基于间接引用机制的科学文献被引频次优化模型构建与实证研究*

刘运梅　李冉　秦佳佳
（上海大学文化遗产与信息管理学院，上海 200444）

摘要：为客观评价科学文献的真实影响力、弱化学术评价中的马太效应问题，通过降低间接引用关系的计数权重，构建科学文献被引频次优化模型，提出引用优化比例指标，并通过实证研究检验该模型的有效性。研究发现：第一，优化间接引用频次权重后，所得结果能相对客观地反映科学文献的真实被引情况与真实影响力；第二，该被引频次优化模型可用于识别潜在的高质量、高被引论文；第三，该被引频次优化模型能够在较大程度上缓解学术评价中被引的时间累积性和马太效应问题，对年轻文献具有较高的评价公平性、较好的筛选能力。

关键词：科学文献；间接引用；被引频次；学术评价
中图分类号：G254　DOI：10.3772/j.issn.1673-2286.2024.10.004
引文格式：刘运梅，李冉，秦佳佳. 基于间接引用机制的科学文献被引频次优化模型构建与实证研究[J]. 数字图书馆论坛，2024，20（10）：33-41.

科学文献是学者在科技创新活动中的主要产出形式，是反映一个学科领域基础研究和应用研究的创新成果，同时也是学者、机构、期刊等科学参与主体学术水平与科研能力的重要标志[1]。基于科学文献间引用、被引关系建立的引文分析方法揭示了引用的数量特征与内在规律，已被广泛应用于学术评价、科学前沿探测、学科知识演化分析等各个研究领域[2-6]，并服务于科技创新的战略情报研究和科技文献的情报分析工作。

文献引用制度中存在着一些不可忽视的重要问题。第一，以引文为基础的量化评价方法隐含了一个理想化的前提和依据：被引频次的高低等同于学术质量或影响力的高低。但是，施引者引用某篇文献并不一定表示对该文献学术质量的认可。第二，论文被引频次两极分化严重，大量的引用集中在少量的论文上，使得学术资源分布极不均衡。那么，是否存在一种隐性的引用机制，导致这样的分布失衡？积累了较高被引频次的论文就一定具有极高的学术价值吗？相应地，那些低被引，甚至零被引论文毫无学术贡献和影响力吗？实际上，文献的引用关系与引用行为具有高度复杂性[7-8]，一篇论文引用不同参考文献的目的、动机各不相同，不同论文引用同一篇参考文献的动机也是各不相同的。因此，有必要追溯以上问题的本源，探究引用形成过程中文献真实的内在价值，对引文影响力形成准确的理解。

随着国际研究环境的急速变化、研究水平的不断提高，科学家科研成果的产出效率提升，研究文献的总量也大幅度增加[9]。这给学者在科学研究中的文献调研与阅读增加了难度，为引用而引用的现象变得更加普遍，一部分学者并不是在阅读、学习文后所有参考文献原文的基础上对其进行引用。因此，在科学研究中，通过间接引用行为产生的虚假引文常有存在[10]，这些间

收稿日期：2024-04-23
*本研究得到国家自然科学基金青年项目“基于全文本引文解构的引用失范行为识别与生成机理研究”（编号：72304181）、国家智能评价与治理实验基地2024年创新评价开放基金资助。

接引用掩盖了被引文献的真实价值，使高被引文献更容易成为再次被引的对象，而低被引、零被引文献更容易无人问津。除此之外，间接引用造成引文分析的开展建立在虚假的数据资料基础之上，从而影响期刊评价、论文影响力评价以及人才评估等文献情报工作的正常开展。以存在间接引用的论文总被引频次评价科技人才会引发不良的社会效应，造成学术界不真实的论文引用计数越来越多[11-12]。因此，优化科学文献间接引用机制中的被引频次计数，客观、公正反映文献真实的内在价值是一项重要且必要的研究工作。

1 相关研究

引文的影响力评价既可以从定量角度建立被引评价指标，也可以从引用文本的定性角度建立评价方法。在基于引用关系的定量评价中，Radicchi等[13]构建作者引用网络，根据PageRank算法获得了学者的影响力排序。王菲菲等[14]在互引、共被引、文献耦合三维引文关联融合视角下，综合使用社会网络分析、主成分分析、熵权法和天际线算法，建立了学者引文影响力评价框架。Min等[15]基于Bass扩散模型量化引文扩散机制，提出了Saturation Level指标，用于粗略估计一篇科学文献在生命周期的当前阶段及未来被引用的潜力。俞立平等[16]从期刊过度自引与人为降低载文量两个方面入手，对篇均自引率在30%以上、载文量与平均载文量之比低于0.5的期刊的影响因子进行修正，实验结果表明该方法能够有效降低过度自引对被引频次计数的影响。综上，基于被引频次的相关研究主要运用数学模型、文献外部特征等建立学术评价指标，但仍囿于传统引文评价的固有缺陷，如引文时间滞后、数据方法粗糙、未对施引动机和引用重要性加以区分等。

