论文早期下载量可否预测后期被引量？

——以图书情报领域期刊为例

Can Downloads Predict Subsequent Citations: A Case Study on Journals of Library and Information Science

熊泽泉[1,2,3]　段宇锋[2,3]

(1.华东师范大学图书馆,上海,200241;　2.华东师范大学经济与管理学部,上海,200241;
3.华东师大学术评价与促进研究中心,上海,200241)

[摘要]　以2006—2008年出版的图情领域期刊论文作为研究对象,采用两步聚类法对单篇论文在7年内的绝对下载量与被引量进行聚类分析。在此基础上,基于动态视角对不同下载模式和不同引文模式下论文的下载量与被引量的相关性进行了研究。结果发现,累积下载量与累积被引量存在线性相关性,且两者相关性随时间的增长而增强;平均下载量最高及平均下载量最低的下载模式比平均下载量居中的下载模式,下载量与被引量相关性更高,且这一现象在发文初期更加显著;不同引文模式下也有类似发现。本研究认为:随着时间的增长,论文的下载量与被引量在表征论文的使用和有用程度上逐渐达到统一,读者的下载行为更多地转化成为引用行为;论文发表初期高下载及低下载论文的下载量可作为后期被引量预测的依据之一;而论文发表初期下载量居中的论文,下载量与被引量相关性较差,不适合利用其下载量数据对被引量进行预测。

[关键词]　被引量　下载量　相关性　图书情报领域　下载模式　被引模式

[中图分类号]G354.2　[文献标识码]A　[文章编号]1003-2797(2018)04-0032-11　**DOI**:10.13366/j.dik.2018.04.032

[**Abstract**]　Taking papers from 2006 to 2008 in the field of library and information science journalsasresearch objects, this paper conducts a two-step cluster analysis on the downloads and citations of every paper. Then, the correlation between downloads and citations in various patterns has been studied in a dynamic perspective. It has been found that there is a linear relationship between cumulative downloads and cumulative citations, which increases with time. Stronger correlation between downloads and citations has been observed in the pattern with the most and fewest average downloads, and this phenomenon is more prominent at the early stage of publication. Similar findings were found in different citation patterns. The result suggests that downloads and citations tend to be unified in the evaluation of the use and usefulness of articles and readers' download behavior is likely to convert into citation behavior; the highest and lowest downloads of papers can be used as a predictor for the latter citations, while isn't applicable in papers of medium downloads.

[**Keywords**]　Citation; Downloads; Correlation; Library and information science; Download pattern; Citation pattern

[基金项目]　本文系2017年度长三角地区高校图书馆联盟—图书情报研究基金项目资助成果之一。

[作者简介]　熊泽泉,硕士,馆员,研究方向:信息管理与科学计量,Email:zqxiong@Library.ecnu.edu.cn;段宇锋,博士,教授,研究方向:信息资源管理,Email:yfduan@infor.ecnu.edu.cn。

1 前言

被引量作为论文学术影响力的主要评价指标，在人才评审、科研立项、科研奖励等过程中发挥着重要的作用，并衍生出影响因子、H指数等一系列指标。但是其时滞性、片面性及地域差异性等问题也逐渐受到学者的诟病[1]。随着互联网的发展，学术论文的电子化日渐普及，几乎所有的期刊论文都能够通过网络数据库被获取[2,3]，人们对学术文献的使用得以被服务器记录，这使得研究者开始关注一个新的学术文献计量指标——Usage Metrics[4,5]，即使用量指标。使用量指标能够即时反映论文被使用的情况，亦能在一定程度上反映在科学研究中被使用但未被体现在引用上的价值。随着使用量指标被众多学者所接受，一些数据库厂商也紧跟步伐，推出了基于自身平台的使用量指标，如Web of Science平台的Usage指标[6-9]，Springer的Download指标[10]，Nature的文章页面浏览量指标（Article Page Views）[11,12]，PLOS的Article Level Metrics[13]，以及中国知网的总下载量指标[14,15]、热度指标等。

一般而言，学术论文在被引用前，对其的使用行为包括浏览、下载、阅读等[16]。以论文为载体，知识/信息刊出后，首先被读者浏览发现，其中一部分读者被某一论文的标题或文摘信息所吸引，进而会进行下载、阅读，获取该部分知识/信息，其中更小一部分读者会在其撰写的论文中进行引用，然后经同行评议后发表，知识/信息进入一个新的使用－引用的循环中[17]。

在这个循环过程中，浏览行为夹杂着太多的随意性，阅读行为则难以统计，而下载行为则更具针对性也易于记录[18]。虽然读者可以通过共享、文献传递等方式获得所需论文，但是从总体上看，从数据库下载仍为互联网时代获取论文最主要的途径，下载量也是最接近、且最易获取的反映论文实际使用量的指标。因此，在已有研究中，一些学者将论文的使用量等同于下载量[19-21]，更多的学者直接采用下载量作为主要的使用量指标，来探讨其合理性[22]、影响因素[23-26]以及与被引量的相关性[27-31]。上述研究为我们了解论文下载量数据的特性等方面提供了丰富的信息，且基本上都认为下载量与被引量之间存在着某种程度的相关性，一些学者甚至提出可以用论文早期下载量来预测其后期被引量，以弥补被引量的时滞性问题[32-35]。

但是，上述研究在数据的选择和处理上仍有不足之处，从而导致不同研究结果中论文下载量与被引量相关性的显著水平存在较大差异[36]，使得利用论文早期下载量预测后期被引量的可行性存在一定的争论。

