doi:10.11835/j.issn.1008-5831.2015.02.014

# 基于特征因子算法改进的作者影响力评价研究

马瑞敏[a,b],韩小林[a,b]

(山西大学 a.管理与决策研究所;b.科学评价研究中心,山西 太原 030006)

**摘要**:特征因子算法是评价期刊质量的一种重要方法,文章在特征因子算法基础上通过改进构造出一种作者影响力评价的新算法。首先对特征因子算法原理进行简单介绍。然后通过分析作者引用相较期刊引用的特殊性,对特征因子算法进行了改进,并对其实现步骤进行了详细说明。最后,选择国内图情学作者引用网络进行应用研究,得到了这些作者的影响力排名,并与传统的被引次数进行了比较。

**关键词**:作者影响力;特征因子算法;改进

**中图分类号**:G350 **文献标志码**:A **文章编号**:1008-5831(2015)02-0106-04

随着信息经济时代向知识经济时代的发展,知识沟通显得尤为重要,文献是知识传播的主要载体,而作者是文献的创作者,所以作者已经成为知识创造的力量源泉和知识传播的主要承载者。作者影响力的评价一直是科研管理界和学者们关注的焦点问题,不少学者对此进行了一系列研究,产生了众多影响力较大的成果,主要集中在如下三个方面:一是被引次数。美国信息学家 E. Garfield 曾编制《科学引文索引》,首次系统化地对作者之间的引用进行研究,并提出用被引次数对作者的影响力进行评估[1];二是 h 指数。美国物理学家 J. E. Hirsch 将作者的发文量和被引次数进行综合考量,提出 h 指数对作者进行评价研究[2];三是改进的 page-rank 算法。如美国印地安那大学的 Ying Ding 考虑到作者引用与网页链接之间的相关性,于是提出对 page-rank 算法进行改进以对作者影响力进行评价[3]。以上研究中前两方面都是仅围绕作者的绝对被引次数展开,第三个方面的研究虽然在考虑作者被引次数的同时也将施引作者的影响力考虑进去,但对一些参数的处理方面还有可改进之处。随着对期刊评价的一种新算法——特征因子算法的提出,该算法在考虑期刊引用之间的被引次数和施引期刊的影响力之外,将其引用的方向性和多次引用情况都考虑进去[4],这样对于评价期刊的影响力更具有科学性和说服力。D. Jevin 考虑到期刊评价与作者评价之间有非常明显的相似性,开始尝试将该算法运用到对作者影响力评价中[5],而国内有不少学者只是针对特征因子算法自身的原理等进行探究[6-8],目前尚未有学者将该算法应用到对作者的评价研究中。

本文试图将特征因子算法拓展到对作者影响力的评价研究中,并且根据作者影响力评价研究自身的特殊性对该算法进行改进,提出一种对作者影响力评价的新算法,这样不仅是对作者影响力评价研究领域的补充,也为该方面的研究提供了一个新的视角。

## 一、特征因子的基本算法

影响因子在计算期刊的引用次数时,对于不同期刊的引用都平等对待,只统计引用次数,而事实上,不同期刊的价值有很大的差别,如 Nature 和 Science 这样影响力非常大的期刊,显然不能和一些普通期刊的引用份量同等对待,两篇文章分别被 Nature 引用和被一个不知名的期刊引用,则这两篇文章的质量相差很大。基于这样的现实情况,于是就引入特征因子(Eigen-factor)这个指标,该指标的制定考虑了引用该期刊的期刊源的权重,通过期刊之间的引用情况构建期刊引用网络,从而对期刊的重要性进行评价。特征因子算法工作原理具体如下:首先选择一个期刊,并任意选择该期刊中一个参考文献链接到另一个期刊,然后在之前链接到的那个期刊中再任意选出一个参考文献,再继续链接到对应的下一个期刊,依此类推,不停地重复这个

**修回日期**:2014-09-25

**基金项目**:国家社会科学基金青年项目"作者引用网络模式与功效研究"(12CTQ026)

**作者简介**:马瑞敏(1983-),男,山西榆次人,山西大学科学评价研究中心执行主任,副教授,硕士研究生导师,主要从事信息计量与科学评价研究。

行为,于是发现被链接到次数越多的期刊,其影响力越大,链接到该期刊的概率值的百分位数就是该期刊的特征因子值。

特征因子算法主要包括两大步骤,首先对期刊引用矩阵进行规范化处理,即:$M_{ij} = \frac{Z_{ij}}{\sum_{k} Z_{kj}}$,其中 $Z_{ij}$ 表示期刊 $j$ 来自期刊 $i$ 的被引次数,$\sum_{k} Z_{kj}$ 表示期刊 $j$ 的总被引次数;然后构建过渡矩阵,即:$P = \alpha M' + (1-\alpha)A$,其中 $\alpha$ 表示期刊引用过程中的阻尼系数,一般取 0.85,$M'$ 表示矩阵 $M$ 经悬点处理后的随机矩阵,$A$ 表示期刊发文量向量对应的单位向量,即 $A = a \cdot e^T$;最后通过计算过渡矩阵的最大特征值得到期刊的特征因子得分[5]。

