基于引文分析的科学家投入产出绩效算法研究

郭 强[1]，陈清文[1]，刘建国[2*]

(1. 上海理工大学复杂系统科学研究中心 上海 杨浦区 200093；2. 上海财经大学会计学院 上海 杨浦区 200433)

【摘要】该文提出了一种考虑投入和产出的科学家绩效算法。考虑到科学家的沟通、时间等投入成本，该算法以科学论文中目标科学家的合作作者数和机构数作为输入变量，以合作发表的文章及其被引数作为输出变量。基于输入和输出数据，建立科学家投入产出绩效评价模型。在实证数据上的实验结果显示，相对于发表文章数、总引用量、I10 指数和 H 指数等指标，该方法可以更准确地识别出获诺贝尔奖的科学家，算法的 AUC 值为 0.795 7，比总引用量指标的准确度提高了 8.77%。此外还发现大部分科学家获奖前的投入产出绩效高于获奖后科学家的投入产出绩效。该工作对科学地评价科学家的绩效具有重要意义。

关 键 词 引文分析； H 指数； 投入产出绩效； 科研合作； 总引用量

中图分类号 N949 **文献标志码** A doi:10.12178/1001-0548.2018236

Modeling of Input-output Performance of Scientists Based on the Analysis of Citation

GUO Qiang[1], CHEN Qing-wen[1], and LIU Jian-guo[2*]

(1. Complex Systems Science Research Center, University of Shanghai for Science and Technology Yangpu Shanghai 200093;
2. School of Accountancy, Shanghai University of Finance and Economics Yangpu Shanghai 200433)

Abstract This paper presents a model to evaluate input-output performance of scientists. With consideration of the input cost of scientists' communication and time, this model takes the number of co-authors and the number of institutions of target scientists in scientific papers as input variables, and the number of co-published articles and their cited number as output variables. The experiments results show the scientists who won Nobel Price are ranked higher than the sciences who did not win Nobel Price. The experimental results also show that the AUC values of input-output performance model could reach 0.795 7 for the APS data set, which is better than the results generated by h-index, i10-index, total number of papers, and total number of citations. Furthermore, The experimental results indicate that most input-output performances of scientists before winning award is higher than the input-output performances of scientists after winning award for the APS data set and the web of science data set. The proposed model also provides an effective tool for policy makers to quantify the input-output performances of sciences.

Key words citation analysis; H-index; input-output performance; scientific cooperation; total number of citations

引文网络的建模与分析已经被广泛用于评价科学家、科研单位甚至地区或国家的学术影响力。论文的应用次数对科学家、科学家的职称评定、科研奖励等方面都具有重要意义[1-3]。引文网络的分析结果已经被应用于科研管理政策的制定、科研激励等措施，对学科发展具有重要意义[4]。

当前，基于科研引文网络分析方法主要归为两类：基于统计和基于网络结构的评价方法。基于统计的评价方法包括基本科学指标数据库 (ESI)[5-7]、总引用次数、总论文发表数、H 指数[8]、G 指数[9]、I10 指数[10] 等指标。2001 年，美国科技信息所 (ISI) 提出 ESI 指标用来度量科学研究绩效[5-6]。ESI 是从论文发表总数、引文次数、平均被引频次等多个方面对国家/地区科研水平、机构学术声誉以及期刊

收稿日期：2018－09－06；修回日期：2019－10－09
基金项目：国家自然科学基金 (61773248, 71771132)；国家社科重大项目 (18ZDA088, 20ZDA060)
作者简介：郭强 (1975－)，女，教授，主要从事知识图谱、知识管理方面的研究.
通信作者：刘建国，E-mail：liujg004@ustc.edu.cn