首先，由于受数据库供应商的限制，早期的一些研究只能从不同的数据库获取下载量和被引量数据。如Moed以期刊Tetrahedron Letters为例，其下载量来源于ScienceDirect，而其被引量来源于SCI数据库，结果显示论文发表25个月后两者的Spearman相关系数仅有0.220[37]；Brody等则分别以arXiv.org和Citebase作为其下载量和被引量数据的来源，来探讨利用早期下载量预测后期被引量的可行性，发现两者的相关系数从论文发表1个月后的0.270上升到24个月后的0.440[38]；Guerrero-Boteh和Moya-Anegon[39]从ScienceDirect和Scopus获取下载量和被引量数据来研究两者之间的相关性，发现在期刊水平上两者的相关系数为0.780，而在论文水平上两者的相关系数仅为0.480；Schloegl等利用ScienceDirect提供的下载量数据，结合JCR或Scopus提供的被引量数据，进行了一系列相关研究，相关系数范围为0.600—0.800[40-42]。这些研究虽然通过数据处理，使得每一篇论文的下载量与被引量能够一一对应，但由于不同数据库平台的使用者重合度无法测量，这样得出的结果可能存在偏差。

其次，在单篇论文的水平上，大多数数据库只提供即时的累积下载量数据，并未提供分年下载数据，这使得研究者要么只能对某一时间剖面的下载量与被引量的相关性进行分析[43]，要么只能从期刊水平上进行相关性的动态分析，而从单篇论文水平上的动态相关性分析则不多见。而事实上，读者使用的是论文本身，并非期刊整体，一本期刊所刊发的论文不会集中于完全一模一样的主题和对象，因此在同一期刊上，也存在着不同的下载模式[44]和引文模式，这些具有不同下载模式和引文模式的论文在下载量和引用量的相关性上有何异同尚未见研究报道。

因此，本文拟通过对图书情报领域中文学术期刊

论文下载量与被引量相关性的动态变化过程进行研究，来探寻不同下载模式和引文模式下，下载量与被引量相关性的变化规律。不同于已有文献，本研究的下载量和被引量数据均来源于同一数据库——中国学术期刊网络出版总库(China Academic Journal Network Publishing Database，CAJD)，这一世界上最大的连续动态更新的中国学术期刊全文数据库。本研究拟研究的问题如下：

(1)采用来源于同一数据库的论文下载量与被引量的相关性，是否高于采用不同数据来源的论文下载量与被引量的相关性？不同下载模式和被引模式下，论文的下载量与被引量的相关性是否存在差异？

(2)论文的早期下载量，在不同下载模式和被引模式下，是否都能用于预测论文后期被引量？

2　数据和方法

2.1　数据来源与处理

以中国学术期刊网络出版总库作为数据源，选择其中的11种图书情报领域期刊在2006—2008年发表，且在2015年12月31日前获得过被引和下载的9042篇论文作为研究对象，选择依据主要是由于这些期刊创刊时间较长，在数据库中收录完整，且其出版日和上线日基本一致，从而能够获得较为真实的下载量及被引量数据。而《图书情报工作》、《中国图书馆学报》等期刊因为出版到上线的滞后期较长，未选择其作为研究对象。将该原始数据集命名为DataSet1。

DataSet1中，每篇论文所涉及的数据包含论文的基本题录信息以及该论文在2006—2015年每一自然年的下载量和被引量，分别加总每一自然年的下载量和被引量，得到每篇论文自出版时到2015年12月31日的总下载量和总被引量；由于不同论文出版月份不同，有的在年初出版，有的在年末出版，因此出版月份较晚的论文在出版当年的下载量和被引量无法体现其真实数量，为了更加准确地呈现论文在出版后1年内的下载量和被引量，本文假设每篇论文下载量和被引量在一年的不同月份不存在差异，首先采用如下公式计算绝对下载量：

$$D'_{Y+1}=D_Y+\frac{D_{Y+1}}{12}\times(12-M)$$

出版后第2年的绝对下载量为：

$$D'_{Y+2}=\frac{D_{Y+1}}{12}\times M+\frac{D_{Y+2}}{12}\times(12-M)$$

其中M表示论文出版月份，并以此类推获得每篇论文出版后3—7年内的绝对下载量；然后采用同样的处理方法获得了每篇论文出版后1—7年内的绝对被引量，汇总获得新数据集DataSet2(由于2008年发表的论文截止至2015年12月31日只有7年的下载及被引数据，所以将所有论文统一统计年限为7年)。以2008年5月发表于《情报科学》的论文“自然语言检索中的中文分词技术研究进展及应用”为例，在DataSet2中，各值计算方法如下：

$$D'_{2008+1}=D_{2008}+\frac{D_{2008+1}}{12}\times(12-7)=137+\frac{177}{12}\times5=210.75$$

$$D'_{2008+2}=\frac{D_{2008+1}}{12}\times7+\frac{D_{2008+2}}{12}\times(12-7)=\frac{177}{12}\times7+\frac{96}{12}\times5=143.25$$

$$C_{2008+1}=C_{2008}+\frac{C_{2008+1}}{12}\times(12-7)=3+\frac{10}{12}\times5=7.17$$

$$C'_{2008+2}=\frac{C_{2008+1}}{12}\times7+\frac{C_{2008+2}}{12}\times(12-7)=\frac{10}{12}\times7+\frac{14}{12}\times5=11.67$$

其在DataSet1、DataSet2中，各数据项如表1所示。

2.2　分析方法

2.2.1　聚类分析

采用IBM SPSS Statistics 23提供的两步聚类法(Two-Step Cluster)，分别根据DataSet2中每篇论文出版后每年的绝对被引量和绝对下载量进行聚类分析，获得不同的下载模式和引文模式。具体聚类步骤为：选择DataSet2中每年的绝对下载量作为连续变量，聚类准则采用施瓦兹贝叶斯准则(BIC)，由于之前对数据已经进行了清理，因此对离群值不再使用噪声处理，评估字段采用唯一的文件识别号，并勾选创建聚类成员变量，最终得到每一篇论文所属下载模式。采用同样的步骤获得每一篇论文所属引文模式。聚类质量通过内聚与分离轮廓测量指标(Silhouette measure of cohesion and separation)进行评价[45]，该值大于