从引文内容角度，国内外相关的文献影响力评价研究也已经取得一定成果。Bonzi[17]根据引文文本的表述方式和在施引文献中出现的次数，判断被引文献的影响力程度，发现大部分被引文献只被作者简单提及，对确定施引文献的主题研究并没有实质贡献或帮助。Macroberts等[18]比较基因学领域引文文本与施引文献、被引文献的主题相关性，并将论文的引用划分为有影响力、无影响力两类，发现施引作者会存在一定的隐瞒行为，如以平淡的语言描述对研究过程具有重要作用的文章，以掩盖被引文献的真实影响力。Wan等[19]认为科学文献中的引文并不同等重要，将引文文本按引用重要性程度分为5个等级，并将其设为因变量，以引文次数、位置、间隔发表时间、平均密度等为自变量进行回归分析，实证研究证明了引用强度值的有用性。综上，基于引文内容信息如引用位置、引用功能或引用情感等构建的学术评价方法能够细粒度区分引用的重要性，但相关研究涉及文献全文内容的抽取与计算，模型复杂度和数据处理难度较高，同时引用动机和功能划分受主观因素影响较大，因而在学术评价实际工作中的实用性与可行性较差。

基于以上研究的优势与不足，本文将从引用结构角度，对不同类型的引用关系加以区分，降低间接引用机制产生的引用频次计数权重，构建文献被引频次优化模型，以解决传统引文评价将所有引用视为等同、数据方法粗糙等问题，同时避免引文内容评价方法中全文数据处理复杂的问题。此外，提出引用优化比例指标，基于实证研究对比间接引用频次优化前、后的计算结果，评估该优化模型的评价效果。

2 基于间接引用机制的被引频次优化模型构建

两级传播理论是传播学四大先驱之一——美国传播学家Lazarsfeld的一大发现，也是大众传播研究领域的重要理论基石[20]。该理论将信息传播过程凝练为两步：第一步是从信息源，通过大众媒介传递到意见领袖，即信息传播的第一个阶段；第二步是从意见领袖到追随者的人际传播，即信息传播过程的第二个阶段。在两级传播过程中，意见领袖起着重要的中介作用。大众传播和人际传播在人们的信息获取和决策中也发挥着不同的作用[21-22]：大众传播影响面广，在人们的认知阶段具有重要作用；人际传播渗透力强、有针对性，在说服和决策阶段人际传播的影响更显著，对大众传播的信息有进一步的整合作用。

类比两级传播理论，科学文献间的知识传播过程中同样存在两级传播机制，如图1所示。原始文献A首先将知识直接传递给中间文献B，此为直接传播阶段。追随文献C在引用中间文献B的同时，出于某些主观上的负面引用动机，如外文文献阅读障碍、文献全文获取权限限制、刻意增加参考文献数量等，通过中间文献B中关于文献A的引文内容信息，不直接阅读文献A而对原

始文献A施加引用[23-24]，此为间接传播阶段。在基于间接引用结构的知识传递过程中，文献A在发表后通过直接传播渠道影响了文献B的作者，将知识传递给文献B；而文献B又充当中间人角色，进一步将知识传递给追随文献C。文献C同时引用了原始文献A与中间文献B。

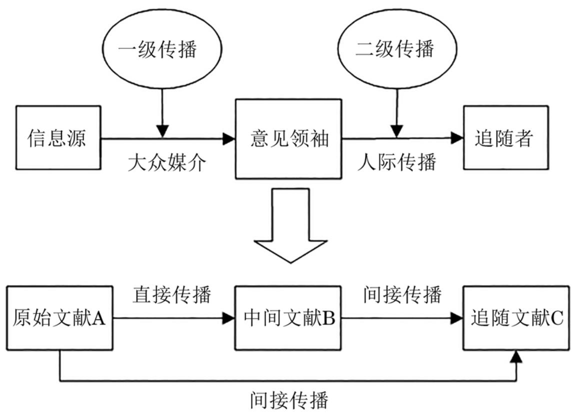

图1　两级传播与间接引用机制示例

从文献的学术影响力视角看，正如两级传播模型中两个阶段的角色和作用不同，文献A在文献B与文献C的两次引用中产生的影响力并不能完全等同。文献B对文献A的引用属于文献A自身影响力产生的直接结果，而文献C对文献A的引用则来自中间文献B的中介影响力。因此，在由文献B向文献C的传播中，文献A的影响力减弱。综上可知，间接引用行为的存在会影响引文数据的准确性与真实性，同时也会削弱引文分析作为科学评价工具的权威性和可信度。因此，有必要对间接引用结构中由间接引用关系产生的文献A的被引频次进行优化，以客观、公正、真实地反映科学文献A的学术价值与学术影响力。

