虑文献的内容、期刊、作者和时间等反映文献价值的因素对文献进行排序的 PageRank 方法;张瑜等[5]基于文献引用关系分析了科学文献与其参考文献之间的相似性和统计特征,对参考文献进行了评估. 李长玲[6]认为评估一篇文献的质量需要了解引用该文献的其他文献的价值;王丹[7]基于 Lucene 排序机制及 PageRank 算法,提出了一种新的文献搜索排序算法. 纵观大量关于文献价值排序的论文,新旧文献由于发表时间因素而导致的排序不合理现象依旧存在,本文基于 PageRank 改进算法,利用发表时间不同来决定新旧文献价值评估的侧重点,同时引入被引下载比来刻画自身固有价值.

1 PageRank 算法介绍

1998 年,Brin 和 Page[1]首次提出 PageRank 算法当即引起了广泛的关注,该算法当时不仅成为了谷歌的关键技术,而且还应用到了其他不同领域. PageRank 的基本思想是通过网页的被链接数量来进行排序,一个页面之所以有指向另一个页面的链接,是因为该页面比较权威,内容真实可信,在相关领域有一定知名度,同时提出 PageRank(PR)值的概念. PageRank 算法中的核心是其除了计算页面的入链数量之外,还将指向目标页面的其他页面自身的 PR 值考虑在内. 因此,如果网页 A 被一个重要的页面链接,网页 B 被很多普通的页面链接,那么很有可能 A 的 PR 值将不低于 B 的 PR 值[8].

假设在上网时,点击网页后,会随着链接的引导一直点击进入下一层页面,直到完成任务关闭页面,又或者随机打开了一个新的页面. 于是,提出了一个阻尼因子 (d),一般取值为 0. 85,表示用户继续点击链接的概率,同时,$1-d$ 将作为用户跳出链接点击一个新的页面的概率. 因此,得到如下 PageRank 计算公式

$$\mathrm{PR}(p_i) = (1-d) + d \times \sum_{k=1}^{n} \frac{\mathrm{PR}(p_j)}{C(p_k)}, \tag{1.1}$$

收稿日期:2019-07-30

* 国家自然科学基金(11671275);科技创新服务能力建设-基本科研业务费(科研类)(025185305000/204)

** 通信作者: zhoujie@ amss. ac. cn

式中 $\mathrm{PR}(p_i)$ 表示页面 p_i 的 PR 值; $C(p_k)$ 表示由页面 p_j 链出的页面 p_k 总数; d 为阻尼因子,有时还用来解决某些特殊情况导致的个别页面 PR 值因无法收敛而难以计算情况的发生.

2 文献价值排序改进算法

类比于网页排序问题,王向阳和马军[9]提出将 PageRank 算法应用在科技文献的排序上,同时将文献价值定义为自身固有价值与被引用后获得的价值的权重加和. 在考虑文献固有价值的时候,由文献所发表的刊物或会议的级别及作者的权威性决定[9],同时用发表年限作为衡量参考文献重要性的标准以刻画价值传递的权重,后者难免有失偏颇.

相比用单一的期刊影响因子来刻画文献固有价值,本文用更有说服力的"文献传播力",即被引下载比,来描述文献自身固有价值,同时用某篇参考文献的价值与参考文献的价值总和的比作为价值传递的权重,提出新的文献排序算法(literature value ranking,LVR).

2.1 算法描述

本文基于文献[9]对科技文献价值排序算法的框架,利用文献自身固有价值与被引获得价值二者的加权求和来刻画某一篇文献的价值,其中自身固有价值由其本身"文献传播力"决定,另一价值由引用文献传递,同时,2 个价值之间的权重关系由发表时间所决定. 因此,提出如下文献价值排序算法公式

$$\mathrm{LVR}(u_i) = (1 - d(t_i)) \times \mathrm{Trans}(u_i) + d(t_i) \times \mathrm{Gain}(u_i), \tag{2.1}$$

式中 $\mathrm{LVR}(u_i)$ 表示文献 u_i 的新算法价值; $\mathrm{Trans}(u_i)$ 表示文献 u_i 的传播力,用来刻画自身固有价值; $\mathrm{Gain}(u_i)$ 表示文献 u_i 被引用后获得的价值; $d(t_i)$ 表示阻尼函数,决定新旧文献的评价权重.