0 表明聚类质量较好。

2.2.2 相关性分析

采用 Spearman 相关系数从单篇论文的角度对总下载量与总被引量的相关性进行分析，同时分别对单篇论文的下载量及被引量进行排序，分析两者的秩序相关性。然后分别研究了总样本集，以及不同下载模式和不同引文模式下，论文下载量与被引量相关性的变化趋势。

表 1 论文在不同数据集中的表示方式的示例

DataSet 1								
	2008(年)	2009	2010	2011	2012	2013	2014	2015
下载量	137	177	96	73	84	102	123	99
被引量	3	10	14	10	13	10	10	11
DataSet 2								
	第 1 年	第 2 年	第 3 年	第 4 年	第 5 年	第 6 年	第 7 年	
下载量	210.75	143.25	86.42	77.58	91.50	110.75	113.00	
被引量	7.17	11.67	12.33	11.25	11.75	10.00	10.42	

3 结果分析

3.1 下载模式

样本中论文总下载量为 2,735,208 次，篇均下载量为 302.50 次，单篇最高下载量为 5087 次。从下载量的分布区间来看，“1≤下载量＜1000”的论文数量为 8838 篇，占样本总量的 97.74%，“1000≤下载量＜2000”的论文数量为 170 篇，“2000≤下载量＜3000”的论文数量为 27 篇，“3000≤下载量 4000”的论文数量为 6 篇，下载量达到 4000 次以上的论文数量为 1 篇，可以看出，下载量的分布遵循布拉德福分散定律。

在前期研究中，作者对学术论文的下载模式进行了分析[46]。结果表明，样本论文基于绝对下载量的下载模式可聚类为如图 1 所示的 4 种。其中，模式 D1、模式 D2 和模式 D3 的下载量均为第一年最高，然后呈逐年下降趋势，三者变化趋势基本一致，拟合曲线均为负幂函数形式，主要是绝对数量上的差异；模式 D4 的下载量则呈现先降低后上升的趋势，下载量在第四年到达最低点后又逐渐上升，到第七年的下载量接近第一年的下载量，其函数关系拟合度最高的为二项式。

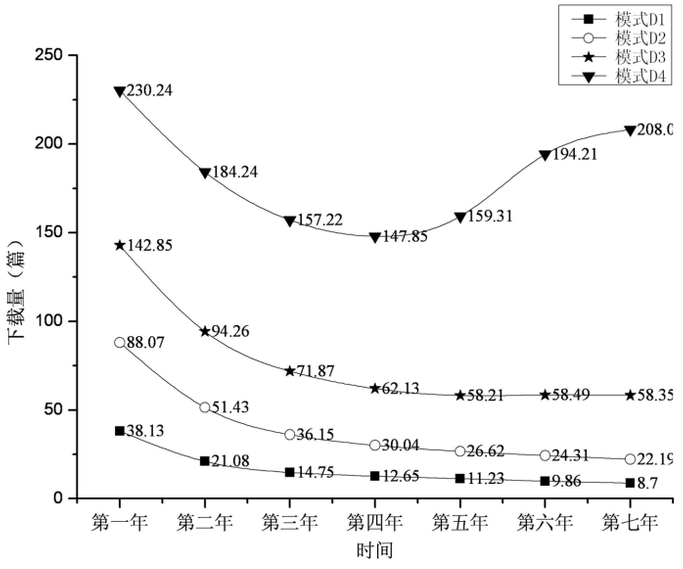


图 1 四种下载模式平均下载量变化趋势

从总体上看，模式 D1、模式 D2 和模式 D3 可以认为是常态的下载模式，反映了学术期刊论文在使用上的一般性规律，即读者倾向于使用最新出版的期刊论文，以获得最新的研究动态。模式 D4 呈现一种先降后升的特殊下载模式，考虑到其平均下载量也显著高于其他下载模式，表明这一类下载模式可能包含了更加丰富的下载动机，多重下载动机的叠加一方面使得该模式的论文获得了更高的下载量，另一方面也改变了其常规的老化模式。Moed 和 Schloegl 等在对其他学科的期刊论文的研究中也有类似发现，并且认为被引量的增加对于下载量的再次上升具有直接作用[47-49]。

3.2　引文模式

样本总被引量为 111，790 次，篇均被引量为 12.36 次，单篇最高被引量为 620 次，与下载量最高的论文为同一论文。从被引量的分布区间来看，“1≤被引量＜50”的论文数量为 8805 篇，占样本总量的 93.65%，“50≤被引量＜100”的论文数量为 190 篇，“100≤被引量＜150”的论文数量为 36 篇，“150≤被引量＜200”的论文数量为 9 篇，被引量达到 250 次以上的论文数量为 2 篇。被引量大体遵循布拉德福分散定律，稍有偏离。

基于绝对被引量变化趋势，作者发现样本具有 3 种引文模式（见图 2）。这 3 种模式的变化趋势基本一致，均呈先上升后下降的规律，但总被引量相对高的论文（模式 C1）达到其被引峰值较晚（3 年），模式 C2 和模式 C3 更早地达到了其被引峰值。

