开放同行评议视角下学术论文同行评议得分与被引频次的关系

■谢维熙　张光耀　王贤文*

收稿日期:2021-09-10
修回日期:2021-11-09

大连理工大学科学学与科技管理研究所暨 WISE 实验室,辽宁省大连市甘井子区凌工路 2 号　116024

摘　要　**【目的】**在开放同行评议的大背景下,探讨会议论文同行评议得分与其被引频次的关系,从而分析同行评议结果与传统文献计量指标在科研评价中的关系,为完善科研评价体系提供一定的参考。**【方法】**基于 OpenReview 平台提供的 ICLR 会议论文的公开评审数据,将全部论文划分为口头报告、海报展示和拒收论文三类,运用文献计量和统计分析方法探究论文的同行评议得分与被引频次之间的关系。**【结果】**三类论文在评审得分和被引频次方面均存在显著差异,同行评议得分与被引频次存在较显著的正相关性。**【结论】**同行评议与传统文献计量指标在科研评价方面的一致性较高,但并非相互替代的关系,文献计量指标应是对同行评议的重要补充。科研评价体系应该是建立在定性同行评议的质量评价基础上,融合定量文献计量指标,形成一种主客观相结合的融合评价模式。

关键词　论文评分;OpenReview;被引频次;开放同行评议
DOI:10.11946/cjstp.202109100717

同行评议(本研究只讨论学术论文发表同行评议,不涉及科研项目和奖项评审情况)和引文分析是科研评价中常用的两种方法。同行评议是科技期刊对论文进行评价和遴选以保证发表论文质量的过程,由期刊邀请同行专家对投稿论文提出修改意见和作出评判,并将其评价作为判断文章是否能发表的主要依据[1]。引文是作者选择支撑其学术研究的理论、观点、数据和方法等研究资料,引文分析则是利用引文与学术成果的互依性进行学术评价[2]。同行评议与引文分析在学术评价中各有优缺点,随着同行评议数据的不断开放,科学工作者们可以从实证角度对同行评议和文献计量间的关系展开研究,但其研究仍受同行评议数据开放程度的限制。

本研究旨在开放同行评议的背景下,对三类论文在评审得分和被引频次方面的差异以及论文同行评议得分与被引频次的相关性进行分析,探讨论文同行评议结果与传统文献计量指标的关系,从而验证同行评议的有效性以及分析同行评议结果与传统文献计量指标在科研评价中的关系,为提高学术评价的科学性以及完善学术评价体系提供一定的参考。

同行评议最早可追溯到 17 世纪,一直以来在学术期刊质量控制和科研评价方面发挥着不可替代的作用[3]。传统的同行评议在实施过程中存在诸多问题[4-5],如审稿人和作者之间信任缺失、由个人利益和喜好导致不公正评价等问题[6]。

随着开放科学运动的不断推进[7],开放同行评议(Open Peer Review,OPR)以其公正、透明的优势日益受到关注并在全球多种学术期刊上得到实践,比如 *PLoS ONE*、*PeerJ*、BMJ 等[8],与国外相比,国内关于开放同行评议的研究和实践仍处于初级阶段。开放同行评议向大众开放审稿信息,包括审稿人和作者身份信息、审稿人的建议、作者的回复以及评审结果等信息[9],评审过程的开放性使得审稿人在评审时会更加谨慎公正,这对提高审稿意见的质量和客观公正性、缩短审稿时间、完善评议过程的监督机制以及促进知识交流等都有一定的促进作用[10]。关于同行评议的开放性是否会对稿件的被引频次产生影响,Zong 等[10] 和 Ni 等[11] 分别对 *PeerJ* 和 *Nature Communications* 的同行评议数据进行分析,得出了不一致的结论:前者认为开放同行评议提高了论文被引频次,而后者则并没有发现这一效果。

Bornmann 等[12] 以 *Atmospheric Chemistry and*

基金项目:国家自然科学基金面上项目"地理与网络二维空间及其交互影响视角下的科学论文扩散研究"(71673038);国家自然科学基金面上项目"科学文献全景大数据下的研究热点及研究前沿探测"(71974029)。

作者简介:谢维熙(ORCID:0000-0003-2330-7980),硕士研究生,E-mail:wrsjcycdfdsnt@163.com;张光耀,博士研究生。

* **通信作者**:王贤文(ORCID:0000-0002-7236-9267),博士,教授,博士生导师,E-mail:xianwenwang@dlut.edu.cn。

Physics 上的 1111 篇接收论文为研究对象，并提取论文发表 3 年后的被引频次，结果发现论文在各个数据库中的被引频次随着同行评议评分级别的降低而减少。Ragone 等[13]调研了 10 本计算机领域的会议论文集，发现同行评议评分等级与被引频次呈正向弱相关。王一华[14]将 IF（JCR）、CiteScore（Scopus）、h 指数、SJR 值、SNIP 值与同行评议结果进行 Spearman 非参数相关分析，发现同行评议结果与这 5 个文献计量指标的测量结果之间呈显著正相关。Bornmann[15]研究了 PLoS 或 F1000 专家推荐评审等级与传统文献计量指标的相关性，结果发现 FFa（F1000 论文因子）与被引频次之间的正相关性显著。

