基于 PageRank 的机构科研影响力评价*

李 勇 安新颖 赵迎光 范少萍

（中国医学科学院医学信息研究所 北京 100020）

〔摘要〕 提出基于 PageRank 算法的机构科研影响力评价方法，针对 PageRank 经典算法中存在的平分网页权值问题，应用篇均被引频次来配置利用各机构的初始权重。应用 2015 年部分高校发表的医学类 SCI 论文的被引用情况构建机构引用网络矩阵，分别计算去除自引和包含自引的机构 PageRank 值，与机构被引频次及 *H* 指数的排序进行对比分析。该方法可同时从质量和数量两个维度评价机构科研影响力。

〔关键词〕 机构评价；PageRank；引文网络

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2017.06.012

Evaluation on the Scientific and Research Influence of Institution Based on PageRank *LI Yong, AN Xin－ying, ZHAO Ying－guang, FAN Shao－ping, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China*

〔**Abstract**〕 The paper puts forward the method for evaluation on the scientific influence of institution based on PageRank algorithm, configures the initial weight of each institution based on the problems of dividing equally the webpage weight number in the classical PageRank algorithm and cited frequency of the application chapter, constructs the network matrix cited by institution through the citation situation of medical SCI papers published by partial colleges and universities in 2015, respectively calculates the PageRank values of institution when self－citation is excluded or included, and conducts comparative analysis on the sequence of cited frequency indicators of the institution as well as *H* index. This method can be used to evaluate scientific influence of institution from two dimensions of quality and quantity.

〔**Keywords**〕 Institution evaluation; PageRank; Citation network

1 引言

机构科研影响力是指某一机构产出的科研成果对其他机构造成的影响，科研影响力分析对科研机构的战略定位、学科规划、资源流量、科技评价和人才引进等方面具有重要的意义。当前的科研机构影响力评价研究与实践常采用文献计算学方法，其中，SCI 发文量与被引频次是应用最为广泛的 2 个指标，基于该指标提出了众多衍生指标，如机构 *H* 指数[1]、H_m 指数[2]等。但是 SCI 发文量仅计算论文的数量，无法体现论文的质量；机构的被引频次虽然一定程度上通过被引用次数表现了机构的影响力，但该指标的基本原理认为所有被引用的情况都是等同的，但在实际中，被一篇重要的论文引用与被一篇相对不重要的论文引用相比，对论文产生的影响力是不同的。*H* 指数存在着对评价对象的个别

〔收稿日期〕 2017－01－11

〔作者简介〕 李勇，博士，助理研究员，发表论文 10 余篇。

〔基金项目〕 中央级公益性科研院所基本科研业务费专项“面向医学科技评价的多源异构数据处理机制研究”（课题编号：2016ZX330027）。

论文并不敏感、只能正向增长、仅适用于规模相当的评价对象等问题。而近年来提出的 Altmetrics[3]，其实质相对于引用指标，是补充而不是替代[4]。因此，这些文献计量指标均不能较全面地既通过数量又能通过质量的计算来评价机构的科研影响力表现。

伴随着互联网的发展，网络信息呈几何级指数增长，面对海量信息，用户在查找有用信息时所付出的时间成本也越来越高，传统搜索引擎是基于关键词匹配的，查询效果不理想。1998 年斯坦福大学的 Brin 和 Page 借鉴文献计量学中经典的引文分析思想，提出了 PageRank 算法，通过分析网页间的链接结构以获得网页排序权重，其基本思想为：将所有网页以及各网页直接的关联视为一个有向图。在这个网络中，每个网页被视为一个节点，网页间的联系由其所链出的网页之间的链接关系来定义。节点重要性由链接该节点的其他节点的重要性和数量决定，即网页的重要性由其被链接的网页的重要性和数量决定。自 Pagerank 算法提出以来，在搜索引擎、文献数据库、期刊评价[5]等领域得到了广泛的应用。机构之间的论文引用网络与网页链接网络的基本原理是相同的，均为有向图。在机构之间的论文引用有向图中，一个节点代表一个机构，节点间的连线代表机构发表论文间的引用关系。因此，PageRank 可以应用于科研机构的科研影响力评价，可同时反映评价对象的数量与质量表现，相比仅考虑被引频次或发文量更为全面。鉴于此，本文拟基于 SCI 论文的引文网络，应用 PageRank 算法来计算中国医学高等院校的科研影响力。