2.1.1 自身固有价值

被引量是衡量文献自身固有价值的重要指标,相对来说是比较合理的,但是作为一篇领域里最新的论文,与发表多年的论文相比,被引量偏低,就会被低估了价值. 根据上述原因,本文提出一个新的概念——文献传播力. 结合被引量与下载量,提出用"文献传播力",即被引下载比,来刻画文献自身固有价值,指文献在被下载之后被同行或业界认可并引用的能力. 公式为

$$\mathrm{Trans}(u_i) = \frac{\mathrm{Cited}(u_i)}{\mathrm{DL}(u_i)}, \tag{2.2}$$

式中 $\mathrm{Trans}(u_i)$、$\mathrm{Cited}(u_i)$ 与 $\mathrm{DL}(u_i)$ 分别表示为文献 u_i 的被引下载比、被引量与下载量.

2.1.2 被引获得的价值

类比于网页之间的链接关系,文献之间的引用关系也存在价值的传递. 被引用文献会获得其他文献对其的"肯定",这个"肯定"就用价值的传递来表示,每篇文献就会将自身价值分别传递给引用的参考文献,于是被引获得的价值表达式为

$$\mathrm{Gain}(u_i) = \sum_{i,j} w(u_i, u_j) \times \mathrm{LVR}(u_j). \tag{2.3}$$

按照 PageRank 算法,价值的传递权重由参考文献的数量决定,但是参考文献本身价值不同,因此均匀分配的形式行不通. 本文提出用"影响力系数"来决定权重,系数的大小由该参考文献价值与所有参考文献价值总和的比值所决定:

$$w(u_i, u_j) = \frac{\mathrm{LVR}(u_i)}{\sum_{k \in B_j} \mathrm{LVR}(u_k)}, \tag{2.4}$$

式中 $w(u_i, u_j)$ 为影响力系数,用于决定分配多少论文的价值给予参考文献; $\mathrm{LVR}(u_i)$ 表示 u_i 用新算法计算得到的文献价值; B_j 表示为文献 u_j 的参考文献集合.

2.1.3 阻尼函数

对于 PageRank 算法中的 d, 本文对其进行了调整. 设想一下,近期发表的文献与若干年前发表的文献虽然都没有很高被引量,但是前后两者相对比,可以分析出后者确实是因为价值不高才不被重视,而新论文还未经过时间的检验,因此将原有 d 与时间因子相结合,得到新的阻尼函数,如下所示:

$$d(t_i) = d \times \frac{t_0 - t_i}{\sum_k (t_0 - t_k)}, \tag{2.5}$$

式中 t_i 表示文献 u_i 发表时间, t_0 表示当前时间, $\sum_k (t_0 - t_k)$ 表示所有文献的发表年限和.

阻尼函数的提出,可根据发表时间的不同,给予文献自身固有价值与被引后获得价值不同的权重. 利用新旧文献不同的时间积淀,使用不同的方法刻画其价值,优化了文献价值排序的结果.

2.2 算法可行性分析

文献价值评估改进算法中利用了文献自身的被引量、下载量以及发表时间等因素.

对于文献自身固有价值的衡量,采用"文献传

播力”,用被引量与下载量的比值表示. 不论是最新论文,还是年份久远的论文,只要是有价值的,那么其被引下载比一定会趋近于 1,同时对于所有文献而言,又能减少发表时间带来的不公平因素.

改进原有 PageRank 算法中的 d,加入时间因子,使得文献根据发表时间获得不一样的权重. 从 $d(t_i)$ 中可以看到,越新的论文,其阻尼函数就越小. 因为新论文与几年前的论文相比,新论文由于没有时间的积累无法获得较高的价值,因此对于他的价值刻画更多的是着眼于自身固有价值而不是被引用而获得的价值,因此算法前部分的权重会高于后半部分,反之,对于发表时间更早的文献,其评价会更加倾向于被引用后其他文献给予的价值,其权重自然也会稍微高一些. 给予自身固有价值与获得的价值不同的权重,让新旧文献在价值评估的时候能够相对公平.