檀旦[16]以医学信息学和糖尿病为主题，分析 F1000 与传统文献计量学指标的相关性后发现两者具有一定的正相关性。万昊等[17]通过对 120 多篇实证研究进行元分析，比较同行评议与文献计量在科研评价中的作用，结果发现两者仅存在适度的正相关性，从而提出建构在定量辅助基础上的知情同行评议模式。黄明睿[18]基于《2014 年版中国科技期刊引证报告（核心版）》，采用多种计量统计方法研究期刊评价指标载文量、总被引频次、影响因子和综合评价总分之间的相互关系，结果表明总被引频次、影响因子、综合评价总分三者之间相互影响，在学术评价中起主要作用。现有的大部分实证研究表明：同行评议结果与以被引频次为基础的传统文献计量指标存在正相关关系，而且大部分研究结果显示两者的相关系数并不高。

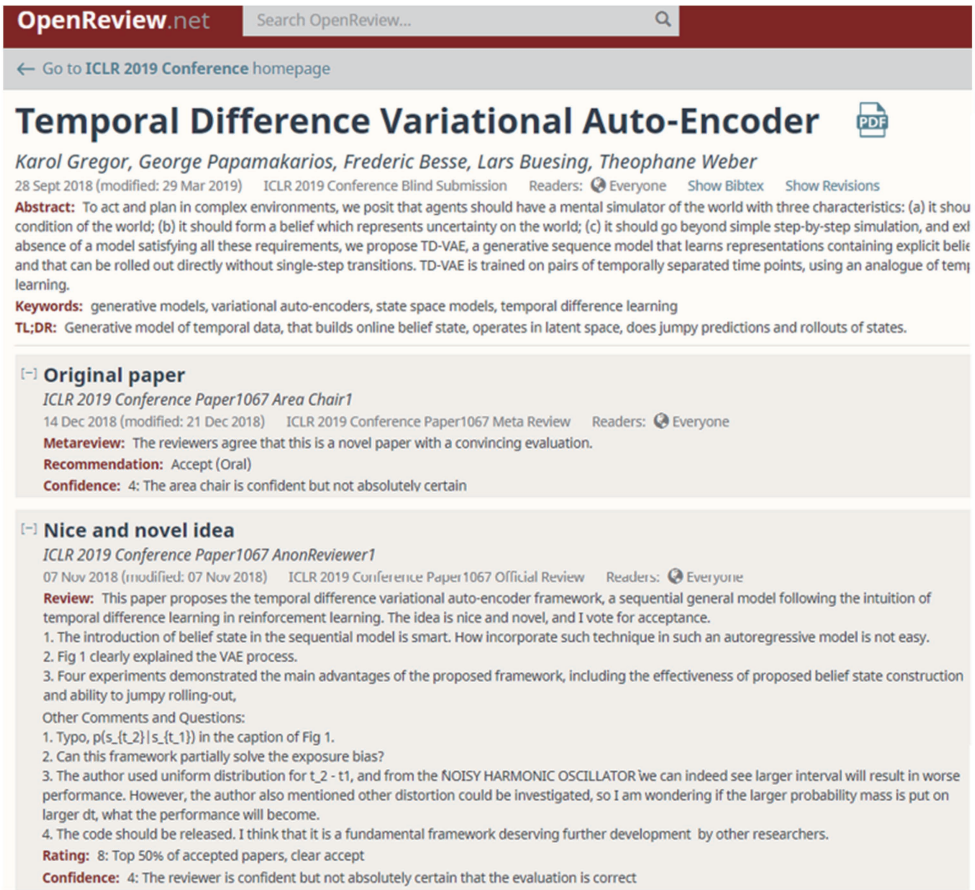
传统同行评议背景下，审稿过程数据的封闭状态限制了同行评议实证研究的开展。随着开放同行评议的推进，大量的关于审稿数据供科研人员进行研究。本研究基于 ICLR 会议论文的开放同行评议数据，使用同行评议的评分来定量测度同行评议的结果，相较于以往的定性研究具有一定的优势，而且 ICLR 数据集除了录用论文外，还包括拒稿，这使得研究更加充实和全面。

1 数据与方法

OpenReview 是一个会议论文公开评审网站，其中 ICLR（International Conference on Learning Representations）的全称为“国际学习表征会议”，是深度学习领域影响力最大的顶级会议之一，虽然成立较晚（2013 年成立），但是其作为深度学习的顶级会议已经得到了学术界的广泛认可。ICLR 备受关注的原因不仅是其在学术上具有较高的影响力，还在于它采取了开放同行评议制度，其公开的同行评议数据有原文题目、作者、摘要、下载链接、评审意见、作者与审稿专家以及参会人的讨论过程、审稿结果即评审得分（Rating）。在 ICLR 论文审稿中，会议主席对其负责的投稿作出录用与否的决策。会议主席考虑的信息包括审稿专家的评分、审稿过程中提供的证据、作者和审稿专家之间的讨论以及自己对论文的评估等等①。一些实证研究已经探索了这一数据集在研究中的可靠性，如基于 ICLR 的评审意见文本数据，对审稿意见情感以及评审中存在的制度偏见进行分析，还有学者提出将融合定性评价的论文质量评价模型用于定性评价文本的定量化研究[19-21]。在本研究中，将 ICLR 系列会议在 OpenReview 平台中的同行评议数据和文献计量指标数据作为研究数据，ICLR 的公开审稿意见（示例）如图 1 所示。