## 二、改进的特征因子算法

作者之间的引用与期刊之间的引用情况非常相似,但两者也有一定的区别。通常,一个期刊的被引用情况与其所承载的论文数量有很大关系,承载论文数量越多的期刊越有机会得到其他期刊的引用,而作者之间的引用受作者自身被引次数的多少影响较大,受作者的发文量的影响相对较小。因此在如下两方面进行改进:第一,将特征因子算法中的 $A$ 改进为作者被引次数向量对应的单位向量;第二,在构建随机矩阵 $M'$ 时,用作者被引次数向量代替悬点向量,从而对矩阵 $M$ 进行改进后的悬点处理。改进后作者影响力算法的具体步骤如下。

(1)构建作者引用网络矩阵。考虑作者引用与期刊引用的相似性,可以根据特征因子对期刊评价的原理与思路,模仿期刊引用网络矩阵构建的方法来构建作者引用网络矩阵。矩阵中第一行的作者表示被引作者,第一列的作者表示引用作者,矩阵中的元素表示被引次数。由于排除了自引,所以矩阵对角线上全为 0,矩阵 $Z$ 即为 $n$ 个作者的作者引用网络矩阵,元素 $C_{i,j}$ 表示矩阵中作者之间的引用次数。

$$Z = \begin{pmatrix} 0 & C_{12} & C_{13} & \cdots & C_{1n} \\ C_{21} & 0 & C_{23} & \cdots & C_{2n} \\ C_{31} & C_{32} & 0 & \cdots & C_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ C_{n1} & C_{n2} & C_{n3} & \cdots & 0 \end{pmatrix}$$

(2)对第(1)步中所构建的作者引用网络矩阵进行规范化处理,即用每列被引次数除以该列被引次数的总和,矩阵中相应元素可以用如下的公式表示:$H_{i,j} = \frac{C_{i,j}}{\sum_{i=1}^{n} C_{i,j}}$

(3)悬点的处理。由于发现有些作者从来没有引用过其他作者,因此在所构建的作者引用网络矩阵中就会有这些作者对应的列全为 0 的现象出现,于是称之为悬点。用 1 对应该矩阵中悬点所在的列,用 0 对应该矩阵中非悬点所在的列,则由 1 和 0 组成的行向量就可以表示该矩阵。假设第二个作者从来没有引用过其他作者,则在作者引用矩阵中第二列的值全为 0,如下所示的向量 $d$ 即可以表示该作者引用矩阵:$d = (0 \quad 1 \quad 0 \quad \cdots \quad 0)$。

(4)计算作者被引次数向量。每个作者的被引次数除以所有作者总的被引次数,这样的一个列向量即为作者被引次数向量,则作者被引次数向量中的元素可以用如下的公式表示:$a_{i,1} = \sum_{j=1}^{n} C_{i,j} \Big/ \sum_{i=1}^{n}\sum_{j=1}^{n} C_{i,j}$。

(5)将所构建的作者引用网络矩阵中的悬点用被引次数向量代替,构建出一个随机矩阵,该随机矩阵对应作者在科学文献引用中的随机漫游过程。此时与特征因子有所不同,特征因子是用文章向量来代替期刊引用网络矩阵中的悬点,而本文在对作者的引用网络进行研究时选择了用被引次数向量来代替悬点,这里主要考虑了在对期刊的引用时其随机性主要受期刊中文章数量的影响,而对作者的引用主要是受作者的被引次数的影响,与作者所发文章数的关系相对较小,故在本文中选择用被引次数向量来替代作者引用网络矩阵中的悬点。如第(3)步中作者引用矩阵中第二列的元素可以用如下公式表示:$H_{i,2} = \sum_{j=1}^{n} C_{i,j} \Big/ \sum_{i=1}^{n}\sum_{j=1}^{n} C_{i,j}$。

(6)构建过渡矩阵。定义过渡矩阵 $P = \alpha M' + (1-\alpha)A$,即 $P = \alpha M' + (1-\alpha)a \cdot e^T$,其中 $\alpha$ 为阻尼系数,仍然取 0.85,$M'$ 为第(5)步中构建出的随机矩阵,$a$ 为第(4)步中的作者被引次数向量。在此基础上,再定义 $\pi^*$ 为作者影响力向量,$\pi^*$ 由过渡矩阵的最大特征值所对应的那个向量表示。