直接引用、共被引与文献耦合结构中仅有两两文献间的直接引用关系，无冗余的间接引用关系，上述3种引用结构中所有引用均计数为1，如图2所示。图中箭头方向由施引文献指向被引文献，箭头上数字表示该引用关系的引用频次计数，括号内数字表示该文献收获的所有引用频次计数。

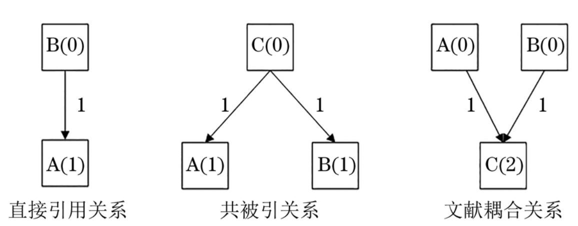

图2　无冗余引用结构的被引频次计数

当文献之间存在间接引用联系时，文献C对文献A的引用来自中间文献B的影响力。此时，文献A收获的来自文献C的间接引用频次不能记作1，将文献A产生的间接引用影响力权重设置为一半，即0.5，以区别直接影响力和间接影响力。A、B、C三方文献的被引频次计数如图3所示，其中虚线表示间接引用关系。

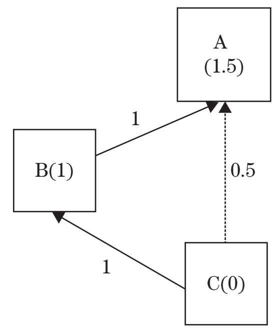

图3　间接引用结构的被引频次计数

在间接引用结构的基础上迭代一层后，共有以下3种引用结构，分别为：文献D引用文献B、C；文献D引用文献A、C；文献D同时引用文献A、B、C。按照直接引用与间接引用关系的被引频次权重计数法，文献A、B、C、D的被引频次计数如图4所示。

图2~图4的7种引用结构包含了复杂引文网络中科学文献之间的全部引用形式。基于这些计数机制，建立引文网络中任意一个节点的被引频次计数算法，即计算直接引用频次与间接引用频次的和。直接引用频次计数P与间接引用频次计数Q的计算公式如式（1）~式（2）所示。

$$P = p \times 1 \tag{1}$$

$$Q = \sum_{i=1}^{q} \frac{1}{d_i - 1} \tag{2}$$

式中：p表示某节点文献的所有无冗余链接数量；q表示直接指向节点文献，且具有其他冗余链接的链接数量；d_i表示第i个冗余链接中，冗余链接的层级数。例如，在间接引用结构中，文献C指向文献A的冗余链接C→B→A，层级数为3。

节点文献被引频次在被优化后，其被引频次计数总和表示为N_f，即直接引用频次计数P与间接引用频次计数Q的和，见式（3）。

$$N_f = P + Q \tag{3}$$

为直观判断一篇科学文献间接引用关系的多少，量化被引频次优化的效果，建立被引频次的优化比例指标。科学文献优化前的原被引频次N_b与优化后的被引频次计数N_f之差为被优化数量，将其除以优化前的原

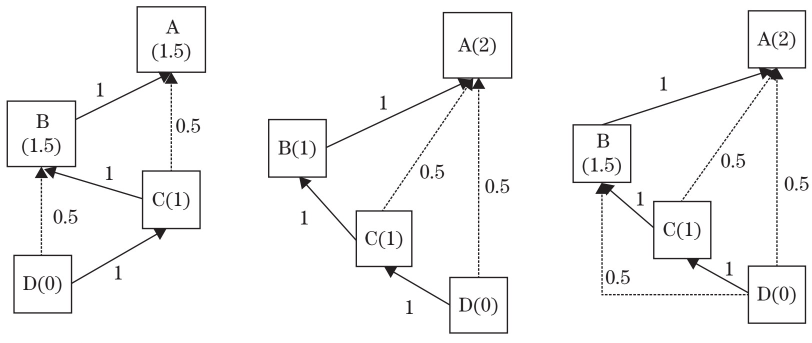

图4　间接引用结构迭代一层后的被引频次计数

被引频次N_b即为引用优化比例指标R，具体计算公式如式（4）所示。

$$R = \frac{N_b - N_f}{N_b} \tag{4}$$

3　基于间接引用机制的被引频次优化实验

以WoS（Web of Science）数据库为数据来源。考虑到学科差异，根据WoS学科分类体系选取医学与生物学、心理学、管理学、化学、物理学、数学、计算机科学、图书情报科学共8个学科的数据。为保证数据样本多样性，文献类型同时包含Article、Review、Proceedings Paper。从以上8个学科的数据样本中各随机抽取5篇样本文献，共获取40篇样本文献作为被引频次优化模型的评价对象。