本研究选取 OpenReview 平台上 ICLR 会议论文集在 2018—2019 年公布的 2220 篇论文（排除审稿意见缺失的 1 篇论文和谷歌学术上查询不到的 8 篇论文，以及 14 篇数据出现异常的论文）作为研究对象，包括 42 篇口头报告论文（Oral Presentation Papers，以下简称“OP 论文”；难度最大，录用率约为 1.35%）和 780 篇海报展示论文（Poster Presentation Papers，以下简称“PP 论文”；录用率约为 22.65%）以及 1398 篇被拒收论文（Rejected Papers，以下简称“RP 论文”）。其中，用于数据分析的变量主要包括同行评议过程中审稿专家对每篇论文给出的评分，

① 来源于作者与 ICLR 项目主席的邮件通信，ICLR 项目主席的邮件原文为：“Within the ICLR review process, Area Chairs make an accept recommendation for each submission in their respective batch. Area Chairs are asked to take into account several sources of information, including the reviewer scores and certainty, the evidence provided in the reviews, discussion between authors and reviewers, and the Area Chair's own assessment of the paper. As such, there is no hard and fast rule on whether a paper will be accepted given a specific score. In addition, the Program Committee work with Area Chairs to calibrate acceptance decisions across Area Chairs, to account for factors such as the fact that some Area Chairs may be more conservative than others in their acceptance decisions. All calibration happens online and asynchronously, i.e., there is no single meeting where decisions are made.”


OpenReview.net Search OpenReview...

← Go to ICLR 2019 Conference homepage

Temporal Difference Variational Auto-Encoder

Karol Gregor, George Papamakarios, Frederic Besse, Lars Buesing, Theophane Weber

28 Sept 2018 (modified: 29 Mar 2019) ICLR 2019 Conference Blind Submission Readers: Everyone Show Bibtex Show Revisions

Abstract: To act and plan in complex environments, we posit that agents should have a mental simulator of the world with three characteristics: (a) it shou condition of the world; (b) it should form a belief which represents uncertainty on the world; (c) it should go beyond simple step-by-step simulation, and exl absence of a model satisfying all these requirements, we propose TD-VAE, a generative sequence model that learns representations containing explicit belie and that can be rolled out directly without single-step transitions. TD-VAE is trained on pairs of temporally separated time points, using an analogue of temp learning.

Keywords: generative models, variational auto-encoders, state space models, temporal difference learning

TL;DR: Generative model of temporal data, that builds online belief state, operates in latent space, does jumpy predictions and rollouts of states.

Original paper

ICLR 2019 Conference Paper1067 Area Chair1

14 Dec 2018 (modified: 21 Dec 2018) ICLR 2019 Conference Paper1067 Meta Review Readers: Everyone

Metareview: The reviewers agree that this is a novel paper with a convincing evaluation.

Recommendation: Accept (Oral)

Confidence: 4: The area chair is confident but not absolutely certain

Nice and novel idea

ICLR 2019 Conference Paper1067 AnonReviewer1

07 Nov 2018 (modified: 07 Nov 2018) ICLR 2019 Conference Paper1067 Official Review Readers: Everyone

Review: This paper proposes the temporal difference variational auto-encoder framework, a sequential general model following the intuition of temporal difference learning in reinforcement learning. The idea is nice and novel, and I vote for acceptance.

1. The introduction of belief state in the sequential model is smart. How incorporate such technique in such an autoregressive model is not easy.
2. Fig 1 clearly explained the VAE process.
3. Four experiments demonstrated the main advantages of the proposed framework, including the effectiveness of proposed belief state construction and ability to jumpy rolling-out,

Other Comments and Questions:

1. Typo, p(s_{t_2}|s_{t_1}) in the caption of Fig 1.
2. Can this framework partially solve the exposure bias?
3. The author used uniform distribution for t_2 - t1, and from the NOISY HARMONIC OSCILLATOR we can indeed see larger interval will result in worse performance. However, the author also mentioned other distortion could be investigated, so I am wondering if the larger probability mass is put on larger dt, what the performance will become.
4. The code should be released. I think that it is a fundamental framework deserving further development by other researchers.

