基于 Web of Science 和 ESI 数据库高被引论文的界定方法*

刘雪立

收稿日期:2012-07-30
修回日期:2012-10-08

河南省科技期刊研究中心,新乡医学院期刊社《眼科新进展》编辑部,453003　河南新乡,E-mail:liueditor@163.com

摘　要　随着 Science Citation Index(科学引文索引,SCI)和 Essential Science Indicators(基本科学指标,ESI)数据库在全球科学评价中的普遍应用,越来越多的学者利用引文分析工具对高被引论文进行统计分析。系统介绍了应用 Web of Science 和 ESI 数据库确定高被引论文的标准和方法,如限定被引频次法、百分比法、普赖斯定律法、共时法和历时法等。对各种确定高被引论文的标准和方法进行了评价。

关键词　Web of Science　Essential Science Indicators　高被引论文　热点论文

近年来,国内外科学评价和情报学研究中越来越重视对高被引和高影响力论文的研究。科技期刊研究人员从事高被引论文研究,主要是评价同类期刊的学术影响力、辅助期刊选题、组稿和约稿[1-3],所用数据库多数为 CNKI 的中国引文数据库或学术文献总库[1-7],也有人应用中文社会科学引文索引[8]和万方数据库[9]。随着 Science Citation Index(科学引文索引,SCI)和 Essential Science Indicators(基本科学指标,ESI)数据库在全球科学评价中的普遍应用[10-12],越来越多的学者利用美国汤森路透科技信息集团开发的系列引文分析工具(包括 Web of Science、ESI、InCites 等),通过论文数、被引频次、篇均被引频次、高被引论文、热点论文等一系列论文被引指标,对国家和地区、机构(大学)和科学家进行学术评价[10-20],尤其是高被引论文的统计分析受到了国内外学者的广泛关注[1-7,21-23]。有鉴于此,有必要对 Web of Science 和 ESI 数据库中高被引论文的界定方法进行详细介绍。

1　Web of Science 数据库中高被引论文的界定

1.1　确定检索对象和主题

通常情况下,对高被引论文进行统计分析必须首先确定检索对象和主题。可以选择对某一学科的高被引论文进行分析,如 SCI 数据库中内科学高被引论文的统计分析;也可以选择对某一主题高被引论文进行研究,如 SCI 数据库中纳米材料高被引论文的统计分析。对一个学科进行高被引论文统计分析时,确定的学科范围不要过大,如把整个自然科学作为一个学科进行研究,一是没有必要,二是没有实际意义,三是检索策略不容易实现。对科技期刊编辑工作者来讲,通常需要跟踪了解同类期刊的学术影响力、高被引和高影响力论文的分布特征,以指导期刊选题、组稿及约稿。因此,选择一个学科期刊进行高被引论文统计分析更有意义。

1.2　论文检索方法

以期刊为研究对象和单纯以论文作为研究对象,其检索方法略有不同,现以眼科学期刊和论文为例分述如下。

以眼科学期刊为研究对象,其检索步骤和方法为:(1)确定检索期刊。最好是检索 SCI 数据库收录的所有眼科学期刊(当然,为了特定目的,有时候也可以选择性检索本学科中一些重要的期刊,通常以影响因子高低作为选择标准)。登录 ISI Web of Science 数据平台(http://www.isiknowledge.com),选择“其他资源”选项,打开页面后选择 Journal Citation Reports(JCR),点选“JCR Science Edition”,选择按学科浏览期刊,通过下拉滚动条找到“Ophthalmology”,排序方式选择按影响因子排序,点击“Submit”。这样就能获得 SCI 数据库收录眼科学期刊的 ISSN 号。(2)确定文献时间区间。文献时间区间的选择应根据研究者的研究目的、文献数量、习惯等自行确定,一般选择近 10 年或者近 5 年。(3)编制高级检索式。利用期刊的 ISSN 号编制检索式:IS = (0002 - 9394 OR 0003 - 9950 OR 0004 - 2749 …… OR 1040 - 5488) AND PY = (2003 OR 2004 OR 2005 OR 2006 OR 2007 OR 2008 OR 2009 OR 2010 OR 2011 OR 2012)。式中,IS 为 ISSN 检索字段,PY 为出版年检索字段,OR 和 AND 为布尔逻辑运算符“或”、“和”。