2 评价对象与数据来源

根据教育部公布的《全国普通高等学校名单》[6]，选取 29 所具有医学类专业的 211 高校作为评价对象。由于各机构 2016 年所发表论文的被引用情况尚不充分，无法充分反映在被引频次和 H 指数方面的表现，同时为便于与 PageRank 指数进行对比，本文在 SCIE 数据库中选取 2015 年各高校发表的 SCI 论文，下载其施引文献题录信息，构建机构 SCI 相互引用网络。各机构发表 SCI 论文的检索方式为：按 WOS 研究方向，选择生命科学与生物医学（Life Sciences & Biomedicine）下与医学相关的 57 个类，如过敏症（Allergy）、血液学（Hematology）、解剖与形态学（Anatomy & Morphology）、免疫学（Immunology）、麻醉学（Anesthesiology）、传染病（Infectious Diseases）、整合与补充医学（Integrative & Complementary Medicine）、法医学（Legal Medicine）、病理学（Pathology）、小儿科学（Pediatrics）等，文献类型为 Article 和 Review，共计 37 597 篇。

3 方法设计

3.1 Pagerank 算法的基本原理

PageRank 算法建立在随机冲浪者模型上，假设网络用户跟随网页间的链接进行了若干步的浏览后转向一个随机的起点网页，再重新开始跟随网页链接进行浏览，一个网页的价值程度由该网页被随机访问的概率决定。PageRank 算法的计算公式[7]为：

$$R(u) = \lambda \sum_{i=1}^{N} \frac{R(v_i)}{C(v_i)} \quad (1)$$

公式（1）中，u 代表一个网页，代表指向 u 的网页，是网页的向外链出的网页的总链接数。λ 为规范化因子，取值范围为 0 - 1，一般取 0.85。

该算法为网页排序提供了一种全新的思路，但却存在 PageRank 值沉淀现象，即网络链接结构是无序的，存在一组网页之间彼此链接，但没有组外网页的链出，造成 PageRank 值的组内沉淀。为了解决这一问题，引进衰退因子 E（u）以补充网页的 PageRank 值，算法改进为：

$$R(u) = c \sum_{i=1}^{N} \frac{R(v_i)}{C(v_i)} + cE(u), \quad ||R(u)|| = 1 \quad (2)$$

随后，Sergey Brin 和 Lawrence Page 又将公式修正为：

$$R(u) = (1-d) + d \sum_{i=1}^{N} \frac{R(v_i)}{C(v_i)} \quad (3)$$

其中 d（$0 < d < 1$）为衰减系数，可理解为网络用户在当前网页感到厌烦的程度，如取值为 0 时，

一直停留在当前网页。

PageRank 算法提出后被广泛关注，但其仍然存在着许多问题并被持续改进：（1）主题漂移现象，仅利用网络的链接结构，无法判断网页内容的相似性，即无论主题是否相关，被检索网页的重要性相同。为此，Haveliwala 提出了基于网页分类的思想来解决该问题[8]，Richardson 提出了结合链接和内容信息的方法[9]。（2）偏重旧网页问题，网页存在的时间越长，获得的链接数量越多，搜索排序的名次越靠前。为此王德广[10]引入了一个与网页权值呈反比的时间权值函数来解决时效性问题。（3）平分网页权值：即一个网页被一个权威网页引用和被一个普通网页所引用，其意义与价值是不同的。为此，田甜等[11]提出了一种权威值不均衡分配的方法（IPR）。（4）忽视用户浏览兴趣，该算法被设计时，未考虑用户浏览偏好和行为。为此，王小玲等[12]提出了基于个人兴趣和反馈技术的 PageRank 算法研究。

3.2 Pagerank 机构影响力评价

在 PageRank 算法中，节点的重要性可通过递归的方法计算得出，如果一个网页被其他重要网页引用的次数更多，则该网页更重要。基于以上思路，本文提出基于 PageRank 的机构科研影响力评价方法，将所有机构以及各机构之间通过 SCI 论文引用关系构建的关联视为一个无标度网络，在这个网络中，每个机构被视为一个节点，机构间的联系由其所发表的 SCI 之间的引用关系来定义，机构的重要性由引用该机构发表论文的其他机构的重要性和数量决定。这种重要性可通过递归的方法计算得出，如果一个机构被其他重要机构引用的次数更多，则该机构更重要，即科研影响力表现更佳。用公式表示为：

$$R\ (u)\ =\ (1-\lambda)\ +\lambda\sum_{i=1}^{N}\frac{R\ (v_i)}{C\ (v_i)} \qquad (4)$$