Rating: 8: Top 50% of accepted papers, clear accept

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

图 1 ICLR 公开的审稿意见示例

用来判断单篇论文的非共识程度的得分方差,以及每篇论文发表至今(2021 年 6—7 月查询)在谷歌学术上的总被引频次。考虑到会议论文数据集在单个数据库中无法保证检全,因此选择谷歌学术上的被引频次作为研究要素。需要说明的是,考虑到同行评议数据的完整性以及统计被引频次时保证两年的被引时间窗口,本研究只选取 2018 年和 2019 年的数据作为研究对象。

2 结果

2.1 OP、PP 与 RP 论文的被引频次差异

为了比较 OP 论文与 PP 论文以及 RP 论文在同行评议结果和引文指标方面的差异,选取同行评议得分与论文发表至今的被引频次这两个指标进行比较分析,结果如表 1 和图 2 所示。由表 1 可知:OP 论文的评审得分和被引频次均明显高于 PP 论文,而 PP 论文的评审得分和被引频次又明显高于 RP 论文;单因素方差分析结果显示,不同类型论文之间的平均得分与平均被引频次差异有统计学意义。由于数据分布不符合正态分布,使用 K-S 检验进一步对三类论文的评审得分和被引频次进行检验,P 值均<0.001,说明 OP 论文、PP 论文和 RP 论文三者之间的评审得分和被引频次均存在显著差异。

表 1 OP 论文、PP 论文与 RP 论文的评价指标对比

评价指标	指标类型	OP 论文	PP 论文	RP 论文
评审得分	观测数	42	780	1398
	平均值	7.59	6.53	4.75
	中位数	7.67	6.67	4.67
	方差	0.21	0.36	0.79
被引频次	观测数	42	780	1398
	平均值	361.81	112.87	21.28
	中位数	153	53	2
	方差	251880.90	41839.94	5424.46

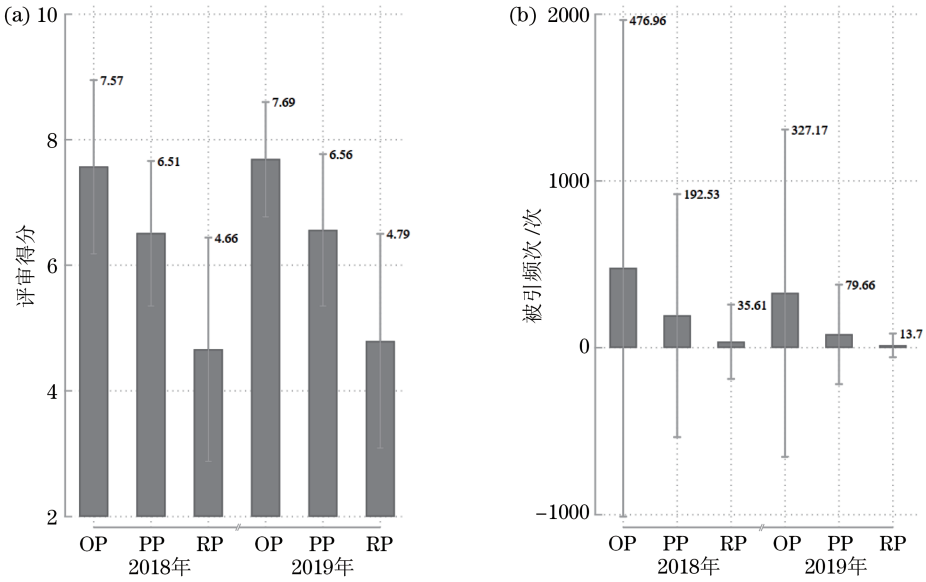


图 2　OP、PP 与 RP 论文评审得分和被引频次分布

(a)评审得分;(b)被引频次

2.2　论文评审得分与被引频次的相关性分析

经过 K-S 检验,论文的评审得分与被引频次不符合正态分布,因此采用 Spearman 秩相关分析方法对各类论文的评审得分与被引频次进行相关性分析。由表 2 可知,对于全部论文(OP、PP 和 RP 论文),相关系数为 0.625,表现出较高的相关性;对于 OP 论文,相关系数为 0.134,即评审得分与被引频次呈显著正相关(P<0.01);对于 PP 论文,相关系数为 0.160,即评审得分与被引频次呈显著正相关;对于全部录用论文(OP、PP 论文),相关系数为 0.209,全部录用论文的评审得分与被引频次呈显著正相关。

表 2　论文评审得分与被引频次的相关性分析结果

论文类别	相关系数	观测数
全部论文	0.625***	2220
OP 论文	0.134***	42
PP 论文	0.160***	780
全部录用论文	0.209***	822

注:*** 表示在置信度(双侧)为 0.01 时显著相关。

对评审得分与被引频次之间的关系进行进一步分析,首先探究全部录用论文(OP 和 PP 论文)中不同得分水平论文的被引频次差异是否有统计学意义。由于被接收论文中只有一篇低于 4 分,其余均分布在 4~10 分范围内,因此剔除一篇最低分论文,将 821 篇论文按得分分到 3 个区间里([4,6)、[6,8)、[8,10]),对这三组论文进行非参数检验,发现不同得分水平论文之间的被引频次差异具有统计学意义(P=0.002)。其次,探究对于不同被引频次水平的论文评审得分对被引频次的影响规律。本研究分析了高被引论文和极高被引论文的得分情况,将所有论文按被引频次降序排列,取前 5%为高被引论文,前 1%为极高被引论文,结果如表 3 所示。可以看到,极高被引论文的评审得分均值(6.91)>高被引论文的评审得分均值(6.57)>非高被引论文的评审得分均值(5.37)。