以眼科学论文为研究对象,检索方法相对复杂一点。除了眼科学期刊发表的论文,还必须考虑综合性医学期刊(如

* 河南省科技发展计划软科学项目(编号:112400450118)

New Engl J Med,*Lancet*)或综合性自然科学期刊(如 *Nature*,*Science*)上发表的眼科学论文。非眼科学期刊发表的眼科学论文不可能通过期刊检索而实现,因为我们根本不可能知道哪些非眼科学期刊发表了眼科学论文。但是,发表眼科学论文的通常都是各眼科机构(主要包括眼科医院、综合性医院眼科及专门的眼科学研究机构)眼科工作者。因此,可以通过地址字段进行检索,其检索式为:AD = (Ophthal * OR Eye *) AND PY = (2003 OR 2004 OR 2005 OR 2006 OR 2007 OR 2008 OR 2009 OR 2010 OR 2011 OR 2012)。

式中,AD 为地址检索字段,* 为截词符号。把上述两个检索式的检索结果以布尔逻辑运算符"OR"组配,就得到近 10 年间 SCI 数据库收录的所有眼科学论文。

另外,Web of Science 数据库还提供了学科类别检索字段(SU)和 Web of Science 分类字段(WC)该数据库 SU 检索字段中把学科类别划分为生命科学与生物医学(75 个次级学科)、自然科学(17 个次级学科)、应用科学(21 个次级学科)、艺术与人文科学(14 个次级学科)、社会科学(24 个次级学科)等五大类。以上可供检索的次级学科名称(英文)在 Web of Science 数据库的高级检索界面,点击"学科类别"的超级链接就可以查阅。而 WC 检索字段给出的是 Web of Science 数据库标准的学科分类,包括自然科学、社会科学、艺术与人文科学的 249 个学科,可供检索的学科更全。

1.3 确定高被引论文

将上述检索结果按照被引频次降序排列(SCI 数据库提供了很多自动排序方式),可以选择以下几种方法确定高被引论文。

1.3.1 限定被引频次法

通常根据被引频次的分布状况,在保证有充足样本数的基础上,选择 10 或 5 的整数倍作为高被引频次的标准。

1.3.2 百分比法

通常根据文献计量学中的"二八现象",把被引频次较高的前 20% 的论文作为高被引论文[24]。如果文献量很大,可以把高被引的标准提高,按照统计学的一般规律,可以取被引频次较高的前 10%、5%,甚至 1% 作为高被引论文。

1.3.3 普赖斯定律法

在文献计量学中,普赖斯定律是用来确定高产和高影响力作者的[25]。多数情况下,高产作者和高被引论文的分布具有相同或相似的规律。因此,可以借用普赖斯定律确定高被引论文[1,26-27]。普赖斯定律的数学表达式为:

$$m = 0.749\sqrt{n_{\max}}$$

式中,$n_{\max}$为被引频次最高的论文的被引频次,m 为确定的高被引论文的最低被引频次。

上述确定高被引论文的方法显然没有考虑论文发表的时间先后,把研究时间区间内不同年代发表的论文等同对待,只考虑论文被引频次高低。这种确定高被引论文的方法属于共时法。通常情况下,论文刚发表时很少被人引用,经历的时间越长,累计的被引频次越高。基于这一引证规律,可以考虑采用历时法确定高被引论文。具体做法是,分别检索不同年代发表的论文,按照上述方法确定不同年代的高被引论文。最后把不同年代的高被引论文集合起来,作为要研究的高被引论文。

2 ESI 数据库中高被引论文的界定

2.1 ESI 数据库共时法高被引论文的界定

ESI 数据库是美国汤森路透科技集团于 2001 年开发的基于 Web of Science 数据库权威的、专门的引文分析工具。该数据库共有 4 个模块:Citation Ranking(引文排序)、Most Cited Paper(最高被引论文)、Citation Analysis(引文分析)和 Science Commentary(科学评论)。最高被引论文模块给出了 Highly Cited Papers(高被引论文)和 Host Papers(热点论文),可以直接确定高被引论文。需要注意的是,ESI 数据库中的高被引论文和热点论文是有严格定义的。高被引论文是指最近 10 年来被引频次排在前 1% 的论文;热点论文是指最近 2 年内的论文,在最近 2 个月被引频次排在前 0.1% 的论文[28]。