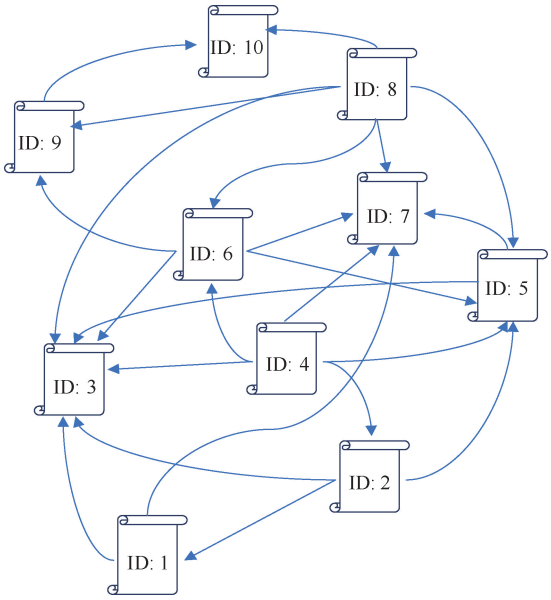

图 1　10 篇文献的互引网络图

3　实证分析

为了更加客观地验证提出算法的有效性,下面展开实证分析.

本研究在中国知网搜集了关于 PageRank 有关的文献信息(搜索时间为 2019 年 3 月 12 日),利用知网的计量可视化分析功能得到总共 1 975 条结果的文献互引网络.

由于互引网络十分庞大,在可视化分析筛选条件中选取了关系强度为 8 的 10 篇文献,文献之间的互引网络图如图 1 所示. 分别计算这10 篇文献的平均被引量、经典 PR 值、LVR 值以及被引下载比等数据(表 1),其中平均被引量

$$\overline{\mathrm{Cited}(u_i)} = \frac{\mathrm{Cited}(u_i)}{t_0 - t_i}. \tag{3.1}$$

为了验证结果,将改进的 LVR 算法与 PageRank 算法及只按照平均被引量排序的结果进行对比.

若按照平均被引量排名,不考虑其他文献对参考文献的价值贡献,如果 2 篇发表时间相同的文献(A 和 B),文献 A 被 10 篇价值较低的文献引用,而文献 B 则被 5 篇价值较高的文献引用,以平均被引量比较,A 比 B 更有价值,而引入新算法后,其他文献给予的价值也引入影响因素体系,因此,排名有了很大的调整. 如排在平均被引量前 2 名的 2 篇综述型文献,相对来说容易被引用,只要是相关领域的文献就有可能引用其,但文献价值高低不一,因此在 LVR 算法中的排名大幅下跌. 分析 LVR 算法进行排序后的结果,前 3 名文献都具有技术革新,更容易被高质量的文献所引用,同时这些文献对搜

表 1　不同方法文献评估结果

文献标题	编号	平均被引		PageRank 算法		LVR 算法		被引下载比
		量	排名	PR 值	排名	LVR 值	排名	
PageRank 算法研究综述	1	13.00	2	0.020 2	7	1.962	7	0.108 9
具有时间反馈的 PageRank 改进算法	2	4.64	7	0.018 2	8	0.701	2	0.080 0
加速评估算法:一种提高 Web 结构挖掘质量的新方法	3	5.80	6	0.048 8	1	1.090	1	0.023 6
基于网页内容和时间反馈的网页排序 PageRank 算法研究	4	2.00	9	0.015 0	9	0.300	8	0.129 3
PageRank 算法研究	5	13.15	1	0.029 4	4	2.011	6	0.011 2
搜索引擎排序算法的研究	6	2.17	8	0.020 7	6	0.334	1	0.049 9
对网页 PageRank 算法的改进	7	7.94	4	0.045 3	2	1.376	3	0.033 8
基于搜索引擎网页排序算法研究	8	1.33	10	0.015 0	10	0.200	9	0.218 0
HITS 算法与 PageRank 算法比较分析	9	6.80	5	0.021 1	5	1.030	5	0.137 7
Google 的 PageRank 技术剖析	10	10.12	3	0.035 5	3	1.555	4	0.035 0

寻者也更有价值.