表 3　高被引论文和极高被引论文的平均评审得分与平均被引频次的比较

论文类别	论文总量/篇	非高被引论文			高被引论文			极高被引论文		
		N_1	R_1	C_1	N_2	R_2	C_2	N_3	R_3	C_3
OP	42	28	7.55	112.96	14	7.67	859.50	7	7.71	1342.86
PP	780	700	6.52	65.23	80	6.64	529.68	12	6.67	1236.64
RP	1398	1381	4.74	15.55	17	5.29	487.24	3	6.11	1204.00
All	2220	2109	5.37	33.33	111	6.57	564.77	22	6.91	1239.91

注:N 为论文量,R 为平均评审得分,C 为平均被引频次。

为了更清晰地展示评审得分与被引频次的关系,绘制了全部 2220 篇论文的评审得分与被引频次的散点图。从图 3(a)可以看出,总体上评审得分与被引频次的相关性并不显著。本研究同时考虑了评审存在分歧的论文即非共识论文的被引频次分布情况。国家自然科学基金委员会管理科学部副主任杨列勋

指出,评审专家在某一项研究项目的评审上两种意见几乎各占一半,且双方均有一定的论据,那么这项研究就是非共识研究[22];刘文波和钮晓鸣[23]认为,非共识研究是指具有不确定性和创新性且在初期评审专家难以对研究成果达成一致意见的研究行为或活动。虽然目前学术界尚未对非共识研究形成统一的界定,但是关于非共识研究同样存在研究价值和创新价值这一观点已经得到学术界的广泛认可。本研究使用一篇论文评审得分的方差来表示该论文的整体非共识度,方差越大,表示论文非共识的离散或者说审稿人意见相左的程度越大,即非共识度越大,或者说对论文评审结果的分歧越大[24]。图 3(b)展现了论文非共识度与被引频次的关系,统计结果显示论文非共识度与被引频次呈正相关,但两者的相关性并不显著。

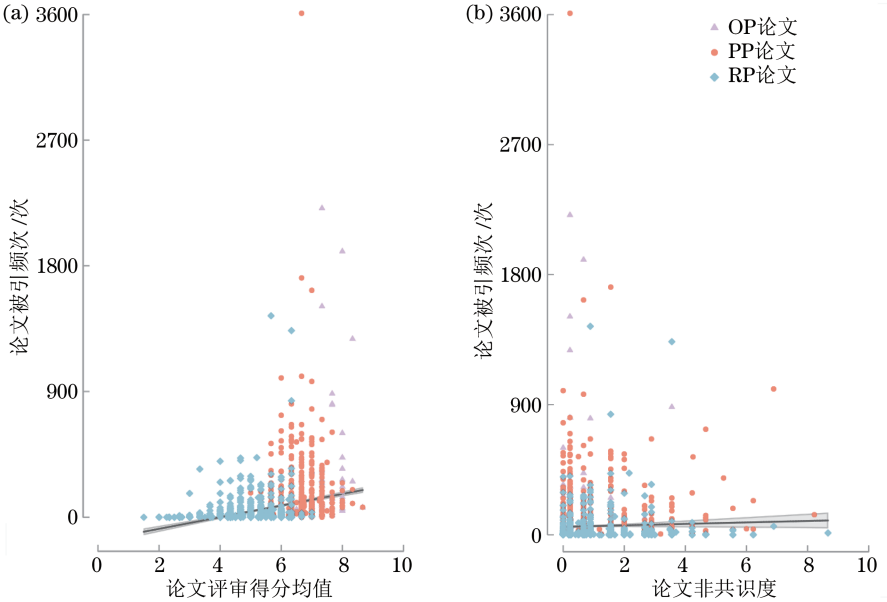


图 3　论文评审得分与非共识度散点图

(a)所有论文评审得分与被引频次散点图;(b)论文非共识度与被引频次散点图

2.3　回归分析

上述内容中的统计检验结果显示了论文录用状态、评审得分以及论文非共识程度和被引频次之间的关系。基于上述分析,本研究拟通过回归分析(OLS 和 mlogit)来进一步检验论文评审得分和被引频次之间的关系。模型设定为

$$Y_i = X_i\beta + \varepsilon_i \quad (1)$$

式中:Y_i 为论文的被引频次;X_i 为解释变量;β 为回归系数;ε_i 为误差项。

$$U_{ij} = X_i\beta_j + \varepsilon_{ij} \quad (2)$$

式中:U_{ij} 表示第 i 篇论文在第 j 种评审状态下的随机效用;β_j 为不同评审状态下对应的回归系数;ε_{ij} 为误差项。变量的描述统计结果和相关系数矩阵如表 4 和表 5 所示,论文评审得分与被引频次的回归结果如表 6 所示。

为避免极端值的影响,在回归前将 Citations、Rating、Variance 在 99 分位作截尾处理。方差膨胀系数(Variance Inflation Factor,VIF)最大值为 2.24,平均值为 1.51,表明不存在严重的共线性。在模型 1 中加入了所有变量,评审得分的回归系数为正且在 0.001 水平上显著相关,录用论文的系数在 0.001 水平上显著正相关,意味着录用论文的被引频次相比于 RP 论文更高。进一步将样本拆分成两部分,在模型2中纳入RP论文样本,在模型3中纳入全