确定目标文献后，获取样本文献对应的被引文献集合、二级被引文献集合。首先，基于目标文献与被引文献集合、被引文献集合与二级被引文献的引用关系，分别建立40篇目标文献的引文网络。其次，根据式（1）~式（2）计算40篇目标文献的间接引用频次计数Q与直接引用频次计数P。最后，将两方相加，计算优化后的被引频次N_f，以及引用优化比例指标R。各项指标计算结果如表1所示。

表1　被引频次优化过程中各项指标计算结果

文献序号	学科	DOI	被引频次/次	Q	P	N_f	R/%
1	医学与生物学	10.1371/journal.pone.0080633	468	113.5	241	354.5	24.25
2	医学与生物学	10.1097/00001888-199306000-00002	421	138.0	145	283.0	32.78
3	医学与生物学	10.1016/j.fertnstert.2008.09.018	220	80.0	60	140.0	36.36
4	医学与生物学	10.1073/pnas.1508520112	270	26.0	218	244.0	9.63
5	医学与生物学	10.1016/S0272-7358（03）00031-X	184	46.5	91	137.5	25.27
6	心理学	10.1037/h0035592	647	235.5	176	411.5	36.40
7	心理学	10.1037/a0031808	565	124.5	316	440.5	22.04
8	心理学	10.1037/0003-066X.38.4.399	268	64.5	139	203.5	24.07
9	心理学	10.1037//0022-0663.92.2.316	415	174.5	66	240.5	42.05
10	心理学	10.1016/j.dr.2016.06.004	366	20.5	325	345.5	5.60
11	管理学	10.1002/smj.439	528	190.0	148	338.0	35.98
12	管理学	10.1016/S0883-9026（99）00054-3	402	130.5	141	271.5	32.46
13	管理学	10.1016/S0149-2063（99）00008-2	357	87.0	183	270.0	24.37
14	管理学	10.1016/j.ijpe.2008.07.008	331	122.5	86	208.5	37.01
15	管理学	10.1016/j.ijpe.2011.05.011	265	68.5	128	196.5	25.85
16	化学	10.1039/c0cc02990d	477	222.5	32	254.5	46.65
17	化学	10.1021/acs.accounts.5b00369	682	272.5	137	409.5	39.96
18	化学	10.1126/sciadv.aar3208	154	66.0	22	88.0	42.86
19	化学	10.1002/adma.200800030	307	147.5	12	159.5	48.05
20	化学	10.1021/ja0751781	495	231.0	33	264.0	46.67
21	物理学	10.1103/PhysRevD.49.6410	862	412.5	37	449.5	47.85

续表

文献序号	学科	DOI	被引频次/次	Q	P	N_f	R/%
22	物理学	10.1103/RevModPhys.89.025003	379	107.5	164	271.5	28.36
23	物理学	10.1088/0004-637X/724/2/1044	587	276.0	35	311.0	47.02
24	物理学	10.1063/1.366523	437	160.5	116	276.5	36.73
25	物理学	10.1016/j.cpc.2006.11.008	480	223.5	33	256.5	46.56
26	数学	10.1088/0031-8949/82/06/065003	339	124.5	90	214.5	36.73
27	数学	10.1023/A:1025384832106	393	156.5	80	236.5	39.82
28	数学	10.1016/j.crma.2004.08.006	643	312.5	18	330.5	48.60
29	数学	10.1002/（SICI）1097-0207	550	235.5	79	314.5	42.82
30	数学	10.1016/j.apm.2012.04.004	419	168.5	82	250.5	40.21
31	计算机科学	10.1109/JBHI.2016.2636665	392	70.5	251	321.5	17.98
32	计算机科学	10.1038/nature14541	435	44.0	347	391.0	10.11
33	计算机科学	10.1109/TIE.2007.906994	146	49.5	47	96.5	33.90
34	计算机科学	10.1016/j.cell.2009.04.048	298	129.0	40	169.0	43.29
35	计算机科学	10.1109/COGINF.2002.1039280	125	41.5	42	83.5	33.20
36	图书情报科学	10.1002/asi.23071	131	46.0	39	85.0	35.11
37	图书情报科学	10.1016/j.joi.2014.09.005	176	55.0	66	121.0	31.25
38	图书情报科学	10.1002/asi.23456	88	17.5	53	70.5	19.89
39	图书情报科学	10.1016/j.joi.2016.12.002	43	11.0	21	32.0	25.58
40	图书情报科学	10.1007/BF02017249	300	119.0	62	181.0	39.67