对比 PageRank 与 LVR 算法,其不同在于 d 的变化、自身价值评价以及被引用时所得到的价值权重. PageRank 算法中 d 统一为 0.85,这对新发表的论文不公平. 对该类论文而言,没有较多的被引量,从外界获得的价值也不高,因此,更多的评价权重应该放在前者"文献传播力"上面. 同时,"文献传播力"也是本文的一个创新点,排除时间的影响来衡量文献自身价值,如文献 2 在 PageRank 算法排名为第 8 名,而在 LVR 算法中却跃升到了第 2 名,查看文献数据得知,其"文献传播力"很强,仅次于文献 3. 解决了新文献在排序时被低估的不公平现象.

4 结 论

对于文献的价值评估,本文以 PageRank 算法的形式,将文献自身价值和文献获得价值 2 个方面相结合. 相对于先前的研究,本文将 d 变成了随时间变化的阻尼函数,又提出"文献传播力"的概念,将被引量与下载量结合来描述文献本身的价值,尽最大可能将时间因素排除,让新发表的论文能够更快的被发现,优化了排序算法. 在实证分析当中,本研究也对 LVR 算法的有效性进行了验证.

此外,本文在数据集的选取环节还存在不足,无法做到将整个网络的文献进行排序,只截取了相对代表性的文献进行验证. 对于很多年前的文献,由于网络并不发达,下载量会被低估. 在接下来的研究当中,将进一步扩大数据研究范围,继续选取更加合适的指标来描述文献自身固有价值,同时考虑在不同主题、不同关键字的情况下文献的排名情况,使得文献检索更加准确.

参 考 文 献

[1] BRIN S, PAGE L. The anatomy of a large-scale hypertextual Web search engine[J]. Computer Networks and ISDN Systems, 1998(30): 107 – 117.

[2] 刘大有,薛锐青,齐红. 基于作者权威值的论文价值预测算法[J]. 自动化学报,2012,38(10): 1654 – 1662.

[3] CHEN P, XIE H, MASLOV S, et al. Finding scientific gems with Google's PageRank algorithm[J]. Journal of Informetrics, 2007(1): 8 – 15.

[4] 刘欣. 基于阅读价值的科技文献排序方法研究[D]. 大连:大连理工大学,2010.

[5] 张瑜,顾进广,张铭晖,等. 科学文献中参考文献影响力评估方法研究[J]. 小型微型计算机系统,2018(10): 2176 – 2179.

[6] 李长玲. 基于 PageRank 的引文分析方法探讨[J]. 情报理论与实践,2007(1): 122 – 124.

[7] 王丹. 基于 PageRank 改进的文献排名算法研究[J]. 计算机时代,2019(1): 59 – 62 + 66.

[8] 李青淋,邵家玉. PageRank 算法的研究与改进. [J]. 工业控制计算机,2016,29(5): 117 – 118.

[9] 王向阳,马军. 一个基于 PageRank 的科技文献质量评价算法[J]. 广西师范大学学报(自然科学版),2009,27(1): 165 – 168.

PageRank-based Improved Literature Value Ranking Algorithm

SUN Zefeng ZHOU Jie LI Zhongyi

(School of Mathematical Sciences, Capital Normal University, Beijing 100048)

Abstract: Scholars have done a lot of the ranking of literature values work through a single indicator or a combination of factors, but the inequality of the evaluating between new and old papers still exists. In order to optimize the ranking problem of the literatures, this paper proposes an improved literature value ranking algorithm, which combines the classic PageRank idea with the commonly used citation and download volume, and introduces the time factor into the damping coefficient (d) to make the new and old literatures have different emphasis in value evaluation, which improves the fairness. The empirical analysis shows that the ranking of the new algorithm is more in line with the needs of the searcher than the traditional sorting results.

Keywords: literature value; rank; PageRank; damping coefficient