表 4　变量的描述统计结果

变量名	类型	定义	平均数	标准差	最小值	最大值
Citations	Count	被引频次	65.59	204.28	0.00	4204.00
Rating	Count	评审得分	5.44	1.21	1.50	9.00
Variance	Count	评审得分方差	0.86	1.03	0.00	10.89
Accept	Dummy	接受=1,拒绝=0	0.37	0.48	0.00	1.00
PP	Dummy	PP=1,OP=0	2.60	0.53	1.00	3.00
Reviewer	Count	审稿人数量	3.04	0.23	1.00	5.00
Year	Dummy	年份	2018.63	0.48	2018.00	2019.00

表 5 变量的相关系数矩阵

变量名	Citations	Rating	Variance	Accept	PP	Reviewer	Year
Citations	1						
Rating	0.26	1					
Variance	0.08	0.03	1				
Accept	0.28	0.74	−0.01	1			
PP	−0.32	−0.75	0.01	−0.96	1		
Reviewer	−0.01	−0.00	0.06	−0.01	0.01	1	
Year	−0.16	−0.00	−0.06	−0.05	0.06	0.09	1

表 6 论文评审得分与被引频次的回归结果

变量	模型 1（所有论文）	模型 2（拒收论文）	模型 3（录用论文）	模型 4（OP 论文）	模型 5（PP 论文）	模型 6（RP 论文）
Rating	0.510***	0.544***	0.344***	0.045***	0.224***	−0.269***
	(0.034)	(0.041)	(0.065)	(0.006)	(0.006)	(0.002)
Variance	0.040	0.019	0.034	0.003	0.009	−0.006
	(0.032)	(0.046)	(0.041)	(0.004)	(0.007)	(0.006)
Accept	1.477***					
	(0.091)					
Reviewer	0.092	0.149	−0.089	0.005	−0.013	0.008
	(0.131)	(0.169)	(0.161)	(0.012)	(0.031)	(0.029)
PP			−0.413*			
			(0.155)			
截距项	−0.812	−1.192*	2.801***			
	(0.433)	(0.551)	(0.737)			
样本量	2178	1387	791		2220	
Year	Y	Y	Y		Y	
R^2	0.46	0.10	0.13			

注：括号里的数字表示标准差；*** 表示在置信度（双侧）为 0.001 时显著相关；** 表示在置信度（双侧）为 0.01 时显著相关；* 表示在置信度（双侧）为 0.05 时显著相关；Y 表示在模型中控制了时间效应。

部录用论文样本，这两个模型的评审得分系数仍然为正，且在 0.001 水平上显著相关。从模型 3 可以看出，PP 论文的系数为负，且在 0.05 水平上显著相关，意味着相比于 PP 论文，OP 论文有着更高的被引频次。模型 4~6 为使用 mlogit 估计的结果，表 6 中呈现的是评审得分的边际效应，其中对于 OP 和 PP 论文，评审得分的边际效应为正，对于 RP 论文，评审得分的边际效应为负。

2.4 低得分-高被引论文和高得分-低被引论文

上述分析结果得出被 ICLR 接收的论文，其 Spearman 秩相关系数 $r=0.209$，因此对这种弱相关背后的一些例外情况进行分析。运用案例分析方法，选取 6 篇评审得分低-被引频次高的论文和 6 篇评审得分高-被引频次低的论文作为案例，对这两组案例的评审得分、被引频次、得分方差、预印本（arXiv）存档以及文献内容进行分析，以发现同行评议结果与被引频次相悖的文献特征，对评审得分低-被引频次高以及评审得分高-被引频次低的论文进行统计，结果如表 7 所示。

在会议集对论文做出接收或拒绝的决定之前，有部分论文已经发布到 arXiv 平台上，表 7 统计的低得分-高被引论文都在被接收之前发布在 arXiv 平台上，这就使得这些论文较其他未发布到 arXiv 的论文有更长的被引窗口。以往的研究表明，arXiv 论文在许多数据库中都具有显著的引用优势[25]。

由表 7 可知，这些低得分-高被引论文具有一些共同的特征：评审专家给分均不高、存在较低分导致平均得分较低、大部分论文都发布在 arXiv 平台。高得分-低被引论文也具有一些共同的特征：大部分论文并未发布到 arXiv 平台，并且这些论文的评审得分方差普遍较低，说明评审专家对这些论文的评价分歧较小。分析上述论文的原文和审稿意见后发现：低得分-高被引论文的创新性通常较低，或者属于综述性研究；而高得分-低被引论文通常具有较高的创新性，因而得到审稿人的高度认可。

表 7 案例论文统计结果

类别	ID[②]	平均得分	被引频次/次	方差	得分 1	得分 2	得分 3	评审结果	是否发布到 arXiv
低得分-高被引论文	HyxnZh0ct7	4.67	290	4.22	5	2	7	PP	是
	Hk99zCeAb	5.67	3162	10.89	8	1	8	OP	是
	Bkg6RiCqY7	6.00	1258	0.67	6	7	5	PP	是
	rkZvSe-RZ	6.00	1373	0.00	6	6	6	PP	是
	r1Ddp1-Rb	6.33	2014	0.22	6	7	6	PP	是
	rJXMpikCZ	6.00	4204	0.67	6	5	7	PP	是
高得分-低被引论文	HktK4BeCZ	8.67	17	0.89	10	8	8	OP	是
	HJx54i05tX	8.33	21	0.22	9	8	8	OP	否
	H1lqZhRcFm	8.00	6	0.00	8	8	8	PP	是
	B1gstsCqt7	7.67	1	1.56	6	9	8	PP	否
	S1JHhv6TW	8.00	16	0.67	8	9	7	OP	是
	BkltNhC9FX	8.00	19	0.67	8	9	7	PP	否