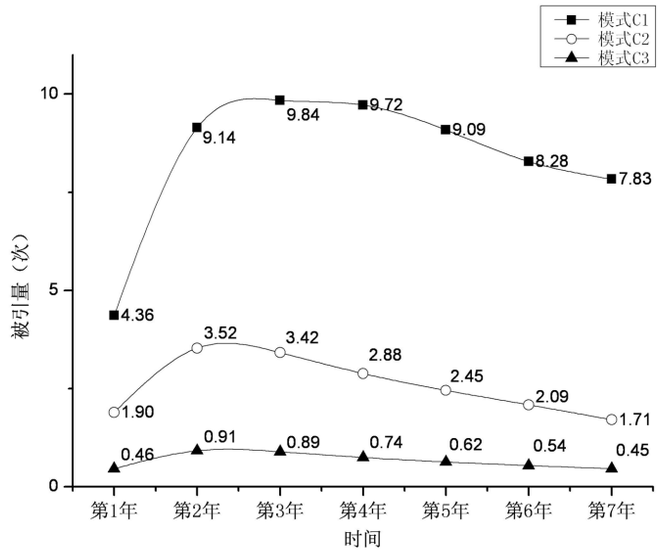


图 2　三种引文模式平均被引量变化趋势

这 3 种引文模式都属于“经典引文曲线”[50]，反映了文献老化规律的普遍存在，同时发现在同一学科中，平均被引量越高，其老化趋势越缓慢。在其他关于引文模式的研究中，发现除“经典引文曲线”外，还存在睡美人型、双峰型、波型等不规则引文模式[51]，这些不规则引文曲线的被引量在某一特殊时期，受到外在因素的影响，突然增加或者减少，从而产生了特殊的波动。譬如 Mazloumian 等人发现，诺贝尔得主的标志性论文被引次数呈爆炸式增长，同时也会带动该科学家其他论文被引次数的增加[52]，引用动机理论的研究也证实了这一点[53]，但是李江等人的研究却发现被引次数的增长并不能归因于获诺贝尔奖，引文曲线的类型与被引用对象的质量没有直接关联[54]。究竟是哪些因素在引文模式的形成和变化中起主要作用，还有待进一步研究。

3.3　下载量与被引量的总体相关性

样本平均下载/被引比为 42.54，最高下载/被引

比为522,最低下载被引比为3.33。单篇论文下载量与被引量的相关系数为0.712,秩序相关系数为0.743,总体上表现出了较高的相关性。

在前期研究中,我们发现下载量呈对数正态分布,因此我们将下载量与被引量分别进行对数转换后绘制两者的散点图,结果如图3所示,可以直观地观察到两者之间的相关性,通过曲线拟合,两者关系可表示为线性函数 $y=1.0595x-1.6432$($R^2=0.517$)。

同时对论文出版1－6年后累积下载量与累积被引量进行对数转换后分别作出散点图(见图4),可以看出在论文发表初期,线性关系较弱,并且论文的下载/被引比较高,随着出版时间的增加,线性关系逐渐加强(相关系数从0.284逐渐上升至0.673),且下载/被引比逐渐上升。由于论文出版初期下载具有较大的随意性,与被引动机差异较大;随着时间的延续,下载动机与引用动机的契合度增强。

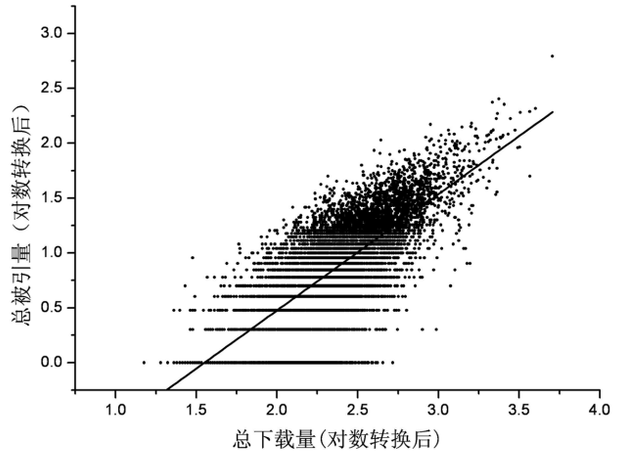


图3　总下载量与总被引量对数散点图

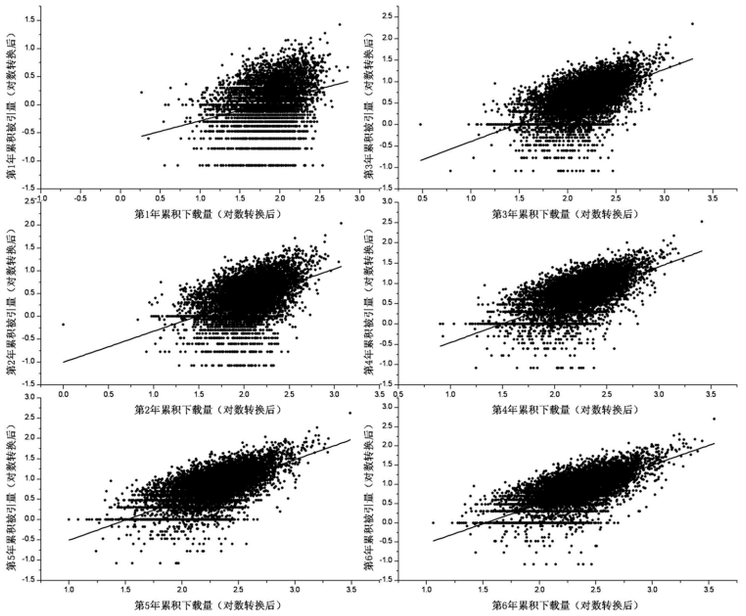


图4　论文出版1—6年后累积下载量与累积被引量对数散点图

3.4　下载量与被引量的变化趋势及两者相关性的动态变化

利用双 Y 轴图可以直观地呈现下载量和被引量不同的变化趋势。如图 5 所示，下载量在论文出版后的第 1 年即达到峰值，然后缓慢下降；被引量的峰值稍有滞后，在第 2 年达到峰值，第 3 年开始直线下降。两者在 7 年的时间窗口内均表现出老化现象，下载量在初期老化速度更快。

