科技论文被引频次与下载频次的相关性分析

丁佐奇　郑晓南　吴晓明*

收稿日期:2009-11-16
修回日期:2010-03-31

中国药科大学《中国天然药物》编辑部,210009　南京市童家巷24号,E-mail: zqding1028@yahoo.com.cn

摘　要　通过清华大学中国知网CNKI数据库下载和引证《中国天然药物》和《中国药科大学学报》的数据分析,研究了科技论文被引频次与下载频次的相关性,结果显示文章出版后2-4年引证达到峰值,而网络下载近2年出版的文章占较高比例;专栏及综述的下载率高,而研究性论文的引用率高;单篇论文的被引频次与下载频次的相关性较差,用先期的下载次数对后期的被引次数进行预测行不通。

关键词　科技论文　被引频次　下载频次

科技期刊是传播科技成果的重要载体,它承载了传播科技创新、引导科研方向的功能。期刊被引频次是学术质量以及学术影响力的重要评价指标;网络下载频次可测度上网期刊的扩散速率,是研究期刊在网络环境下传播效率的一个新指标。所以,对论文或是期刊进行评价,包括对核心期刊的确定,被引次数与被下载次数同时作为文献价值的表征而共同被纳入到评价指标体系。引用率越高,学术影响力越大;WEB即年下载率越高,说明读者对期刊的兴趣越大。

由于大多数情况下文献的被引用次数与该文献质量的高度正相关,使得引文分析作为科学评价的方法具有一定的可行性,但是作者引用的文献往往仅占其在研究工作中所阅读过的文献的一部分,那么其中未被引用的文献的价值该如何去体现,下载次数是一个日渐公认的评估指标,下载次数在直观上能够与该文献的被阅读次数相对应,而且数据源以及数据统计目前已可提供[1]。

文章的被引频次与下载频次是否正相关?普遍认为,文章被阅读的次数越多,被引用的可能性越大。刘筱梅等利用Person(皮尔松)相关系数公式计算证明了期刊下载行为与引用行为存在正相关关系[2]。庞景安也得出"期刊总被引频次与总下载频次指标间存在较强的相关关系"的结论[3]。万锦堃等探讨了期刊论文下载频次与被引频次的排序情况,答案是并不完全一致,建议将两者都作为独立指标在综合评价中予以适当考虑[4]。但这些研究均针对期刊的总被引频次与总下载频次,单篇论文的下载频次和被引频次是否也存在正相关?相关研究很少,赵大良统计《西安交通大学学报》等4种期刊在中国知网的下载量与被引频次的关系,发现网络访问量(包括WEB下载量)与文章的被引频次存在着负相关[5]。笔者发现笔者所在期刊——《中国天然药物》论文的被引频次与下载频次显示很差的相关性。本文选取中国药科大学主办的两种医药学核心期刊《中国天然药物》(*CJNM*)和《中国药科大学学报》(*JCPU*),对被引频次与下载频次的相关性进行深入的分析和探讨。

1　数据来源和处理方法

根据《中国天然药物》和《中国药科大学学报》2003~2008年发表的论文在《中国知网》CNKI《中国学术期刊文献评价统计分析系统》下载数据库与引证数据库中的下载频次与被引频次,对两种期刊下载与被引频次最高的前20(Top20)篇文章进行分析,探讨影响科技论文被引频次和下载频次相关性的各种因素。

2　结果分析与讨论

2.1　两种期刊被引频次Top20论文分析

对*CJNM*和*JCPU*被引频次Top20论文进行分析,统计文章发表后0~6年各年的被引频次,如表1、表2所示,两种期刊高被引论文多数是发表后2~4年被引达到高峰。根据论文产生周期(查阅文献、论文成文、发表时滞)计算,推测*N*年用户的检索行为可以在*N*+2年发表的论文中得到体现,也就是说论文出版后第2年起应该比较广泛地被引用;医药学类论文的半衰期比较长,本文的研究结果与上述相符。

2.2　两种期刊被引/下载Top20论文相关性分析

对比*CJNM*被引/下载Top20论文的被引频次和下载频次,发现了一个有意思的现象:被引/下载Top20论文中只有4篇是相重的,也就是说80%的高被引论文未被高下载,