利用 ESI 数据库可以直接检索 22 个学科领域的高被引论文和热点论文,包括农业科学、生物学与生物化学、化学、临床医学、计算机科学、经济与贸易、工程学、环境科学/生态学、地球科学、免疫学、材料科学、数学、微生物学、分子生物学与遗传学、多学科、神经与行为科学、药理与毒理学、物理学、植物与动物学、精神病学与心理学、综合性社会科学、空间科学等。具体检索方法是:登录 ISI Web of Science 数据平台(http://www.isiknowledge.com),选择"其他资源"选项,打开页面后选择 Essential Science Indicators 数据库,然后点击 Most Cited Papers 模块的 Highly Cited Papers,就进入了高被引论文检索页面(图 1)。除了按学科检索高被引论文外,还可以检索某科学家、某机构、某国家和地区、某期刊的高被引论文。当然,也可以文题中一个主题词或关键词进行检索(研究某一主题的高被引论文可以使用此方法进行文献检索)。热点论文的检索与高被引论文的检索完全一样,在此不再赘述。

2.2 ESI 数据库历时法高被引论文的界定

ESI 数据库把学科系统分为 22 个大的学科体系,但多数情况下我们需要了解的是更微观学科的高被引论文分布情况。比如,怎样利用 ESI 数据库确定眼科学的高被引论

文呢?

2.2.1 确立学科基准线

登录 ESI 数据库,点击 Citation Analysis 模块的 Baseline(基准线),页面打开后点击"View the percentiles table",得到 22 个学科的百分比基准线。现将临床医学的百分比基准线列于表 1。

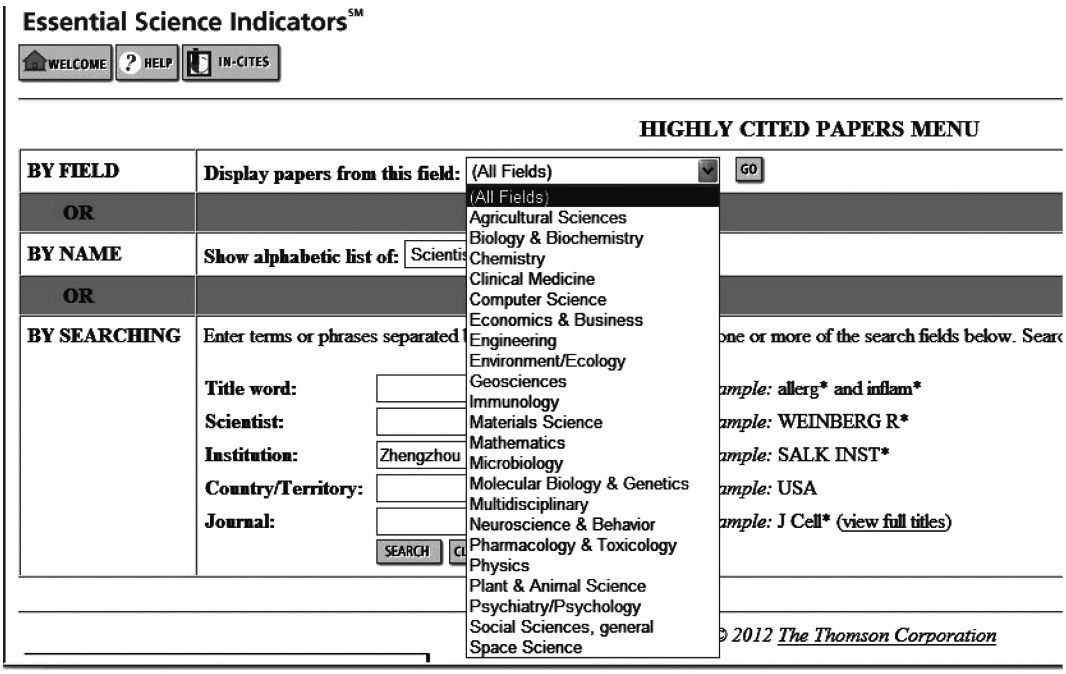

图 1 ESI 数据库高被引论文检索界面

表 1 ESI 数据库中临床医学领域被引论文百分比基准线

基准线%	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	All Years
0.01	2367	1545	1545	1464	973	868	647	508	257	99	10	1079
0.10	645	624	547	485	415	343	244	174	98	32	6	385
1.00	206	192	174	156	129	108	81	56	33	11	3	118
10.00	55	52	48	44	37	31	23	16	10	4	1	30
20.00	33	31	30	27	23	19	14	10	6	2	1	17
50.00	11	11	11	10	9	7	5	4	2	1	0	5