为了研究下载量与被引量的相关性在论文出版后不同年份的差异，本文分别计算了论文发表后每一年的下载量与被引量之间的相关系数，同时考虑到两者的交互作用可能存在滞后性，又对第 N 年下载量与第 N+1 年被引量的相关系数，以及第 N 年被引量与第 N+1 年下载量的相关系数进行了计算，结果如图 6 所示。可以看出，在论文发表初期，第 N 年下载量与第 N+1 年被引量的相关系数较高，暗示前一年的下载量可能对后一年的被引量具有一定的促进作用；而随着时间的发展，这种下载量效用的滞后性便逐渐消失，表现为第 N 年下载量与第 N 年被引量的高度相关性。

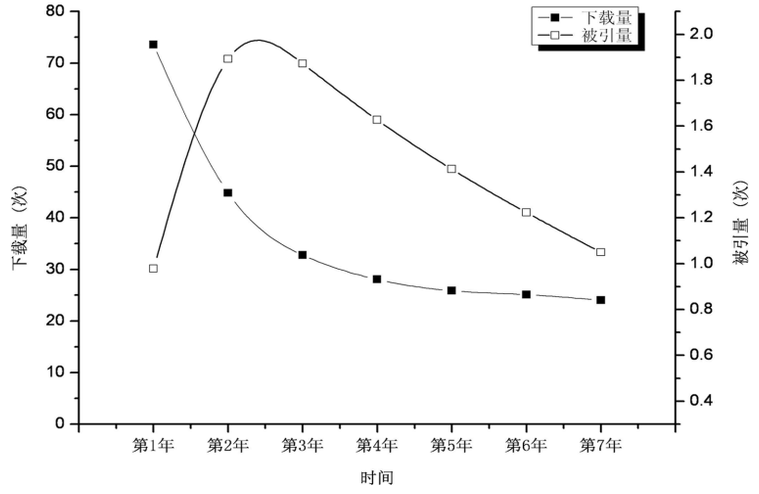


图 5　分年下载量与被引量变化趋势

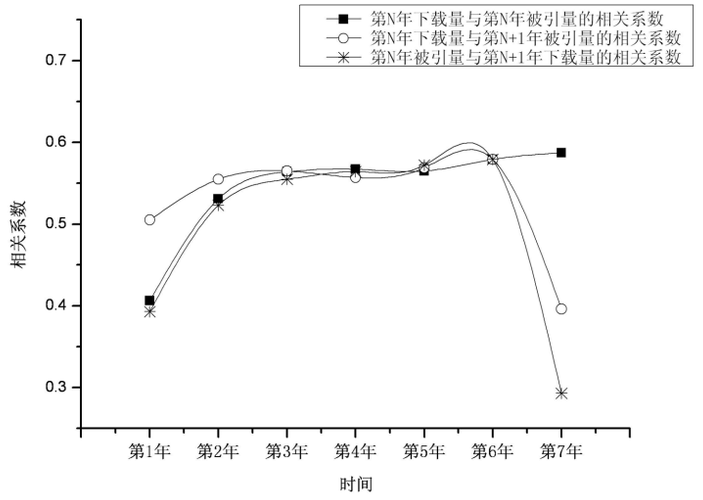


图 6　分年下载量与分年被引量相关系数变化趋势

考虑到读者多数以论文已有的累积下载量与累积被引量来对论文的影响力进行评价，因此本文对累积下载量与累积被引量的相关性也进行了动态分析，结果如图 7 所示。可以看出，虽然在论文发表初期，第 N 年累积下载量与第 N+1 年累积被引量的相关系数在三类相关系中最高，但随着时间的发展，这三类相关系数最终都达到同一水平。

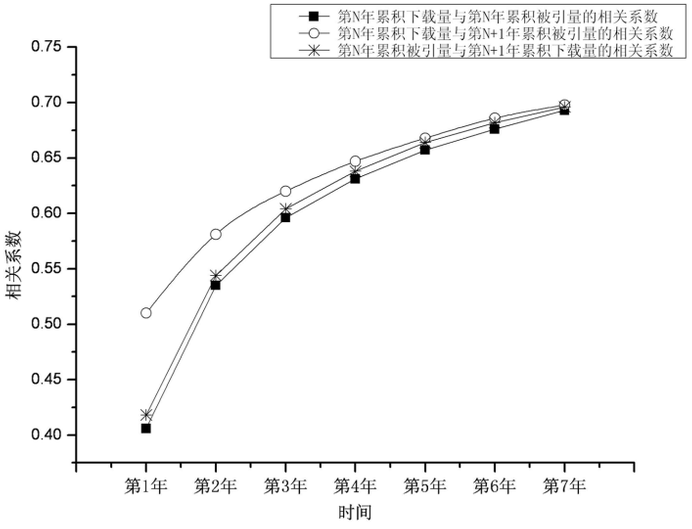


图 7　累积下载量与累积被引量相关系数变化趋势

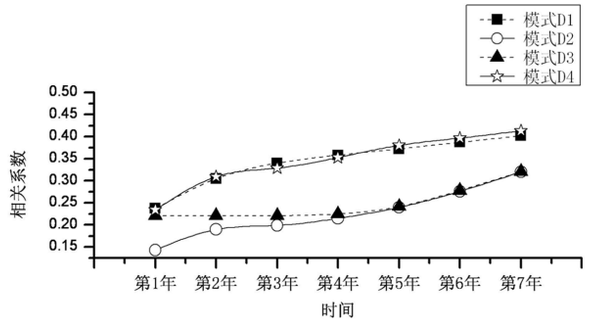
论文相关性的变化可以从读者下载及引用的动机来进行分析。在论文刚发表时，论文的下载量与被引量都接近于零，读者主要基于论文的内容特征及所在期刊来决定是否下载[55]，此时下载量迅速积累，其被引的概率也随之增大，但是由于施引文献从撰写到投稿，再到出版，需要经历较长的一段时间，此时被引量仍处于较低水平，而后在出版后的 2—3 年逐渐达到被引高峰，因此，此时下载量与被引量的相关系数仅为 0.4 左右。在此之后，下载量与被引量均有了不同程度的分化，高下载量论文与高被引论文凭借其“累积优势”得到更多下载和被引，两者在后期下载及引用行为中的权重及相关性得到进一步加强，相关系数上升到 0.7 左右，并逐渐趋于稳定，此时无论是下载量对于被引量的推动作用，还是被引量对于下载量的牵引作用，效用都已充分发挥。