* 通讯作者:吴晓明,E-mail: xmwu@cpu.edu.cn

表 1　*CJNM* 被引频次 Top20 论文分布

序号	发表年	总被引频次	N_0	N_1	N_2	N_3	N_4	N_5	N_6	N_X
1	2004	60	1	3	10	19	20	7	↙	0
2	2003	42	0	4	4	11	9	7	2	5
3	2005	40	1	5	15	19	0	↙		0
4	2003	37	0	6	6	8	13	3	1	0
5	2003	31	0	4	5	10	4	7	1	0
6	2003	30	0	1	3	9	8	6	3	0
7	2004	30	0	2	9	10	6	2		1
8	2003	29	0	5	3	7	6	6	1	1
9	2005	27	0	4	8	9	6	↙		0
10	2006	26	2	13	10	1	↙			0
11	2004	26	0	4	7	6	7	2	↙	0
12	2003	26	0	1	4	7	8	5	0	1
13	2003	25	0	5	0	8	7	3	0	2
14	2004	24	0	1	7	8	7	1	↙	0
15	2003	21	0	0	4	3	7	4	3	0
16	2004	21	0	3	8	4	5	1	↙	0
17	2004	21	0	8	3	3	6	1	↙	0
18	2003	20	0	0	1	3	9	6	0	1
19	2004	20	0	0	4	8	7	1	↙	0
20	2003	19	0	3	5	1	7	2	0	1
平均值			0.2	3.6	5.8	7.7	7.1	3.2	0.6	0.6

注：N_0 表示出版当年的被引频次，$N_1 \sim N_6$ 表示出版后第 1 至第 6 年的被引频次，N_x 表示出版年未知的被引频次

表 2　*JCPU* 被引频次 Top20 论文分布

序号	发表年	总被引频次	N_0	N_1	N_2	N_3	N_4	N_5	N_6	N_X
1	2003	46	1	8	6	8	14	6	3	0
2	2003	38	1	11	3	9	6	5	1	2
3	2003	38	3	7	7	10	7	3	0	1
4	2004	37	0	1	6	15	12	3	↙	0
5	2003	35	0	1	6	7	6	8	5	2
6	2004	30	0	3	2	16	9	0	↙	0
7	2003	26	0	1	3	5	13	3	0	1
8	2003	26	0	2	6	4	7	4	0	3
9	2003	26	0	0	0	9	8	8	0	1
10	2003	25	0	2	2	4	7	8	2	0
11	2003	25	0	1	6	5	6	4	2	1
12	2004	24	0	0	8	4	9	3	↙	0
13	2004	23	0	3	5	7	7	0	↙	1
14	2004	23	0	3	7	4	8	1	↙	0
15	2005	21	0	1	10	10	0	↙		0
16	2003	20	0	3	3	7	4	3	0	1
17	2005	20	0	4	4	7	4	↙		1
18	2003	19	1	1	2	6	5	3	0	1
19	2004	19	0	1	5	7	5	0	↙	1
20	2004	19	0	1	7	6	5	0	↙	0
平均值			0.3	2.7	4.9	7.5	7.1	3.1	0.7	0.8

注：N_0 表示出版当年的被引频次，$N_1 \sim N_6$ 表示出版后第 1 至第 6 年的被引频次，N_x 表示出版年未知的被引频次

80% 的高下载论文没有高被引。为了检验这种现象是否为个例，又研究了 *JCPU*，发现被引/下载 Top20 论文中只有 3 篇相重，即 85% 的高被引论文未被高下载，85% 的高下载论文没有高被引。对 *CJNM* 和 *JCPU* 被引 Top20 论文的被引次数和下载次数进行相关性分析，发现相关系数分别为 0.331 和 0.012，如图 1 所示；*CJNM* 和 *JCPU* 下载 Top20 论文的下

载次数和被引次数的相关性同样较差，如图 2 所示，相关系数分别为 0.0233 和 0.084。这方面的研究尚未见系统报道，下面从论文出版年份和论文类型两方面分析。