学术水平进行衡量。但是 ESI 只考虑编入 Thomson Reuters 索引的期刊中发表的书籍、文献[7]，限制了其客观性。文献 [8] 提出既考虑引用量又考虑发文章数的 H 指数。一个人或组织的 H 指数定义为其发表的所有文章中被引次数大于等于 H 次的论文超过 H 篇。一名科学家的 H 指数越高，他的论文影响力越大。但是 H 指数无法对只发表了少数几篇重要文献的科学家的工作进行评价。文献 [9] 在 H 指数上做出改进，提出了 G 指数。G 指数是一种基于学者以往贡献的科学家影响力评估方法。此外，一些用来完善或优化 H 指数的指标也相继被提出。2011 年，谷歌提出了 I10 指数[10]，即科学家发表文章中被引次数大于等于 10 次的文章数。基于网络结构的评价方法包括基于科学家合作[11-12] 和引用网络的 PageRank 算法[13-14]。基于合作网络的 PageRank 算法是指基于合作网络中科学家之间的合作关系对科学家进行评价，该方法主要反映了科学家在合作网络中的影响力。基于引用网络的 PageRank 算法则是基于文献之间的引用关系和科学家之间的引用网络对科学家的学术水平进行评估。但是，上述全部方法都只考虑了科学家发表文章数、文章引用量，没有考虑到科学家的沟通、时间等投入成本。因此，本文提出一种考虑输入和输出变量的投入产出模型，对科学家的绩效进行综合评价。

假设有甲乙两位科学家，科学家甲与多名科学家合作发表了一篇文章，而科学家乙与一名科学家合作也发表了一篇文章，同时他们文章的引用量也相同。用 H 指数等指标计量甲乙两名科学家的投入产出绩效是相同的。但是，甲比乙投入的多，占用的社会资源更多。如果乙和甲拥有相同的社会资源，乙就可能有更多的产出。综合考虑科学家的投入和产出要素，本文工作主要是提出了一种考虑投入和产出的科学家绩效算法。算法在考虑科学家的科研产出的同时，也考虑了科学家的沟通、时间等投入成本，从投入和产出的视角对科学家的绩效进行建模评价。在 APS 实证数据集上的实验结果表明，本文提出的方法可以更准确地识别出获诺贝尔奖的科学家，其中本文算法的 AUC 值为 0.795 7，比只考虑总引用量的评价方法的准确度提高了 8.77%。此外，对于 APS 数据集，64.29% 的科学家获得诺奖前的投入产出绩效高于获得诺奖后的投入产出绩效。对于 Web of science 数据集，81.25% 的科学家获得杰青前的投入产出绩效高于获得杰青后的投入产出绩效。

1　科学家投入产出绩效算法

1.1　科学家投入产出绩效算法的建立

合理的投入能够最大限度地增加文章的发表数和文章影响力，因此科学家的投入产出绩效算法应该满足两个要求：科学家产出最大化和科学家投入最小化。其中，$J=\{1,2,\cdots,n\}$表示科学家的集合，$I=\{1,2,\cdots,s\}$表示投入指标的集合，$R=\{1,2,\cdots,t\}$表示产出指标的集合，$X_j=\{x_{1j},x_{2j},\cdots,x_{sj}\}$表示科学家 j 的投入要素，$Y_j=\{y_{1j},y_{2j},\cdots,y_{tj}\}$表示科学家 j 的产出要素，v_i为 i 个投入指标的权重，u_r为 r 个产出指标的权重，则第 j 个科学家的投入的综合值为$\sum_{i=1}^{s}v_ix_{ij}$，产出的综合值为$\sum_{r=1}^{t}u_ry_{rj}$，则科学家j的投入产出绩效为：

$$h_j=\sum_{r=1}^{t}u_ry_{rj}\Big/\sum_{i=1}^{s}v_ix_{ij} \tag{1}$$

本文限定科学家的投入产出绩效h_j不超过 1，即$\max h_j\leqslant 1$，这意味着，若第j位科学家$h_j=1$，则第j位科学家相对于其他科学家而言，他的投入产出绩效最高；若$h_j<1$，则说明第 j 位科学家相对于其他科学家而言，他的投入产出绩效有待提高。科学家 $j^*(j^*\in j$，且 j^*为 j 中任意一个科学家) 的投入产出绩效经 Charnes-Cooper 变换，可得[15]：

设$\sum_{i=1}^{s}v_ix_{ij^*}=\frac{1}{c}$，$\mu_r=cu_r$，$\omega_i=cv_i$，则：

$$\max\sum_{r=1}^{t}\mu_ry_{rj*}$$

$$\text{s.t.}\begin{cases}\sum_{i=1}^{s}\omega_ix_{ij}-\sum_{r=1}^{t}\mu_ry_{rj}\geqslant 0\\ \sum_{i=1}^{s}\omega_ix_{ij^*}=1\\ \omega_i\geqslant 0\quad i=1,2,\cdots,s\\ \mu_r\geqslant 0\quad r=1,2,\cdots,t\end{cases} \tag{2}$$