3.5　不同下载和引文模式下论文下载量与被引量相关性动态变化

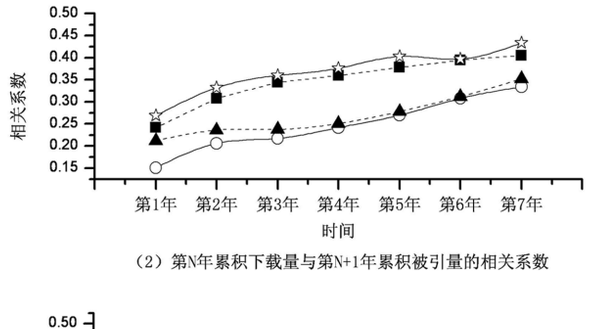
在本部分研究中，作者进一步对不同下载模式下论文下载量与被引量的相关系数的变化规律进行了研究（见图 8）。可以看出，无论是否考虑下载量或被引量作用的滞后效应，4 种下载模式的相关系数均随出版时间的增长而增加，同时明显地归为两个集群：相关系数较高的模式 1 和模式 4，分别代表了平均下载量最高和平均下载量最低的两种下载模式；相关系数较低的模式 2 和模式 3，则分别代表了下载量居于中间水平的两种下载模式。但是，由于样本数据的变异系数较大，各模式下的下载量与被引量相关系数均不高。

本文对不同引文模式下论文下载量与被引量的相关系数的变化规律进行了探讨（见图 9）。与不同下载模式下的研究结果类似，不同引文模式下的 3 类相关系数均随出版时间的增长而增加；在论文发表初期，属于引文模式 C2 的论文（被引量居于中间水平），其累积下载量与累积被引量相关系数显著低于引文模式 C1 和 C3 的论文，而在论文发表 6—7 年后，各引文模式下论文累积下载量与累积被引量相关系数达到几乎同一水平。

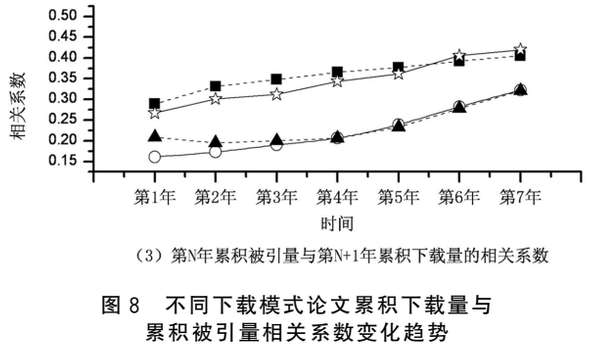
已有学者研究发现，论文下载量与被引量的相关性存在学科差异[56]，暗示在利用早期下载量预测后期



（1）第N年累积下载量与第N年累积被引量的相关系数

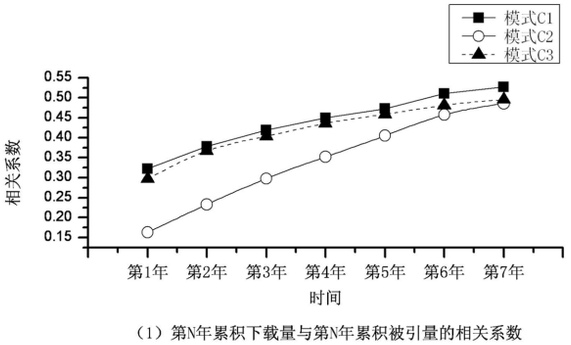


（2）第N年累积下载量与第N+1年累积被引量的相关系数

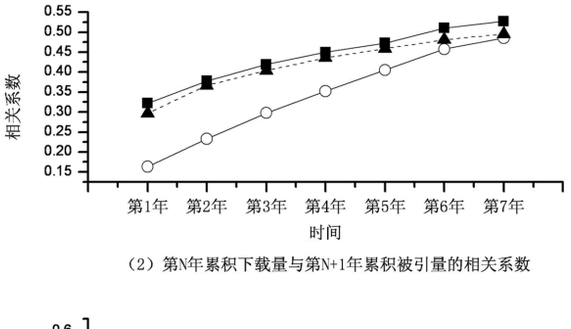


（3）第N年累积被引量与第N+1年累积下载量的相关系数

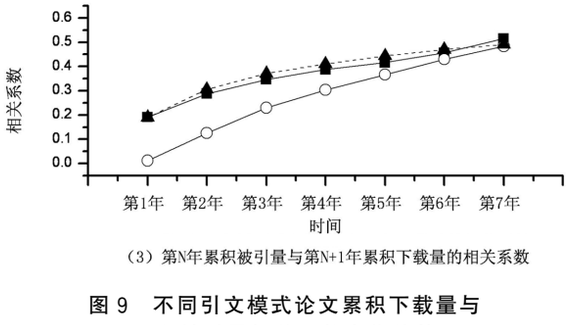
图 8　不同下载模式论文累积下载量与累积被引量相关系数变化趋势



（1）第N年累积下载量与第N年累积被引量的相关系数



（2）第N年累积下载量与第N+1年累积被引量的相关系数



（3）第N年累积被引量与第N+1年累积下载量的相关系数

图 9　不同引文模式论文累积下载量与累积被引量相关系数变化趋势

被引量时，不同学科之间的预测准确度也必然存在较大的学科差异。本研究进一步扩展了这一观点：就算排除了学科差异，对于具有不同早期下载量和下载模式的论文，其预测准确度也会存在差异。