4　被引频次优化模型效果评估

4.1　使用数量与被引频次优化结果的关系分析

WoS数据库中论文的使用数量（Usage）是平台所有用户访问论文全文链接或保存记录的次数，捕获了用户试图获取全文的各种操作[25-26]。在间接引用结构中，追随文献C的作者在未阅读文献A原文的情况下，通过中介文献B的影响力和原文线索对文献A施加间接引用，这导致文献A的被引频次存在冗余的间接引用计数，但并不会对文献A的使用数量产生影响。因此，使用数量指标能够相对真实地记录文献A的被阅读情况和影响力。为了对比被引频次优化实验前后的评价效果，分别对优化前、后的被引频次计数结果与文献在WoS中获得的使用数量指标进行Pearson相关性分析[27]。优化前被引频次N_b、优化后被引频次计数N_f与使用数量的Pearson相关系数计算结果如表2所示。

表2　优化前、后被引频次计数值与使用数量的相关性矩阵

指标	使用数量		N_b		N_f	
	Pearson相关系数	显著性（双尾）	Pearson相关系数	显著性（双尾）	Pearson相关系数	显著性（双尾）
使用数量	1.000		0.029	0.860	0.072	0.659
N_b	0.029	0.860	1.000		0.900**	0
N_f	0.072	0.659	0.900**	0	1.000	

注：**在0.01水平（双尾）上显著相关。

表2计算结果显示：优化前被引频次N_b、优化后被引频次N_f与使用数量的相关系数分别为0.029、0.072。显然，在优化文献的间接引用频次后，文献的被引频次计数与使用数量的相关性提高。相比优化前未加处理、直接统计的论文被引频次，通过被引频次优化模型计算的结果能够较客观地反映科学文献的真实被引情况和真实影响力。本文建立的被引频次优化方法在一定程度上削弱了科学工作中文献间接引用机制带来的负面影响，在科学评价与引文分析工作中具有一定的实际应用价值。

4.2　ESI高被引论文的被引频次优化结果分析

通过样本文献的引用优化比例指标R计算结果，比

较高被引论文与非高被引论文影响力在被引频次优化实验中的变化。WoS中的高被引论文标识是指同一年同一个ESI学科中发表的所有论文按被引频次由高到低排序，排在前1%的论文。将40篇样本文献按8个学科类别进行分类，并将8个学科的文献分别按照引用优化比例由小到大排序。表3显示了每篇论文是否为高被引论文及其对应的引用优化比例。

根据统计结果，与所属学科全部样本文献相比，ESI前1%高被引论文的引用优化比例普遍相对较低。这表明高被引论文中存在的间接引用关系比较少，高被引论文大部分的被引频次来自施引文献真实的、直接的引用。例如，在医学与生物学、化学、计算科学、心理学等学科，引用优化比例较低的文献均是ESI前1%的高被引论文。引用优化比例越低，优化前、后文献被引频次的差异越小，也就意味着对应文献的间接引用关系越少。

表3　引用优化比例指标与ESI高被引论文的关系

学科	论文情况	引用优化比例/%	学科	论文情况	引用优化比例/%
医学与生物学	Y	9.63	物理学	Y	28.36
医学与生物学	Y	24.25	物理学	N	36.73
医学与生物学	N	25.27	物理学	N	46.56
医学与生物学	N	32.78	物理学	Y	47.02
医学与生物学	N	36.36	物理学	N	47.85
心理学	Y	5.60	数学	Y	36.73
心理学	Y	22.04	数学	N	39.82
心理学	N	24.07	数学	Y	40.21
心理学	N	36.40	数学	N	42.82
心理学	N	42.05	数学	N	48.60
管理学	N	24.37	计算机科学	Y	10.11
管理学	Y	25.85	计算机科学	Y	17.98
管理学	N	32.46	计算机科学	N	33.20
管理学	N	35.98	计算机科学	N	33.90
管理学	N	37.01	计算机科学	N	43.29
化学	Y	39.96	图书情报科学	Y	19.89
化学	Y	42.86	图书情报科学	N	25.58
化学	Y	46.65	图书情报科学	Y	31.25
化学	N	46.67	图书情报科学	N	35.11
化学	N	48.05	图书情报科学	N	39.67

注：Y表示高被引论文，N表示非高被引论文。

ESI前1%高被引论文的低引用优化比例验证了高被引论文多获得直接的一次引用关系，而非复杂的、冗余的多级引用。因此，科学文献的质量越高，其对应的引用关系越简单、直接和真实。此外，ESI前1%高被引论文的低引用优化比例也表明本文建立的文献被引频次优化模型能够用于识别潜在的高质量、高被引论文。具体地，在科学论文获得一定数量的引用后，可根据其引文网络进行间接引用频次优化。引用优化比例越低，则说明该论文的被引质量越好，未来成为高被引论文的潜能越大。