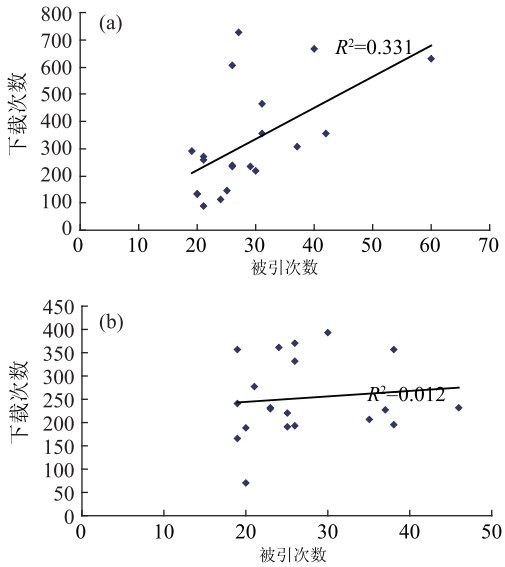

图 1　*CJNM*(a)和 *JCPU*(b)被引 Top20 论文被引和下载次数的相关性

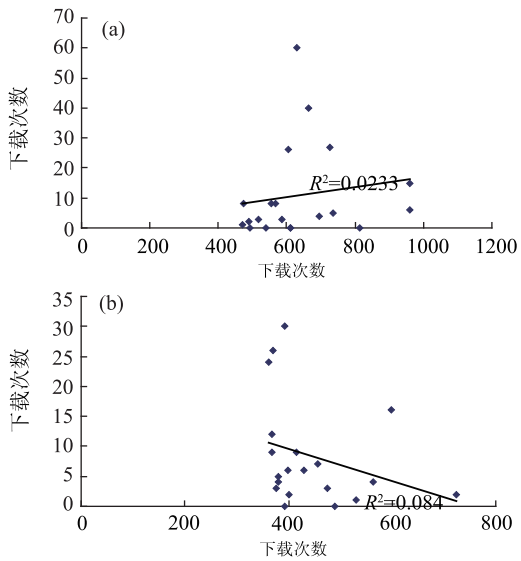

图 2　*CJNM*(a)和 *JCPU*(b)下载 Top20 论文下载和被引次数的相关性

2.2.1　两种期刊被引/下载 Top20 论文出版年分析

从表 3 和表 4 可见，高被引/下载文章与出版年之间有一定规律。文章出版后 2 ~ 4 年被引次数达到峰值，CNKI 引证数据库中，*CJNM* 在 2003 ~ 2005 年出版的文章占 Top20 的 95%，*JCPU* 在 2003 ~ 2005 年出版的文章占 Top20 的 100%。网络下载近两年出版的文章占较高比例，在 CNKI 下载数据库中，*CJNM* 2007 年和 2008 年出版的文章占到 Top20 的 50%，*JCPU* 2007 年出版的文章占 6 篇，是历年中高下载篇数最多的，说明数字化传播比起传统传播方式具有即时性。

表 3　*CJNM* 被引/下载 Top20 论文出版年分析

数据库	2003	2004	2005	2006	2007	2008
CNKI 下载	1	3	3	3	5	5
CNKI 引证	10	7	2	1		

表 4　*JCPU* 被引/下载 Top20 论文出版年分析

数据库	2003	2004	2005	2006	2007	2008
CNKI 下载	4	3	5	2	6	0
CNKI 引证	11	7	2	0		

2.2.2　两种期刊被引/下载 Top20 论文类型分析

表 5 和表 6 分析了两种期刊被引/下载 Top20 论文的类型。*CJNM* 高下载 Top5 中，有 4 篇专论“思路与方法”、1 篇综述；*JCPU* 高下载 Top5 中有 3 篇专论“药学前沿”，1 篇研究论文和 1 篇信息，说明策划名牌栏目，能够加强学术期刊的下载率和影响力。*CJNM* 高被引 Top5 中，有 2 篇专论(组稿)，2 篇综述(组稿)，1 篇研究论文，说明《中国天然药物》的组稿质量较高，要提高文章的被引次数，应多组高质量、权威性的论坛、评述、综述类文章。根据表 7 的统计结果，*CJNM* 下载 Top20 中专论和综述有 17 篇，占 85%，而引证 Top20 中研究论文 13 篇，占 65%；*JCPU* 下载 Top20 中专论、综述及信息占半数，而引证 Top20 中研究论文 18 篇，占 90%。可知专论、综述及信息的下载率高，而研究论文的引证率高。

表 5　*CJNM* 被引/下载 Top20 论文类型分析

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
CNKI 下载	C	R	C	C	C	C	R	P	P	R	P	R	C	C	C	C	C	C	C	C
CNKI 引证	P	R	C	R	C	P	R	P	C	P	P	P	P	P	R	P	P	P	P	P

注：C-Column，R-Review，P-Paper

表 6　*JCPU* 被引/下载 Top20 论文类型分析

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
CNKI 下载	C	P	C	C	I	R	P	P	C	C	C	I	P	P	P	P	P	C	P	P
CNKI 引证	P	P	C	P	P	P	P	P	P	P	P	P	P	P	R	P	P	P	P	P