3 结论与讨论

3.1 研究结论

论文的评审得分反映的是审稿人对研究的主观评价,而且大多数都是定性评价,被引频次反映的是学术同行对科研劳动成果的认可程度,在一定程度上反映了科研产出的质量,是一种定量评价。上述研究结果表明用这两种方法对科研成果进行评价得到的结果并不总是一致的。

ICLR 通过同行评议决定论文是否录用以及录用为口头汇报还是海报展示,通过对 OP 论文、PP 论文和 RP 论文进行描述统计和方差分析,发现这三类论文的评审得分和被引频次是有差异的,进行两两比较后发现差异均有统计学意义($P<0.05$),这个结果从一定程度上反映了同行评议的有效性和同行评议结果与传统计量指标的一致性。

通过对论文评审得分和被引频次进行相关性分析和回归分析,发现 PP 论文、录用论文、全部论文的评审得分与被引频次存在显著的正相关关系,这一结果与以往关于同行评议结果和被引频次的研究结果类似。本研究结果表明:虽然同行评议和被引频次从不同角度反映科学研究的学术影响力,但是两者在一定程度上呈正相关,证明了同行评议和被引频次在科研评价中的有效性和一致性;同行评议能够选出具有价值的论文,并在发表之后具有更高的影响力,证实了同行评议的有效性。

录用论文的评审得分与被引频次的相关性不高,可能是因为同行评议与传统计量指标是从不同维度对文章进行评价,同行评议具有主观性和封闭性等特点,引用具有偏性和引用动机复杂性等特点。对这种弱相关性背后的一些个例进行统计,对低得分-高被引和高得分-低被引论文进行分析发现,前者是事先发布到 arXiv 平台的微创新性研究论文或综述性文章,后者则大多是非共识度低、但创新程度高的研究论文或学术争议文章。这一结果从一定程度上反映了以引用为代表的定量指标和同行评议定性评价指标是相辅相成的,可将定量和定性两种评价工具结合起来进行相对有效、全面的科研评价。

3.2 讨论与启示

同行评议的结果是从评审专家的角度来评估论文的质量,而以被引频次为基础的传统计量指标是从作者的角度来判断论文的质量及影响力。同行评议作为科学研究的"守门人",虽然存在主观偏见可能导致结果有失公允,但是其作为控制科研质量的首要机制,对科研评价体系的建设起到至关重要的作用。被引频次作为传统文献计量评价的基础,虽然存在引用的不完备性和有偏性,但是被引频次可以作为一种定量化工具,在一定程度上反映同行对研究质量及影响力的评价。本研究结果发现虽然同行评议结果与引文度量指标之间呈正相关,但是同行评议和文献计量指标之间是不可相互替代的:同行评议仍然是目前科研评价体系最重要的一环;相比于同行评议的精英评价,文献计量指标能够提供更大范围内公开的同行评价参考。

从期刊评价实践的角度来看,文献计量指标是对同行评议的重要补充。期刊评价体系应该是建立在定性同行评议的质量评价基础上,融合定量文献计量指标,形成一种主客观相结合的评价模式。

[②]论文示例来自 https://openreview.net/forum?id=HyxnZh0ct7。

4 局限

本研究存在一些局限:首先,本研究使用的开放同行评议数据,其开放透明的特点给研究带来了极大的便利,但是由于目前采取开放同行评议模式的期刊和会议集较少,而且开放程度也不尽相同,本研究仅选取了公布全部投稿论文的评审得分数据的 ICLR 数据集进行分析,论文样本量较小,可能会限制研究的开展;其次,本研究的对象是计算机领域的会议论文,可能存在学科差异,结论外推时需谨慎;最后,本研究对同行评议结果和文献计量指标的相关性进行分析,提出应将定性、定量两种评价工具结合起来才能进行有效的科研评价,但如何实现二者的融合评价是亟需解决的问题,需要后续进一步研究。

参考文献

[1] 郭碧坚,韩宇. 同行评议制:方法、理论、功能、指标[J]. 科学学研究,1994,12(3):63-74.

[2] 李冲. 引文分析的本质与学术评价功能的条件性[J]. 科学学研究,2013,31(8):1121-1127.

[3] 孟美任,张晓林. 中国科技期刊引入开放同行评议机制的思考与建议[J]. 中国科技期刊研究,2019,30(2):149-155.

[4] Shatz D. Peer review: Acritical inquiry [M]. Lanham: Rowman & Littlefield,2004.

[5] Seeber M, Bacchelli A. Does single blind peer review hinder newcomers? [J]. *Scientometrics*,2017,113(1):567-585.

[6] Rennie D. Let's make peer review scientific[J]. *Nature*,2016,535(7610):31-33.