式中 $R\ (u)$ 代表机构 u 的重要性；v_i 代表引用机构 u 所发表论文的机构；$C\ (v_i)$ 代表机构 v_i 所发表论文的所有引用论文的机构数量；λ 为规范化因子，取值范围为 0 ~ 1，在本文中取 0.85。基本思路和权重传递示例，见图 1。

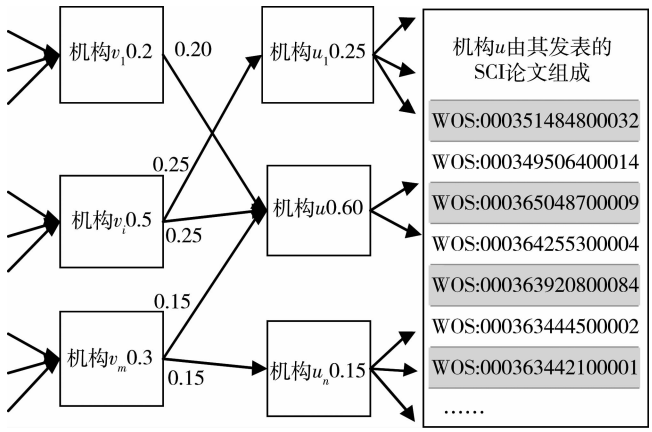

图 1 基于 PageRank 的机构评价

为了解决 PageRank 算法中的平分网页权值问题，本文将初始权重的分配方法改进为：

$$R\ (u)_0 = \frac{1}{N}\cdot\frac{C_u}{T_u} \qquad (5)$$

式中：$R\ (u)_0$ 表示机构 u 的初始权重；N 为机构总数；T_u 表示机构 u 的 SCI 发文总数；C_u 表示机构 u 的总被引频次。即用每个机构的篇均被引频次去干预其初始权重。

此外，网页不存在自我链接情况，而在科技文献中，自引是一种常见的现象。因此，本文在计算中，同时考察计算自引和排除自引两种情况。

3.3 计算步骤

3.3.1 构建机构引用网络矩阵 基于本文选取的 29 所具有医学类专业的 211 高校所发表的 SCI 论文，按研究方向筛选后，下载其施引文献，在施引文献集中识别高校间的相互引用情况，构建高校 SCI 相互引用网络，见图 2。

University	PEKING U	BEIJING U	SECOND MIL MED U	FOURTH MIL MED U	SOUTHEAST U	FUDAN U	HUAZHONG SCI TECH U	JILIN U	JINAN U	LANZHOU U	NANCHANG U
PEKING UNIV	1	1	1	1	1	1	1	1	1	1	1
BEIJING UNIV CHINESE	1	1	1	1	1	1	1	1	1	1	1
SECOND MIL MED UNIV	1	1	1	1	1	1	1	1	1	1	1
FOURTH MIL MED UNIV	1	1	1	1	1	1	1	1	1	1	1
SOUTHEAST UNIV	1	1	1	1	1	1	1	1	1		1
FUDAN UNIV	1	1	1	1	1	1	1	1	1	1	1
HUAZHONG UNIV SCI TECH	1	1	1	1	1	1	1	1	1	1	1
JILIN UNIV	1	1	1	1	1	1	1	1	1	1	1
JINAN UNIV	1	1	1	1	1	1	1	1	1	1	1
LANZHOU UNIV	1		1	1	1	1	1	1	1	1	1
NANCHANG UNIV	1	1	1	1	1	1	1	1	1	1	1
NANJING UNIV	1	1	1	1	1	1	1	1	1	1	1
NANKAI UNIV	1	1	1	1	1	1	1	1	1	1	1
TSINGHUA UNIV	1	1	1	1	1	1	1	1	1	1	1
XIAMEN UNIV	1	1	1	1	1	1	1	1	1		1
SHANDONG UNIV	1	1	1	1	1	1	1	1	1	1	1
SHANGHAI JIAO TONG U	1	1	1	1	1	1	1	1	1	1	1
SICHUAN UNIV	1	1	1	1	1	1	1	1	1	1	1
SOOCHOW UNIV	1	1	1	1	1	1	1	1	1	1	1
TIANJIN MED UNIV	1		1	1	1	1	1	1	1		1
TONGJI UNIV	1	1	1	1	1	1	1	1	1	1	1