2.2.2 根据基准线确定高被引论文

论文被引百分比基准线的意义在于,某年发表的论文被引频次排在前百分之几需要的被引频次。如 2008 年发表的一篇临床医学论文,被引频次达到 81 次才能排在前 1%(见表 1)。结合 ESI 数据库中高被引论文的定义,1% 基准线可以被确定为高被引论文的标准。有了这一标准,就可以确定临床医学领域任一学科某年度的高被引论文。如检索 2003 年眼科学论文,按被引频次降序排列,被引频次≥192 者均为临床医学领域的高被引论文。逐年检索眼科学论文,按被引频次排序,与论文被引 1% 基准线比对,就可以把每年的高被引论文筛选出来。

这一确定高被引论文的方法显然属于历时法。如果想利用该基准线确定共时法高被引论文,则以"All Year"的 1% 基准线(118)作为标准。

3 对确定高被引论文标准和方法的评价

根据确定高被引论文的标准不同,我们把高被引论文确定的方法分为 3 种,一是固定被引频次法,二是百分比法,三是普赖斯指数法。第一种方法主观性太强,缺乏理论依据,因而学术性较差;第二种方法以文献计量学和统计学理论为依据,具有较强的科学性和广泛是适用性,建议列为首选;第三种方法虽然有文献计量学理论基础,但是样本数量不足或者引文分布不符合普赖斯定律时,不适合用此方法确定高被引论文。因此,该方法具有明显的局限性。

根据文献检索和数据处理方法不同,我们把高被引论文确定的方法分为 2 种,一是共时法,二是历时法。共时法由

于不考虑论文发表时间的早晚,用统一的被引频次标准界定不同年度的论文,数据处理比较简单,但是近期发表的高水平论文不容易成为高被引论文。因此,该方法确定高被引论文必然会遗漏部分近期发表的高影响力论文。由于历时法确定高被引论文需要逐年统计论文被引频次,各年度论文高被引论文需建立不同的被引频次标准,因而文献检索和数据处理相对较复杂,但科学性较强,极大地克服了共时法的局限性。

在高被引论文研究中,究竟采用什么标准、应用哪种方法确定高被引论文? 建议根据各自研究目的、文献检索条件、数据分布特征等,科学规划,灵活掌握,以达到最佳研究效果和预期目标。

4 结语

目前,高被引论文和热点论文的统计分析逐渐成为文献计量学、期刊评价和科学评价领域的研究热点,论文被引数据主要来源于中国引文数据库或学术文献总库、中文社会科学引文索引、中国科学引文数据库和万方数据库等。随着SCI数据库及其系列引文分析工具在国内科学评价领域的广泛应用,将会有越来越多的学界同仁应用 Web of Science 和 ESI 数据库对高被引论文进行分析。系统介绍基于 Web of Science 和 ESI 数据库确定高被引论文的方法,将有助于今后高被引论文研究及科技期刊的文献计量学评价。

参考文献

1 刘雪立,王兆军. 2004 ~ 2008 年我国情报专题研究高被引论文的统计与分析. 情报杂志,2010,29(1):64 - 67

2 聂兰英,王钢,金丹等. 我国 11 种医学影像学核心期刊的高被引论文分析. 中国科技期刊研究,2011,22(3):377 - 380

3 张坤,赵粉侠,曹龙. 我国林业类核心期刊高被引论文统计分析. 中国科技期刊研究,2011, 22(4):549 - 554

4 徐剑,赵征南. 中国版权研究的历史回顾与反思——基于 CNKI (1979 ~ 2009) 的高被引论文分析. 上海交通大学学报:哲学社会科学版,2011,19(2):52 - 59