Brody 等认为，当下载量与被引量具有较好的相关性时，前期下载量可作为后期被引量的一个预测依据；而当下载量与被引量的相关性较差时，下载量可作为一个独立的“使用影响力”指标，弥补被引量的不足[57]。在本研究中，作者发现，对于高下载及低下载论文，下载量与被引量具有中度的相关性；而下载量居中的论文，下载量与被引量相关性较差；不同引文模式下的研究也有类似发现。

因此，虽然在长期时间窗口内，下载量与被引量具有较强的相关性，但是在短期内，下载量与被引量的相关性并不高，特别是下载量与被引量处于中间水平的论文（分别占样本总量的 49% 和 32%）相关性更弱，因此不建议采用论文出版后 5 年内的数据来进行后期被引量预测。同时，由于累积下载量与累积被引量的相关性强于分年下载量与分年被引量的相关性，作者建议在后期被引量预测中采用累积数据而非分年数据。

从知识/信息的传递过程来看，对于单篇论文，下载行为早于引用行为。一般认为如果下载量与被引量之间存在正相关性，那么就应该是下载为因，引用为果，先期的下载量对于后期的被引量具有某种程度上的决定作用。因此，在预测被引量的研究中，部分学者基于先期的下载量来预测后期的被引量[58-61]。

但实际上,以论文为载体的知识/信息是处在一个使用—引用的循环中,两者的相关性也与因果性无关,论文后期被引量可能与先期的下载量有关,同样先期的被引量也可能直接或间接影响后期的下载量[62,63],单篇论文的下载量与被引量的关系可能类似于DNA的双螺旋结构——两者通过某种函数关系紧密相关,同时相互促进,螺旋式上发展,而驱动两者向上发展的原始动力,还在于论文本身的质量。因此,如果要利用论文的早期下载量与被引量来对长期的被引量进行预测,必须同时考虑到论文内容特征,期刊特征及学科等因素,作者将在下一步工作中开展此方面研究。

4 结论与展望

通过上述分析,本研究主要获得如下结论:

第一,下载量与被引量是分别从不同角度对论文的测度,下载量是从读者的角度,测量论文被使用的程度;被引量是从作者的角度,评价论文对其有用的程度。因此,两者既具有一定的相关性,又具有不同的变化规律。本研究发现累积下载量与累积被引量存在线性相关性,且两者相关性随时间的增长而增强(从出版后第一年的0.4左右上升到第7年的0.7左右),表明这两个指标随着时间的增长在表征论文的使用和有用程度上逐渐达到统一,即随着时间的增长,对于论文更多的使用是有效的使用,读者的下载行为更多地转化成为其引用行为。

第二,不同下载模式下,下载量高或低的论文累积下载量与累积被引量的相关系数高于下载量居中的论文;不同引文模式下,被引量高或低的论文累积下载量与累积被引量的相关系数同样高于被引量居中的论文,但是这一差距随着时间的增长逐渐消失,表明随着时间的延续,下载动机与引用动机的契合度增强。因此,在后期被引量预测时,建议选择出版后5年及以上的累积数据进行预测。

本研究基于论文下载量与被引量相关性的动态分析,对能否利用论文早期下载量预测后期被引量进行了探讨,认为早期的高下载和低下载论文更具有可预测性。因此,在后续的研究中我们将对不同下载模式下论文的早期下载量和后期被引量进行回归分析,以便更好地回答本研究中的问题。

随着信息时代的发展,下载量等基于学术文献使用的新型数据越来越受到人们的重视,其数据的即时性在领域热点分析、读者行为分析等方面具有引文分析无法比拟的优势,但在学术文献影响力评价方面仍存在一定的局限性,如数据不透明、易被人为操纵等。如何合理地利用这些新型数据,综合引文指标、补充计量学指标来构建学术文献影响力多维评价体系,将成为科学计量学领域的研究热点之一。

致谢:本研究数据由中国知网(CNKI)提供,在此表示感谢!

支撑数据

支撑数据由作者自存储,Email:zqxiong@library.ecnu.edu.cn。

1 熊泽泉.原始下载量与被引量.xlsx. 所获取的CNKI每篇论文分年下载量与被引量原始值.

2 熊泽泉.转化后的下载量与被引量.xlsx. 对原始下载量和被引量数据进行转换后的绝对下载量与被引量.

3 熊泽泉. 不同被引模式的相关性.spv. 聚类结果及相关性分析结果.

参考文献

1 李勇,邵钟钰,赵星. Altmetrics背景下的期刊多维度测评指标研究[J]. 情报学报, 2017(02): 190-196.

2,16 Kurtz M J, Bollen J. Usage Bibliometrics[J]. Annual Review of Information Science and Technology, 2010, 44: 3-64.

3,27,40,48 Schloegl C, Gorraiz J, Gumpenberger C, et al. Comparison of Downloads, Citations and Readership Data for Two Information Systems Journals[J]. Scientometrics, 2014, 101(2): 1113-1128.

4 Glaenzel W, Gorraiz J. Usage Metrics Versus Altmetrics: Confusing Terminology? [J]. Scientometrics, 2015, 102(3): 2161-2164.

5 赵星. 学术文献用量级数据Usage的测度特性研究[J]. 中国图书馆学报, 2017(3): 44-57.

6 Wang X, Fang Z, Sun X. Usage Patterns of Scholarly Articles On Web of Science: A Study On Web of Science Usage Count [J]. Scientometrics, 2016, 109(2): 917-926.

7 Wang X, Xu S, Peng L, et al. Exploring Scientists' Working Timetable: Do Scientists Often Work Overtime? [J]. Journal of Informetrics, 2012, 6(4): 655-660.