图 2 机构 SCI 论文相互引用网络（部分）

3.3.2 配置初始权重 应用式（5）配置各机构

的初始权重。N 为 29，T_u 和 C_u 的分别从其发表 SCI 论文和被引论文中统计得到。

3.3.3 计算机构 PageRank 值 应用式（4）计算各机构的 PageRank 值，规范化因子 λ 取值为 0.85。分别考虑计算自引和排除自引两种情况。其中考虑排除自引的方法可描述为：

$\forall u \epsilon U_N$: $R(u)_0 = \frac{1}{N} \cdot \frac{C_u}{T_u}$; #初始化每个机构的 PageRank 值，为待评价机构的集合；

While ($|R(u)_i - R(u)_{i-1}| > \delta$); #PageRank 的迭代计算的收敛判断条件，取 $\delta = 000\ 01$；

for: $\forall u \epsilon U$ *and* $u \neq v$; #排除自引情况；

$R'(u)_m = R'(u)_{m-1} + (1-\lambda) + \lambda \sum_{i=1}^{N} \frac{R'(v_i)}{C(v_i)}$; #迭代计算机构 u 的 PageRank 值。

计算自引的方法可描述为：

$\forall u \epsilon U_N$: $R(U)_0 = \frac{1}{N} \cdot \frac{C_u}{T_u}$; #初始化各机构的 PageRank 值，$U_N$ 为待评价机构的集合；

While ($|R(u)_i - R(u)_{i-1}| > \delta$); #PageRank 的迭代计算的收敛判断条件，取 $\delta = 000\ 01$；

for: $\forall u \epsilon U$:

$R'_{sc}(u)_m = R'_{sc}(u)_{m-1} + (1-\lambda) + \lambda \sum_{i=1}^{N} \frac{R'(v_i)}{C(v_i)}$; #迭代计算机构 u 的 PageRank 值。

4 结果与讨论

对比分析 29 所具有医学类专业的 211 高校的 H 指数、SCI 被引频次和 PageRank 值（去除自引或包含自引），见表 1、表 2（仅显示 PageRank 值前 10 的机构）。其中 H 指数通过各机构在本文所选的 57 个 WOS 研究方向中 2015 年所发表论文的被引用情况计算得出。

表 1 去除自引的机构 SCI 发文数、SCI 被引频次、H 指数、PageRank 值对比

序号	大学	去除自引的被引	发文量（篇）	H 指数	去除自引的 PageRank
1	复旦大学	9 196	3 192	26	0.039 16
2	浙江大学	7 716	2 853	23	0.039 11
3	上海交通大学	10 788	4 089	25	0.039 08
4	南京大学	4 023	1 179	20	0.039 07
5	华中科技大学	4 601	1 815	20	0.039 06
6	中南大学	5 268	1 824	19	0.039 05
7	中山大学	8 564	2 925	23	0.038 99
8	四川大学	5 038	2 272	21	0.038 94
9	第二军医大学	3 148	1 295	16	0.038 88
10	暨南大学	1 254	655	12	0.038 87

表 2 包含自引的机构 SCI 发文数、SCI 被引频次、H 指数、PageRank 值对比

序号	大学	包含自引的被引	发文量（篇）	H 指数	包含自引的 PageRank
1	复旦大学	10 093	3 192	26	0.038 67
2	浙江大学	8 503	2 853	23	0.038 58
3	中山大学	8 695	2 925	23	0.038 58
4	上海交通大学	11 785	4 089	25	0.038 54
5	南京大学	4 354	1 179	20	0.038 53
6	中南大学	5 770	1 824	19	0.038 51
7	华中科技大学	5 120	1 815	20	0.038 51
8	第二军医大学	3 391	1 295	16	0.038 42
9	四川大学	5 532	2 272	21	0.038 34
10	西安交通大学	2 860	1 211	15	0.038 31

从表 1 和表 2 中均可看出，无论是否考虑自引，基于 PageRank 算法得出的排名均与被引或发文量的排名有较大差异。按机构总发文量的排名结果前 10 为：上海交通大学、复旦大学、中山大学、浙江大学、北京大学、山东大学、四川大学、中南大学、华中科技大学和吉林大学，其中有 7 个跟 PageRank 前 10 是重复的，在包含自引的情况下，也有 7 个跟 PageRank 前 10 重复。在去除自引的情况下，按被引频次排序的前 10 为：上海交通大学、复旦大学、中山大学、浙江大学、北京大学、中南大学、四川大学、山东大学、华中科技大学和南京大学，其中有 8 个跟 PageRank 前 10 是重复的。在计算自引的情况下，也有 8 个结果与 PageRank 前 10 是重复的。排名出现差异的原因为机构发文总量和 SCI 被引频