注:C-Column,R-Review,P-Paper

表 7　*CJNM* 和 *JCPU* 被引/下载 Top20 论文类型统计

	CJNM			*JCPU*			
	C	R	P	C	R	P	I
CNKI 下载	13	4	3	7	1	10	2
CNKI 引证	3	4	13	1	1	18	0

注:C-Column,R-Review,P-Paper,I-Information

3　讨论

期刊被引频次是学术质量以及学术影响力的重要评价指标,网络下载频次从某种意义上说,可以更直接地显示期刊文献被读者使用的情况,从而避免一些人为因素(如引文行为和引文动机)对引文分析数据的干扰,而且,从论文上网到被读者下载,其间相对较短的延时使得下载次数能够作为论文价值的快速反映[1]。下载频次虽然反映了文章被社会关注的程度,但它仅仅是文章被引的前奏,并不是所有的下载都会引用。

根据本文的研究结果,文章出版后 2 ~ 4 年被引次数达到峰值,而网络下载近两年出版的文章占较高比例,引用和下载有时间差(至少两年),这在一定程度上能够解释高下载和高被引论文不对应的现象,但本研究中两种期刊在 2003 ~ 2005 年发表的高下载论文到 2008 年被引次数还很少,从本文的研究结果看,这在一定程度上与论文类型有关。例如,*CJNM*,"思路与方法"专栏的下载率特别高,有可能是因为浏览量和下载量受论文选题影响很大,这类启迪科研思路、总结科研方法的评述和综述,实用性较强,对科研人员的实际工作有很大的启发和指导作用,但不一定会转化成作者的引用行为。同样,*JCPU* 中"药学前沿"专栏和信息的下载量高,也是由于专栏和信息的实用性较强。

有观点[6]认为,用"下载率"来衡量期刊的质量,已有相当的市场,现已引起出版界及图书情报界的重视,据统计,已达到与影响因子相当的效果。被引用的文章只是说明此文章所起的部分效果,有些人下载文章只是为了学习,并不发表文章,这和以上所说的"实用性"意思相近。该观点还认为,引文率高,与引用的作者的写作条件和习惯有关,可能作者手头恰好有某本期刊,于是引用了其中的相关文献,这种引用就没有下载这个过程。

还有一种情况,高下载有可能是由于文章的题目和摘要吸引人,但内容并无多少可借鉴之处,或者可能是作者由于种种原因故意不引或引其他不相关文献,这就涉及到引文行为和引文动机,引用除了自引和课题组引,还存在否定引用,而且很多引用并无实际价值,情况比较复杂[7]。

引文分析是建立在必要的假设基础之上,其包括对文献的引用意味着引用作者在其研究工作中对该文献的使用,文献被引用是该文献的质量、重要性以及影响力的体现,作者实际引用的文献在所有可能被引用的文献中具有最优性,以及引用文献与被引用文献在内容上相互关联等,尽管这些假设与实际的引用行为以及引用动机并不完全一致,但这不妨碍引文分析作为有价值的文献计量方法而有其广泛且重要的应用。但引文分析有其相对的滞后性,而下载次数作为反映文献价值的早期指标,使得科学评价活动可以有所提前[1]。

既然下载次数与被引次数共同作为文献价值的体现,那么两者是否应该具有共性,是否可以利用先期的下载次数对后期的被引次数进行预测[7],根据本文的研究结果,这点似乎行不通,值得商榷。当然,本文仅考察了 2 种期刊,下载次数与被引次数两者之间统计相关性的规律还有待在大样本期刊中进一步考察。

参考文献

1　郭强,赵瑾,刘思源等. 科技论文下载次数的统计性质研究. 情报科学,2009,27(5):690 - 694

2　刘筱梅,张建勇. 数字获取资源对科学研究的影响——电子期刊全文下载与引用分析. 大学图书馆学报,2009,(1):60 - 63

3　庞景安. 中文科技期刊下载计量指标与引用计量指标的比较研究. 情报理论与实践,2006,29(1):44 - 48

4　万锦堃,花平寰,孙秀坤. 期刊论文被引用及其 Web 全文下载的文献计量分析. 现代图书情报技术,2005,(4):58 - 62

5　赵大良. 不可思议的现象:网络传播与被引频次的关联分析. [2009-01-11 16:23:43]. http://zhaodal.blog.163.com

6　许昌淦. 改进对我国科技期刊的评估. 中国科技期刊研究,2009,20(5):862 - 865

7　Hari P. Sharma. Download plus citation counts — a useful indicator to measure research impact. *Current Science*, 2007, 92(7):873