8 丁佐奇. 基于Web of Science的论文使用次数和被引频次的相关性分析[J]. 中国科技期刊研究, 2017(12): 1166-1170.

9 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(1): 68-73.

10,56 Wang X，Wang Z，Mao W，et al. How Far Does Scientific Community Look Back?［J］. Journal of Informetrics，2014，8(3)：562-568.

11 Wang X，Liu C，Mao W，et al. The Open Access Advantage Considering Citation，Article Usage and Social Media Attention［J］. Scientometrics，2015，103(2)：555-564.

12 Wang X，Mao W，Xu S，et al. Usage History of Scientific Literature：Nature Metrics and Metrics of Nature Publications［J］. Scientometrics，2014，98(3)：1923-1933.

13 杨思洛，袁庆莉，韩雷. 中美发表的国际开放获取期刊论文影响比较研究［J］. 中国图书馆学报，2017(01)：67-88.

14,43 Vaughan L，Tang J，Yang R. Investigating Disciplinary Differences in the Relationships Between Citations and Downloads［J］. Scientometrics，2017，111(3)：1533-1545.

15,18 Wan J，Hua P，Rousseau R，et al. The Journal Download Immediacy Index（DII）：Experiences Using a Chinese Full-Text Database［J］. Scientometrics，2010，82(3)：555-566.

17,19,32,38,57,58 Brody T，Harnad S，Carr L. Earlier Web Usage Statistics as Predictors of Later Citation Impact［J］. Journal of the Association for Information Science and Technology，2006，57(8)：1060-1072.

20,28,41,49 Schloegl C，Gorraiz J. Comparison of Citation and Usage Indicators：The Case of Oncology Journals［J］. Scientometrics，2010，82(3)：567-580.

21,29,42 Schloegl C，Gorraiz J. Global Usage Versus Global Citation Metrics：The Case of Pharmacology Journals［J］. Journal of the Association for Information Science and Technology，2011，62(1)：161-170.

22 Bollen J，de Sompel H V，Smith J A，et al. Toward Alternative Metrics of Journal Impact：A Comparison of Download and Citation Data［J］. Information Processing & Management，2005，41(6)：1419-1440.

23 Jamali H R，Nikzad M. Article Title Type and its Relation with the Number of Downloads and Citations［J］. Scientometrics，2011，88(2)：653-661.

24,39 Guerrero-Bote V P，Moya-Anegon F. Relationship Between Downloads and Citations at Journal and Paper Levels，and the Influence of Language［J］. Scientometrics，2014，101(2)：1043-1065.

25 Subotic S，Mukherjee B. Short and Amusing：The Relationship Between Title Characteristics，Downloads，and Citations in Psychology Articles［J］. Journal of Information Science，2014，40(1)：115-124.

26 牛昱昕，宗乾进，袁勤俭. 开放存取论文下载与引用情况计量研究［J］. 中国图书馆学报，2012(4)：119-127.

30,37,47 Moed H F. Statistical Relationships Between Downloads and Citations at the Level of Individual Documents within a Single Journal［J］. Journal of the American Society for Information Science and Technology，2005，56(10)：1088-1097.

31 陆伟，钱坤，唐祥彬. 文献下载频次与被引频次的相关性研究——以图书情报领域为例［J］. 情报科学，2016(1)：3-8.

33 Lokker C，Mckibbon K A，Mckinlay R J，et al. Prediction of Citation Counts for Clinical Articles at Two Years Using Data Available within Three Weeks of Publication：Retrospective Cohort Study［J］. Bmj，2008，336(7645)：655-657.

34,62 Stegehuis C，Litvak N，Waltman L. Predicting the Long-Term Citation Impact of Recent Publications［J］. Journal of Informetrics，2015，9(3)：642-657.

35,59 郭强，赵瑾，刘新新，等. 利用期刊下载次数估计后期被引次数的研究［J］. 图书馆理论与实践，2010(11)：45-49.

36 谢娟，龚凯乐，成颖，等. 论文下载量与被引量相关关系的元分析［J］. 情报学报，2017(12)：1255-1269.

44,46 Duan Y，Xiong Z. Download Patterns of Journal Papers and their Influencing Factors［J］. Scientometrics，2017，112(3)：1761-1775.

45 Tkaczynski A. Segmentation Using Two-Step Cluster Analysis［M］. Segmentation in Social Marketing：Process，Methods and Application，Dietrich T，Rundle-Thiele S，Kubacki K，Singapore：Springer Singapore，2017，109-125.

50,51,54 李江，姜明利，李玥婷. 引文曲线的分析框架研究——以诺贝尔奖得主的引文曲线为例［J］. 中国图书馆学报，2014(2)：41-49.

52 Mazloumian A，Eom Y H，Helbing D，et al. How Citation Boosts Promote Scientific Paradigm Shifts and Nobel Prizes［J］. PLoS One，2011，6(5)：e18975.

53 Willett P. Readers' Perceptions of Authors' Citation Behaviour［J］. Journal of Documentation，2013，69(1)：145-156.

55 Fu L D，Aliferis C F. Using Content-Based and Bibliometric Features for Machine Learning Models to Predict Citation Counts in the Biomedical Literature［J］. Scientometrics，2010，85(1)：257-270.

60 Zavos C，Kountouras J，Zavos N，et al. Predicting Future Citations of a Research Paper From Number of its Internet Downloads：The Medical Hypotheses Case［J］. Medical Hypotheses，2008，70(2)：460-461.

61 Perneger T V. Relation Between Online "Hit Counts" and Subsequent Citations：Prospective Study of Research Papers in the BMJ［J］. Bmj，2004，329(7465)：546-547.

63 Ponomarev I V，Williams D E，Hackett C J，et al. Predicting Highly Cited Papers：A Method for Early Detection of Candidate Breakthroughs［J］. Technological Forecasting and Social Change，2014，81：49-55.

（收稿日期：2018-03-12）