4.3　文献发表时间与被引频次优化结果的关系分析

为了比较文献发表时间对优化结果的影响，根据40篇样本文献的发表年份、引用优化比例指标计算结果，构建散点图，如图5所示。同时，为了进一步细致对比不同优化结果与文献发表时间的联系，根据引用优化比例计算结果的分布，将40篇科学文献的引用优化比例分为[0，0.1）、[0.1，0.2）、[0.2，0.3）、[0.3，0.4）、[0.4，0.5）5个区间，并统计各个区间中文献的发表时间分布。不同引用优化比例区间对应的文献数量及发表年份等信息如表4所示。

在图5中，根据样本文献的发表时间与引用优化比例构建线性趋势。随着文献发表时间推进，文献的引用优化比例逐渐降低。在2012年以后发表的文献形成了一个比较明显的簇团，引用优化比例分布在30%以下。因此，发表时间越晚的文献，其引用优化比例越低；发表时间越早的文献，其对应的引用优化比例越高。这是由于间

接引用关系的形成至少需要两次连续引用——样本文献在被施引文献引用之后、施引文献继续被二级施引文献引用，这两次连续引用需要的时间较长。发表时间较早的文献已积累复杂的引用联系，从而产生较多的间接引用关系。而发表时间较晚的新文献还未建立复杂的间接引用关系，引用关系中存在的间接引用频次较少。

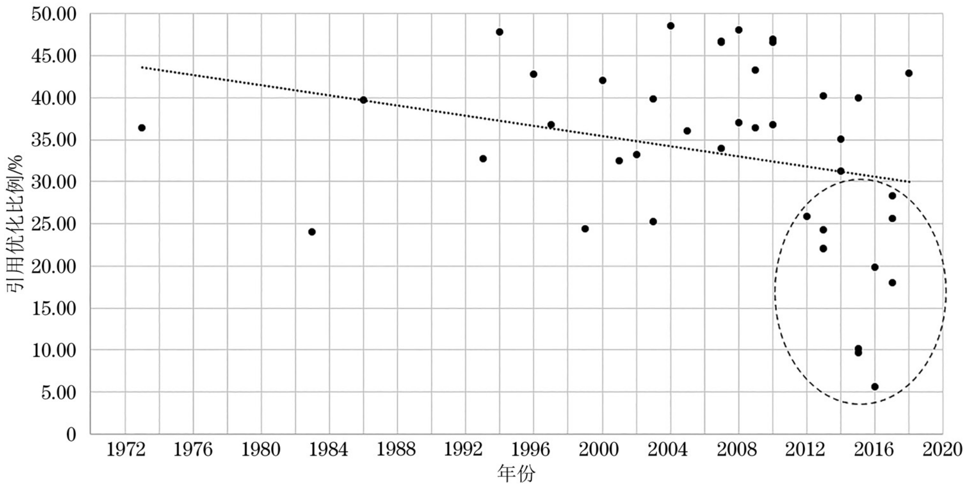

图5　引用优化比例指标与文献发表时间的关系

表4　引用优化比例指标分布区间及文献发表时间分布

引用优化比例	文献数量/篇	发表年份	文献序号	学科
[0，0.1）	2	2016	10	心理学
		2015	4	医学与生物学
[0.1，0.2）	3	2017	31	计算机科学
		2016	38	图书情报科学
		2015	32	计算机科学
[0.2，0.3）	8	2017	39	图书情报科学
		2017	22	物理学
		2013	7	心理学
		2013	1	医学与生物学
		2012	15	管理学
		2003	5	医学与生物学
		1999	13	管理学
		1983	8	心理学
[0.3，0.4）	15	2014	37	图书情报科学
		2015	17	化学
		2014	36	图书情报科学
		2010	26	数学
		2009	3	医学与生物学
		2008	14	管理学
		2007	33	计算机科学
		2005	11	管理学
		2003	27	数学
		2002	35	计算机科学
		2001	12	管理学
		1997	24	物理学
		1993	2	医学与生物学
		1986	40	图书情报科学
		1973	6	心理学

续表

引用优化比例	文献数量/篇	发表年份	文献序号	学科
[0.4，0.5）	12	2018	18	化学
		2013	30	数学
		2010	16	化学
		2010	23	物理学
		2009	34	计算机科学
		2008	19	化学
		2007	25	物理学
		2007	20	化学
		2004	28	数学
		2000	9	心理学
		1996	29	数学
		1994	21	物理学

从表4可以看到，引用优化比例在0.2以下的文献均是2015年以后发表的新文献。在[0.2，0.3）区间中，大部分文献也是发表于2010年之后。而在[0.3，0.4）和[0.4，0.5）两个区间中，2010年之后的新文献占比骤降，2010年之前发表的文献明显占绝大多数。相比发表时间久远的文献，发表时间较晚的新文献在时间上未能有机会形成较多、较复杂的引文网络，大部分获得无间接联系的一次引用关系，间接引用关系较少，引用优化比例较低。