5 张诗博. 2004 ~ 2008 年国内图书馆学研究高被引论文的统计与分析. 情报科学,2011,19(3):387 - 390

6 方红玲. 2003 ~ 2008 年眼科学高被引论文统计分析. 中国科技期刊研究,2010,21(2):197 - 200

7 陈勤,姬晓云. 寄生虫学相关期刊 2003 ~ 2010 年高被引论文分析. 中国科技期刊研究,2011, 22(4):559 - 562

8 张士靖,姚强,杜建. 基于 CSSCI 的知识服务领域高被引作者的可视化研究. 情报杂志,2010,29(9):45 - 48,59

9 张建合,任长江.《中国科技期刊研究》高被引论文特征分析. 中国科技期刊研究,2011,22(2):207 - 210

10 Huang Y, Wang J. A bibliometric study of the trend in articles related to eutrophication published in Science Citation Index. *Scientometrics*, 2011, 89(3): 919 - 927

11 Singleton A. Bibliometrics and Citation Analysis; from the Science Citation Index to Cybermetrics. *Learned Publishing*, 2012, 23(3): 267 - 268

12 Fu HZ, Chuang KY, Wang MH, et al. Characteristics of research in China assessed with Essential Science Indicators. *Scientometrics*, 2011,88(3):841 - 862

13 Han W, Leydesdorff L. Korean journals in the Science Citation Index: What do they reveal about the intellectual structure of S & T in Korea? *Scientometrics*, 2008, 75(3): 439 - 462

14 Andreis M, Jokic M. An impact of Croatian journals measured by citation analysis from SCI-expanded database in time span 1975 ~ 2001. *Scientometrics*, 2008, 75(2): 263 - 288

15 Sotudeh, H. How sustainable a scientifically developing country could be in its specialties? The case of Iran's publications in SCI in the 21st century compared to 1980s. *Scientometrics*, 2012, 91(1): 231 - 243

16 Collazo-ReyesF, Luna - Morales ME, Russell JM, et al. Publication and citation patterns of Latin American & Caribbean journals in the SCI and SSCI from 1995 to 2004. *Scientometrics*, 2008,75(1): 145 - 161

17 王璞,刘雪立,刘睿远. SSCI 数据库中 3 种编辑出版类期刊的分析与评价. 中国科技期刊研究,2012,23(3):363 - 368

18 丁君,王兰英,郑艳丽等. 基于 Web of Science 数据库的第二语言习得文献计量学分析与评价. 现代情报,2012,32(6):101 - 106

19 邱均平,杨瑞仙. 基于 ESI 数据库的材料科学领域文献计量分析研究. 情报科学,2010,18(8):1121 - 1126

20 王璞,刘子扬,刘雪立 . 2001 ~ 2010 年 Nature 和 Science 发表我国科研论文及其学术影响力——基于 SCI 数据库的综合分析. 中国科技期刊研究,2011,22(6):844 - 847

21 Chuang KY, Wang MH, Ho YS. High-impact papers presented in the subject category of water resources in the essential science indicators database of the institute for scientific information. *Scientometrics*, 2011, 87(3):551 - 562

22 Miyairi N, Chang HW. Bibliometric characteristics of highly cited papers from Taiwan, 2000 ~ 2009. *Scientometrics*, 2012, 92(1): 197 - 205

23 Madhan M, Chandrasekar G, Arunachalam S. Highly cited papers from India and China. *Current Science*, 2010, 99(6): 738 - 749

24 刘雪立,方红玲,苗媛等. 五种综合性眼科学期刊论文下载量与被引量的关系及部分论文的量引背离现象. 中国科技期刊研究,2010,21(5):629 - 632

25 邱均平. 信息计量学. 武汉:武汉大学出版社,2007

26 魏瑞斌,陈丹丹. 基于引证网络的高被引文献实证分析——以知识服务为例. 现代情报,2011,31(3):117 - 121

27 苏君华. 中国档案学核心期刊影响力分析——以 2000 ~ 2009 年所载论文为研究对象. 档案学通讯,2010(5):15 - 20

28 Thomson Reuters. Citation thresholds. [2012-07-30]. http://www.sciencewatch.com/ about/met/thresholds/