[7] 张光耀,姜春林,王贤文. 即时开放获取论文在时间和地理空间上的使用优势分析:以《新英格兰医学期刊》为例[J]. 中国科技期刊研究,2019,30(8):890-896.

[8] 贺颖,付江阳. 透明性同行评议:产生、内涵与建构[J]. 中国科技期刊研究,2021,32(3):332-336.

[9] 刘丽萍,刘春丽. 开放同行评议利弊分析与建议[J]. 中国科技期刊研究,2017,28(5):389-395.

[10] Zong Q J, Xie Y F, Liang J C. Does open peer review improve citation count? Evidence from a propensity score matching analysis of *PeerJ*[J]. *Scientometrics*,2020,125(1):607-623.

[11] Ni J, Zhao Z Y, Shao Y P, *et al*. The influence of opening up peer review on the citations of journal articles [J]. *Scientometrics*,2021:1-12.

[12] Bornmann L, Marx W, Schier H, *et al*. From black box to white box at open access journals: Predictive validity of manuscript reviewing and editorial decisions at *Atmospheric Chemistry and Physics*[J]. *Research Evaluation*,2010,19(2):105-118.

[13] Ragone A, Mirylenka K, Casati F, *et al*. A quantitative analysis of peer review [C]//13th International Society for Scientometrics and Informetrics Conference, Durban. Belgium: International Conference on Scientometrics & Informetrics,2011:724-736.

[14] 王一华. 基于 IF(JCR)、IF(Scopus)、H 指数、SJR 值、SNIP 值的期刊评价研究[J]. 图书情报工作,2011,55(16):144-148.

[15] Bornmann L. Interrater reliability and convergent validity of F1000Prime peer review [J]. *Journal of the Association for Information Science and Technology*,2015,66(12):2415-2426.

[16] 檀旦. F1000 与传统文献计量学指标的相关性研究[J]. 中国科技期刊研究,2016,27(1):111-115.

[17] 万昊,谭宗颖,朱相丽. 同行评议与文献计量在科研评价中的作用分析比较[J]. 图书情报工作,2017,61(1):134-152.

[18] 黄明睿. 期刊评价 4 个核心指标之间关系的探讨[J]. 农业图书情报学刊,2017,29(12):151-155.

[19] 张明阳,王刚,彭起,等. 学术论文公开评审平台数据分析[J]. 计算机科学,2021,48(6):63-70.

[20] Tran D, Valtchanov A, Ganapathy K, *et al*. An open review of openReview: A critical analysis of the machine learning conference review process[EB/OL]. (2020-10-11)[2021-08-20]. https://arxiv.org/abs/2010.05137.

[21] 林原,王凯巧,丁堃,等. 学术论文的定性评价定量化研究[J]. 情报理论与实践,2021,44(8):28-34.

[22] 杨列勋,汪寿阳,席酉民. 科学基金遴选中非共识研究项目的评估研究[J]. 科学学研究,2002,20(2):185-188.

[23] 刘文波,钮晓鸣. 非共识项目的立项决策研究[J]. 上海管理科学,2017,39(4):61-65.

[24] 西桂权,王冠宇. APF 组合分析法在非共识项目评价中的运用[J]. 科技智囊,2020(5):35-39.

[25] Chen Y, Wang Z, Tan J, *et al*. The position of preprint in scholarly communication: A biliometric and empirical study of arXiv[C]//Proceedings of the 16th Conference on International Society of Scientometrics and Informetrics, October 16-20, 2017, Wuhan, China. Belgium: International Conference on Scientometrics & Informetrics,2017:799-809.

作者贡献声明:

谢维熙:处理数据,撰写论文;

张光耀:处理数据,修改论文;

王贤文:设计研究思路,指导论文修改。

Relationship between peer review score and cited frequency of conference papers under the background of open peer review

XIE Weixi, ZHANG Guangyao, WANG Xianwen*

WISE Lab, Institute of Science of Science and S&T Management, Dalian University of Technology, 2 Linggong Road, Ganjingzi District, Dalian 116024, China

Abstract: [**Purposes**] Under the background of open peer review, this paper discusses the relationship between the peer review score of conference papers and their cited frequency, and further analyzes the relationship between peer review results and traditional bibliometric indicators in scientific research evaluation, which is expected to provide a reference for improving scientific research evaluation system. [**Methods**] The public review data of International Conference on Learning Representations (ICLR) papers were retrieved from OpenReview, and the papers were classified into oral reports, poster presentations, and rejected papers. Then, bibliometric method and statistical method were employed to explore the relationship between the peer review score and cited frequency. [**Findings**] Both the review score and cited frequency were significantly different among the three types of papers, and the review score was highly correlated with cited frequency. [**Conclusions**] Peer review and traditional bibliometric indicators have high consistency in scientific research evaluation, but they are not substitutes for each other. Bibliometric indicators supplement peer review. Scientific research evaluation system should be based on the quality evaluation of qualitative peer review and integrate quantitative bibliometric indicators, thereby combining both subjective and objective assessment.

Keywords: Paper scoring; OpenReview; Cited frequency; Open peer review

(本文责编:梁永霞)