一篇论文发表的年份越早，其被引频次可能越多。在现有的学术影响力评价工作中，由于被引频次的时间累积性问题，一些发表年份较早、资历较老的文献在被引频次上具有较大的时间累积优势[28-29]。因此，目前大

部分的科学定量评价工作忽略了一些质量较高，但受发表时间影响，未积累足够被引频次的年轻文献。发表时间较久远、较经典的文献提供了重要的研究基础，但反映科研新动态、学科新进展的年轻文献的价值仍不可埋没。科学研究正是需要通过这些新文献、新研究来对一个学科领域进行不断的探索和突破。因此，文献发表时间与引用优化比例的负向线性关系表明本文构建的科学文献被引频次优化模型能够在较大程度上过滤掉发表时间较久远文献的间接引用频次，缓解科学文献被引的时间累积性问题和马太效应问题，避免新近发表文献在被引频次计数中处于劣势地位，相对较科学、公正、真实地对不同发表时间段、不同年龄的科学文献进行评价。同时，本文构建的被引频次优化模型对高质量、年轻的科学文献具有较好的筛选能力，评价结果能够使高质量的新文献、新成果、新发现尽快被发现。

5 结语

科学文献的学术影响力反映了一篇论文在学术界和同行中的被认可程度，也是学者、科研机构、学术期刊等科研主体学术水平和科研能力的重要标志。随着科学技术快速发展，科学论文的数量持续、爆发式增长，随之而来的引用不当、引用不规范问题广泛出现。如何客观、合理地评价文献的学术影响力是一个值得深入研究的问题。

本文基于科学文献的间接引用机制，将文献的被引频次划分为直接引用频次与间接引用频次，并降低间接引用频次的计数权重，构建在复杂引文网络中具有普适性的被引频次优化模型。选择WoS数据库中多个学科、多种文献类型的40篇科学文献作为实证研究的样本，指标计算与结果分析显示：①通过被引频次优化模型计算的优化后被引频次与使用数量指标的相关系数显著提高，优化后的被引频次能够较客观地反映科学文献的真实被引情况和真实影响力；②ESI高被引论文的间接引用关系较少，其大部分引用为施引文献真实的、直接的引用，本文建立的被引频次优化模型能够用于识别潜在的高质量、高被引论文；③发表时间越晚的文献，引用优化比例越低；发表时间越早的文献，其对应的引用优化比例越高。因此，本文构建的科学文献被引频次优化模型能够在较大程度上弱化发表时间较久远文献的间接引用关系的重要性，缓解学术影响力评价中的马太效应和时间累积性问题，对高质量的、年轻的学术成果具有较好的筛选能力。

本研究存在以下不足之处。首先，基于间接引用机制的文献被引频次优化模型将所有存在冗余关系的间接引用视为等同，并统一降低其被引频次的计数权重。然而，在实际引用情境中，部分间接引用关系可能是由文献A真实影响力产生的，这种情况下降低其权重有失公正。在未来工作中，将针对不同引用情境对被引频次优化模型进一步精细化。其次，对于单篇科学文献的影响力评价工作，还需要不断地扩大数据样本量来保证研究结论的全面性，同时还须深入文献具体的内容结构进行文本语义分析，挖掘科学文献的潜在价值。

参考文献

[1] KI E J，PASADEOS Y，ERTEM-ERAY T. The structure and evolution of global public relations：a citation and co-citation analysis 1983–2019[J]. Public Relations Review，2021，47（1）：102012.

[2] HUANG Y，BU Y，DING Y，et al. Exploring direct citations between citing publications[J]. Journal of Information Science，2021，47（5）：615-626.

[3] SONG Y H，WU L J，MA F. A study of differences between all-author bibliographic coupling analysis and all-author co-citation analysis in detecting the intellectual structure of a discipline[J]. The Journal of Academic Librarianship，2021，47（3）：102351.

[4] 温芳芳. 国际化背景下我国图书情报学与世界各国 研究相似性的测度与比较：基于1999—2018年Web of Science论文的耦合分析[J]. 情报学报，2020，39（7）：687-697.

[5] 逯万辉，谭宗颖. 基于深度学习的期刊分群与科学知识结构测度方法研究[J]. 情报学报，2020，39（1）：38-46.

[6] RIVKIN A. Manuscript referencing errors and their impact on shaping current evidence[J]. American Journal of Pharmaceutical Education，2020，84（7）：877-880.

[7] 段庆锋，潘小换. 文献相似性对科学引用偏好的影响实证研究[J]. 图书情报工作，2018，62（4）：97-106.

[8] 刘宇，李武. 引文评价合法性研究：基于引文功能和引用动机研究的综合考察[J]. 南京大学学报（哲学·人文科学·社会科学），2013，50（6）：137-148，157.

[9] 卫军朝，蔚海燕. 科学结构及演化分析方法研究综述[J]. 图书与情报，2011（4）：48-52.