(7)计算作者影响力值。作者影响力值向量的计算与特征因子值向量的计算相似,是对应作者引用网络矩阵和第(6)步中 $\pi^*$ 的点积,经过规范化处理后乘以 100,换算为百分值所得。

(8)用 Matlab 软件编写程序进行迭代计算,计算出最终结果。

## 三、应用研究

### (一)数据的收集与处理

本文选择中国图书情报领域内的所有作者在 2010 - 2012 年的引用情况作为研究对象。为了保证收集

数据的可靠性与科学性,我们选择中国社会科学引文索引(CSSCI)数据库作为本次研究的数据来源。另外,在高级检索处,选择发文年代:2010 - 2012 年;文献类型:论文;学科类别:图书馆、情报与文献学;学位类别:图书馆、情报与档案管理(一级),其他都为默认的选择。最终得到文献记录为 24 041 条,作者数为 27 036 个。构建 27036 ×27036 矩阵,然后利用 Matlab 自编程序进行数据清理和计算。

(二)结果分析

基于改进算法,得到这 27 036 位作者的影响力得分。首先,对所有作者的影响力的分布进行分析,结果呈现出非常明显的偏斜现象(图 1),符合长尾分布特征。

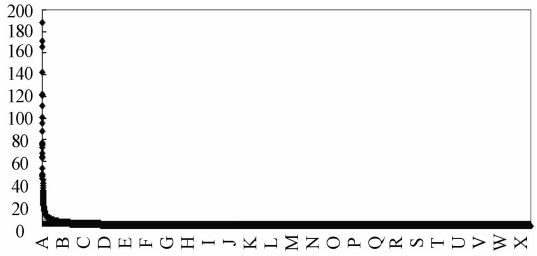


图 1 作者影响力的偏斜分布图

从图 1 中可以看出,只有极少数作者的影响力较高,大部分作者的影响力都很低,并且有不少作者的影响力为 0。另外,从图中作者影响力的偏斜程度看,影响力较高的作者之间波动也相对较大,呈现出明显的下滑趋势,可见该领域内高影响力的学者偏少,作者间影响力强弱差距较大。也从一个侧面可以看出本文所提出的方法能够较好地区分作者之间的影响力。

为了更好地观察结果,下面对前 50 名作者进行研究。被引次数是当前评价作者影响力非常重要的指标,虽然 h 指数提出这么多年,但是仍然不能取代被引次数在作者影响力评价中的独特地位。下面就本文所提方法结果和被引次数进行比较,表 1 给出的是前 50 名作者的影响力与作者被引次数的具体分布情况。

表 1 前 50 名作者的影响力和作者被引次数分布情况表

| 排名 | 作者 | 影响力得分 | 被引次数排名 | 排名 | 作者 | 影响力得分 | 被引次数排名 |
|---|---|---|---|---|---|---|---|
| 1 | 邱均平 | 187.99 | 1 | 26 | 肖珑 | 39.84 | 49 |
| 2 | 范并思 | 170.73 | 2 | 27 | 刘国钧 | 37.19 | 24 |
| 3 | 程焕文 | 142.32 | 3 | 28 | 刘军 | 37.12 | 27 |
| 4 | 于良芝 | 122.27 | 6 | 29 | 陈峰 | 35.36 | 54 |
| 5 | 李国新 | 120.91 | 9 | 30 | 沈祖荣 | 33.04 | 28 |
| 6 | 张晓林 | 111.37 | 5 | 31 | 王余光 | 32.16 | 29 |
| 7 | 蒋永福 | 100.91 | 4 | 32 | 刘兹恒 | 31.18 | 26 |
| 8 | 马费成 | 94.96 | 7 | 33 | 叶继元 | 30.30 | 34 |
| 9 | 王知津 | 87.94 | 8 | 34 | 程亚男 | 30.24 | 31 |
| 10 | 邱冠华 | 76.69 | 25 | 35 | 任树怀 | 30.02 | 30 |
| 11 | 黄宗忠 | 75.99 | 10 | 36 | 赖茂生 | 29.82 | 56 |
| 12 | 吴慰慈 | 75.92 | 11 | 37 | 刘则渊 | 28.55 | 36 |
| 13 | 王子舟 | 74.35 | 12 | 38 | 胡小菁 | 28.30 | 45 |
| 14 | 吴建中 | 73.07 | 18 | 39 | 赵蓉英 | 27.01 | 38 |
| 15 | 柯平 | 67.69 | 13 | 40 | 庞景安 | 26.86 | 43 |
| 16 | 初景利 | 64.68 | 14 | 41 | 来新夏 | 26.57 | 37 |
| 17 | 肖希明 | 64.08 | 17 | 42 | 孟广均 | 26.32 | 64 |
| 18 | 包昌火 | 62.98 | 15 | 43 | 杜定友 | 25.71 | 35 |
| 19 | 陈传夫 | 53.00 | 16 | 44 | 盛小平 | 25.39 | 40 |
| 20 | 李东来 | 47.80 | 39 | 45 | 黄晓斌 | 24.23 | 50 |
| 21 | 胡昌平 | 46.55 | 20 | 46 | 永瑢 | 24.19 | 32 |
| 22 | 苏新宁 | 46.20 | 19 | 47 | 谢灼华 | 23.82 | 41 |
| 23 | 刘炜 | 45.98 | 22 | 48 | 吴稌年 | 23.75 | 36 |
| 24 | 冯惠玲 | 43.93 | 21 | 49 | 万锦堃 | 23.71 | 51 |
| 25 | 王世伟 | 42.75 | 23 | 50 | 叶鹰 | 23.45 | 57 |