1.2　投入要素、产出要素的选取

当前，科研合作是科研人员进行科学研究的主要方式。科研合作伙伴之间技能互补、相互信任，有助于科学家双方科研事业长期可持续发展。其中，科研论文合作是科研合作的重要形式，论文的质量是度量科研产出的重要指标。已有的文献显示，论文作者越多，则论文被引用次数越多[16]。也

有学者发现一篇论文的署名机构越多，则论文被引用次数越高。因此，本文假定合作科学家数量和合作机构数量可以作为投入产出模型的输入变量[17-18]。

科学家间的合作能够促进科研产出[19-20]。图 1 给出了科学家发文量和平均被引用次数与合作科学家数量，以及合作机构数之间的关系。从中可以发现，合作科学家数量和机构数对于提高论文数量和平均被引次数具有促进作用。

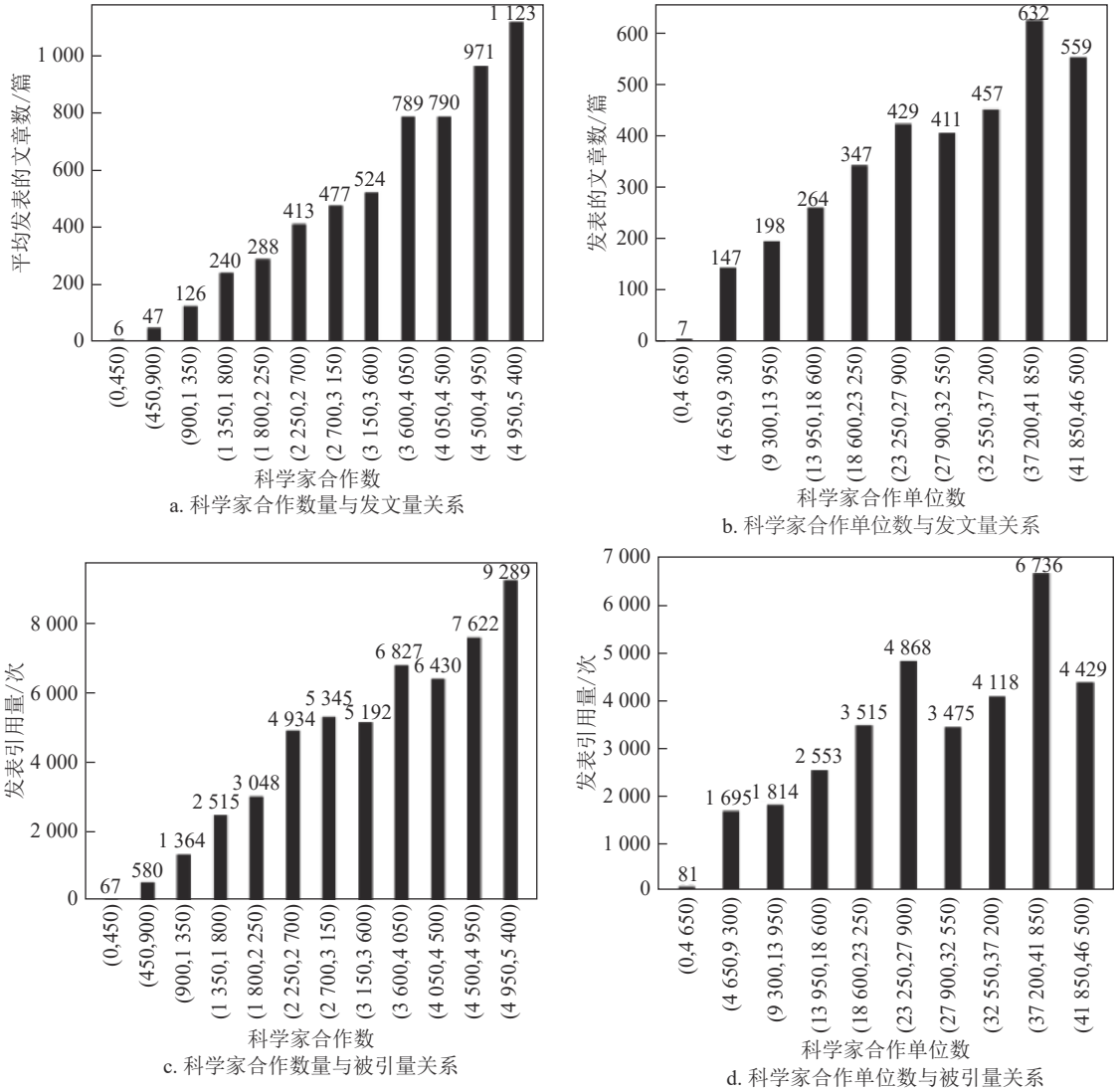


图 1 论文平均被引次数与合作科学家数和合作机构数的关联关系

1.3 科学家投入产出绩效算法的计算示例

假如科学家甲和乙都发表了 1 篇论文，其被引次数都为 0，科学家甲与 3 人合作，分别隶属于与科学家甲不同的 3 所科研机构，而科学家乙与 1 人合作，隶属于与科学家乙不同的 1 所科研机构，则甲、乙科学家的 H 指数、发文章量、引用量也一样。此时，如果不考虑科学家合作的科学家数量以及科学家合作的机构数量，则无法准确地判定出哪一位科学家的绩效更高。根据投入产出绩效算法可以计算得出：

$$\max \mu$$
$$\text{s.t.}\begin{cases} h_{甲}=\mu \leqslant 1 \\ h_{乙}=3\mu \leqslant 1 \\ \omega_1+\omega_2=1 \\ \omega_1 \geqslant 0, \omega_2 \geqslant 0, \mu \geqslant 0 \end{cases} \tag{3}$$