次仅考虑了数量，没有考虑论文质量，而 PageRank 可同时反映评价对象的数量与质量表现，比仅考虑被引频次或发文量更为全面。*H* 指数的排名前 10 的机构分别为：复旦大学、上海交通大学、北京大学、中山大学、浙江大学、四川大学、同济大学、华中科技大学、南京大学和中南大学。与去除自引和包含自引的 PageRank 排名前 10 相比均有 8 个重复，但排序各不相同。*H* 指数虽然可兼顾数量和质量，但其对个别高被引论文不敏感，例如机构 A 和机构 B 的 *H* 指数均为 15，但 A、B 这 15 篇论文的被引频次分布规律是不同的，被其他机构的引用情况也各不相同，因此应用 PageRank 计算可得到不同的排名。此外，*H* 指数中存在的对评价对象的个别论文并不敏感、只能正向增长的缺点在 PageRank 算法中是不存在的，因为其基本原理是机构间的论文引用链接，只要存在引用频次高的论文，那么在 PageRank 算法的迭代计算中就会体现出来；与此同理，如果一个机构的高被引论文随时间增长而减少，*H* 指数不会减少，但其 PageRank 值是一定会发生变化的。在包含自引与去除自引的前 10 排名中，有 9 个是重复的，但是 7 个机构的排名顺序出现了变化。考虑到以上各机构的平均自引率为 0.09，可认为自引对 Pagerank 值的排名会产生一定影响，但这种影响程度还需要更大样本数据量的计算来评估。

5 结语

本文通过试验证明，作为一种非常流行的网页链接分析算法，PageRank 也可以应用于机构科研影响力评价的研究与实践中。该算法可以为现有基于文献计量的机构科研影响力评价提供一种新的指标，PageRank 不仅考虑了评价对象的数量、还反映了评价对象的质量，相对于机构的 SCI 发文总数、被引频次、*H* 指数等，该方法提供了一种不同的分析角度。在本文中仅选取小样本的数据进行实证分析，为了得到更理想的分析结果，未来将针对更明确的研究方向、面向更多的机构进行 PageRank 分析。此外，本文仅针对 PageRank 算法中的平分初始权值问题进行了改进，未来还将对该算法存在的偏重旧节点、主题漂移现象等问题进行改进。

参考文献

1 Mitra P. Hirsch – type Indices for Ranking Institutions Scientific Research Output [J]. Current Science, 2006, (91): 1439.

2 Molinari J F, Molinari A. A New Methodology for Ranking Scientific Institution [J]. Scientometrics, 2008, 75 (1): 163 – 174.

3 Piwowar H. Altmetrics: value all research products [J]. Nature, 2013, (493): 159.

4 杨柳，陈贡. Altmetrics 视角下科研机构影响力评价指标的相关性研究 [J]. 图书情报工作，2015，59 (8)：106 – 114.

5 苏成，潘云涛，袁军鹏，等. 基于 PageRank 的期刊评价研究 [J]. 中国科技期刊研究，2009，20 (4)：614 – 617.

6 教育部. 2015 年全国高等学校名单 [EB/OL]. [2016 – 12 – 04]. http://www.moe.edu.cn/publicfiles/business/htmlfiles/moe/moe_229/201505/187754.html.

7 Brin S, Page L. The Anatomy of a Large – scale Hypertextual Web Search Engine [J]. Computer Networks and ISDN Systems, 1998, 30 (1 – 7): 107 – 117.

8 Haveliwala T H. Topic – sensitive PageRank [C]. Hoho Lulu Hawaii: Proceedings of the Eleventh International World Wide Web Conference, 2002.

9 Richardson M, Domingos P. The Intelligent Surfer: probabilistic combination of link and content information [J]. PageRank Advances in Neural Information Processing Systems, 2002, (14): 673 – 680.

10 王德广，周志刚，梁旭. PageRank 算法的分析及其改进 [J]. 计算机工程，2010，36 (11)：291 – 293.

11 田甜，倪林. 基于 PageRank 算法的权威值不均衡分配问题 [J]. 计算机工程，2007，33 (9)：53 – 55.

12 王小玲，胡平. 基于个人兴趣和反馈技术的 PageRank 算法研究 [M]. 合肥：合肥工业大学出版社，2006.