对作者影响力排名和作者被引次数排名做 Spearman 等级相关分析,发现这两个指标之间的相关系数为 0.896,可以看出运用新算法所得的作者影响力评价结果的排名与作者被引次数的排名呈现非常明显的正相关性。从表 1 中也可以看出运用新算法评价出的作者影响力排名中前 3 名作者的被引次数排名完全相同,其余大部分作者运用新算法所得的排名和被引次数排名的差距也基本在 ±5 名内。由此可见,本文提出的方法所得结果和被引次数很相关,是被引次数的有益补充。

另外,从原理看,改进的特征因子算法不仅考虑了作者的绝对被引次数,而且将施引作者的影响力也考

虑进去,使对作者影响力的评价更有说服力,这在对图情学学者的评价中也有所体现。如邱冠华、赖茂生、孟广均等作者,他们都是该学科领域内的精英或者某个方面的带头人,通过查阅原始数据发现他们的被引次数相对不是很高,排名稍靠后,但是施引作者的影响力都相对较强,那么这些作者的影响力排名靠前是可以解释通的。从这点出发,本文提出的方法在原理上有一定的优越性,得到的结果也符合实际。

## 四、结语

作者影响力评价是当前科学计量学研究的热点,不同学者提出了不同的解决方案。本文受特征因子这一期刊质量评价方法的启迪,对其进行了改进,使其更符合作者引用网络的特征。文章详细介绍了实现新算法的步骤,并将该方法应用在中国图情学学者影响力评价上,发现该方法能够较好地区分作者的影响力,其分布符合长尾分布特征。与作者引用次数——一种经典的作者影响力评价指标相比,本文提出的新方法不仅原理上较为科学,而且在结果呈现上和作者被引次数所得排序高度等级相关,但两者也有一定差别。通过实例分析可证实本方法较符合实际,切实可行,能够成为作者影响力评价方法的有益补充。

**参考文献:**

[1]邱均平. 信息计量学[M]. 武汉:武汉大学出版社,2007.
[2]HIRSCH J E. An index to quantify an individual's scientific output[J]. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102 (46):16569 - 16572.
[3]YING D. Applying weighted rage-rank to author citation networks[J]. Journal of the American Society for Information Science and Technology, 2011, 62(2):236 - 245.
[4]BERGSTROM C T, WEST J D, et al. The eigen-factor metrics[J]. The Journal of Neuroscience, 2008, 28(45):11433 - 11434.
[5]JEVIN D W. Author-level eigen-factor metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community [J]. Journal of the American Society for Information Science and Technology, 2013 (4):787 - 801.
[6]刘艳华,华薇娜. 期刊评价新指标——特征因子[J]. 情报杂志,2010(7):122 - 126.
[7]米佳,濮德敏. 特征因子原理及实证研究[J]. 大学图书馆学报,2009(6):63 - 68.
[8]任胜利. 特征因子( Eigen-factor):基于引证网络分析期刊和论文的重要性[J]. 中国科技期刊研究,2009,20(3):415 - 418.

# An Evaluation Research of Author Influence Based on the Improvement of the Eigen-factor Algorithm

MA Ruimin[a,b], HAN Xiaolin[a,b]

(*a. Institute of Management and Decision-making; b. Center for Science Evaluation, Shanxi University, Taiyuan 030006, P. R. China*)

**Abstract:** The Eigen-factor algorithm is an important method for journal quality evaluation. This paper constructs a new algorithm to evaluate the author influence based on the improvement of the Eigen-factor algorithm. This paper first introduces the basic principle of the Eigen-factor algorithm, and then improves the Eigen-factor algorithm by analyzing the particularity of author citation compared with journal citation. Next it introduces the basic steps of the new algorithm. Finally it makes an empirical research based on author citation network of library and information science in China and the ranking of the authors is obtained. At the same time, the ranking result is compared with traditional citation counts.

**Key words:** author influence; eigen-factor algorithm; improvement

(责任编辑　傅旭东)