可以看到，$h_{甲}=0.333<h_{乙}=1$，虽然科学家甲和乙的 H 指数、发文章数、总被引量都是一样的，但是由于科学家乙合作的科学家数量和合作的机构数量少，因而拥有较高的投入产出绩效。而科学家甲合作的科学家数量和合作的机构数量多，所以影响了科学家甲的投入产出绩效。此外，还可以得出，

如果科学家甲和乙二人一起申报职称、基金，用 H 指数将难以做出取舍，而甲乙两位科学家的投入产出绩效各不相同，用投入产出绩效就可以解决问题。

2　数值实验

2.1　数据集

本文采用美国物理学会 (APS)1893～2009 年的数据。为了研究科学家的投入产出绩效，最终处理的 APS 数据集包含超过 247 889 名科学家 (包括 35 名获得诺贝尔物理学奖的科学家)、451 034 篇论文和 462 145 次引用。此外，本文采用了 Web of science 数据集包括 2011−2015 年国家杰出青年科学基金 (NSFDYS) 管理科学部的资助者在 Web of science 数据库发表的所有论文。数据集包含标题、出版年份、科学家名称、每位科学家的隶属机构以及每篇论文的引用次数。为了研究获奖者获奖前后科学家的投入产出表现，本文手动处理了科学家获奖前后论文的引文量，筛选出获奖前后都有数据的科学家为实验对象。Web of science 最终处理的数据集包含 32 位管理学科的获杰青的科学家、1 680 篇论文和 22 335 次引用，APS 的最终处理数据集包含 28 位获得诺奖的科学家、2 433 篇论文和 6 949 次引用。

2.2　实验结果

在 APS 数据集中，获诺贝尔奖的 35 名科学家占总科学家数的 1.4‰，本文分别计算诺贝尔奖科学家和非诺贝尔奖科学家的投入产出绩效，其中投入产出绩效值在 0～1 之间，1 代表科学家的投入产出绩效最高，0 代表科学家的投入产出绩效最低，结果分布如图 2 所示。在投入产出绩效为 0～0.2 时，非诺贝尔科学家的绩效累积分布的趋势急剧上升，而诺贝尔科学家的上升趋势比较平缓。总体上，在同一投入产出绩效下，非诺贝尔科学家的绩效累积分布比获诺贝尔科学家的累积分布高。