[10] LIU Y M，CHEN M. Applying text similarity algorithm to

analyze the triangular citation behavior of scientists[J]. Applied Soft Computing，2021，107：107362.

[11] FARYS R，WOLBRING T. Matthew effects in science and the serial diffusion of ideas：testing old ideas with new methods[J]. Quantitative Science Studies，2021，2（2）：505-526.

[12] 姚建文，黄筱玲，吴丽萍. 论去除论文引用泡沫：基于客观公正评价科技人才的视角[J]. 情报理论与实践，2013，36（8）：11-14，5.

[13] RADICCHI F，FORTUNATO S，MARKINES B，et al. Diffusion of scientific credits and the ranking of scientists[J]. Physical Review E，2009，80（5）：056103.

[14] 王菲菲，王筱涵，刘扬. 三维引文关联融合视角下的学者学术影响力评价研究：以基因编辑领域为例[J]. 情报学报，2018，37（6）：610-620.

[15] MIN C，DING Y，LI J，et al. Innovation or imitation：the diffusion of citations[J]. Journal of the Association for Information Science and Technology，2018，69（10）：1271-1282.

[16] 俞立平，王龙华. 一种基于惩罚函数的降低影响因子操纵方法：合理影响因子[J]. 情报理论与实践，2022，45（6）：67-73.

[17] BONZI S. Characteristics of a literature as predictors of relatedness between cited and citing works[J]. Journal of the American Society for Information Science，1982，33（4）：208-216.

[18] MACROBERTS M H，MACROBERTS B R. Quantitative measures of communication in science：a study of the formal level[J]. Social Studies of Science，1986，16（1）：151-172.

[19] WAN X J，LIU F. Are all literature citations equally important? automatic citation strength estimation and its applications[J]. Journal of the Association for Information Science and Technology，2014，65（9）：1929-1938.

[20] 罗杰斯. 创新的扩散[M]. 唐兴通，郑常青，张延臣，译. 北京：电子工业出版社，2016：323.

[21] 刘强. 传播学受众理论论略[J]. 西北师大学报（社会科学版），1997，34（6）：97-101.

[22] 张益明. 基于两级传播理论的卷烟品牌口碑传播[J]. 中国烟草学报，2015，21（1）：112-118.

[23] 刘运梅，马费成. 面向全文本内容分析的文献三角引用现象研究[J]. 中国图书馆学报，2021，47（3）：84-99.

[24] 刘运梅，张帅，司湘云，等. 基于内容标注的三角引用动机研究方法探析[J]. 图书情报工作，2021，65（10）：48-55.

[25] 付中静. WoS数据库收录论文文献级别用量指标与被引频次的相关性[J]. 中国科技期刊研究，2017，28（1）：68-73.

[26] 段鑫龙. Web of Science-5.19更新介绍[EB/OL]. [2024-10-10]. http://v.qq.com/x/page/n0168 gbqol0.html? ptag=biog_sciencenet_cn.

[27] 蔡智澄，何立民. 相关性分析原理在图书情报分析中的应用[J]. 现代情报，2006，26（5）：151-152，156.

[28] 韩毅，夏慧. 时间因素视角下科研人员评价的Pt指数研究[J]. 中国图书馆学报，2015，41（6）：73-85.

[29] 邱均平，缪雯婷. 文献计量学在人才评价中应用的新探索：以“h指数”为方法[J]. 评价与管理，2007，5（2）：1-5.

作者简介

刘运梅，女，博士，讲师，研究方向：科学计量与知识发现，E-mail：1972165675@qq.com。
李冉，女，硕士研究生，研究方向：信息计量学。
秦佳佳，女，硕士研究生，研究方向：信息计量学。

Construction and Empirical Research of Citation Frequency Optimization Model of Scientific Literature Based on Indirect Citation Mechanism

LIU YunMei LI Ran QIN JiaJia
(School of Cultural Heritage and Information Management, Shanghai University, Shanghai 200444, P. R. China)

Abstract: In order to objectively evaluate the real influence of scientific literature and weaken the Matthew effect in academic evaluation, this paper constructs a citation frequency optimization model of scientific literature by reducing the counting weight of indirect citation relationships, proposes a citation optimization ratio index, and tests the effectiveness of this model through empirical research. The findings are as follows. Firstly, the results after optimizing the indirect citation frequency weights can objectively reflect the real citation and real influence of scientific literature. Secondly, the citation frequency optimization model can be used to identify potentially high-quality and highly cited papers. Thirdly, the citation frequency optimization model can weaken the time accumulation and Matthew effect of citations in academic evaluation to a large extent, and has high evaluation fairness and better screening ability for young literature.

Keywords: Scientific Literature; Indirect Citation; Citation Frequency; Academic Evaluation

（责任编辑：王玮）