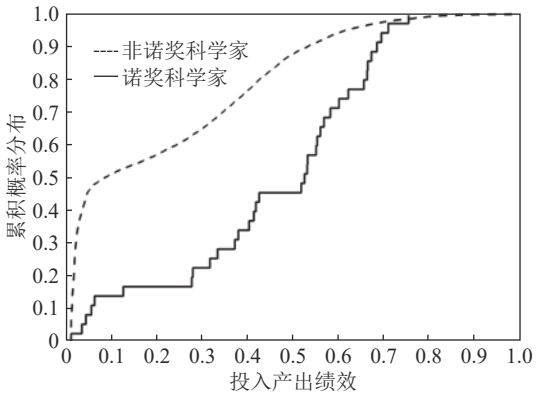


图 2　科学家的投入产出绩效累积分布图

为了直观看出本文提出的投入产出绩效算法的准确性[21]，图 3 给出了投入产出绩效算法与其他指标结果的对比图，子图展示了绩效排名前 1 000 名的科学家中获诺贝尔奖的科学家数分布状态。从中可以发现本文提出的投入产出绩效算法对科学家排名的准确性比其他指标高。

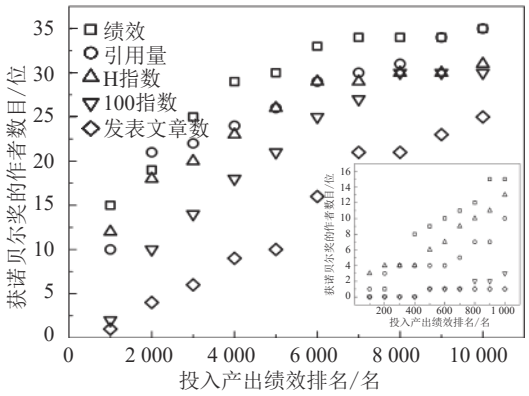


图 3　投入产出绩效算法与其他指标结果对比

本文采用 AUC 指标评价投入产出绩效算法的准确性。具体定义过程如下：分别从测试集合和非测试集合中随机选取一位科学家，比较其投入产出绩效。进行 n 次抽样后，如果测试集合中的科学家投入产出绩效高于非测试集合中的科学家绩效，则记为 n_1。如果两者相同，则记为 n_2，AUC 值定义为：

$$\mathrm{AUC}=\frac{n_1+0.5n_2}{n} \tag{4}$$

当 AUC＝1 时表示所有测试集中的科学家绩效均高于非测试集中的结果；AUC＝0.5 则表示结果与随机抽样的结果相同。抽样次数 n 越大，结果越可靠，本文取 $n=10^5$。表 1 给出了不同指标的 AUC 值，从中可以发现本文方法的结果为 0.795 7，比其他指标中最高的总引用量指标提高了 8.77%。

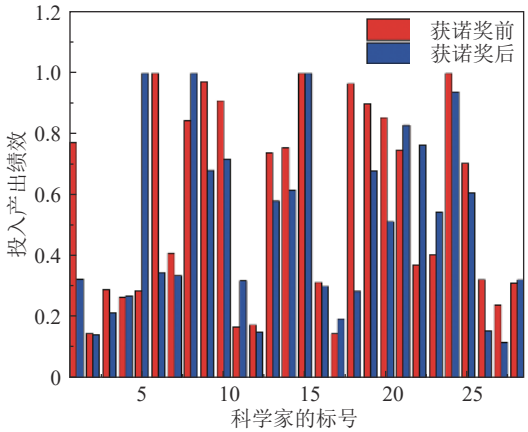
表 1　各指标的 AUC 值

	投入产出模型	引用量	H指数	I10指数	发表文章数
AUC	0.795 7	0.708 0	0.675 9	0.557 2	0.427 9

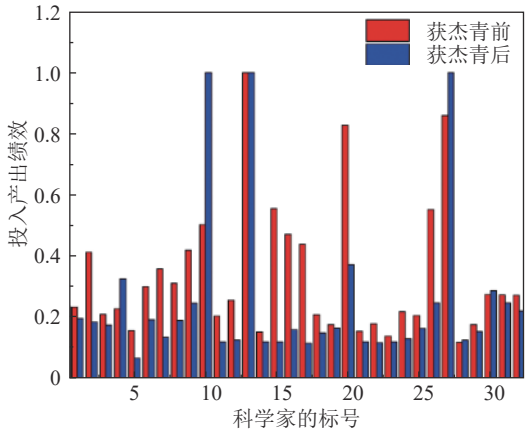
2.3　获奖前后的投入产出绩效

本文研究了杰出青年基金获得者和诺贝尔奖获得者两个数据集的科学家投入产出绩效：APS 数据集和 web of science 数据集。图 4a 是 28 位科学家获诺贝尔奖前后投入产出绩效柱状图。其中，红色代表科学家获得诺贝尔奖前的投入产出绩效，蓝色

代表获得科学家诺贝尔奖后的投入产出绩效。从图 4a 可以看出 18 位科学家的获奖前的投入产出绩效比获奖后的投入产出绩效高，1 位科学家的投入产出绩效不变。图 4b 的 2011−2015 年获得国家杰出青年科学基金的管理学部的 32 位科学家投入产出绩效柱状图。其中，红色代表获得杰青基金前的投入产出绩效，蓝色代表获得杰青基金后的投入产出绩效。从图 4b 可以看出 26 位科学家获奖前的投入产出绩效比获奖后的投入产出绩效高，1 位科学家的投入产出绩效不变。



a. 诺贝尔获得者前后绩效对比



b. 杰青基金获得绩效对比

图 4 科学家获奖前后投入产出绩效柱状图

3 结 束 语

本文提出了一种考虑科学家投入和产出信息的绩效评价算法。在评价科学家绩效的时候，除了要考虑科学家的发表论文和论文影响力等产出绩效，还需要考虑科学家的投入精力因素。如科学家需要花大量的时间进行沟通、协商才能够彼此合作。因此，本文考虑了合作科学家数和合作机构数等投入因素，对科学家的投入产出绩效进行综合评价。在包含近百年数据的美国物理学会上的实验结果表明，本文提出方法的 AUC 值为 0.795 7，相比于总引用量的评价结果，准确率提高了 8.77%。此外，科学家在获奖前后的投入产出绩效实验结果表明，大部分科学家获奖前的投入产出绩效高于获奖后科学家的投入产出绩效。

科学家投入产出绩效算法取决于投入要素和产出要素的选取，因此可以研究更多投入要素，使科学家的排名更准确。如科学家投入产出绩效在一定程度上取决于科学家研究的主题，而本文方法并没有考虑到研究主题这个投入变量。同时，具有意义的研究主题可能会有更多的产出(发表的论文数)，在未来的工作里会考虑加入研究主题来研究科学家的投入产出绩效[22-23]。除此之外，获奖科学家获奖前后绩效的差异的原因很多，如得奖的年龄很大，得奖后文章的价值还没有完全发挥出来等，而本文的方法中并没有考虑到这些影响因素。

参 考 文 献

[1] HICKS D, WOUTERS P, WALTMAN L, et al. The Leiden manifesto for research metrics[J]. Nature, 2015, 520(7548): 429.

[2] 刘浏, 王东波. 引用内容分析研究综述[J]. 情报学报, 2017, 36(6): 637-643.
LIU Liu, WANG Dong-bo. A review of citation content analysis research[J]. Journal of Information, 2017, 36(6): 637-643.

[3] 胡小军, 郭强, 杨凯, 等. 基于相对熵的多属性作者学术影响力排名研究[J]. 电子科技大学学报, 2018, 47(2): 281-285.
HU Xiao-jun, GUO Qiang, YANG Kai, et al. Multi-attribute researcher academic influence ranking based on relative entropy[J]. Journal of University of Electronic Science and Technology of China, 2018, 47(2): 281-285.

[4] VAN H B A, PHELPS J, BARNES M, et al. Evaluating scientific impact[J]. Environmental Health Perspectives, 2000, 108(9): A392.

[5] FITZPATRICK R B. Essential science indicators[J]. Medical Reference Services Quarterly, 2005, 24(4): 67.

[6] 曹志梅, 刘伟辉, 杨光. 高校 ESI 潜势学科排名提升策略探讨[J]. 情报探索, 2017(4): 44-47.
CAO Zhi-mei, LIU Wei-hui, YANG Guang. Discussion on the strategy of improving the ESI potential discipline in colleges and universities[J]. Information Research, 2017(4): 44-47.

[7] CSAJBOK E, BERHIDI A, VASAS L, et al. Hirsch-index for countries based on essential science indicators data[J]. Scientometrics, 2007, 73(1): 91-117.

[8] 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.
[9] EGGHE L. Theory and practise of the G-index[J]. Scientometrics, 2006, 69(1): 131-152.
[10] DELGADO L C E, ROBINSON G N, TORRES S D. The Google scholar experiment: How to index false papers and manipulate bibliometric indicators[J]. Journal of the Association for Information Science and Technology, 2014, 65(3): 446-454.
[11] BRIN S, PAGE L. The anatomy of a large-scale hypertexual web search engine[J]. Computer Networks and ISDN Systems, 1998, 30(1-7): 107-117.
[12] 王露, 郭强, 刘建国. 基于加权方法的节点重要性度量[J]. 计算机应用研究, 2018(5): 1426-1428.
WANG Lu, GUO Qiang, LIU Jian-guo. Node importance measure based on weighting method[J]. Journal of Computer Applications, 2018(5): 1426-1428.
[13] 顾亦然, 许梦馨. 基于 PageRank 的新闻关键词提取算法[J]. 电子科技大学学报, 2017, 46(5): 777-783.
GU Yi-ran, XU Meng-xin. News keyword extraction algorithm based on PageRank[J]. Journal of University of Electronic Science and Technology of China, 2017, 46(5): 777-783.
[14] 陈仕吉, 史丽文, 左文革. 科学合作网络中节点合作效果评测与分析[J]. 图书情报工作, 2012, 56(10): 61-143.
CHEN Shi-ji, SHI Li-wen, ZUO Wen-ge. Evaluation and analysis of node cooperation effect in scientific cooperation network[J]. Library and Information Service, 2012, 56(10): 61-143.
[15] CHARNES A, COOPER W W, RHODES E. Measuring the efficiency of decision making units[J]. European Journal of Operational Research, 1978, 2(6): 429-444.
[16] DE S P D J, BEAVER D. Collaboration in an invisible college[J]. American Psychologist, 1966, 21(11): 1011.
[17] 苏芳荔. 科研合作对期刊论文被引频次的影响[J]. 图书情报工作, 2011, 55(10): 144-148.
SU Fang-li. The influence of scientific research cooperation on the citation frequency of journal papers[J]. Library and Information Service, 2011, 55(10): 144-148.
[18] 何海燕, 李芳. 高校科研合作对论文产出质量的影响—基于国家重点实验室分析[J]. 北京理工大学学报 (社会科学版), 2017, 19(5): 162-167.
HE Hai-yan, LI Fang. The influence of scientific research cooperation on the output quality of papers-Based on the analysis of national key laboratories[J]. Journal of Beijing Institute of Technology (Social Science Edition), 2017, 19(5): 162-167.
[19] 王卫, 史锐涵, 潘京华. 基于期刊论文的作者学术合作与科研产出关系研究—以图书情报领域为例[J]. 情报杂志, 2017, 36(3): 191-195.
WANG Wei, SHI Rui-han, PAN Jing-hua. Research on the relationship between academic cooperation and scientific research output based on journal papers—Taking the field of library and information as an example[J]. Journal of Information, 2017, 36(3): 191-195.
[20] BROWN S A, DENNIS A R, VENKATESH V. Predicting collaboration technology use: Integrating technology adoption and collaboration research[J]. Journal of Management Information Systems, 2010, 27(2): 9-54.
[21] SHEN H W, BARABASI A L. Collective credit allocation in science[J]. Proceedings of the National Academy of Sciences, 2014, 111(34): 12325-12330.
[22] 刘静, 马建霞. 我国管理科学研究进展分析—以国家自然科学基金立项项目及论文产出为分析数据[J]. 科技管理研究, 2015, 35(326): 249-258.
LIU Jing, MA Jian-xia. Analysis of the progress of management science research in China —Analysis of national natural science foundation projects and paper outputs as analysis data[J]. Science and Technology Management Research, 2015, 35(326): 249-258.
[23] ZHANG Song-tao, GUAN Zhong-cheng. Education experience of scientific workforce—A case study on the winners of NSFDYS in CAS[J]. Forum on Science and Technology in China, 2015(12): 132-137.

编　辑　叶　芳